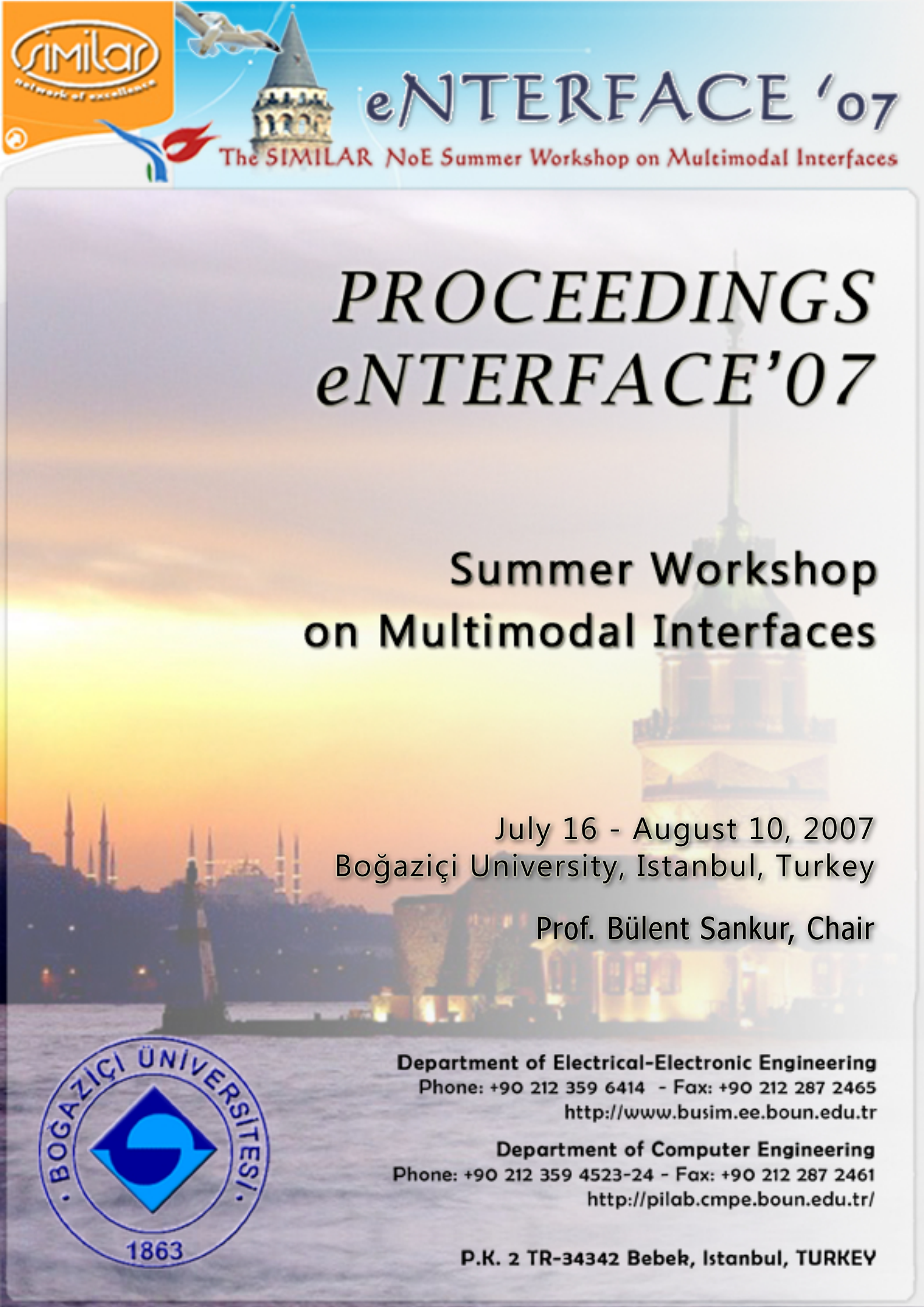


eNTERFACE '07

The SIMILAR NoE Summer Workshop on Multimodal Interfaces

PROCEEDINGS eNTERFACE'07

Summer Workshop on Multimodal Interfaces

July 16 - August 10, 2007
Boğaziçi University, Istanbul, Turkey

Prof. Bülent Sankur, Chair

Department of Electrical-Electronic Engineering
Phone: +90 212 359 6414 - Fax: +90 212 287 2465
http://www.busim.ee.boun.edu.tr

Department of Computer Engineering
Phone: +90 212 359 4523-24 - Fax: +90 212 287 2461
http://pilab.cmpe.boun.edu.tr/

P.K. 2 TR-34342 Bebek, Istanbul, TURKEY

Published by:
TELE (Communications and Remote Sensing Lab)
Université catholique de Louvain
Louvain-la-Neuve, Belgium
http://www.tele.ucl.ac.be

ISBN: 978-2-87463-105-4

Credits:
Cover design: Hamdi Dibeklioğlu (BUMM) & Colin Michel (UCL-TELE)
LaTeX editors: Christian Frisson (UCL-TELE) & Rémy Lehembre (UCL-TELE)
using LaTeX's 'confproc' class by Vincent Verfaille

Printed in Louvain-la-Neuve by Ciaco — December 2007

Foreword from Bülent Sankur, Conference Chair

eNTERFACE'07 took place in İstanbul, at the campus of the Boğaziçi University during July 15 - August 12 period. This one month long workshop was attended by 140 people overall, including short-term advisors and invited speakers. The workshop was organized around 12 projects, for which there was a six-month long definition phase. The projects were formed with the contributions not only the originators but also with the consequent inputs from the invited scientists. The projects were classified into four categories:

1. **Synthesis & Analysis Category :** Multimodal Speaker Identity Conversion; Audio-Driven Human Body Motion Analysis and Synthesis; Audiovisual Content Generation Controlled by Physiological Signals for Clinical and Artistic Application; Real-time and Accurate Musical Control of Expression in Singing Synthesis

2. **Multimodal Interaction Category :** Advanced Multimodal Interfaces for Flexible Communications; Event Recognition for Meaningful Human-Computer Interaction in a Smart Environment: Mobile-phone Based Gesture Recognition

3. **Biometric Interfaces Category :** 3D Face Recognition Performance under Adversarial Conditions; Benchmark for Multimodal Biometric Authentication

4. **Medical Interfaces Category :** Multi-Approach DT-MRI Data Analysis & Visualization Platform; IMAG Tool: A Software for Real-time Elastography and Tensorial Elastography; Multimodal Framework for the Communication of Disabled

The workshop hosted several international speakers from Stanford University, USA to University of Genoa, Italy, from University of Cambridge, UK to Berkeley, USA, from National Council of Research, Italy to Télécom Paris. The workshop had a very active social program that extended from historical city tour to a weekend in the Prince's Island, to increase social contacts and cement long-term interactions between diverse groups.
This workshop was enabled with the financial aid from three sources:

1. SIMILAR FP6 project

2. TUBITAK (Turkish National Research Council)

3. Boğaziçi University Foundation

The funds from SIMILAR Project, about 55% of the 40.000 Euro budget was spent to build the infrastructure, to provide scholarships to undergraduates and for social organization. The funds from TUBITAK were spent entirely for invited speakers and for additional aid to selected eNTERFACE participants. The aid from Boğaziçi University Foundation enabled us to appropriate and to use 1000 m^2 of the Engineering complex with all its associated services. The participants were lodged in the comfortable suites of Superdorm, a modern luxury dormitory facility of Boğaziçi University.
The polls we conducted at various levels, from experts to the M.Sc students indicated unanimously that workshop has provided them with invaluable experiences in project cooperation and teamwork and has been instrumental in increasing their research motivation. Almost all participants expressed a desire to participate themselves to future workshop and/or to exhort their younger cohorts in this direction.

Bülent Sankur
Conference Chair

eNTERFACE Workshops Motivations Reflexions

by Bülent Sankur, BUMM - Boğaziçi University Multimedia Group

Boğaziçi University Multimedia (BUMM) students and faculty have been participating to the previous eNTERFACE activities. Their first-hand impressions show that the eNTERFACE workshops have been instrumental in increasing the research motivation of our students and have provided them with invaluable experiences in project cooperation and teamwork. We have been therefore very enthusiastic about SIMILAR project in general and about eNTERFACE'07 in particular, and we strongly believe in the benefits of such workshops. At eNTERFACEs, seeds of long lasting partnerships can be planted.

We were involved in the planning stage of the third workshop in the triad of eNTERFACEs in Mons, Dubrovnik and Istanbul. Among our endeavors we had planned to extend the invitation for participation beyond SIMILAR to all other FP6 project groups with concordant themes. We have realized projects on novel themes, such as man-machine interfaces in finance, bio-informatics and security.

by Igor S. Pandžić, Organizer of eNTERFACE'06

eNTERFACE'06 took place in Dubrovnik, Croatia, at the peak of summer tourist season. With the venue 500 m from the historic city center and two minutes walk from the beach, the big question we were asked was: "how will you get people to work there?". In fact, the 65 participating researchers and students were so motivated that such questions quickly became pointless and days passed in a busy and creative working atmosphere in the two classrooms filled with computers and electronics. Yet no one could accuse the eNTERFACE'06 participants of being mere geeks - in the evenings and weekends a lively and versatile social scene developed. With the organized social program serving as a basis, participants have organized numerous activities: sports tournaments, excursions, cultural events as well as a spicy nightlife featuring beach parties. All together, eNTERFACE'06 turned out to be a well-rounded mix of research, learning and fun. With such basis, I hope and believe that some of the seeds planted in Dubrovnik will grow into full-blown research collaborations of the future. As the memories of the more mundane organizational chores fade from memory, what is left is pure satisfaction of being part of the unique series of events that is eNTERFACE.

eNTERFACE Workshops Motivations Reflexions

by Thierry Dutoit, Initiator of the eNTERFACE concept and organizer of eNTERFACE'05

I take this opportunity to mention how the idea of eNTERFACE came to life in mid 2003, when the core committee of SIMILAR (the European Network of Excellence on Multimodal Interfaces), was busy establishing its list of workpackages. It is basically a three acts play.

Act 1. I have been a researcher since 1987. It is therefore becoming hard to navigate in the "conferences" directory of my laptop. Small to big, short to long, close to far away, I think I have tried them all. One of them, however, will last forever in my memory as the most productive meeting I have ever attended. It was a summer school on Prosody, in July 1993, organized by the ELSNET (already a scientific network). I spent two weeks there, at UCL London, attending lectures and, more importantly, taking labs with my fellow PhD students from all over the world. I must say this is simply the place where I met most of my friends for life!

Act 2. In 1996, I had the opportunity to work for AT&T at Bell Labs for 1.5 years, in the TTS group. This was set about 2 years after I finished my PhD (i.e., 2 years after I had signed with Kluwer for writing the "3-months-of-work" book in TTS I took 3 years to complete; I finished it at AT&T...). It was clear to me that I was then about to meet the greatest gurus in speech processing (yet I had underestimated the number of famous people who were working in this lab), and that I would work with the best maintained software archive in the world (you snap your finger, and you get what you were looking for; this, I had overestimated..). I did meet all these people, and the atmosphere was such that meeting each other was really easy, but I also realized something I had never imagined: research in the US is a huge network thing. Network in terms of "You seldom work on your own on a problem", but also in terms of "Be with the network; the network will take care of you". In other words, research is very much advertised and supported by your employer, by all sorts of professional organizations, and even among the general public. Hence its dynamics.

Act 3. I was aware of the successful DARPA workshops on speech recognition organized yearly by Prof. Fred Jelinek at Johns Hopkins University. Funded by the Defense Agency (which implies a strong financial support), these workshops have progressively become a "must" for researchers in the field, who come from all around the world to participate. One of our researchers took part to it, and my ex-colleague Hervé Bourlard, now the Director of IDIAP in Switzerland, was an active member of it. I have always envied this event, and dreamt of finding money to organize something SIMILAR. Thanks to EU financing, and with special care from SIMILAR, this dream has come true. With its 55 researchers from 15 countries all around the world working together for four weeks on seven pre-selected projects, eNTERFACE'05 in Mons, Belgium, has been a great success. eNTERFACE'06, gathering in Dubrovnik, Croatia, 63 researchers from 12 countries organized in 9 teams, has been even more successful. eNTERFACE'06 in Dubrovnik, Croatia, gathering 63 researchers from 12 countries organized in 9 teams, has been even more successful. Long life to eNTERFACE workshops!

Organizers at Boğaziçi University

Medical Interfaces

Burak Acar
Dept. of Electrical-Electronic Engineering
Email: acarbu@boun.edu.tr

Brain and Face Interfaces

Bülent Sankur
Dept. of Electrical-Electronic Engineering
Email: bulent.sankur@boun.edu.tr

Gestural Interfaces

Lale Akarun
Department of Computer Engineering
Email: akarun@boun.edu.tr

Speech Interfaces

Levent Arslan
Dept. of Electrical-Electronic Engineering
Email: arslanle@boun.edu.tr

Social Activities

Oya Aran
Department of Computer Engineering
Email: aranoya@boun.edu.tr

Speech Interfaces

Murat Saraçlar
Dept. of Electrical-Electronic Engineering
Email: murat.saraclar@boun.edu.tr

Accommodation

Oya Çeliktutan
Dept. of Electrical-Electronic Engineering
Email: oyaxceliktutan@yahoo.com

Technical Support

Cağlayan Dicle
Dept. of Electrical-Electronic Engineering
Email: cdicle@gmail.com

Webmaster

İsmail Arı
Department of Computer Engineering
Email: ismailar@boun.edu.tr

Technical Support

Cem Keskin
Department of Computer Engineering
Email: keskinc@cmpe.boun.edu.tr

Publicity

Hamdi Dibeklioğlu
Department of Computer Engineering
Email: hamdimail@gmail.com

Projects

Arman Savran
Dept. of Electrical-Electronic Engineering
Email: arman.savran@boun.edu.tr

Workshop Reports

Team 1 (Leaders: Burak Acar, Roland Bammer, Marcos Martin-Fernandez, Ali Vahit Sahiner, Suzan Üskudarli)

Team 2 (Leader: Ana C. Andrés del Valle)

Team 3 (Leaders: Lale Akarun, Murat Saraçlar, Dimitrios Tzovaras)

Team 4 (Leaders: Thierry Dutoit, Ferran Marqués, Igor S. Panžić, Murat Saraçlar, Yannis Stylianou)

Team 5 (Leaders: Lale Akarun, Tanju Erdem, Engin Erzin, Ferda Ofli, A. Murat Tekalp, Yücel Yemez)

Team 6 (Leaders: Lale Akarun, Ramon Morros, Albert Ali Salah, Ben Schouten)

Team 7 (Leaders: Lale Akarun, Bülent Sankur, Tevfik Metin Sezgin, İlkay Ulusoy)

DTI APPLICATION WITH HAPTIC INTERFACES

Murat Aksoy [1], *Neslehan Avcu* [2], *Susana Merino-Caviedes* [3], *Engin Deniz Diktaş* [4], *Miguel Ángel Martín-Fernández* [3], *Sıla Girgin* [4], *Ioannis Marras* [5], *Emma Muñoz-Moreno* [3], *Erkin Tekeli* [4], *Burak Acar* [4], *Roland Bammer* [1], *Marcos Martin-Fernandez* [3], *Ali Vahit Sahiner* [4], *Suzan Üskudarli* [4]

[1] Stanford University, School of Medicine, Dept. of Radiology, LUCAS MRS/I Center, USA
[2] Dokuz Eylül University, EE Dept., Izmir, Turkey
[3] University of Valladolid, Valladolid, Spain
[4] VAVlab, Boğaziçi University, İstanbul, Turkey
[5] AIIA lab, Aristotle University of Thessaloniki, Greece

ABSTRACT

Diffusion Tensor Magnetic Resonance Imaging (DTI) is a rather new technique that allows in vivo imaging of the brain nervous structure. The DTI data is a 3D second-order positive semi-definite tensor field. DTI analysis and visualization is a challenging field due to the lack of standards and high-dimensionality of the data. This project focused on 1) implementation of a base DTI tool compatible with the tensor field standard (STAC) as proposed by the SIMILAR NoE WP10 group, 2) developing haptic interfaces for effective visualization and investigation of aforementioned 3D tensor fields. Both goals have been achieved, yet their integration could not be completed with the Enterface2007 workshop. Howeever, the know-how built during the workshop and the codes generated are invaluable resources for developing the final application. VAVlab (Boğaziçi University, EE Dept., İstanbul, Turkey) is currently working on the integration of these components into a complete application.

KEYWORDS

DTI – Tensor Fields – Tractography – Tensor Standards – Haptic Interfaces – 3D Interactive Interfaces for Radiology.

1. INTRODUCTION

Diffusion Tensor Magnetic Resonance Imaging (DTI) is a rather new technique that enables researchers and physicians to image fiberous structures, such as nerves in brain, in vivo. Previously, such investigations could only be done in vitro, with the unavoidable consequences of dissection of the brain. Furthermore, in vivo imaging of fiber networks in brain is likely to facilitate the early diagnosis, surgery planning and/or follow-up screening [1, 2, 3].

DTI is based on measuring the MR signal attenuation due to random walk (diffusion) of water molecules in restricted media, namely inside the myelin coated nerve fibers, in response to a special diffusion weighting gradient magnetic fields. Thus measured 3D data is called the Diffusion Weighted Imaging (DWI). Several DWI sets can be combined to compute the DTI datasets, which is a second order, symmetric positive semi-definite tensor field in 3D. Each tensor represents the local physical diffusion process upto a second order approximation.

The challenge in DTI is not only in data acquisition, which is an active research area, but also in post-processing and visualizing tensor fields and user interfaces for effective communication of the information content.

The tensor data is a high dimensional volumetric data which is significantly different then volumetric scalar fields. For conventional 3D scalar fields, one has to be careful to set the relation between the data grid and the world (patient) reference frame correct. However, for a tensor field, one has to be careful in managing the relation between the DTI reference frame (the reference frame with respect to which the tensors are **defined**) and the world reference frame as well. Furthernore, the application of simple geometric transformations to a tensor field can be tricky as one has to transform the data itself together with the data grid. Such difficulties lead to significant problems in the absence of a tensor standard. Consequently, the SIMILAR NoE WorkPackage10 had initiated an effort to define a general tensor standard for use with DTI data as well in other fields. The standard proposed is called STAC (Similar Tensor Array Core) and is designed to incorporate the essential components to define a tensor field without ambiguity [4].

Fiber tractography is a well-known and widely used approach for visualization and analysis of DTI data. As will be explained in more detail below, it consists of numerical integration of the principal diffusion direction (PDD) as represented by the tensor data. The output of fiber tractography is a set of 3D streamlines representing fiber bundles in brain. The required interface for fiber tractography should enable the users to answer questions like "*Where do the fibers passing through region A end?*", "*Are there any fibers connecting region A and region B?*", "*What is the shape of the fiber bundles around this lesion?*", etc. All these questions require an intuitive, easy to navigate user interface that allows the users to see volumetric cross-sections they choose and the to shape and position region-of-interests (ROIs).

DTInteract is a proof-of-concept application developed during eNTERFACE 2007, İstanbul , Turkey. Its major goals are *i)* to incorporate the STAC into the volumetric analysis and visualization framework (VAVframe) being developed at Boğaziçi University, VAVlab, İstanbul, Turkey, and *ii)* to develop and test the use of haptic interfaces for DTI visualization and analysis.

2. DTI DATA

2.1. Theory of DTI Tensor Calculation

It is of utmost importance to understand what the DT-MRI signal represents in order to develop adequate analysis and visualization methods. DT-MRI measures the average signal attenuation within a small subvolume (i.e. a voxel) due to water molecules spinning out-of-phase.

The basis of MRI is to perturb the water molecules (the dipoles) that were aligned in a constant magnetic field (B_0 approx 1-7 Tesla) and let them re-orient themselves with B_0 during which the dipoles rotate around B_0 according to the Bloch's Equation. This rotation causes a temporal change in total magnetic field which induces a time-varying current at the receiving coils of the MR scanner. The time it takes for the dipoles to fully relax depends on the environment (i.e. the tissue). Thus, the amount of current induced is proportional to the number of dipoles (i.e water concentration) and the tissue type. These current measurements are transformed to monochromatic images in which each pixel value is also a function of water concentration and tissue type

In DT-MRI, extra spatially varying magnetic fields, the so called Diffusion Weighting Gradients (DWG), G, are applied together with B_0. Due to this added G field, the water molecules under continuous brownian motion experience different total magnetic field at different locations. This causes them to rotate at different frequencies, i.e. to be out-of-phase. The effect of out-of-phase dipoles on the induced current is attenuation of the MR signal. So, the amount of attenuation in the received signal (equivalently in the diffusion weighted MR images) is a function of the brownian motion (i.e diffusion) of water molecules and the applied G field.

This phenomenon is described by the superposition of the Bloch's Equation and the Diffusion Equation whose solution is

$$M_k = M_0 \exp\left(-\sum_{i=x,y,z}\sum_{j=x,y,z} \mathbf{b}_k^{i,j}\mathbf{D}_{i,j}\right) \quad (1)$$

Here, M_k is the pixel intensity in the case when diffusion weighting is applied using the k_{th} diffusion encoding direction, M_0 is the pixel intensity when there is no diffusion weighting, $\mathbf{D}$ is the 3×3 diffusion tensor, $\mathbf{b}_k$ is a 3×3 symmetric matrix that is determined by the applied diffusion weighting direction and strength and i,j are matrix indices. The six independent components of the symmetric diffusion tensor $\mathbf{D}$ can be computed using at least 6 linearly independent equations of this form where

$$\mathbf{b}_k = b_{\text{nom}}\mathbf{r}_k\mathbf{r}_k^T \quad (2)$$

[1]

b_{nom} is a user-determined nominal b-value which is used to adjust the amount of diffusion weighting. This can be obtained from the header of the DICOM files containing the diffusion weighted images (DICOM tag (0019,10b0)). An example set of diffusion encoding directions ($\mathbf{r}_k$) is

$$\begin{array}{llll} & & \mathbf{r}_0 = & [0\ 0\ 0]^T \\ \mathbf{r}_1 = & [0.707\ 0.707\ 0]^T & \mathbf{r}_2 = & [0\ 0.707\ 0.707]^T \\ \mathbf{r}_3 = & [0.707\ 0\ 0.707]^T & \mathbf{r}_4 = & [-0.707\ 0.707\ 0]^T \\ \mathbf{r}_5 = & [0\ -0.707\ 0.707]^T & \mathbf{r}_6 = & [0.707\ 0\ -0.707]^T \end{array}$$

The diffusion weighted MR images are stored in `dwepi.*.grads` files.

Taking the natural logarithm and using linear algebra, for each pixel, a linear system can be obtained from multiple diffusion weighted measurements:

$$\underbrace{\begin{pmatrix} \ln M_1 \\ \ln M_2 \\ \cdots \\ \ln M_n \end{pmatrix}}_{\mathbf{S}} = \underbrace{\begin{pmatrix} 1 & -\mathbf{b}_1^{xx} & -\mathbf{b}_1^{yy} & -\mathbf{b}_1^{zz} & -2\mathbf{b}_1^{xy} & -2\mathbf{b}_1^{xz} & -2\mathbf{b}_1^{yz} \\ 1 & -\mathbf{b}_2^{xx} & -\mathbf{b}_2^{yy} & -\mathbf{b}_2^{zz} & -2\mathbf{b}_2^{xy} & -2\mathbf{b}_2^{xz} & -2\mathbf{b}_2^{yz} \\ \cdots & & & & & & \\ 1 & -\mathbf{b}_n^{xx} & -\mathbf{b}_n^{yy} & -\mathbf{b}_n^{zz} & -2\mathbf{b}_n^{xy} & -2\mathbf{b}_n^{xz} & -2\mathbf{b}_n^{yz} \end{pmatrix}}_{\mathbf{M}} \underbrace{\begin{pmatrix} \ln M_0 \\ \mathbf{D}_{xx} \\ \mathbf{D}_{yy} \\ \mathbf{D}_{zz} \\ \mathbf{D}_{xy} \\ \mathbf{D}_{xz} \\ \mathbf{D}_{yz} \end{pmatrix}}_{\mathbf{d}}$$

[1] $\mathbf{b}_k^{i,j}$ is the (i,j) component of b-matrix $\mathbf{b}_k$

Then, the unknown $\mathbf{d}$ vector can be found using pseudo-inverse:

$$\mathbf{d} = (\mathbf{M}^T\mathbf{M})^{-1}\mathbf{M}^T\mathbf{S} \quad (3)$$

2.2. STAC: Similar Tensor Array Core

Diffusion Tensor (DT) images are a relatively new kind of images, that requires special features for their storage. Nowadays, there is not a standard format for this sort of data, and therefore specific writers and readers must be implemented to manage data acquired by different equipments. In order to deal with this, a standard has been proposed in [4] that is valid not only for diffusion tensor images, but for general tensor fields. It is named the Similar Tensor Array Core (STAC).

The standard consists of two parts: the general tensor standard array core (STAC) that is the basic type of storage for tensor fields, and the extensions that provides more specific functionalities for specific tensor types. By now, we have implemented the core header that is the main part of the tensor standard. The definition and implementation of an appropriate extension for DT-MRI will be considered as a future task.

The standard core header contains the following fields:

- `array_dimensionality`: It is the number of dimensions of the field.
- `array_size`: It is the size of the field.
- `array_index_names`: It contains the name of the index of the field, therefore is an array n values, where n is the `array_dimensionality`. It is an optional field that can be useful for processing issues.
- `array_tensor_metric`: It is a metric tensor that describes the metric in the array space. Thus, diagonal elements represents the scaling factor in each dimension. It is a $n \times n$ metric tensor, where n is the `array_dimensionality`. It is not a mandatory field, by default is set to the identity.
- `tensor_order`: It is a scalar number describing the order of the tensor.
- `tensor_index_types`: Each tensor index can be contravariant or convariant. In this field, an array of m elements, where m is the `tensor_order`, describes what kind is each of the index.
- `tensor_index_names`: The name of the index of the tensor elements. It is an optional field, useful for processing tasks.
- `tensor_name`: A name can be given to the tensor. It is an optional field.
- `description`: Description of the tensor field can be included. It is an optional field.

We have implemented both reader and writer for tensor standard data storage. Both of them are included as methods in the VavDTI class. Two files are required to store the data: the first, that has extension `.stach`, contains the header, whereas the second contains the raw data and it has `.stacd` extension. Both files have the same name, they only differs in the extension.

2.2.1. The Reader

The `readerSTAC` method receives as an input the file name (without extension), and it assigns the appropriate values to the VavDTI attributes. The `array_size` is assigned to the `dimArray` variable, the `spacingArray` elements are set as the values of the diagonal of the metric tensor, and a VavTensor is built reading the data file. A value of '0' is returned by the function if no error happens.

2.2.2. The Writer

The `writerSTAC` method has as input the file name (without extensions) where the VavTensor data will be stored. First of all, the header file (`*.stach`) is written:

- `array_dimensionality`: Since VavDTI class is implemented for 3D arrays, it is set to 3.
- `array_size`: The size of the array is obtained from the `dimArray` value.
- `array_index_names`: This is an optional field, if no value is given for the array names is set to '`None`' by default.
- `array_tensor_metric`: The metric information that provides the VavDTI class is the `spacingArray` variable. Therefore, the tensor metric is represented by a 3x3 diagonal matrix. Each diagonal element represents the spacing in each direction.
- `tensor_order`: The diffusion tensors are second order tensors, so this field is set to 2.
- `tensor_index_types`: The diffusion tensor is a contravariant tensor, so both index are of contravariant type.
- `tensor_index_names`: This is an optional field, if no value is given for the array names is set to '`None`' by default.
- `tensor_name`: The tensor is named DT (Diffusion Tensor)
- `description`: We include a brief description of the tensor field in this field.

After writing the header file, the data are written in the `*.stacd` file in raw format.

2.3. VAVframe Data Structure

In VavFrame, three data classes are generated for three main data types. These are: VavScalar, VavVector and VavTensor for scalar, vectoral and tensor fields. And VavData class is generated as an abstract data class for these three data types of VavFrame. All data classes are inherited from this class.

VavData class includes 3-dimensional vectors for dimension, space and origin information for each direction x,y,z. And two vtk objects are member of it. They are used to store data fields. These variables are common for three data classes, but the size of data components are not the same for all of them. There are one component for VavScalar, three components for VavVector, nine components for VavTensor classes for each voxel in the field.

Data are stored in VtkImageData type object. It is one dimensional array, but data is in three dimension. So, we hold data in order of x,y,z. And, the common reference frame with respect to the anatomy is as follows:

- 1st dimension (X): Left → Right
- 2st dimension (Y): Posterior → Anterior
- 3st dimension (Z): Inferior → Superior

VavDTI class is created to hold DTI specific data and manage DTI operations. It includes dimension, space and origin information; diffusion tensor field, eigenvectors and eigenvalues of this tensor field, mean diffusivity, B_0 images, fractional anisotropy, lattice index, coherence index, diffusion weighted images, bmatrices and a header information. These variables are instances of VavScalar, VavVector and VavTensor classes according to their types. There are reader functions for diffusion tensor data from different formats in VavDTI class, such as Stanford DTI data and PISTE DTI data. Also it includes a conversion function from DWI to DTI and an eigen decomposition function for DTI data.

2.4. Stanford DTI Data

There are two types of DTI data format that is being used at Stanford University. The first one is the diffusion weighted image DICOM (DWI-DICOM) format, and the second one is the tensor raw data.

The DWI-DICOM data simply consists of the diffusion weighted images belonging to each slice, diffusion weighted direction and scan repetition. All these files are in DICOM format, which is currently the standard file format used in the area of medical imaging [2]. These DICOM files are ordered consecutively and the file format is `Ixxxx.dcm`, where `xxxx` is a number which starts with 0001 and goes up to the total number of images. The DICOM file ordering, from the outermost to the innermost loop, is as follows: scan repetition → diffusion weighting direction → slice (Figure 1).

The second file format consists of pre-calculated raw tensor data [3]. This consists of two sets of files: The first set is the `TensorElements.float.xxx` files, where `xxx` denotes the slice number. These contain the tensor element in the order: $\mathbf{D}_{xx}$, $\mathbf{D}_{yy}$, $\mathbf{D}_{zz}$, $\mathbf{D}_{xy}$, $\mathbf{D}_{xz}$, $\mathbf{D}_{yz}$. In this case, the pixels are the innermost loop, i.e., for an image resolution of $n \times n$, the n^2 $\mathbf{D}_{xx}$ elements are written first, followed by n^2 $\mathbf{D}_{yy}$ elements, and so on.

The second set is the `Tensor.float.xxx` files. Here, the following $n \times n$ images are written to the file consecutively: Mean diffusivity (units x10^6 mm^2/s), maximum eigenvalue, medium eigenvalue, minimum eigenvalue, x component of the maximum eigenvector (values between -1 and 1), y component of the maximum eigenvector, z component of the maximum eigenvector, Fractional Anisotropy (FA, multiplied by 1000) and b=0 image (arbitrary units).

2.5. PISTE DTI Data

DT-MRI based tractography techniques have been proposed to propagate fiber trajectories in diffusion tensor fields. For the accuracy and acceptability of these tracking methods, there must be a general data set to evaluate the algorithms and their results. PISTE, developed on behalf of the ISMRM Diffusion/Perfusion Study Group, following an initiate at the 2003 Toronto meeting, is intended as a resource for researchers interested in Diffusion Tensor Magnetic Resonance Imaging and Tractography. The aim is to provide a general database of simulated common fiber tract trajectories that can be used for testing, validating and comparing various tractography algorithms.

To evaluate the performance of tractography algorithms and analyze the influence of several factors on tracking, PISTE includes several datasets differing with respect to

- signal-to-noise ratio,
- tensor field anisotropy,
- fiber geometry,
- interpolation and
- step-size.

For each setup there are 5 datasets provided:

- A T2-weighted image;
- The 6 elements of the diffusion tensor;

[2] http://medical.nema.org/
[3] http://rsl.stanford.edu/moseley/tensorcalc/

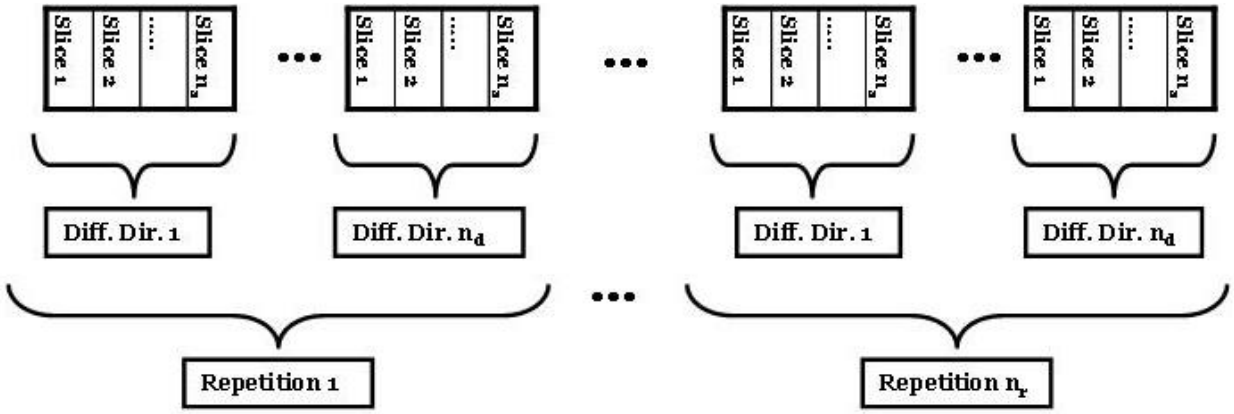


Figure 1: *Structure of DICOM files for DWI data. The data is acquired n_r times for each one of the n_d diffusion gradient directions. n_d must be at least 6 to recover the diffusion gradient and n_r is usually selected to be 4. The repeated data acquisition is performed for noise suppression.*

- The 3 eigenvalues and eigenvectors;
- An image containing the tract initiation point or region of interest;
- A calibration image showing the positive X,Y and Z directions.

3. DTINTERACT

3.1. System Components

DTInteract application is a proof-of-concept C++ application built on the VAVframe infrastructure. It consists of three main components, the data standardization, fiber tractography, haptic interfaces for tractography. The first component has been explained in Section 2. Its major concern is to load 3D tensor fields (and related data) into a common workplace, which is chosen to be the VAVframe environment. We have currently dealt with two datasets, the DTI data provided by Stanford University, LUCAS MRS/I Center, CA[4] and the PISTE phantom DTI dataset. The second component is composed of 4^{th} order Runge-Kutta based fiber tractography (PDD tracking). This is the most well-known and widely used method for tractography. Despite its drawbacks, we have decided to use it in this proof-of-concept application. The final component is the haptic interface. It is used to position orthogonal slices (axial, sagittal, coronal) in 3D, as well as to shape and locate 3D ROIs (region-of-interest). These will be discussed in more detail in the following.

3.2. Base Application

The programming language used for the interface development is C++, given the availability of libraries of classes written in this language for visualization, haptics and user interface development. In addition, C++ performance, although not as efficient as C, is comparable to it.

Three C++ libraries have been used: VTK (Visualization Toolkit) [5], CHAI-3D [6] and Qt [7], which will be described below. VTK and CHAI-3D are open source and free. The open source version of Qt was employed for this project.

VTK is a widely used C++ class library for visualization tasks and image processing, and it provides high level classes for computer graphics. It makes use of the underlying low level graphics library, like for example OpenGL. VTK provides tools for creating complex visualization scenes, and is multiplatform.

The application uses a Sensable Phantom Omni as haptic input device for the interaction with the user. In order to control it, the library CHAI-3D is employed, which provides classes for the management of a wide variety of haptic devices and real-time simulations.

Qt is a C++ object library with a rich variety of resources for user interfaces development. It also provides a signal-slot paradigm for communication between objects. Signals and slots can be connected, so that when a signal is sent, the slots it is attached to are executed. This mechanism provides the fundamental communication system of our interface.

The main classes that intervene in the application are VavDTI, TestApplication, VavDTIViewer and the classes involved in haptic management, which will be explained here below.

3.2.1. VavDTI

This class is able to load and save DTI data in Stanford and STAC formats, extract some of the tensor characteristics. The methods for computing tractographies and visualizing the different DTI data are also included in this class.

3.2.2. TestApplication

This class provides the interface management and interconnectivity of the different parts of the application. It gives functionality to the interface and manages the methods provided by VavDTI. It also provides functions for connecting the haptic classes for user interaction.

3.2.3. VavDTIViewer

This class inherits from the class vav3DViewer, which has as a consequence that it encapsulates both a QVTKWidget and a vtkRenderer object. This class is used to provide a display for the visualization of the DTI components and provide haptic interaction to the application, along with the class TestApplication.

The application uses orthogonal slices of scalar voumes for the fractional anisotropy, mean diffusivity and B0 coefficients. The major eigenvector field is visualized using lines at each point of the slices, and the tracts are visualized as tubes (Figure 2).

[4] Courtesy of Prof. Roland Bammer, PhD.

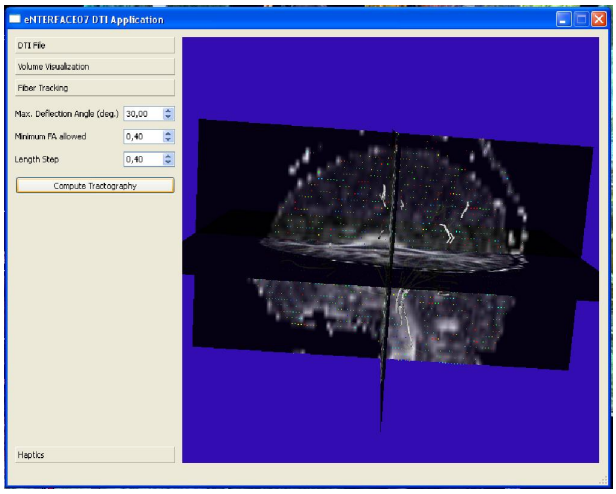


Figure 2: *GUI for the base DTI Application.*

3.3. DTI Tractography

Since water diffusion in brain is constrained by the myelin covering of the axons, it is possible to assess the fiber structure by tracking the direction given by the DT major eigenvector. This eigenvector represents the maximum diffusion direction, and for this reason, it is supposed to be tangent to the fiber bundles at each voxel. In this way, the integration of the vector filed will provide the diffusion path. Some numerical methods can be used in order to integrate the vector fields, such as Euler or Runge-Kutta integration methods. In this implementation, we have considered a 4^{th} order Runge-Kutta integration method, which is detailed just below.

3.3.1. 4^{th} order Runge-Kutta method

The integration method must start from an initial or seed point r_o, from which the diffusion path is defined. The points that belongs to the path are iteratively computed according to the next equation [8]:

$$\mathbf{r}_{n+1} = \mathbf{r}_n + h\mathbf{V}_{n+1} \tag{4}$$

h stands for the step length parameter, that is a parameter set by the user in our implementation. $\mathbf{V}_n$ is a vector computed by the next equation that is actually the 4^{th} order Runge-Kutta method:

$$\mathbf{V}_{n+1} = \frac{1}{6}(\mathbf{k}_1 + 2\mathbf{k}_2 + 2\mathbf{k}_3 + \mathbf{k}_4) \tag{5}$$

Where $\mathbf{k}_1, \mathbf{k}_2, \mathbf{k}_3, \mathbf{k}_4$ are defined as a function of a continuous **V**ector field $\mathbf{E}(\mathbf{r})$:

$$\begin{aligned}
\mathbf{k}_1 &= \frac{\mathbf{V}_n\mathbf{E}(\mathbf{r}_n)}{\mathbf{V}_n\mathbf{E}(\mathbf{r}_n)} \cdot \mathbf{E}(\mathbf{r}_n) \\
\mathbf{k}_2 &= \frac{\mathbf{V}_n\mathbf{E}(\mathbf{r}_n + \frac{h}{2}\cdot\mathbf{k}_1)}{\mathbf{V}_n\mathbf{E}(\mathbf{r}_n + \frac{h}{2}\cdot\mathbf{k}_1)} \cdot \mathbf{E}(\mathbf{r}_n + \frac{h}{2}\cdot\mathbf{k}_1) \\
\mathbf{k}_3 &= \frac{\mathbf{V}_n\mathbf{E}(\mathbf{r}_n + \frac{h}{2}\cdot\mathbf{k}_2)}{\mathbf{V}_n\mathbf{E}(\mathbf{r}_n + \frac{h}{2}\cdot\mathbf{k}_2)} \cdot \mathbf{E}(\mathbf{r}_n + \frac{h}{2}\cdot\mathbf{k}_2) \\
\mathbf{k}_4 &= \frac{\mathbf{V}_n\mathbf{E}(\mathbf{r}_n + h\cdot\mathbf{k}_3)}{\mathbf{V}_n\mathbf{E}(\mathbf{r}_n + h\cdot\mathbf{k}_3)} \cdot \mathbf{E}(\mathbf{r}_n + h\cdot\mathbf{k}_3)
\end{aligned}$$

This definition of the coefficients ensures that the angle between consecutive vectors is smaller than 90°.

3.3.2. Implementation

The Runge-Kutta method is implemented by the RungeKuttaIntegration method, that has three input parameters, the curvature threshold, the step length and the anisotropy threshold, that are more detailed below. It returns 0 value if there is not error. The resulting streamlines are stored in the streamlineList attribute of the VavDTI class that is a vtkPolyData object. Next, we comment some implementation details.

- `Seed Points`: In our implementation, we consider as seed point every voxel whose anisotropy is higher than a given threshold value set by the user. Since fibers belongs to the white matter, that is the area where anisotropy is higher, it only makes sense to compute streamlines in areas of high anisotropy. Moreover, the voxels that belongs to a previously computed streamline are not used as seed points.
- `Initial conditions`: For each seed point r_o, a streamline is computed following the Runge-Kutta integration method. The initial value of $\mathbf{V}_n$ is $\mathbf{V}_o = \mathbf{e}(\mathbf{r}_o)$, that is, the eigenvector in the initial point.
- `Stop criteria`: The streamline computation from a given seed point stops when one of the following conditions is achieved:
  - The line arrives to a region of lower anisotropy, that is, the fractional anisotropy is lower than the threshold set by the user.
  - The angle between successive streamlines segments is higher than a threshold value set by the user. This criteria is based on a priori anatomical knowledge: the fiber curvature is small.
- `Interpolation`: Let's notice that the Runge-Kutta algorithm considers a continuous vector field, but the eigenvector field is defined in a discrete grid. For this reason, interpolation of the vector field is required. In this implementation, we have computed a linear interpolation among the nearest eigenvectors.
- `Tracking Parameters`: Three parameters are set by the user to carry out the tractography algorithm:
  - Fractional anisotropy threshold: The voxels whose anisotropy is higher than this threshold will be considered as white matter, and the fibers will be computed from them. By default is set to 0.15.
  - Curvature threshold: It represents the maximum angle of curvature allowed for a streamline. If the angle is higher, the streamline tracking is stopped. The default value is $\pi/6$ rad.
  - It is the value of the step length h in the Runge-Kutta algorithm. By default is set to 0.9, that is one half of the voxel size in the smallest dimension.

3.4. Haptic Interface

Haptic devices are capable of enhancing the space comprehension in a user interaction with the virtual environment, by providing many Degrees-of-Freedom (DoF) for the input as well as force-feedback, thus achieving an increased realism for the perception of the scene contents. The primary goal of haptic introduction into the DTI application is to allow the user to interact with both the volumetric and DTI data in a better way.

3.4.1. Application Environment

The haptic device that was used in this project was the Phantom Omni from SensAble. Phantom Omni provides 6-DoF input (position + orientation) and 3-DoF of force-feedback. In order to control the haptic device, the CHAI-3D library was used. In addition, VTK and Qt were used for visualization and GUI, respectively.

Since VTK and CHAI-3D consume each much time for their own main-loops, Qt timers were used to provide the necessary time distribution for both loops. The correlation between the two loops is accomplished by Qt's signal-slot mechanism. Whenever an event occurs in the haptic environment, such as creation or modification of a new ROI, a Qt-signal is sent from the haptic loop, which is later caught by a QVTK widget for further processing. The integration of Qt, VTK and CHAI-3D has been the primal concern in this effort.

3.4.2. Haptic Slicer

In the haptic environment the first problem to be solved is the representation of the three perpendicular and discriminated slices, as they exist in VTK environment. As a result, the user will be able to move those slices in the local coordinate system according to this specific perpendicular schema. The basic model for these slices is stored as an external WRL file. The haptic environment loads the above file during the initialization through a Qt-signal and afterwards reads the position of the slices from the VTK. This position is in the form of a vector $\vec{P}_{VTK} = (p_x, p_y, p_z)$ and the values p_x, p_y, p_z correspond to the position of the plane XY, YZ and XZ, respectively. As mentioned before, the user will be able to rotate the entire scene using the haptic cursor. This means that if the user has already changed the position of the slices in VTK environment before the initialization of the haptic environment or has rotated the entire scene using the haptic device, the application should be able to move the slices in the appropriate motion direction. One should wonder here how can the haptic understand the direction in which should it move the slices. The haptic proxy, during collision detection, returns the 3D point in the local coordinate system. This information is not adequate in order to find the direction of motion for the slice with which the proxy has collided. One solution is to analyze the vector of the haptic cursor in order to find the direction according to which the user "pushes" the current slice. On could ask what happens if we have only collision detection and almost zero vector force in this point? A better solution for this problem would be to replace each one of the prototype slices with two discriminate slices on either side that are extremely close to the prototype slice's position. Thus, if for example the position of the XY plane is p_x then the two slices will have the positions $p_x - 0.005$ and $p_x + 0.005$, respectively. The small distance between the slices makes this trick invisible to the user and enables the application to work properly without the need to store any immediate information regarding the rotations of the scene. It also doesn't depend on the rotations the user has performed through the VTK environment. In this case, the resulting wrl file consists of six slices. Each slice has a specific name p.e. for the XY plane we have two slices: XY_Plane_Back and XY_Plane_Front. The two slices can be moved along the vector $\vec{V} = (0, 0, 1)$ by adding a positive number for the first one while adding a negative number for the second one, to their current positions. In case the user has loaded the haptic environment but has rotated the scene in VTK, a Qt-signal is enough for the calculation of the new positions in order for the application to obtain balance between the two autonomous environments. The range of this position vector depends on the size of the volumetric dataset, which has been loaded in the VTK environment. To overcome this limitation, the wrl file includes slices of size 2, meaning that each slice could be positioned between $[-1, 1]$. Thus, the vector $\vec{P}_{\mathrm{VTK}}$ is transformed in the haptic environment to $\vec{P}_{\mathrm{haptic}}$ as follows:

$$\vec{P}_{\mathrm{haptic}} = (\frac{p_x}{N_z}, \frac{p_y}{N_x}, \frac{p_z}{N_y}) \tag{6}$$

where N_z, N_x, N_y are the z,x,y dimensions of the plane perpendicular correspondingly to each slice. When the user rotates the scene in the local coordinate system of the slices, the coordinates remain unaltered, so the vectors along which the slices that could be moved do not changed.

The haptic slicer application consists of two display windows: the first window is an OpenGL window that contains the haptic scene whereas the second one is a VTK-window that contains the main environment for the DTI application.

In order to control the VTK world, several objects are placed into the haptic scene. Three orthogonal planes in the haptic world represent the slices in the VTK-world. Pushing these slices without pressing a key rotates the scene, while pushing them while pressing the key, translates the planes along that direction so that different slices in the VTK-world can be seen.

3.4.3. Haptic ROI Manipulations

The second aim of the haptic environment was the creation of the ROI's. By default, a ROI is a sphere which is provided to the system as an external wrl file. In this point we should mention that in the beginning the slices consist of a separate and autonomous object in the local coordinate system. When the user loads a ROI then the sphere becomes a child of this object and it is placed in an default position. When the haptic device collides with an ROI and the user presses the button in the haptic device then the current ROI becomes a child of the world object and afterwards it is moved accordingly by the haptic cursor. In that way, the user is able to "capture" a ROI, with the haptic, and can rotate the scene in order to find the desirable without placing the ROI somewhere else in the scene. The realistic perception of the 3D scene is one of the main advantages of the haptic device. When the user chooses an appropriate position for the ROI, he releases the button of the haptic device and thus the ROI becomes again a child of the local coordinate system now having the position the user decided. As mentioned before, the placement the ROI results in a different coordinate system whose child we wish to become, having a specific position.

The next step is to use the haptic to place the ROI mesh in a specific location $\vec{P}_{\mathrm{roi}}$ and afterwards set the ROI mesh as a child of the slice system, by locating it in the same position $\vec{P}_{\mathrm{roi}}$. To achieve that we have to set the local position of ROI mesh, which is relative to the slice system, equal to the global position of the proxy, since the position of an object is relative to its parent. If the slice system has no rotation, the two mentioned positions should be the same., If the slice system has a rotation, we need to compute the position of the proxy relative to slice system's frame and set ROI mesh's position equal to that. The slice system is constantly moving so we need to calculate its position for every iteration. Let $\vec{P}_{\mathrm{roi,final}}$ be the position of the ROI in the slice system local coordinate system, R_{scene} be the global rotation matrix of the scene and $\vec{P}_{\mathrm{device}}$ be the global position of the device. Then the appropriate transformation is defined as:

$$\vec{P}_{\mathrm{roi,final}} = [R_{\mathrm{scene}}]^T * (\vec{P}_{\mathrm{device}} - \vec{P}_{\mathrm{roi}}) \tag{7}$$

where $[.]^T$ denotes the transpose of a matrix.

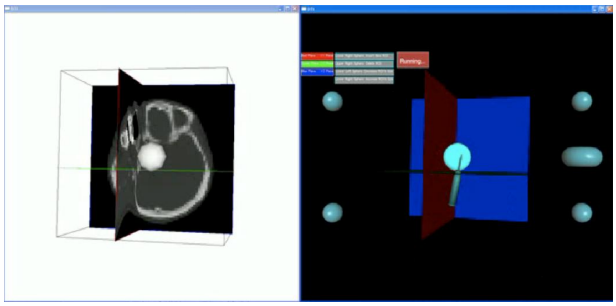


Figure 3: *The snapshot of the haptic interface demonsration. The left panel is the VTK display which works in parallel with the right panel which is a graphical representation of the data volume, three orthogonal slices and the spherical ROI. The tip of the 3D haptic pointer is shown touching (holding) the ROI for positioning.*

Apart from creating a ROI, the user is able to destroy, enlarge or shrink a ROI by selecting it with the haptic cursor and applying the desired procedures. At the right part of the haptic world there are two spheres that can be used for the addition and removal of the sphere-shaped ROIs. When the bottom-right sphere is selected, a ROI is added to the center of the scene. This ROI can be translated by touching it, pressing the haptic button and moving it around the scene. If the ROI is placed inside the bounding box of the volumetric data, it will stay there. But if the user tries to place the ROI outside the bounding box, it will automatically be placed at the center of the scene. In order to remove the ROI is removed, someone should drag and drop it to the upper thus erasing it from the scene.

At the left part of the haptic world there are two additional spheres that are used to increase and decrease the size of the ROI. By selecting the ROI and dropping it at the upper sphere it grows. If it is dropped at the lower sphere, the ROI shrinks.

The step factor for changing the size of a ROI is determined from the user. For the time being, the shape of the ROI's are just spheres and their shapes cannot be modified. Figure 3 shows a screenshot of the haptic interface demonstration.

4. DISCUSSION

The DTI project's goals can be grouped into three: The implementation of a baseline application, the implementation of the STAC standard and the development of haptic interaction tools. The development was done within the VAVframe framework. VAVframe is a C++ library that uses VTK, ITK and Qt. VAVframe has two goals: To provide a unified and standardized environment for code development and documentation, to function as a repository of the R&D at VAVlab (`http://www.vavlab.ee.boun.edu.tr`) for the continuation and progress of research efforts. It is a common problem of research labs that many invaluable contributions that cost significant time and money, become unusable as soon as the individuals who developed them move. VAVframe aims at avoiding this problem. The project team achieved all goals, yet the integration of these components could not be completed during the workshop.

The major problem that we have encountered was the integration of the VTK library, that VAVframe uses extensively, and the open-source CHAI-3D library used for haptic devices. Part of the project team worked exclusively on finding solutions for this integration. The current implementation is a proof-of-concept application which allows the users to rotate 3D volume, slide cross-sectional planes, position a spherical ROI and change its size. One of the unfulfilled goals was to develop haptic interface to modify the shape of 3D ROI. Using FEM for shape-modification seems a promising approach but it will certainly add some overhead due to real-time calculations. This is left for future work. In addition to the haptic controls discussed so far, haptics could also be used to move along individual fibers (like a bead sliding along a space-curve) for enhancing the perception of the topological properties of the fibers by the user, like curvature, torsion of the curve representing the fiber. In such a case, the haptic device could be used as a play-back device that moves the user's hand along the fiber or as a passive device that applies constraint forces whenever the haptic cursor tends to move away from the curve. Another option would be to send high force-magnitudes to the haptic device in regions where the fiber intensity is high so that the user can "feel" regions of high neural activity & concentration in the brain.

The base implementation and integration of the STAC with the software was completed. However, due to the unforeseen delays in the development of haptic interface, we could not demonstrate the final goal of the project. The goal was to demonstrate the use of real-time haptic ROI manipulations (shaping and positioning) with tractography results within a dynamic ROI query paradigm. The target application was envisioned to display the fiber tracts that intersect with a 3D ROI which is manipulated with a haptic device, in real time. Although this goal could not be achieved, the invaluable know-how that was developed during the workshop regarding the integration of different software libraries is a big step forward.

5. CONCLUSION

A proof-of-concept DTI application was aimed during the project. Several issues, including the integration of STAC, implementation of DWI to DTI conversion (tensor estimation), integration of CHAI-3D and VTK through Qt were solved during the project. Although the final application could not be completed, the know-how generated during the workshop is invaluable. Current efforts are concentrated on the integration of these components with the addition of original DTI analysis algorithms.

The project was run by VAVlab, Boğaziçi University, İstanbul, Turkey, in collaboration with Stanford University, School of Medicine, Department of Radiology, LUCAS MRS/I Center, CA, USA, University of Valladolid, Valladolid, Spain and Aristotle University of Thessaloniki, Thessaloniki, Greece.

6. ACKNOWLEDGMENTS

This work is conducted as part of EU 6^{th} Framework Programme, SIMILAR NoE, eNTERFACE 2007 workshop organized at Boğaziçi University, İstanbul, Turkey. The project is in part supported by SIMILAR NoE grants and Tubitak KARIYER-DRESS (104E035) research grants.

7. REFERENCES

[1] P. Basser, "Inferring microstructural features and the physiological state of tissues from diffusion-weighted images", *NMR Biomed*, vol. 8, pp. 333–344, 1995. 1

[2] P. Basser and C. Pierpaoli, "Microstructural and physiological features of tissues elucidated by quantitative diffusion tensor MRI", *J. Magn. Reson.*, vol. B 111, pp. 209–219, 1996. 1

[3] R. Bammer, B. Acar, and M. Moseley, "In vivo MR Tractography Using Diffusion Imaging", *European J. of Radiology*, vol. 45, pp. 223–234, 2002. 1

[4] A. Brun, M. Martin-Fernandez, B. Acar, E. M. noz Moreno, L. Cammoun, A. Sigfridsson, D. Sosa-Cabrera, B. Svensson, M. Herberthson, and H. Knutsson, "Similar Tensor Arrays - A framework for storage of tensor data", tech. rep., Las Palmas, Spain, 2006. 1, 2

[5] W. Schroeder, K. Martin, and B. Lorensen, *The Visualization Toolkit. An Object Oriented Approach to 3D Graphics.* Kitware, Inc., 2002. 4

[6] "The Open Source Haptics Project", August, 8 2002. http://www.chai3d.org. 4

[7] J. Blanchette and M. Summerfield, *C++ GUI Programming with Qt 3*. Pearson Education, 2004. 4

[8] C. R. Tench, P. S. Morgan, M. Wilson, and L. D. Blumhardt, "White Matter Mapping Using Diffusion Tensor MRI", *Magnetic Resonance in Medicine*, vol. 47, pp. 967–972, 2002. 5

8. BIOGRAPHIES

Murat Aksoy is a PhD student at Stanford University LUCAS MRS/I Center, CA, USA.
Email: maksoy@stanford.edu

Neslehan Avcu graduated from Dokuz Eylül University, Department of Electrical and Electronics Engineering, in 2006 and currently she is a MS student in Dokuz Eylül University, The Graduate School of Natural and Applied Science, Electrical and Electronics Engineering Department. She has been a scholarship student, funded by TUBITAK since 2006. She is now a research assistant in Control Laboratory, and researcher in Control and Intelligent Systems Research Laboratory, DEU, Department of Electrical and Electronics Engineering. Her research interests include ANNs and their signal processing applications, cognitive science, fuzzy logic and pattern classification.
Email: neslihan.avcu@deu.edu.tr

Susana Merino-Caviedes graduated in Telecommunications Engineering from the University of Valladolid in 2005, and she is currently a PhD student in the aforementioned university. She is a member of the Image Processing Laboratory at the University of Valladolid. She is a participant of the SIMILAR NoE and a national research project focused on DTI. Her research interests are medical image processing, DTI visualization, user interfaces, image segmentation and level set algorithms.
Email: smercav@lpi.tel.uva.es

Engin Deniz Diktaş was born in İstanbul-Turkey, in 1978 and is currently a PhD student at Computer Engineering Department at Boğaziçi University. He received his BS degree in Mechanical Engineering and MS degree in Systems & Control Engineering from Boğaziçi University. His research interests include computer graphics, computational geometry, haptic rendering & interfaces and real-time simulation.
Email: denizdiktas@gmail.com

Miguel Ángel Martín-Fernández received his M.S. degree in Electronical Engineering from the University of Valladolid in 1999, and currently he is a teaching assistant and PhD student at the same received University. He has participated in several national and international research projects as member of the Image Processing Laboratory at the University of Valladolid. His research work is related to medical image processing, specially registration algorithms and analysis of medical images.
Email: migmar@tel.uva.es

Sıla Girgin was born in Bayburt, Turkey in 1982. She received the BS degree from the Department of Computer Engineering of Boğaziçi University, in 2005. She is currently a researcher and MS student at the same university. She is a member of VAVLab (Volumetric Analysis and Visualization Laboratory) in the Departments of Electrical and Electronics Engineering and MediaLab in the Department of Computer Engineering, at the Boğaziçi University. Her current interests are medical information visualization, computer graphics and computer vision.
Email: silagirgin@gmail.com

Ioannis Marras was born in Karditsa, Greece in 1980. He received the BS degree from the Department of informatics of Aristotle University of Thessaloniki in 2005. He is currently a researcher and teaching assistant and studies towards the Ph.D. degree at the Department of Informatics, in the Artificial Intelligence Information Analysis (AIIA) laboratory, at the Aristotle University of Thessaloniki. His current research interests lie in the areas of medical image analysis, signal and image processing, pattern recognition and computer vision.
Email: imarras@aiia.csd.auth.gr

Emma Muñoz-Moreno received her M.S. degree in Electronical Engineering from the University of Valladolid in 2003, and currently she is a PhD student at the same received University. She has participated in several national and international research projects as member of the Image Processing Laboratory at the University of Valladolid. Her research work is related to medical image processing, specially registration algorithms of DT-MRI and surgical simulation.
Email: emunmor@lpi.tel.uva.es

Erkin Tekeli was born in İzmit, Turkey in 1980. He received the BS degree from the Department of Mechatronics Engineering of Sabanci University, in 2004 and his MS degree from the Department of Electrical and Electronics Engineering and Computer Sciences of Sabanci University, in 2007. He is currently a Phd. student at the Department of Electrical and Electronics Engineering of Boğaziçi University. He is a member of VAVLab (Volumetric Analysis and Visualization Laboratory) in the Departments of Electrical and Electronics Engineering at the Boğaziçi University. His current interests are computer vision, medical image processing, shape spaces, content based image retrieval, texture analysis and segmentation.
Email: erkin.tekeli@boun.edu.tr

Roland Bammer, PhD, is an assistant professor at Stanford University, LUCAS MRS/I Center, CA, USA.
Email: rbammer@stanford.edu

Burak Acar was born in Izmir in 1972. He received his BS, MS and PhD degrees, all in Electrical & Electronics Engineering, from Bilkent University, Ankara, Turkey, in 1994, 1996 and 2000, respectively. He worked on ECG signal processing during his PhD. He worked at University of London, St. George's Hospital Medical School, London, UK, between 1998-1999. He was at Stanford University Medical School, Department of Radiology, 3D Lab, CA, USA, between 2000-2003, where he worked on medical image analysis and computer aided detection/diagnosis. He joined Boğaziçi University, Electrical & Electronics Engineering Department in 2003. His main research interests are 3D data (ranging from images to tensors) analysis and visualization with emphasis on medical applications. Further information can be found at `http://www.vavlab.ee.boun.edu.tr`.
Email: acarbu@boun.edu.tr

Marcos Martin-Fernandez received the Ingeniero de Telecomunicacion degree and the PhD degree from the University of Valladolid, Valladolid, Spain, in 1995 and 2002 respectively. He is Associated Professor at the ETSI Telecomunicacion, University of Valladolid where he is currently teaching and supervising several Master and PhD students. From March 2004 to March 2005, he was a Visiting Assistant Professor of Radiology at the Laboratory of Mathematics in Imaging (Surgical Planning Laboratory, Harvard Medical School, Boston, MA). His research interests are statistical methods for image segmentation and filtering in multidimensional (tensor) signal processing. He is also with the Laboratory of Image Processing (LPI) at the University of Valladolid where he is currently doing his reseach. He was granted with a Fullbright fellowship during his visit at Harvard. He is reviewer of several scientific journals and member of the scientific committee of the IEEE International Conference on Image Processing (ICIP), the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), European Signal Processing Conference (EUSIPCO) and IEEE International Symposium on BioInformatics and BioEngineering (BIBE). He also participates in fellowship evaluation for the National Competence Centre in Biomedical Imaging (NCCBI), Switzerland. He has more than 50 papers in scientific Journals and Conferences.
Email: marcma@tel.uva.es

Ali Vahit Sahiner was born in Ankara, Turkey in 1956. He received his PhD. in Computer Science from University of Westminster, London, UK, in 1991. He is currently working as an instructor at Boğaziçi University, İstanbul, Turkey. His interests include Computer Graphics, Visualization and Animation, Video and TV technologies, Graphic design and Typography.
Email: alivahit.sahiner@boun.edu.tr

Suzan Üskudarli is a faculty member of The Computer Engineering Department at Boğaziçi University. After receiving her doctorate degree at University of Amsterdam, she worked in Palo Alto, California. She joined Boğaziçi University in 2005. Her interests include software engineering and cooperation technologies. She is part of the VAVlab team, where she contributes to the software development aspects.
Email: suzan.uskudarli@boun.edu.tr

MULTIMODAL SERVICES FOR REMOTE COMMUNICATIONS: THE VOTE-REGISTRATION EXAMPLE

Jérôme Allasia [1], *Ana C. Andrés del Valle* [2], *Dragoş Cătălin Barbu* [3], *Ionut Petre* [3], *Usman Saeed* [4], *Jérôme Urbain* [5]

[1] IRISA lab, Rennes, France
[2] Accenture Technoloy Labs, Sofia Antipolis, France
[3] Institute for Research and Development in Informatics (ICI), Bucharest, Romania
[4] Eurecom, Multimedia Communications Department, Sofia Antipolis, France
[5] Circuit Theory and Signal Processing lab (TCTS), Faculté Polytechnique de Mons, Belgium

ABSTRACT

With the development of the Next Generation Networks, participants in meetings will communicate using the medium that best suits them (text, audio, video). Services provided to make meetings more efficient will have to be implemented taking into account the multiple modalities available. The most advanced multimodal/multimedia signal processing techniques will become the core tools that services will exploit. This project aimed at developing a proof of concept for multimodal services for communications. The communication exchange, image, speech and text processing for automatic vote counting in meetings is proposed as an example of a service to be developed over a communications architecture.

KEYWORDS

Telecommunications – Image processing – Speech processing – Vote registration – Multimodal – Multimedia communications

1. INTRODUCTION

Technology has enabled remote synchronous collaboration. It allows employees to work together from distant locations by sharing audio and visual information. Meetings with remote colleagues have extended the work environment beyond the physical office space. Some of the technologies that make communications rich because they are multimodal (e.g. videoconference) are still not widely utilized, despite being considered useful to most people [1]. This is due to the fact that a technological solution is only adopted when the benefits from using it (both economic and in human performance) overrun the hassles of introducing it in the work environment.

Most of the research efforts made by the signal processing and HCI communities have been driven towards enhancing the way humans work. By examining some of the recent works to incorporate advanced signal processing in the workplace, we realize that most approaches focus on developing smart environments. For instance, researchers from European projects (FP6) AMI / AMIDA [2] or CHIL [3] investigate meetings as artifacts to improve inside an intelligent space (meeting room) and they combine different technological solutions, e.g., recording audio, person localization, automatic summarization from audio, animation of avatars, etc. so to help workers perform their tasks [4]. These projects take a technology-driven approach where, from technical challenges related to making technology pervasive to employees, they create practical scenarios to test their solutions.

Meetings and work are no longer restricted to an office environment. Conference calls and most infrequently utilized video conferences enable participants to intervene in meetings wherever they are. Current communication technologies do not completely exploit the multimodal nature of humans. The Next Generation Networks (NGN) [5] [6] standardization initiatives aim at enabling platform-agnostic (any vendor and any provider), heterogeneous (any modality and medium) and asymmetric (no need for both sides of the communication to share the same modality to communicate) communications. Within this new paradigm, services that will enhance communications and thus make meetings more efficient will be more easily developed.

Doing research on multimodal services for synchronous communications implies being at the intersection of three major fields: telecom networking, multimedia/multimodal signal processing and ubiquitous computing for context awareness. The networking community is not only focusing on standardization but also on developing the protocols and algorithms that will facilitate the smooth data exchange for multimodal services to run over any kind of network. Multimedia signal processing concentrates its efforts in creating algorithms that can analyze and synthesize multimodal signals (text, video, audio) to extract or provide useful information that that can be exploited to enhance communications. Ubiquitous computing is looking at multimodal signals from sensors to specifically make technology aware of the context on which humans use computers and other devices (e.g. phones) so to adapt to their circumstances and automatically provide users with, for instance, the most suitable communication solution available.

This project has focused on developing an automatic vote registration service over a simulated communication network as a proof of concept of a novel modular context-aware approach to create technology that enhances communications for meetings. The chosen service enables automatic vote counting from all participants joining a meeting independently of the medium they utilize to communicate. During the project we have focused specifically on the multimodal signal processing needed, the protocol for the exchange of information between the vote registration system and the user and, inside the network. We have assumed that the system already "knew" the best communication option for each participant.

This report is divided as follows. Section 2 overviews the theoretical framework for the context-aware communication platform and services upon which our research is based. Right after, we discuss the communications protocol and the information exchange needed for this kind of platform in the context of NGN and current multimedia streaming protocols. In Section 4 we describe the vote registration service. Sections 5, 6 and 7 explain in depth the real-time signal processing techniques and algorithms created to enable vote counting on each medium as well as how

the natural interaction with the participant has been designed. They also cover some preliminary tests of their algorithmic performance. The prototype implementation for simulating how the service could work cross platforms is discussed in Section 9. We conclude and give some future perspectives in Section 10

2. MULTIMODAL/MULTIMEDIA SERVICE ORIENTED TECHNOLOGY DESIGN FOR COMMUNICATIONS

Developing a context-aware multimodal framework to create advanced services for synchronous communications will lead to improving the efficiency of meetings. Our research objective is to move from a technology-driven approach to improve meetings, to a method that designs technology mainly conducted by user and business needs.

In [7], the authors detail the approach taken to face the challenges; the framework they propose is the foundations of this project's work. In this section we will summarize the most important points. This approach focuses on two key aspects:

1. **Modularity**: the full technology offer will comprise a range of modular services for meetings, each addressing a specific aspect of a meeting, from preparation to follow-up (i.e. services will help users before, during, and after the actual meeting time).
2. **Adaptation to user and business needs**: these services are combined and proposed to a specific meeting and to their participants depending on the context: people's context (previous history, roles of participants, whereabouts), meeting context (tasks, objectives) and the enterprise context (underlying business processes). Fig. 1 depicts the concept.

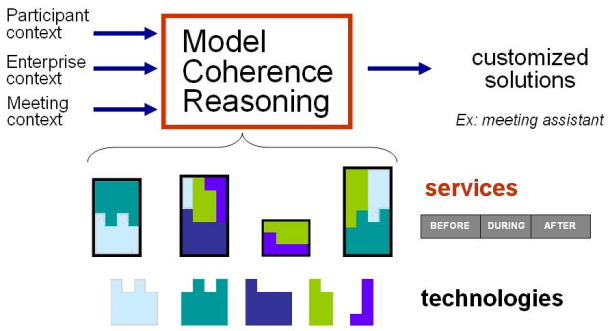


Figure 1: *This illustration conceptually depicts the architecture to create technology for synchronous collaboration. The reasoning piece of the architecture is built based on the participant's context and the task interaction so to provide the services needed to make the meeting more efficient. These services are built on top of modular multimodal/multimedia signal processing blocks.*

The core elements of this platform are:

1. **Service Conception**: a service for synchronous collaboration is a set of technologies devoted to a particular aspect of a meeting. When defining a service, the specific technology that will be used to implement it is not considered. Some examples of services are: vote registration; summarization (minutes); time discipline service; follow-up management; digitalization of analog content; question capture and management; etc.
2. **Modeling Interaction and Building the Reasoning Level**: after analyzing workflow, context and interaction specific to synchronous collaboration, we can build models that will deduce the suitable combination of services for a meeting. These models will determine when certain services will or will not be required. For example, in meetings where team-members know each other well, an "introduction service" where participants are introduced and their identity is present along the entire meeting will only add load to the work environment and no value, therefore, this service might not be proposed.
3. **Modular Design of Technologies**: in this framework, no prior hypotheses are established when developing the technology that seamlessly deploys the services that are needed. Different technologies might be required to deploy the same service; different services might require the same technology. For instance, in meetings where decisions must be taken, a "vote registration service" like the one this project proposed could be included. To automatically register votes, distinct technologies will adapt the service to the context. Those attending the meeting at the office could naturally raise their hand; using video-activity analysis votes could be counted. For those connected through a call, voting by speech recognition is a suitable option. Complementarily, the same video-activity analysis that detects raising hands could be used in an "interruption management service" too.

3. COMMUNICATION PROTOCOL

3.1. Next-Generation Network Services

Today business environments are competitive and complex. Success lies on outstanding customer service. The demand is growing for powerful new communications services as a way to enhance customer service and build a competitive edge. At the centre of these new services is the next-generation network (NGN).

Standardization of multi service network technologies for next generation networks are conducted by European Telecommunications Standards Institute - ETSI [6] and by the International Telecommunication Union - ITU [5].

Providing end users with multi-modal access to availability and location of other end users improves flexibility and efficiency of human-machine communications, and support the user in person-to-person communication.

Concepts like integrating presence and context information with multimodal interaction may influence the use of such services. In order to achieve this, the network has to provide basic interaction functions and dialog control, wherever several devices are used in parallel, like an Instant Messenger as a multimodal application [8].

The next-generation network is the next step in world communications, traditionally enabled by three separate networks: the public switched telephone network (PSTN) voice network, the wireless network and the data network (the Internet). NGNs converge all three of these networks-voice, wireless, and the Internet-into a common packet infrastructure. This intelligent, highly efficient infrastructure delivers universal access and a host of new technologies, applications, and service opportunities.

There are three types of services which drive NGNs: real-time and non real-time communication services, content services, and transaction services. The service-driven NGN gives service providers greater control, security, and reliability while reducing their operating costs.

Built on open modular elements, standard protocols, and open interfaces, the NGN caters to the specific needs of all users

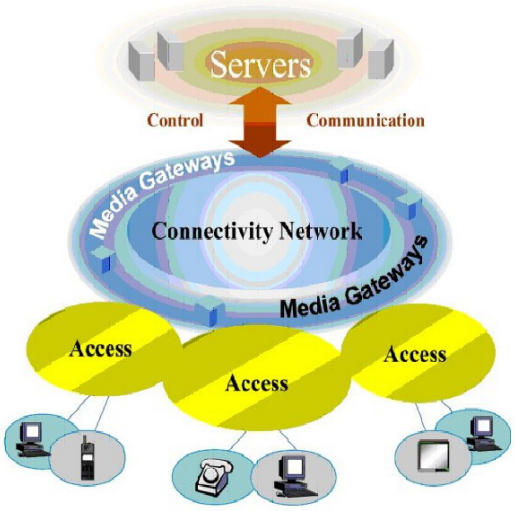


Figure 2: *Multi service networks (illustration courtesy of ITU-T).*

(see Fig. 2). It unites traditional wireline and wireless voice, video, and data using a packet-based transport. The new class of services it enables is more flexible, scalable, and cost-efficient than services that have been offered in the past.

The solutions for the next generation communications are built with open, flexible, standards-based building blocks. Using modular building blocks makes it easy to add new features, services, and value to existing systems.

Some interesting use cases of services and capabilities to be supported in next-generation networks are presented in [9] by the ITU-T Focus Group on NGN.

Multimodal interaction has become one of the driving factors for user interface technologies, since it allows combining the advantages of traditional graphical interfaces that are used in computer environments with speech driven dialogs emerging from the telephony world.

The concepts of dialog presentation for graphical and vocal interfaces, especially for internet based applications, require a new approach to combine interface description languages like Hyper Text Mark-Up Language (HTML) and VoiceXML.

Multimodal interaction has significant advantages:

- User can select at any time the preferred modality of interaction;
- Can be extended to selection of the preferred device (multi-device);
- User is not tied to a particular channel's presentation flow;
- Improves human-machine interaction by supporting selection of supplementary operations;
- Interaction becomes a personal and optimized experience;
- Multimodal output is an example of multi-media where the different modalities are closely synchronized.

3.2. Protocol and architecture

After evaluating current networking possibilities that would enable a vote registration service, we decided to develop a centralized architecture where a server will manage the meeting initialization and vote counting part. The client application will be in charge of adapting the service, in our case the vote registration exchange, to each modality. This implies that the required signal processing will be performed locally and that no media streaming with the vote registration service server will be done. We are letting the server application decide what the appropriate method for establishing the communication is; this decision will be based on the context and present information held in the system. The client application is responsible for the distribution of data to the handling resources, the activation and synchronization of the speech resources and the presentation of information to the participant based as well on its personal context information.

Since the server is not doing the signal processing and only manages the communication and the exchange of information for the service, we found the use of XML files to transfer data a suitable solution.

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<context_profile>
  <meeting_id>100</meeting_id>
  <num_attendees>3</num_attendees>
  <participant>
    <participant_type>Organizer</participant_type>
    <communications_in>video_audio</communications_in>
    <communications_out>video_audio</
        communications_out>
    <communications_id>01</communications_id>
    <firstname>Ana</firstname>
    <lastname>Andres del Valle</lastname>
    <company_name>Accenture</company_name>
    <title>Project Manager</title>
  </participant>
  <participant>
    <participant_type>Atendee</participant_type>
    <communications_in>video</communications_in>
    <communications_out>video_audio</
        communications_out>
    <communications_id>02</communications_id>
    <firstname>Usman</firstname>
    <lastname>Saeed</lastname>
    <company_name>EURECOM</company_name>
    <title>PhD student</title>
  </participant>
</context_profile>
```

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<command_profile>
  <command_name>Start_chat</command_name>
  <num_param>2</num_param>
  <param_1>bosphorus_07</param_1>
  <param_2>bosphorus_08</param_2>
</command_profile>
```

Figure 3: *Example of context and command (start_chat) XML files.*

For a better understanding how the system works we will detail the steps taken by the service:

1. First the participant connects to the meeting server
2. Participant sends an XML file with the context profile
   The context profile has the following information:
   - `meeting_id` - identifier of the meeting to get connected to
   - `participant_type` - participant role (organizer/attendee)
   - `num_attendees` - indicates how many participants are on this communication link
   - `communications_in` - values from {video & audio, text only, audio only}
   - `communications_out` - values from {video & audio, text only, audio only}
   - `firstname` - for attendee

- `lastname` - for attendee
- `company_name` - for attendee
- `title` - for attendee

3. After the server processed the context profile, tracks how many participants are there already and then assign connections to participants
4. The server stores the participant's context information
5. The server analyzes the different contexts to establish the communication strategy
6. The server sends commands to participants to indicate the modality of communications to be used.
   The commands have the following structure:
   - `Command_Name` - values from start chat, start call, start videoconference, start voting mode, end voting mode
   - `No_Parameters` - the number of parameters the command takes;
   - `Parameter_List` - a list with all the parameters.

We refer to Fig. 3 for a couple of examples of the XML used in the project.

3.3. Discussion and future perspectives

A centralized architecture that provides service management ensures minimum exchange of data related to the new service running on top of communications. We expect NGN to be able to manage how each platform will adopt this architecture, most likely through specific gateways. At this point in the evolution of these networks, it is difficult to forecast if service/client architectures will be more efficient that a peer to peer ones. We refer the reader to [10] [11] [12] [13] for multiple discussions on the topic.

Each platform uses different devices to run their communications. If the required signal processing is performed locally, we should be aware that the computing power differs from a mobile telephone to a PDA or a laptop. Therefore, even if the proposed algorithms are working in real-time on a computer, some tuning might be needed to run on other devices. In our project we have only run simulations that use PC or laptop as processing points (see Section 9 for more details). A plausible option to run complicated algorithms on low-computing power devices is externalizing the processing to a more powerful device (i.e. a server). In this case, we would apply the same client/server architecture for the signal processing. The advantage is clear; we can obtain the benefits of complicated algorithms on, for instance, mobile phones, as it is already exploited in [14]. The major drawback is that media streaming (e.g. video and audio) between the device and the server will be required thus loading the network with data and also introducing even more delays to the processing. We clearly recommend separating the service management part, which is common to all platforms and communication styles, from the signal processing part. Although both could be designed using a server/client architecture, we believe that developing them jointly could negatively interfere in the performance of the service.

Future work regarding the architecture will include these three points:

- understand how the vote-registration service would be adopted in a NGN architecture;
- study how to develop the algorithm for each platform, evaluating if a specific media processing server is needed or not;
- use ubiquitous techniques to get participant, business and social context automatically so as to update the XML data that provide critical information to the system.

4. MULTIMODAL VOTE-REGISTRATION SERVICE

During a meeting held among participants that are connected from different points of the world using different communication means (videoconference, chat, telephone, etc.) a decision must be taken. A vote-registration service would enable the automatic count of each of participants' choice. One of the participants, most likely the organizer, will announce the decision to be taken to the system, along with the available options. The system will gather each participant's choice and announce the results from the survey.

This system seems too complex to simply count the votes in a meeting, above all if the meeting is organized among a few participants. The usefulness of a system like this comes when meeting size scales up. If we need to register the vote of many people, traditional around-the-table human strategies to count are not feasible. Therefore, voting among many people is not done during meetings, but off-line through surveys, and meetings where decisions must be taken tend to have a reduced number of participants. A service like the one we propose could help change that, and allow technology to adapt to business needs instead of having people adapting to the restrictions technology imposes.

4.1. Types of questions that have been evaluated

In our project we have evaluated two kinds of questions for vote registration:

- Binary: the expected answer can be either yes or no.
- Multiple-choice: the participant can choose among a close set of options.

We have not analyzed open sets of options where participants are also allowed to vote for non pre-established options.

4.2. Studied modalities

In this project we have studied the following cases:

1. **Text as input and output modality**: we consider the scenario where participants join the meeting through a chat. We do not analyze how the meeting and any of the other available communication modalities (e.g. audio) would be transcribed to text, but only how the vote-registration should be performed.
2. **Speech as input and output modality**: this is the most common scenario in current remote meetings. People join the meeting through the phone.
3. **Video/Audio input and only video output modality**: this case is considered specifically for the vote registration when multiple people want to naturally vote, and they raise their hands in a meeting room that might or not be multicast to other meeting participants.
4. **Video/Audio input and video/audio output**: In a second scenario involving video, if only just one person is facing the camera, the vote registration can be done by analyzing the video to understand what he or she has voted and then also couple the information with the speech analysis. In this case, if one of the two systems is not robust enough the use of multimodality should help us better discern their choice.

This report covers the development of the multimodal signal processing techniques needed to register one single vote coming from a participant over a chat, one single participant facing the camera, multiple participants facing a camera and one single participant speaking over the phone.

As future work we would like to be able to exploit the multimodality coupling between audio and video both are available as input and output.

5. TEXT MODALITY

From all the modalities analyzed to perform the vote registration, text-based ones, i.e., those used for platforms like instant messaging, SMS and chats, are the simplest in terms of signal analysis. Typed characters in sentences are easy to display and analyze as they have always belonged to the most common device interfaces.

Text vote-registration is a quite controlled scenario. Its simplicity makes researchers consider how to build natural communication exchange the real challenge for this modality. Natural language processing (NLP) is the subfield of artificial intelligence (AI) and linguistics that deals with the understanding, transformation and natural synthesis of language (the same principles that apply to text communications are extendable to speech exchange). Text parsing and advanced AI techniques can be used to developed almost "human" interfaces.

5.1. The proposed interaction algorithm

Our approach mainly leads the participant to choose an available option, by double-checking his choice, and making sure that he responds (refer to Fig. 4). Currently, we do not allow the user to avoid voting, unless a blank vote is considered as an option. We treat multiple-choice and binary questions the same way. The implementation of natural language interaction on text should lead to solutions that will not upset or tire users.

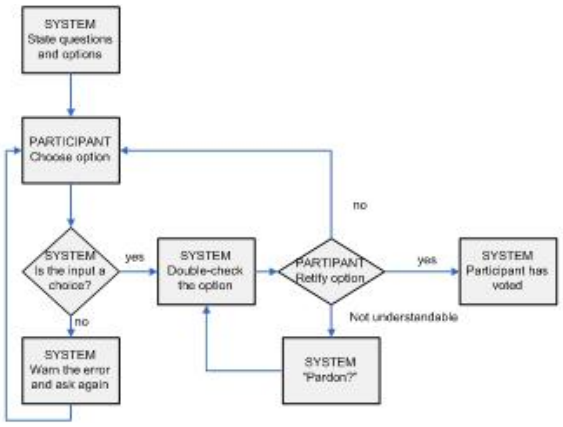


Figure 4: *This schema illustrates the interaction between the participant and the automatic text vote registration.*

5.2. Future perspectives

Although the language processing required for a vote system is not very complex. It would be interesting to design different exchange and interaction schemas applied to binary and multiple-choice questions. Not being this the main goal of the project, we leave the usability analysis of this part of the vote-registration service for future research.

6. SINGLE-FACE VIDEO

Head gesture recognition systems aspire to have a better understanding of subliminal head movements that are used by humans to complement interactions and conversations. These systems vary considerably in their application from complex sign language interpretation to simple nodding of head in agreement. They also carry additional advantage for people with disabilities or young children with limited capabilities.

As part of the project, we focused on a simple yet fast and robust head gesture recognition system to detect the response of users to binary type questions. We did not wish to be limited by using specialized equipment thus we have focused our efforts in using a standard webcam for vision based head gesture recognition.

6.1. State of the Art

Head gesture recognition methods combine various computer vision algorithms for feature extraction, segmentation, detection, tracking and classification so categorizing them based on distinct modules would be overly complicated. We thus propose to divide the current head gesture recognition systems into the following categories.

6.1.1. Holistic Approach

This category of techniques focuses on the head as a single entity and develops algorithms to track and analyze the motion of head for gesture recognition. The positive point of these techniques is that as head detection is the main objective, they are quite robust at detecting it. The main disadvantage is the accuracy in detecting small amounts of motion.

In [15] the authors have embedded color information and a subspace method in a Bayesian network for head gesture recognition but this system fails when slight motion of head is concerned. Similarly [16] uses a color model to detect head and hair and then extracts invariant moments, which are used to detect three distinct head gestures using a discrete HMM. Experimental results have yielded a detection rate of 87%. In [17] the mobile contours are first enhanced using pre-filtering and then transformed into log polar domain thus reducing 2D motion into simple translation. Pan and tilt are then detected by analyzing the energy spectrum.

[18] recognize bodily functions like standing up, sitting down using the motion of head. The head is assumed to be the most mobile object in the scene and detected by frame differencing. The centroid of the head is extracted in each frame and used as a feature vector in a Bayesian classifier. [19] have build a mouse by tracking head pose using a multi-cues tracker, combining color, templates etc. in layers so if one fails the other layer can compensate for it. Then a ray tracing method is used to extend a ray from the tracked face to the monitor plane, representing the motion of cursor.

6.1.2. Local Feature Approach

These algorithms detect and track local facial features such as eyes. The advantage is accuracy in motion estimation but the downside is that local features are generally much difficult and computationally expensive to detect.

[20] propose a "between eye" feature that is selected by a circle frequency filter based on the fact that there exist a prominent region between the dark eyes and bright forehead and nose bridge. This region is then conformed by eye detection and tracked for gesture recognition. [21] have based there gesture

recognition on an IR camera with LEDs placed under the monitor to detect accurately the location of the pupil. These observations are used to detect nods using an HMM. Tests were carried out with 10 subjects and 78% correct recognition was reported.

6.1.3. Hybrid Approach

The aim of these algorithms is to combine holistic and local feature based techniques. Thus in reality trying to find a compromise between robustness of holistic approaches and accuracy of local feature based techniques, but most of them end up being computationally expensive as they combine various different levels of detection and tracking.

[22] have reported a head gesture based cursor system that detects heads using a statistical model of the skin color. Then heuristics are used to detect the nostrils as the darkest blobs in a certain region. Finally nostrils are tracked to detect head gestures. The color model given is overly simplistic and nostrils can be easily occluded thus causing the algorithm to breakdown. In [23] they have combined previous work that has been done in face detection and recognition, head pose estimation and facial gesture recognition to develop a mouse controlled by facial actions. [24] first searches for the head based on skin color histogram and then selects local extremes of luminance distribution for tracking using Lucas-Kanade algorithm. Feature vectors over sequence of images are collected and recognition is finally performed using a neural net based approach.

6.2. Proposed Method

The method proposed builds upon previously developed algorithms that are well accepted like Lucas Kanade for tracking. The specific requirements of our project dictate that the head gesture recognition algorithm should be robust to lighting and scale yet fast enough to maintain a frame rate of 30 f/s. On the other hand scenarios concerning occlusion and multiple heads in the scene have not been handled in the current implementation.

6.2.1. Face Detection

The first module is the face detector, which is based on a cascade of boosted classifiers proposed by [25]. Instead of working with direct pixel values this classifier works with a representation called "Integral Image", created using Haar-like features. The advantage of which is that they can be computed at any scale or location in constant time. The learning algorithm is based on AdaBoost, which can efficiently select a small number of critical visual features from a larger set, thus increasing performance considerably. The final advantage of this classifier is that it is a combination of several simpler classifiers that are applied one after the other to the image until at some stage the object is detected or all the stages have passed.

The classifier has been trained with facial feature data provided along the Intel OpenCV library [26]. The face detection using the above classifier is very robust to scale and illumination but has two disadvantages, first although it can be considered fast as compared to other face detection systems but still it attains an average performance of 15 f/s. Secondly it is not as accurate as local feature trackers. Thus head detection was only carried out in the first frame and results passed on to the next module for local feature selection and tracking.

6.2.2. Feature Selection and Tracking

The next step involves the selection of prominent features within the region of the image where the face has been detected. We have applied the Harris corner and edge detector [27] to find such points. The Harris operator is based on the local autocorrelation function which measures the local changes of the signal with patches shifted by a small amount in different directions.

Tracking of these feature points is achieved by Lucas Kanade technique [28]. It uses the spatial intensity gradient of the images to guide in search for matching location, thus requiring much less comparisons with respect to algorithms that use a predefined search pattern or search exhaustively. Fig. 5 shows the face detection and the other facial features selection.

6.2.3. Yes/No Decision

The final module analyzes the coordinate points provided by the tracking algorithm to take decision whether the gesture is a Yes or a No. First a centroid point is calculated from the tracked points, then the decision is taken based on the amount of horizontal or vertical motion of this centroid. If the amount of vertical motion in the entire sequence is larger than the horizontal a yes decision is generated, similarly for a No.

Figure 5: *Detected face & feature points that will be tracked.*

6.3. Experiments and Results

The development and testing was carried out on a basic 1.5 MHz laptop with 512 MB of RAM, without any specialized equipment. Video input of the frontal face was provided by a standard webcam with a resolution of 320X240 at 30 f/s.

Illumination and scale variability are the two main causes of errors in image processing algorithms, thus we have tried to replicate the possible scenarios most probable to occur in a real life situation. Although the amount of testing was limited to 5 people due to time concerns. Nevertheless, the amount of variability introduced both in environmental conditions and subject characteristics (glasses/facial hair/sex) make these tests quite adequate. A screenshot of the application can be seen in Fig. 6.

6.3.1. Illumination variability (Fig. 7)

As illumination variation is not dealt with explicitly in our algorithm, we defined three illumination scenarios to measure the effect of lighting change on our algorithms.

- **Inside Front (L1)**: The face is illuminated from the top front with normal office lighting.
- **Inside Side (L2)**: The face is illuminated from the side with strong sunlight coming through the window in an office.
- **Outside (L3)**: The face is illuminated by ambient light from the sun in an open environment, with some self shadowing.

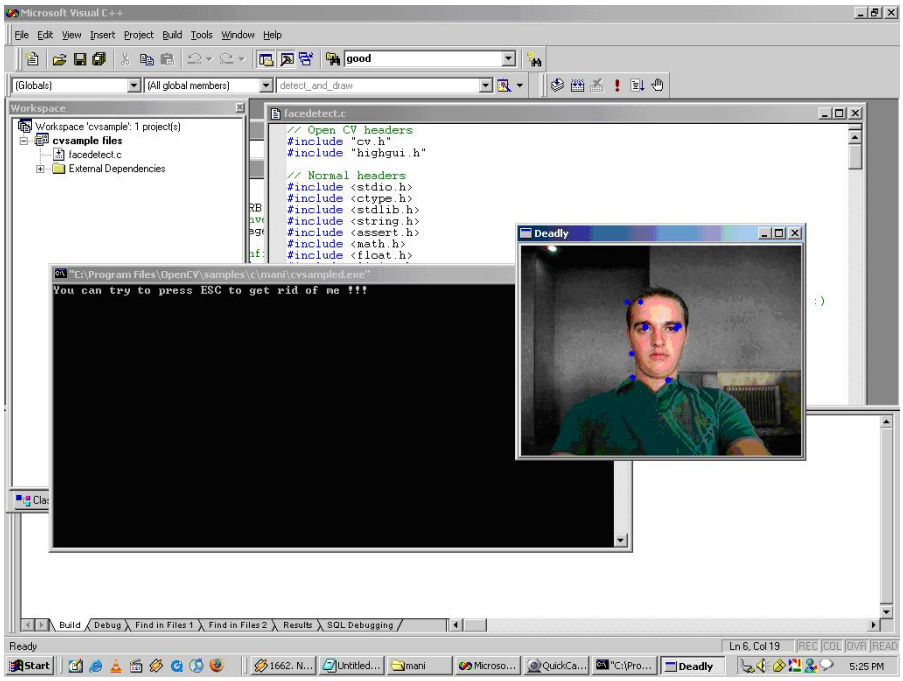


Figure 6: *Screen Shot of the System.*

Figure 7: *Light Variation.*

6.3.2. *Scale variability (Fig. 8)*

The second important source of variability is scale, we have experimented with 3 different scale defined as the distance between the eyes in number of pixels. The 3 measures are S1: 15, S2: 20, S3: 30 pixels.

Figure 8: *Scale Variation.*

6.3.3. *Inter-person variability*

Inter-person variability was both forced and natural, forced in the sense that we tried to include people from both sexes, with-/without glasses and facial hair. The second form of inter-person included comes naturally as no specific instructions were given to the subjects on the how to respond, so they were free to choose the amount and speed of head motion.

6.3.4. *Questions*

The following five questions were selected due to the fact that they can be easily responded to by using head gestures of yes and no.

1. Are the instructions clear?
2. Are you male?
3. Are you female?
4. Are you a student at Boğaziçi University?
5. Do you like chocolate?

6.3.5. *Results*

The system was tested with 5 people who were asked 5 questions each by varying the scale and lighting, we achieved a correct recognition rate of 92 % for the Yes/No gesture. The system did have some problems with detecting the face at large distances when illuminated from the side or in direct sunlight. Refer to Table 1 for detailed results, where P: Person, Q: Question, L: Lighting, S: Scale, CR: Correct Results, Y:Yes, N:No, F: Failure.

6.4. Conclusions and Future Work

In this report we have introduced a real time and highly robust head gesture recognition system. It combines the robustness of a well accepted face detection algorithm with an accurate feature tracking algorithm to achieve a high level of speed, accuracy and robustness. Like all systems, our implementation does have its limitations which were partly enforced by the project definition. The first one is that it cannot handle occlusion of the face; even partial occlusion causes failure. The second is handling head gestures from multiple persons simultaneously in a given scene.

Currently our system only handles binary type questions; a valuable future contribution could be handling multiple choice questions by using facial video. Although several methods can be proposed for this, the one that seems promising to us is using

P	Q	L	S	R	CR
P1	Q1	L1	S2	Y	Y
P1	Q2	L2	S3	F	Y
P1	Q3	L3	S1	N	N
P1	Q4	L3	S2	N	N
P1	Q5	L2	S1	Y	Y
P2	Q1	L1	S3	Y	Y
P2	Q2	L1	S2	Y	Y
P2	Q3	L2	S1	N	N
P2	Q4	L3	S1	N	N
P2	Q5	L3	S2	Y	Y
P3	Q1	L2	S2	Y	Y
P3	Q2	L3	S3	F	N
P3	Q3	L1	S1	Y	Y
P3	Q4	L3	S3	Y	Y
P3	Q5	L3	S1	N	N
P4	Q1	L3	S1	Y	Y
P4	Q2	L2	S2	Y	Y
P4	Q3	L1	S1	N	N
P4	Q4	L1	S2	N	N
P4	Q5	L2	S1	Y	Y
P5	Q1	L3	S2	Y	Y
P5	Q2	L3	S1	N	N
P5	Q3	L2	S3	Y	Y
P5	Q4	L1	S2	Y	Y
P5	Q5	L1	S1	Y	Y

Table 1: *Test results.*

lip reading to recognize a limited vocabulary such as numbers 1, 2, 3, 4. A huge amount of literature already exists on lip reading with complete vocabulary but the results are not so accurate in this case and most of the times video lip reading is assisted with audio information.

7. MULTIPLE-PARTICIPANT VIDEO

The goal of this part of the system is to automatically count votes in a meeting by detecting how many people have raised their hands in a given moment.

To detect a raised hand in a video there are several possible approaches, a review of the state-of-the-art in gesture tracking and recognition algorithms can be found in [29][30]. Most approaches use a movement based method that requires a good frame rate. There are some detecting systems for low-frame rate video like [31] but they deal with only one people at a time.

7.1. Proposed Method

Our algorithm will be integrated into a multiparty videoconferencing system that supports full-motion video, low-frame-rate video and multi-user. In our case we base our algorithm only on segmentation and skin color detection keeping in mind that we need a real time processing system.

The algorithms work as follows (refer to Fig. 10 for the schema of the procedure):

7.1.1. Skin detection

Our method detects pixels that are in the "typical" skin color space, to achieve that we first change color space to YCbCr and then verified that:

$$\begin{aligned} C_r &> 138 \\ C_r &< 178 \\ C_b + 0.6 * C_r &> 200 \\ C_b + 0.6 * C_r &< 215 \end{aligned}$$

A review of the state-of-the-art in skin detection can be found in [32]. The method utilized is not algorithmically complicated but gives acceptable results if the illumination of the scene is not extreme (neither dark nor too bright).

7.1.2. Head detection

Haar based detection, although being a quite naïve approach, allows us to detect the potential heads that we test with de skin mask to decimate wrong detections. This detection is computationally expensive so we fix the number of people to detect. That way this detection is only run a couple of times at the very beginning of the vote session.

7.1.3. Head tracking

We first try to find the head in the same region of the precedent position using the same Haar based detection, and if the tracking fails the movement of the average skin pixel of the zone is used to interpolate the head movement.

7.1.4. Update of research zones

Taking the head position and size into account, we calculate the corresponding Body and Left Zone/Right Zone (search zone for Left and Right hand) (see Fig. 9) It is initialized to a typical hand size.

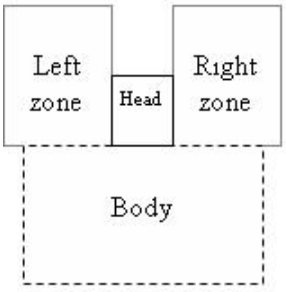


Figure 9: *Search zones.*

7.1.5. Skin mask for hands search

Remove heads and bodies from the Skin mask, so to leave those areas where the hand could potentially be.

7.1.6. Hands detection

Detect all potential hands in the scene using the new skin mask.

7.1.7. Update people's hands

- Find and update hands that can only belong to one person (either left or right hand).
- Associate remaining hands to closest person that does not have any hands detected yet.

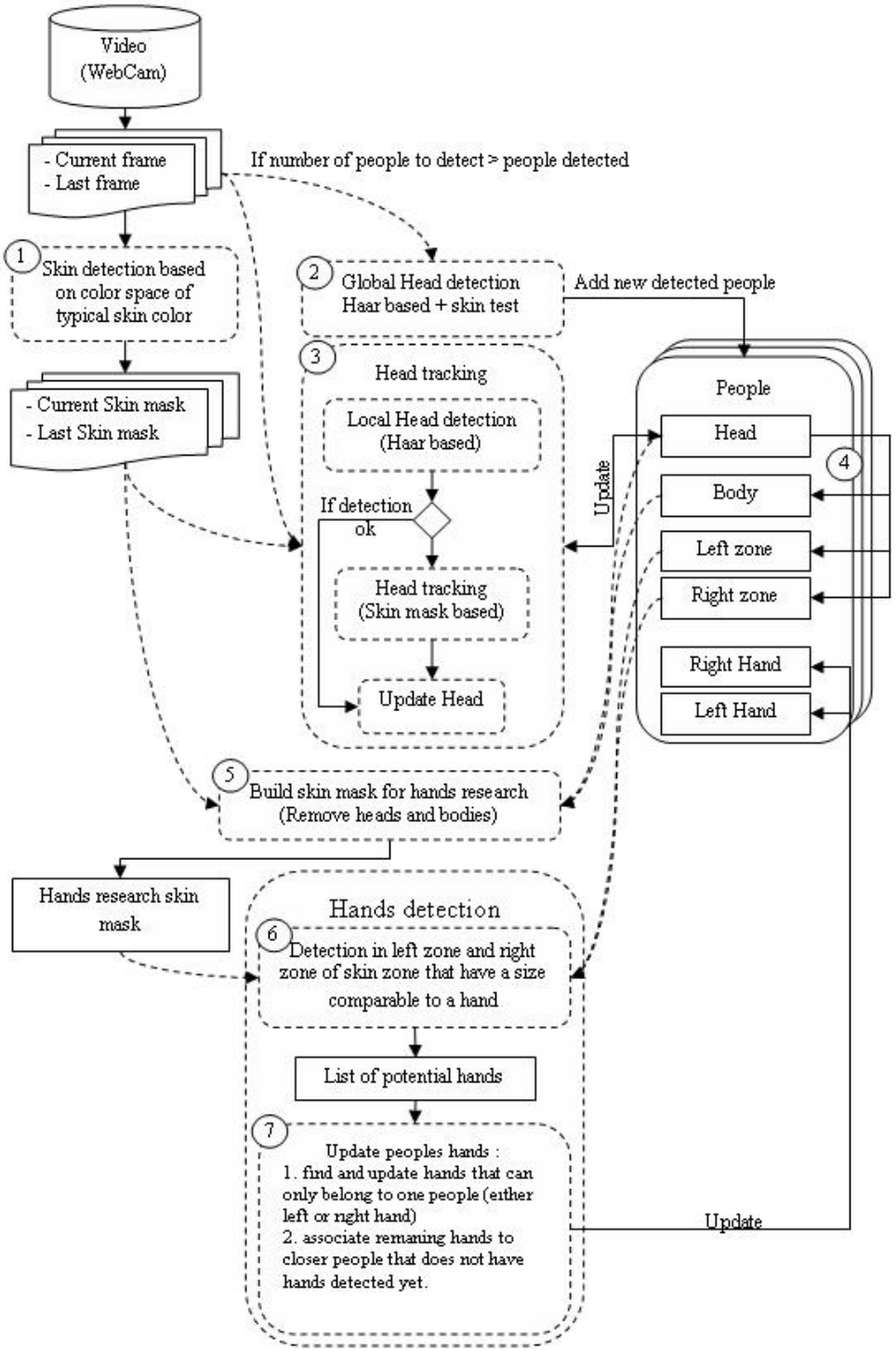


Figure 10: *Schema of the multiple-participant hand raising algorithm.*

7.1.8. Conclusion

Just count the number of people that have at least one hand raised.

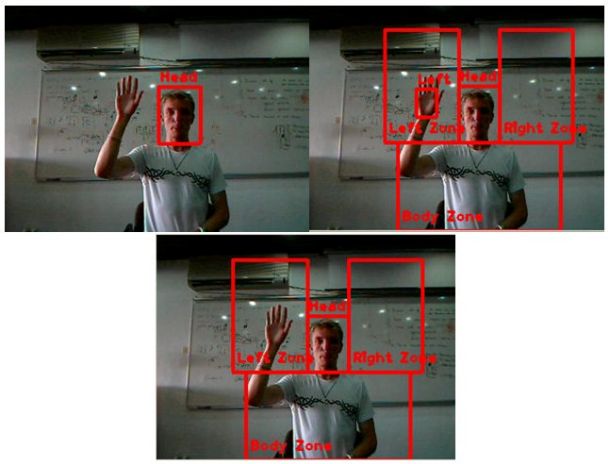


Figure 11: *Single-user: step 2(head), step 4(Research zones) and step 7(Hands).*

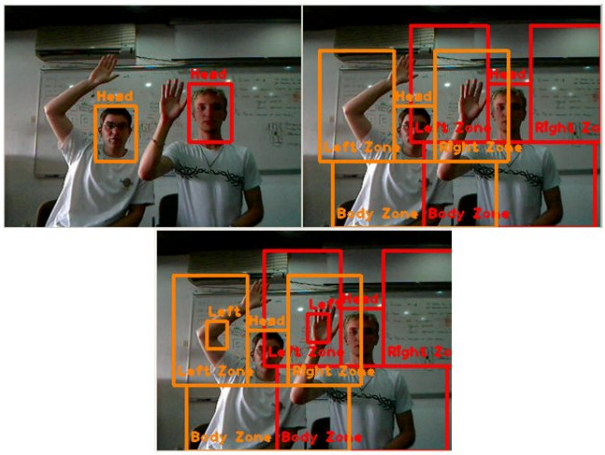


Figure 12: *Multi-user: step 2(head), step 4(Research zones) and step 7 (Hands).*

7.2. Experimental Results

7.2.1. Single person tests

The proposed method was first tested with only one person on the scene. The tracking/detection system works well except if the lighting conditions are extreme. For instance, if the illuminating light is blue, then the skin color detection does not recognize enough skin to detect a hand/head. One possible way to be more robust to color light illumination is to improve the skin color detection by using an adaptive skin detection base on head skin color.

For multiple users, the results are positive, but the system is assuming that each user is raising only one hand, and occlusions are not yet perfectly manage (if 2 hands are really close together then the system will detect only one hand). Nevertheless, the algorithm is able to assign the hands to their owner even if the search zone is shared.

8. SPEECH MODALITY

8.1. Overview of State of the Art

8.1.1. Speech Synthesis

For a large number of years, speech synthesis has been performed through concatenation of units (generally diphones) stored in a large corpus: for each unit, target duration and pitch patterns were computed, and the unit in the database that was closest to these targets while ensuring smooth transitions with its neighbours was selected [33] [34]. This approach enabled to improve speech synthesis a lot. Nevertheless, the generated voices suffer from a lack of naturalness because of discontinuities when the utterance to be played is far from all the targets in the database [35].

In the last years, there has been an increasing interest for statistical methods, enabling a better generalization to units far from the training set. In this approach, speech synthesis is performed through the use of Hidden Markov Models (HMM). To improve the synthesis, classification techniques such as decision trees can even be coupled to these methods, to select different HMMs according to the context and to predict events like pauses [36]. This approach already yielded promising results and much research is done to further improve them.

Since it is possible to synthesize in a (sometimes) non-human but understandable way any sentence for several years, current research is mainly focusing on adding naturalness to the synthesis: avoiding disturbing discontinuities, generating human-like prosody and even expressing emotions.

8.1.2. Speech recognition

For speech recognition, HMMs have been quite utilized. They enable to represent the sequences of units (generally phonemes, which are each associated to a small HMM) and their transition probabilities [35]. HMMs are usually coupled to Gaussian Mixture Models or Multi-Layer Perceptrons to model the emission probabilities for each state of the HMM [33] [36].

The most important problems encountered in speech recognition are the adaptation to the speaker (ideally the system should perform equally with any speaker, but performance is increased when the speaker was involved in training the system), the size of the vocabulary (the recognition task is easier when the vocabulary that can be used by the speaker is limited), the processing of continuous speech (with hesitations, coughs, laughers, influence of emotions,... as opposed to isolated word recognition) and the robustness to noise (recognition is harder in a noisy real-life environment than in a quiet lab) [37].

Current speech recognizers achieve very good results in the best conditions, with errors rates close to zero ($< 1\%$ [38]).

8.2. Speech System Evaluation

We focused our study on systems freely available for us to perform research and development, thus discarding powerful commercial solutions like Loquendo for speech synthesis [36] and Dragon Naturally Speaking for speech recognition [38].

We tested two speech synthesis systems: Festival [39] and Microsoft Speech SDK [40]. Festival gives slightly better voices, but it is not straightforward to integrate it in our project. Due to the limited time we disposed, we thus decided to keep Microsoft Speech SDK to perform the speech synthesis for this project, but we made our code modular so that anyone can replace it, for example with Festival.

For speech recognition, we had access, for research only, to a high-performing grammar-based recognizer named EAR and

based on STRUT (Speech Training and Recognition Toolkit) [41], both developed by the Faculty of Engineering, Mons, and Multitel ASBL [42], while Acapela [43] holds the commercial rights. Among others, EAR gives the possibility to perform recognition between keywords, with is sufficient for our voting application.

8.3. Proposed Method

8.3.1. Speech Synthesis

Like Festival, Microsoft Speech SDK offers the possibility to do the full text-to-speech process: the text given as input is transformed into a series of units to play (with information about duration, pitch, etc.) and then this series of units is synthesized. We used it in this very simple way, without exploiting the possibilities to customize the voice (changing the pitch, word-emphasis, etc.). Unfortunately, very little information about the techniques used by Microsoft Speech SDK to generate the speech waves is available.

8.3.2. Speech Recognition

As said earlier, EAR is a STRUT-based SDK. EAR enables to use STRUT to do online speech recognition (sending the sound buffers in real-time and not waiting for the sound file to be entirely recorded before starting the recognition). STRUT provides a grammar-based recognizer in the classical way described in the state of the art: each word of the vocabulary is mapped to a sequence of phonemes and Multi-Layer Perceptrons (MLP) are used to compute the acoustic emission probabilities of each of the phonemes. When a new acoustic realization is presented, it is decomposed in 30ms vectors which are given to the HMM-MLP combination. The most likely utterance is then selected (it is also possible to look at the second, third,... most probable utterances but this feature was not used in this project). A confidence score is associated to the recognized utterance. We have used this confidence score to process the results.

To control the audio acquisition, we used the Portaudio audio library [44]. Portaudio is used to fill in audio buffers which are given to EAR for word recognition.

STRUT is a grammar-based recognizer and thus needs a grammar, consisting of the list of recognizable words together with their phonetic transcriptions. This grammar enables STRUT to build the HMMs. We wanted the grammar to be automatically generated. Therefore, we developed a method that takes the options of the vote and looks for their phonetic transcription in a phonetic dictionary to build the grammar. We used the BEEP phonetic dictionary of Cambridge University (freely available for research only) [45], slightly modified to be in accordance with the SAMPA phonetic alphabet [46], which was used to build the MLPs of STRUT.

8.3.3. Architecture of our system:

Our system works according to Fig. 13, where a large dotted arrow indicates a time dependence (the latter component does not per se need the former one to have finished its task to accomplish its job, but it waits for it before starting):

1. The question and options are extracted from the Vote object
2. The recognition grammar containing the options is automatically generated
3. The speech recognition devices are initialized (in our case Portaudio and EAR), so that the speech recognition will be ready to start immediately after the user has heard the question
4. The question and vote options are put together with some wrappers ("The options are:", "Please vote now") in a string which is synthesized (in our case by Microsoft Speech SDK)
5. Audio acquisition is launched for a few seconds and speech recognition is performed
6. According to the recognition confidence score, a decision is taken to validate the vote and close the process or
7. To get back to the user to increase the confidence score. It is then decided either to ask the user to repeat his choice or to confirm, answering yes or no, that the recognition was correct. A new round of synthesis and recognition (either on the initial vocabulary or after switching to a binary yes-no choice) is launched. The language interaction with the participant has been designed very similar to the text-based vote user-computer interaction.

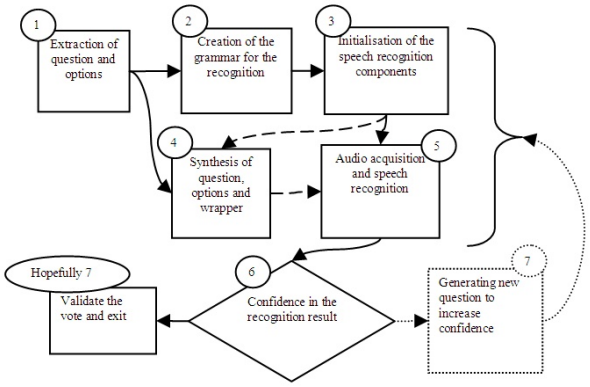


Figure 13: *Architecture of our audio processing system.*

8.4. Experimental Results

In order to have a qualitative evaluation of the performance of the complete system we conducted some experiments over the speech recognition system as we consider it to be the most critical point in terms of performance. We have not evaluated the user likeness of the system by analyzing the voice synthesis or the dialog interaction.

In the experiment, we used a restricted vocabulary (5 words: car, plane, train, boat and bus). We asked 12 different speakers to choose one of the words and we analyzed the recognition in a silent environment, first, and then with music in the background to simulate ambient noise.

We obtained the following results:

- Few mistakes in a quiet environment (7%), as expected some more with the added noise (18%)
- Mistakes are in 85% accompanied with a medium (<0.5) or small (<0.1) confidence level, so it should possible to correct these errors via the confirmation question
- Most confusions between:
  - plane and train; or
  - boat and bus.

Noticeably, words whose phonemes are very similar resulted in pronunciations that were confused. We must point out that most of the people involved in the experiment were not native speakers, which make us think that this system is robust enough to

be integrated in a general international communications framework.

8.5. Future work

The main limitation of the current system is that it only recognizes single words. Indeed, we have implemented EAR and the automatic generation of the grammar to distinguish only single words. This is of course a restrictive aspect, even if any vote process can be reduced to a single word distinction (the easiest way is to associate a number to each option).

However, this should ideally be done automatically and this leads us to the second point where there is clearly room for improvement: developing "artificial intelligence" to manage the voting process (creating the most appropriate recognition vocabulary in relation to the options that could then contain several words, increasing the influence of the past in the decisions, etc.).

Furthermore, none of the solutions we used is perfect or will remain state of the art across the years. This is why we tried to make our solution as modular as possible so that each component can be easily replaced or updated.

Finally, incorporating new languages would only require slight changes.

9. PROTOTYPE IMPLEMENTATION FOR THE SIMULATION

The work presented in this report comprised two different tasks: first, we wanted to analyze and do some research on how to make services for remote meetings multimodal and a reality in Next Generation Networks; second, we wanted to build a simulation where, building on top of current communications infrastructure, we could test the feeling of utilizing such a service. This section covers the description of the prototype implementation.

9.1. Network Simulation

We have implemented a client/server multithread Socket class that will provide threading support which handles the socket connection [47] and disconnection to a peer. This class was built to have a reliable communication between two peers supported with TCP/IP with error handling.

In this class we want to have event detection like: connection established, connection dropped, connection failed and data reception.

The protocol we used in the case of connection oriented (TCP) is the following:

1. Server:
    - Create endpoint
    - Bind address
    - Specify queue
    - Wait for connection
    - Transfer data
2. Client:
    - Create endpoint
    - Connect to server
    - Transfer data

For the transfer of data we use XML files with all the information related to the context and the commands. We have chosen to use the XML files because we can easily enrich it as we develop more ambitious services. The information from the XML files is processed with an XML parser to extract the context information and the commands with all the necessary parameters. Thanks to these data the application develops context awareness part.

9.2. Text and Speech Communication Simulation

Initially, Next Generation Networks will run on top of packet networks so, regardless of the origin of the communication link, at some point, the information is susceptible of going through the Internet. Therefore, to simulate telephone and text interaction we considered using an available commercial (almost standard) VoIP (voice over IP) and instant messaging application a good way to establishe the communication link between participants. After a quick analysis of the systems available for exploitation, we decided to run the simulation over Skype. Its freely available API allowed us to develop a solution that automatically starts a chat or a call using their network. It was very suitable for our initial tests and adapting to other commercial applications or developing our own remains future work.

Skype is a peer-to-peer application [48]. As we pointed in Section 3, communications and meeting service providers can, and from our perspective should, be independent; thus a server/-client architecture to run the service is not incompatible to utilize any other kind of network strategy for communications.

9.3. Videoconference Simulation

The most recent innovations in video-conferencing technology involve making the overall experience more realistic. Video coding and specific algorithms deliver high sustained frame rates, sharp images, and smooth motion. Cameras can now emulate the human eye, through auto-focusing, auto-aperture, and automatic locating of the speaker. High-resolution displays which include CRT displays, plasma flat panels, and large LCD screens, serve to improve the natural feel of video-conferencing. Higher quality audio allows full-duplex transmission, stereo sound, and better sampling rates. Full-duplex also reduces latency, for better lip-synchronization. Microphones with incredible pickup ranges and sensitivities allow better focus on the speaker, and updated circuitry provides echo cancellation and background noise suppression for better quality.

9.3.1. Our solution for the video conference part

In order to develop our application, we have to make a few choices. What is the best camera to use for videoconferencing, and on which protocol should we start our application.

In order to make our system scalable and to avoid implementing video streaming over the Internet, we use IP cameras for videoconferencing. An IP camera is a camera with an IP address; it is a true networking device containing an embedded Operating System, it supports multiple users, and it can be viewed using a suitable Web browser. An IP camera does not require additional hardware to operate and therefore has the flexibility to be located anywhere within a network connection. In mostly all real-world applications there is a need for a stand-alone functionality.

From the IP cameras we acquire m-jpegs that are processed by motion detection, as detailed further in next paragraphs.

The image size, this depends on the resolution and the compression scheme used. An image of (352 x 288 or 352x240) that is compressed using M-JPEG is only about 4-10 Kbytes. Higher

resolution cameras that have a resolution of 1200 x 1024, create file sizes as large as 80Kbytes per frame. This can be improved by using MPEG4, so the compression is better by transferring only the difference between frames. But the frame size is not used in MPEG4 compression. Instead we estimate an average data rate based on the resolution, frame rate and expected activity the camera will see. There is a about 4 times improvement in compression using MPEG4.

M-JPEG is a video format that uses JPEG compression for each frame of video. M-JPEG (Motion JPEG) prioritizes image quality over frame rate, provides and supports very high resolution, has low latency (delay between actual activity and the presented video) and shows graceful degradation at bandwidth constraints and packet loss. M-JPEG guarantees a defined level of picture quality, which is vital in most image analysis applications. As M-JPEG is also a less complicated compression, the number of available third party applications available is higher than for MPEG4.

After deciding to use IP cameras, we determined the minimum data requirements to establish the videoconference; therefore we require information from and for the person that connects to the videoconference, such as name, type of camera, etc. In order to do this, the client sends an XML file to the server, a file which contains his personal data : first name, last name, the IP address of the network camera, resolution parameters(width, height) and the URL, an URL that is specific for different IP cameras.

The structure that the client sends, in XML, looks like the one below:

```
<first_name>Michael</first_name>
<last_name>Carrick</last_name>
<ip>193.230.3.58</ip>
<url>axis-cgi/mjpg/video.cgi</url>
<width>320</width>
<height>240</height>
```

Based upon the data from the clients' XML files, we create a webpage that displays the video from all network cameras used in the videoconference. The webpage is automatically reloaded (refreshed) after a certain period of time, so the information about the clients is being regularly updated.

9.3.2. Discussion

Currently the biggest drawback of using IP Cameras is the latency between the moment the frame is acquired and the moment it reaches the endpoint of the communication. Building specific channels to communicate (virtual circuits) would improve the performance, nevertheless no delivery is guarantee, and no error correction or compression beyond JPEG is done, so although scalability is feasible in a easy way, practical implementations of large system would lead to saturated networks.

Optimal video streaming on different channels is a very active topic of research. IP cameras have been fine solution for simulation purposes, streamed video promises to enable high definition (HD) videoconference over packet networks in the future (as it is almost a reality for HD TV over IP).

10. CONCLUSION

We have presented one-month work related to the project Advanced Multimodal Interfaces that has allowed our team to experiment with the current state of the art in multimodal signal processing to develop services for synchronous communications in a novel way. The service chosen to exemplify the novel concepts exposed in this report was a vote registration for automatically counting votes in remote meetings.

The tests performed in this project have not been exhaustive; nevertheless, the experimental results obtained allow researches to have a first "feeling" of how reliable and how far from reality integrating multimodal signal processing into remote communications is.

From our study we can conclude that basic multimodal services can be implemented with very reasonable level of reliability. Real-time multimedia signal processing on semi-controlled situations (like vote registration in meetings) performs quite well. Nevertheless, integration of multimodal services in current networks is extremely complicated. For a multimodal service to be fully deployed, we must adapt them to many platforms, standards, etc. We really expect NGN to ease integration as a key point to enable multimodal signal processing to enhance communications.

Future research will also have to focus on building human-computer interaction techniques to compensate for not 100% reliable interfaces (NLP, AI, etc.) Multimedia analysis is not an exact science and services will never become 100% reliable. The same way humans are error prone in doing actions like counting votes during a meeting and need to verify and follow a protocol to make sure that no error is made; researchers will have to build protocols for interaction in service management to create fully automatic systems.

11. ACKNOWLEDGEMENTS

This report, as well as the source code for the software developed during the project, is available online from the eNTERFACE'07 web site: http://www.enterface.net.

Corresponding author A. C. Andrés del Valle is a researcher at the Accenture Technology Labs in Sophia Antipolis, France. (e-mail: ana.c.andresdelvalle@accenture.com).

12. REFERENCES

[1] J. Wilcox, *Videoconferencing. The whole picture*. N.Y.: CMP, 3 ed., 2000. ISBN: 1-57820-054-7. 11

[2] "AMI - Augmented Multiparty Interaction", 2007. FP6 EU Project. http://www.amiproject.org. 11

[3] "CHIL - Computers in the Human Interaction Loop", 2007. FP6 EU Project. http://www.chilserver.de. 11

[4] A. Nihjolt and H. J. A. op den Akker, "Meetings and meeting modeling in smart surroundings", in *Workshop in Social Intelligence Design*, pp. 145–158, 2004. 11

[5] "ITU - International Telecommunication Union", 2007. http://www.itu.int/ITU-T/ngn/index.phtml. 11, 12

[6] "ETSI - European Telecommunication Standards Institute", 2007. http://www.etsi.org/. 11, 12

[7] A. C. Andrés del Valle and M. Mesnage, "Making meetings more efficient: towards a flexible, service oriented framework for ubiquitous technologies", in *Workshop of the 9th International Conference on Ubiquitous Computing*, (Innsbruck, Austria), Sept 9-16 2007. 12

[8] J. Sienel, D. Kopp, and H. Rössler, "Multimodal Interaction for Next generation Networks", in *W3C Wokshop on Multimodal Interaction Activity*, (Sophia Antipolis, France), July 2004. 12

[9] M. Carugi, B. Hirschman, and A. Narita, "Introduction to the ITU-T NGN focus group release 1: target environment, services, and capabilities", *IEEE Communications Magazine*, vol. 43, no. 10, pp. 42–48, 2005. 13

[10] J. P. G. Sterbenz, "Peer-to-peer vs the Internet: a discussion of the proper and practical location of functionality", in *Dagshul seminar on service management and self-organization in IP-based networks*, 2005. 14

[11] Basic Networking, "Client-Server vs Peer-to-Peer", 2007. http://www.enterprise-technology.net/network2.htm. 14

[12] McGarthwaite, "Client-server versus peer-to-peer architecture: comparisons for streaming video", in *5th Winona Computer Science undergraduate research seminar*, 2007. 14

[13] "Client-Server versus Peer-to-Peer networking", 2007. http://www.extreme.net.au/Network/server.asp. 14

[14] W. Knight, "Video search makes phone a 'second pair of eyes'", *New Scientist*, October 2007. 14

[15] P. Lu, X. Huang, X. Zhu, and Y. Wang, "Head Gesture Recognition Based on Bayesian Network", in *Iberian Conference on Pattern Recognition and Image Analysis*, p. 492, 2005. 15

[16] N. P. Chi and L. C. D. Silva, "Head gestures recognition", in *Proceedings of International Conference on Image Processing*, vol. 3, pp. 266–269, 2001. 15

[17] A. Benoit and A. Caplier, "Head nods analysis: interpretation of non verbal communication gestures", in *International Conference on Image Processing*, vol. 3, pp. 425–428, 2005. 15

[18] A. Madabhushi and J. K. Aggarwal, "Using head movement to recognize activity", in *Proceedings of 15th International Conference on Pattern Recognition*, vol. 4, pp. 698–701, 2000. 15

[19] K. Toyama, "Look, Ma–No Hands! Hands free cursor control with real-time 3D face tracking", in *Workshop on Perceptual User Interface*, 1998. 15

[20] S. Kawato and J. Ohya, "Real-time detection of nodding and head-shaking by directly detecting and tracking the "between-eyes"", in *4th International Conference on Automatic Face and Gesture Recognition*, pp. 40–45, 2000. 15

[21] A. Kapoor and R. Picard, "A real-time head nod and shake detector", in *Workshop on Perspective User Interfaces*, 2001. 15

[22] V. Chauhan and T. Morris, "Face and feature tracking for cursor control", in *12th Scandinavian Conference on Image Analysis*, 2001. 16

[23] H. Pengyu and T. Huang, "Natural Mouse - a novel human computer interface", in *International Conference on Image Processing*, vol. 1, pp. 653–656, 1999. 16

[24] S. C. Ong and S. Ranganath, "Automatic Sign Language Analysis: A Survey and the Future beyond Lexical Meaning", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 6, pp. 873–891, 2005. 16

[25] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features", in *Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 1, pp. 511–518, 2001. 16

[26] "Open CV". http://www.intel.com/technology/computing/opencv/. 16

[27] C. Harris and M. Stephens, "A Combined Corner and Edge Detector", in *4th Alvey Vision Conference*, pp. 147–151, 1988. 16

[28] B. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision", in *DARPA Image Understanding Workshop*, pp. 121–130, 1981. 16

[29] A. D. Gavrila, "The visual analysis of human movement: a survey", *Computer Vision and Image Understanding*, vol. 8, pp. 82–98, January 1999. 18

[30] V. Pavlovic, R. Sharma, and T. Huang, "Visual Interpretation of Hand Gestures for Human-Computer Interaction: A Review", *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 8, pp. 677–695, July 1997. 18

[31] M. Chen, "Achieving effective floor control with a low-bandwidth gesture-sensitive videoconferencing system", in *Tenth ACM international Conference on Multimedia*, pp. 476–483, 2002. 18

[32] M. A. Akbari and M. Nakajima, "A novel color region homogenization and its application in improving skin detection accuracy", in *3rd international Conference on Computer Graphics and interactive Techniques in Australasia and South East Asia*, pp. 269–272, 2005. 18

[33] M. Rajman and V. Pallota, eds., *Speech and Language Engineering (Computer and Communication Sciences)*. Marcel Dekker Ltd, February 2007. 20

[34] P. Rutten, G. Coorman, J. Fackrell, and B. V. Coile, "Issues in corpus based speech synthesis", in *IEEE Seminar on State of the Art in Speech Synthesis*, 2000. 20

[35] A. Auria, *HMM-based speech synthesis for French*. Master thesis, Faculté Polytechnique de Mons and Universitat Politècnica de Catalunya, 2007. 20

[36] "Loquendo, Vocal Technologies and Services", 2007. http://www.loquendo.com. 20

[37] J. Markhoul and R. Schwartz, "State of the art in continous speech recognition", in *National academy of science*, pp. 9956–9963, 1995. 20

[38] Nuance Communications, "Dragon Naturally Speaking 9", 2007. http://www.nuance.com/talk. 20

[39] "Festival Speech Recognition System", 2007. http://www.csrt.ed.ac.uk/projects/festival. 20

[40] "Microsoft Speech SDK 5.1", 2007. http://www.microsoft.com/downloads. 20

[41] "STRUT project". TCTS Lab, Faculté Polytechnique de Mons, 2007. http://www.tcts.fpms.ac.be/asr/project/strut/. 21

[42] "Multitel ASBL", 2007. http://www.multitel.be. 21

[43] "Acapela Group", 2007. http://www.acapela-group.com/. 21

[44] P. Burk, "Portaudio. Portable cross-platform audio API", 2007. http://www.portaudio.com. 21

[45] A. Hunt, "British English example prononciation (BEEP)", 1997. http://www.eng.com.ac.uk/comp.speech/Section1/Lexical/beep.html. 21

[46] J. Wells, "Speech assessment methods phonetic alphabet (SAMPA)", 2005. http://www.phon.ucl.ac.uk/home/sampa. 21

[47] "Socket Library Functions". http://www.cisco.com/univercd/cc/td/doc/product/software/ioss390/ios390sk/sklibfun.htm. 22

[48] S. A. Baset and H. Schulzrinne, "An analysis of the skype peer-to-peer internet telephony protocol", Tech. Rep. 39 - 2004, Dpt. Of Computer Science at Columbia University, NY, US, 2004. 22

13. APPENDICES

13.1. About current video conferencing systems

Although it has not been part of the project, we would like to include here a brief discussion of what is actually the current start of the art regarding video-conferencing.

A **video conference**, by definition is a set of interactive telecommunication technologies which allow two or more locations to interact via two-way video and audio transmissions simultaneously. Video conferencing uses telecommunications of audio and video to bring people at different sites together for a meeting. Besides the audio and visual transmission of people, video conferencing can be used to share documents, computer-displayed information, and whiteboards.

13.1.1. Challenges of the real-time video conference

Multimedia networking faces many technical challenges like real-time data over non-real-time network, high data rate over limited network bandwidth, unpredictable availability of network bandwidth. They usually require much higher bandwidth than traditional textual applications. The basis of Internet, TCP/IP and UDP/IP, provides the range of services needed to support both small and large scale networks.

This type of multimedia application requires the real-time traffic which is very different from non-real-time data traffic. If the network is congested, real-time data becomes obsolete if it doesn't arrive in time.

Unfortunately, bandwidth is not the only problem. For most multimedia applications, the receiver has a limited buffer, if the data arrives too fast, the buffer can be overflowed and some data will be lost, also resulting in poor quality.

Therefore, considering that a videoconference is a real-time application, it is necessary to have the proper protocol to ensure it.

13.1.2. The necessity of an adequate protocol

In information technology, a protocol consists in a set of technical rules for the transmission and receipt of information between computers. A protocol is the "language" of the network; a method by which two dissimilar systems can communicate. We find protocols in different levels of a telecommunication connection. In practice, there are many protocols, each one governing the way a certain technology works.

For example, the IP protocol defines a set of rules governing the way computers use IP packets to send data over the Internet or any other IP-based network. Moreover, it defines addressing in IP.

13.1.3. Solutions to real-time challenges

From a network perspective, video conferencing is similar to IP telephony, but with significantly higher bandwidth requirements. In practice, the bandwidth requirement for one interactive video conference is in the range of 300 Kbps to 4 Mbps, which includes the audio, video and control signaling. Ultra high-definition telepresence applications can require as much as 15 Mbps to 45 Mbps of bandwidth.

Therefore, the use of both full file transfer and TCP as a transfer protocol is clearly unsuitable for supporting video and audio. To truly support video and audio over the internet, one requires the transmission of video and audio on-demand, and in real time, as well as new protocols for real time data.

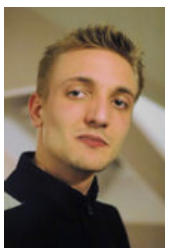
Therefore, protocols for real-time applications must be worked out before the real multimedia time comes. As a solution to this matter the Integrated Services working group in the Internet Engineering Task Force developed an enhanced Internet service model that includes best-effort service and real-time service. The Resource Reservation Protocol (RSVP), together with Real-time Transport Protocol (RTP), Real-Time Control Protocol (RTCP), Real-Time Streaming Protocol (RTSP), Session Initiation protocol (SIP) are used to provide applications with the type of service they need in the quality they choose.

13.2. Software details

Contents of the Source Code folders:

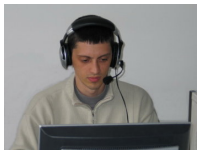
- **Amifc_audio_processing code**: this folder contains the source code of the speech registration vote. It does not include the sources related to the EAR software. If a license or access to these sources is wanted please contact Jérôme Urbain (jerome.urbain@fpms.ac.be)
- **Htm generation for IP cameras**: this folder contains the source code for the automatic generation of HML pages containing access to multiple IP cameras.
- **Risinghands detection**: this folder contains the source code for automatic vote registration over video with multiple people
- **Serversocket_demo**: this folder contains the source code for the server/client architecture of the vote registration system. It also includes the code for the automatic vote registration for the text modality. This latter integrated in the demo that it is run when:
  - Server starts the demo by sending Start in the text field
  - Client starts the automatic text vote registration by sending VM in the text field to the server.
- **Text vote registration**: this folder contains the source code for the class that wraps the question and the options for a vote. It is shared by all vote registration systems.

14. BIOGRAPHIES



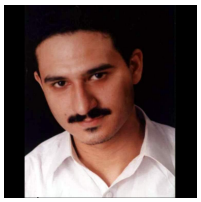
Jérôme Allasia was born in Clamart, France, in 1981. He received the M.Sc. degree in mathematics and computer science from the ENSEEIHT, Toulouse, in 2005. He is currently a research engineer at the IRISA Labs, Rennes, France. He works in projects related to Computer Vision, Image Processing, 3D modeling and network applied to highly streamable/scalable 3D video but also to compression, coding and distributed source coding.
Email: jerome.allasia@irisa.fr

Ana C. Andrés del Valle (M'04) was born in Barcelona, Spain, in 1976. She received the M.Sc. degree in telecommunications from the Technical University of Catalonia, Barcelona, in 1999 and the Ph.D. degree in image and signal processing from Tlcom Paris, Paris, France, in 2003. She is currently a Senior Research Specialist at the Accenture Technology Labs, Sophia Antipolis, France. She works in projects related to Computer Vision and Image Processing domains applied to advanced interfaces for medical purposes and the creation of novel multimodal services for synchronous communications. Prior to this position, she was a Project Leader at the VICOMTech Research Center, Spain. During her Ph.D. studies, she was a Research Engineer at the Eurecom Institut, France, and held an Invited Professor Fellowship from the University of the Balearic Islands, Spain. She began her work in research after being an intern for AT&T Labs-Research, Red Bank. She publishes regularly and has contributed to several books, such as the Encyclopedia of Information Science and Technology (Hershey, PA: Idea Group, 2005). Dr. Andrés del Valle was awarded the "2ème Prix de la Meilleure Thèse Telecom Valley" for the outstanding contributions of her Ph.D. work.
Email: ana.c.andresdelvalle@accenture.com



Dragoş Cătălin Barbu was born in Lehliu-Gara, Romania, in 1979. He received the B.Sc. degree in computer sciences from the Bucharest University, Faculty of Mathematics and Computer Sciences, in 2002 and the M.Sc. degree in theoretical computers science from Bucharest University, Bucharest, Romania, in 2004. He is currently a Research Specialist at the National Institute for Research and Development in Informatics, Bucharest, Romania. He works in projects related to Intelligent Agents in Virtual World and Neural networks.
Email: dbarbu@ici.ro

Ionut Petre was born in Bucharest, Romania, in 1981. He received the B.Sc. degree in communications from the Politehnica University of Bucharest, Faculty of Electronics, Telecommunications and Information Technology, in 2005 and he is a MS student in Medical Electronics and Informatics at Politehnica University of Bucharest.
Email: ipetre@ici.ro



Usman Saeed was born in Lahore, Pakistan in 1981. He received a BS in Computer System Engineering from GIK Institute (Topi, Pakistan) in 2004. After graduation he was associated with the Electrical Engineering dept. of Comsats Institute (Lahore, Pakistan) as a research associate. In 2005, he joined the University of Nice-Sophia Antipolis (Sophia Antipolis, France) for a Master of Research in Image Processing. He is currently a PhD student in the Multimedia Communication department of Institut Eurecom (Sophia Antipolis, France) under the supervision of Prof. Jean-Luc Dugelay. His current research interests focus on facial analysis in video.
Email: saeed@eurecom.fr

Jérôme Urbain was born in Brussels, Belgium, in 1984. He received the Electrical Engineering degree from the Faculty of Engineering, Mons (FPMs), in 2006. He is currently PhD student in the Circuit Theory and Signal Processing (TCTS) Lab of FPMs, where he is working on emotional speech recognition in the framework of EU FP6 Integrated Project CALLAS.
Email: jerome.urbain@fpms.ac.be

A MULTIMODAL FRAMEWORK FOR THE COMMUNICATION OF THE DISABLED

Savvas Argyropoulos [1], *Konstantinos Moustakas* [1], *Alexey A. Karpov* [2], *Oya Aran* [3], *Dimitrios Tzovaras* [1], *Thanos Tsakiris* [1], *Giovanna Varni* [4], *Byungjun Kwon* [5]

[1] Informatics and Telematics Institute (ITI), Aristotle University of Thessaloniki, Hellas, Greece
[2] Saint-Petersburg Institute for Informatics and Automation, Russian Academy of Sciences, Russian Federation
[3] Perceptual Intelligence Lab, Boğaziçi University, İstanbul, Turkey
[4] InfoMus Lab - Casa Paganini, DIST, University of Genoa, Italy
[5] Koninklijk Conservatorium, The Hague, Netherlands

ABSTRACT

In this paper, a novel system, which aims to provide alternative tools and interfaces to blind and deaf-and-mute people and enable their communication and interaction with the computer is presented. All the involved technologies are integrated into a treasure hunting game application that is jointly played by the blind and deaf-and-mute user. The integration of the multimodal interfaces into a game application serves both as an entertainment and a pleasant education tool to the users. The proposed application integrates haptics, audio, visual output and computer vision, sign language analysis and synthesis, speech recognition and synthesis, in order to provide an interactive environment where the blind and deaf-and-mute users can collaborate to play the treasure hunting game.

KEYWORDS

Multimodal interfaces – Multimodal fusion – Sign language analysis – Sign language synthesis – Speech recognition

1. INTRODUCTION

The widespread deployment of novel human-computer interaction methods has changed the way individuals communicate with computers. Since Sutherland's SketchPad in 1961 or Xerox' alto in 1973, computer users have long been acquainted with more than the traditional keyboard to interact with a system. More recently, with the desire of increased productivity, seamless interaction and immersion, e-inclusion of people with disabilities, and with the progress in fields such as multimedia/multimodal signal analysis and human-computer interaction, multimodal interaction has emerged as a very active field of research [1].

Multimodal interfaces are those encompassing more than the traditional keyboard and mouse. Natural input modes are employed [2], [3], such as voice, gestures and body movement, haptic interaction, facial expressions, and physiological signals. As described in [4], multimodal interfaces should follow several guiding principles. Multiple modalities that operate in different spaces need to share a common interaction space and to be synchronized. Also, multimodal interaction should be predictable and not unnecessarily complex, and should degrade gracefully, for instance by providing for modality switching. Finally, multimodal interfaces should adapt to user's needs, abilities, and the environment.

A key aspect in multimodal interfaces is also the integration of information from several different modalities in order to extract high-level information non-verbally conveyed by users. Such high-level information can be related to expressive, emotional content the user wants to communicate. In this framework, gesture has a relevant role as a primary non-verbal conveyor of expressive, emotional information. Research on gesture analysis, processing, and synthesis has received a growing interest from the scientific community in recent years and demonstrated its paramount importance for human machine interaction.

The present work aims to make the first step in the development of efficient tools and interfaces for the generation of an integrated platform for the intercommunication of blind and deaf-mute persons. It is obvious that while multimodal signal processing is essential in such applications, specific issues like modality replacement and enhancement should be addressed in detail.

In the blind user's terminal the major modality to perceive a virtual environment is haptics while audio input is provided as supplementary side information. Force feedback interfaces allow blind and visually impaired users to access not only twodimensional graphic information, but also information presented in 3D virtual reality environments (VEs) [5]. The greatest potential benefits from virtual environments can be found in applications concerning areas such as education, training, and communication of general ideas and concepts [6].

Several research projects have been conducted to assist visually impaired to understand 3D objects, scientific data and mathematical functions, by using force feedback devices [7]. PHANToM™ is one of the most commonly used force feedback device. Due its hardware design, only one point of contact at a time is supported. This is very different from the way that people usually interact with surroundings and thus, the amount of information that can be transmitted through this haptic channel at a given time is very limited. However, research has shown that this form of exploration, although time consuming, allows users to recognize simple 3D objects. The PHANToM™ device has the advantage to provide the sense of touch along with the feeling of force feedback at the fingertip.

Deaf and mute users have visual access to 3D virtual environments; however their immersion is significantly reduced by the luck of audio feedback. Furthermore effort has been done to provide applications for the training of hearing impaired. Such applications include the visualization of the hand and body movements performed in order to produce words in sign language as well as applications based on computer vision techniques that aim to recognize such gestures in order to allow natural human machine interaction for the hearing impaired. In the context of the presented framework the deaf-mute terminal incorporates sign-language analysis and synthesis tools so as to

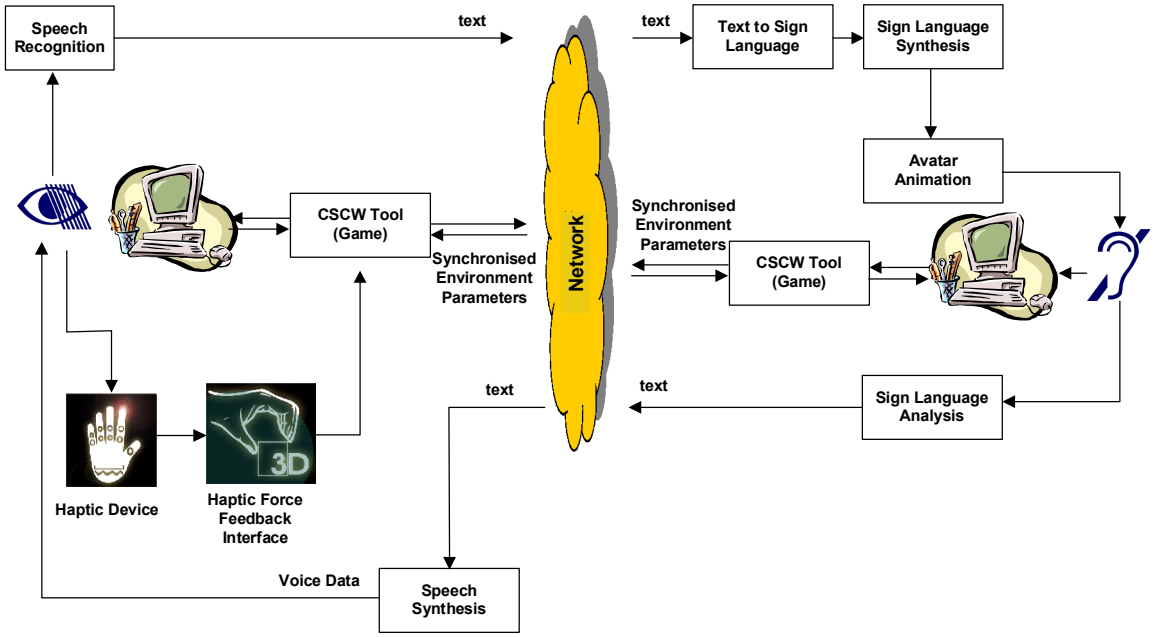


Figure 1: *Architecture of the collaborative treasure hunting game.*

allow physical interaction of the deafmute user and the virtual environment.

The paper is organized as follows: Section 2 presents the overall system architecture, and Section 3 describes the modality replacement framework. Subsequently, Sections 4 and 5 present the audio and visual speech recognition modules, respectively. In Section 6, the audio and visual multimodal fusion framework is described and the employed algorithms are analytically discussed. In the following, Section 7 presents the path sketching module using gesture recognition. Then, Section 8 presents the sign language recognition module. Finally, Section 9 presents the application scenario and conclusions are drawn in Section 10.

2. OVERALL SYSTEM DESCRIPTION

The basic development concept in multimodal interfaces for the disabled is the idea of *modality replacement*, which is defined as *the use of information originating from various modalities to compensate for the missing input modality of the system or the users*.

The main objective of the proposed system is the development of tools, algorithms and interfaces that will utilize modality replacement so as to allow the communication between blind or visually impaired and deaf-mute users. To achieve the desired result the proposed system combines the use of a set of different modules, such as

- gesture recognition
- sign language analysis and synthesis
- speech analysis and synthesis
- haptics

into an innovative multimodal interface available to disabled users. Modality replacement was used in order to enable information transition between the various modalities used and thus enable the communication between the involved users.

Figure 1 presents the architecture of the proposed system, including the communication between the various modules used for the integration of the system as well as intermediate stages used for replacement between the various modalities. The left part of the figure refers to the blind user's terminal, while the right refers to the deaf-mute user's terminal. The different terminals of the treasure hunting game communicate through asynchronous TCP connection using TCP sockets. The interested reader is referred to [8] for additional details.

The following sockets are implemented in the context of the treasure hunting game:

- SeeColor terminal: Implements a server socket that receives queries for translating color into sound. The code word consists of the following bytes, "b; R; G; B", where b is a boolean flag and R, G, B the color values.
- Blind user terminal: Implements three sockets:
  - A client socket that connects to the SeeColor terminal.
  - A server socket to receive messages from the deaf-mute user terminal.
  - A client socket to send messages to the deaf-mute user terminal.

The deaf-mute user's terminal implements:

- A server socket to receive messages from the blind user terminal.
- A client socket to send messages to the blind user terminal.

Also, file sharing is used to ensure consistency between the data used in the various applications.

3. MODALITY REPLACEMENT

The basic architecture of the proposed modality replacement approach is depicted in Fig. 2. The performance of such a system is directly dependent on the efficient multi-modal processing of two or more modalities and the effective exploitation of their complementary nature and their mutual information to achieve

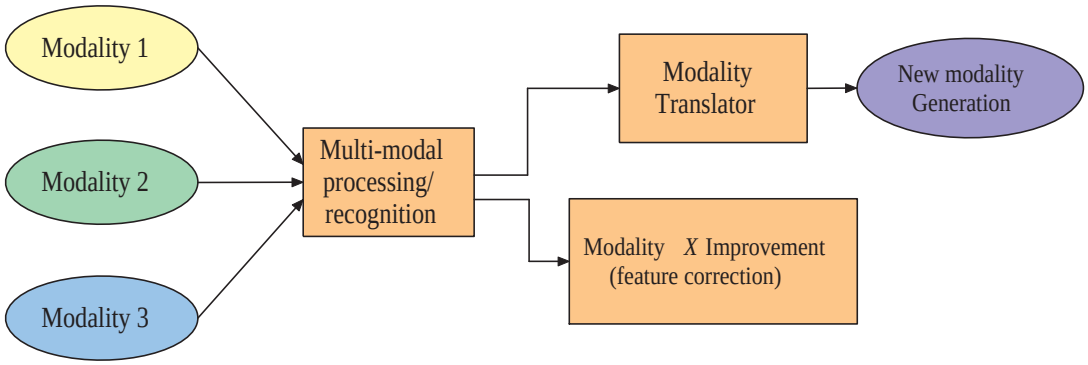


Figure 2: *The modality replacement concept.*

accurate recognition of the transmitted content. After the recognition has been performed effectively, either a modality translator can be employed in order to generate a new modality or the output can be utilized to detect and correct possibly erroneous feature vectors that may correspond to different modalities. The latter could be very useful in self-tutoring applications. For example, if an individual practices sign language, the automatic recognition algorithm could detect incorrect hand shapes (based on audio and visual information) and indicate them so that the user can identify the wrong gestures and practice more on them.

The basic idea is to exploit the correlation between modalities in order to enhance the perceivable information by an impaired individual who can not perceive all incoming modalities. In that sense, a modality, which would not be perceived due to a specific disability, can be employed to improve the information that is conveyed in the perceivable modalities and increase the accuracy rates of recognition. The results obtained by jointly fusing all the modalities outperform those obtained using only the perceived modalities since the inter- dependencies among them are modelled in an efficient manner.

A critical feature of the proposed system is its ability to adaptively assess the reliability of each modality and assign a measure to weight its contribution. There exist different approaches to measure reliability, such as taking into account the noise level of the input signal. The common way of incorporating these reliability values into decision fusion is to use them as weighting coefficients and to compute a weighted average [9]. The proposed scheme aims at maximizing the benefit of multimodal fusion so that the error rate of the system becomes than that of the cases where only the perceivable information is exploited. Modality reliability has also been examined in [10], in the context of multimodal speaker identification. An adaptive cascade rule was proposed and the order of the classifiers was determined based on the reliability of each modality combination.

A modified Coupled Hidden Markov Model (CHMM) is employed to model the complex interaction and interdependencies among the modalities and combine them efficiently in order to recognize correctly the transmitted message. In this work, modality reliability is regarded as a means of giving priority to single or combined modalities in the fusion process, rather than using it as a numerical weight. These issues are discussed in detail in the following sections and the unimodal speech recognition modules, based on audio and visual information, are described to illustrate how they can be combined.

4. AUDIO SPEECH RECOGNITION

Audio speech recognition is one part of the proposed audio-visual speech recognition interface intended for verbal human-computer interaction between a blind person and the computer. 16 voice commands were selected to be pronounced by the blind person. For the demonstration purposes one man was selected to show eyeless human-computer interaction so the automatic recognition system is speaker-dependent. All the voice commands can be divided into two groups: (1) communication with the game process; (2) eyeless interaction with GUI interface of the multimodal system, as illustrated in Table 1.

Voice command	Phonemic transcription	Interaction type
Catacombs	k a t a c o m s	game
Click	k l i k k l i	interface
Door	d o r	game
East	i s t i s	game
Enter	e n t e r	game
Exit	e g z i t e g z i	game
Go	g o u	game
Help	h e l p h e l e l	interface
Inscription	i n s k r i p s i o n	game
North	n o r s n o r	game
Open	o p e n	game
Restart	r i s t a r t r i s t a r	interface
South	s a u s s a u	game
Start game	s t a r t g e i m s t a r g e i m	interface
Stop game	s t o p g e i m	interface
West	u e s t u e s	game

Table 1: *Recognition vocabulary with phonemic Transcription*

HTK 3.4 toolkit [11] was employed to process speech signal. The audio signal is captured by microphone of webcamera Philips SPC900 and sampled at 11025 Hz with 16 bits on

each sample using a linear scale. Mel-frequency cepstral coefficients (MFCC) are computed for the 25 ms overlapping windows (frames) with 10 ms shift between adjacent frames. Audio speech recognizer system uses 12 MFCCs as well as an estimation of the first and second order derivatives that forms a vector of 36 components.

The acoustical modeling is based on Hidden Markov Models (HMMs) with mixture Gaussian probability density functions [12]. HMMs of phonemes have 3 meaningful states and 2 "hollow" states intended for concatenation of the models (Fig. 3). Each word of the vocabulary is obtained by concatenation of context-independent phonemes. The speech decoder uses Viterbi-based token passing algorithm [11]. The input phrase syntax is described in a simple grammar that allows recognizing only one command each time.

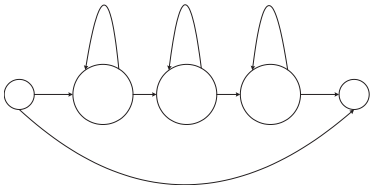


Figure 3: *Topology of the Hidden Markov Model for a phoneme.*

In order to train the speech recognizer an audio-visual speech corpus was collected in an auditorium room using USB webcamera Philips SPC900. 320 utterances were used for training HMMs of phonemes and 100 utterances for the testing purpose. The wave audio files were extracted from the training avi video files using the VirtualDub software.

After expert processing of the utterances, it was found that the SNR for audio signal is quite low (15-20 Db) because of far position (about 1 meter) of the speaker in front of the microphone and usage of the microphone built in a standard webcamera. Thus some explosive consonants (for instance "t" or "k") at the beginnings and endings of phrases are not identified in the wave files. In Table 1, some words have several different variants of transcriptions, it is explained by periodical loss of explosive consonants in the speech signal. 30 % training utterances were manually labeled on phonemes by the software WaveSurfer, and the remaining data were automatically segmented by the Viterbi forced alignment method [11].

The audio speech recognizer was compiled as dynamic link library ASR.dll, which is used by the main executable module that combines the modules for audio and visual speech recognition (Fig. 4).

The audio speech recognizer can work independently or jointly with the visual speech recognizer. In the on-line operation mode the audio speech recognizer uses an energy-based voice activity detector to find the speech frames in audio signal. When any speech activity is found the module sends the message `WM_STARTSPEECH` to the window of the main application as well as when speech frames are changed by pause frames the message `WM_ENDSPEECH` is sent. After receiving one of the messages the visual recognizer should start or finish the video processing, correspondingly. The audio speech recognizer operates very fast so the result of speech recognition will be available almost immediately after the message `WM_ENDSPEECH`. Moreover, the MFCC features, calculated while processing speech, are stored in an internal buffer and can be transferred to the visual speech recognizer in order to fuse these parameters with visual parameters of the lips region.

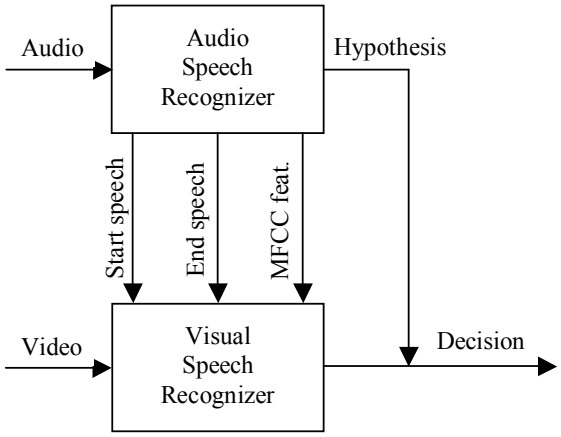


Figure 4: *General data flow in audio-visual speech recognition system.*

5. VISUAL SPEECH RECOGNITION

For the lip shape modality, the robust location of facial features and especially the location of the mouth region is crucial. Then, a discriminant set of visual observation vectors have to be extracted. The process for the extraction of the lip shape is presented in [13], and is described in brief below so that the paper is self-contained.

Initially, the speaker's face is located in the video sequence as illustrated in Fig. 5. Subsequently, the lower half of the detected face is selected as an initial candidate of the mouth region and Linear Discriminant Analysis (LDA) is used to classify pixels into two classes: face and lip. After the lip region segmentation has been performed the contour of the lips is obtained using the binary chain encoding method and a normalized 64x64 region is obtained from the mouth region using an affine transform. In the following, this area is split into blocks and the 2D-DCT transform is applied to each of these blocks and the lower frequency coefficients are selected from each block, forming a vector of 32 coefficients. Finally, LDA is applied to the resulting vectors, where the classes correspond to the words considered in the application. A set of 15 coefficients, corresponding to the most significant generalized eigenvalues of the LDA decomposition is used as the lip shape observation vector.

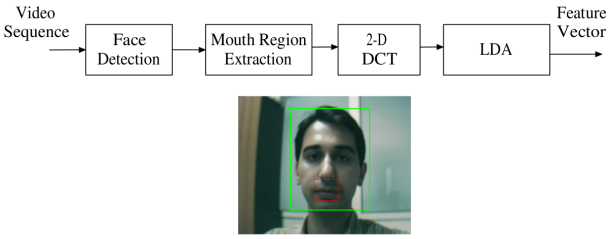


Figure 5: *Lip motion extraction process.*

6. AUDIO-VISUAL SPEECH RECOGNITION

6.1. Multimodal Fusion

The combination of multiple modalities for inference has proven to be a very powerful way to increase detection and recognition

performance. By combining information provided by different models of the modalities, weakly incorrect evidence in one modality can be corrected by another modality. Hidden Markov Models (HMMs) are a popular probabilistic framework for modelling processes that have structure in time. Especially, for the applications that integrate two or more streams of data, Coupled Hidden Markov Models (CHMMs) have been developed.

A CHMM can be considered as a collection of HMMs, one for each data stream, where the hidden backbone nodes at time t for each HMM are conditioned by the backbone nodes at time $t-1$ for all the related HMMs. It must be noted that CHMMs are very popular among the audio-visual speech recognition community, since they can model efficiently the endogenous asynchrony between the speech and lip shape modalities. The parameters of a CHMM are described below:

$$\pi_0^c(i) = P(q_t^c = i) \quad (1)$$
$$b_t^c(i) = P(\mathbf{O}_t^c | q_t^c = i) \quad (2)$$
$$a_{i|j,k}^c = P(q_t^c = i | q_{t-1}^A = j, q_{t-1}^L = k) \quad (3)$$

where q_t^c is the state of the coupled node in the c_{th} stream at time t, $\pi_0^c(i)$ is the initial state probability distribution for state i in c_{th} stream, $\mathbf{O}_t^c$ is the observation of the nodes at time t in the c_{th} stream, $b_t^c(i)$ is the probability of the observation given the i state of the hidden nodes in the c_{th} stream, and $a_{i|j,k,n}^c$ is the state transitional probability to node i in the c_{th} stream, given the state of the nodes at time $t-1$ for all the streams. The distribution of the observation probability is usually defined as a continuous Gaussian Mixture. Fig. 6 illustrates the CHMM employed in this work. Square nodes represent the observable nodes whereas circle nodes denote the hidden (backbone) nodes.

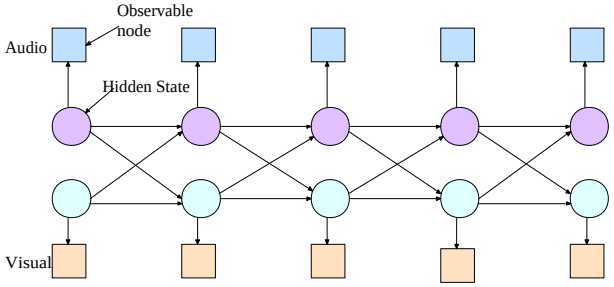


Figure 6: *Coupled HMM for audio and visual information fusion.*

One of the most challenging tasks in automatic speech recognition systems is to increase robustness to environmental conditions. Although the stream weights needs to be properly estimated according to noise conditions, they can not be determined based on the maximum likelihood criterion. Therefore, it is very important to build an efficient stream weight optimization technique to achieve high recognition accuracy.

6.2. Modality Reliability

Ideally, the contribution of each modality to the overall output of the recognition system should be weighted according to a reliability measure. This measure denotes how each observation stream should be modified and acts as a weighting factor. In general, it is related to the environmental conditions (e.g., acoustic noise for the speech signal).

The common way of incorporating these reliability values into decision fusion is to use them as weighting coefficients and to compute a weighted average. Thus, the probability $b_m(\mathbf{O}_t)$ of a feature $\mathbf{O}_t$ for a word m is given by:

$$b_m(\mathbf{O}_t) = w_A \cdot b_A(\mathbf{O}_t^A) + w_L \cdot b_L(\mathbf{O}_t^L)) \quad (4)$$

where $b_A(\mathbf{O}_t^A)$, and $b_L(\mathbf{O}_t^L)$ are respectively the likelihoods for an audio feature $\mathbf{O}_t^A$ and a lip shape feature $\mathbf{O}_t^L$. The parameters w_A and w_L are audio and lip shape weights, respectively, and $w_A + w_L = 1$.

In the proposed method, a different approach is employed to determine the weights of each data stream. More specifically, for each modality, word recognition is performed using a HMM for the training sequences. The results of the (unimodal) word recognition indicate the noise levels in each modality and provide an approximation of their reliability. More specifically, when the unimodal HMM classifier fails to identify the transmitted words it means that the observation features for the specific modality are unreliable. On the other hand, a small word error rate using only one modality and the related HMM means that the corresponding feature vector is reliable and should be favoured in the CHMM.

6.3. Word Recognition

The word recognition is performed using the Viterbi algorithm, described above, for the parameters of all the word models. It must be emphasized that the influence of each stream is weighted at the recognition process because, in general, the reliability and the information conveyed by each modality is different. Thus, the observation probabilities are modified as:

$$b_t^A(i) = b_t(\mathbf{O}_t^A | q_t^A = i)^{w_A} \quad (5)$$
$$b_t^L(i) = b_t(\mathbf{O}_t^L | q_t^L = i)^{w_L} \quad (6)$$

where w_A and w_L are respectively the weights for audio and lip shape modalities and $w_A + w_L = 1$. The values of w_A and w_L are obtained using the methodology of section 6.2.

7. PATH SKETCHING

In this revised version of the Treasure Hunting Game, the engagement of deaf-and mute players is improved by path sketching based on gesture modality. The user can interact with the interface by means of a gesture performed by his/her hand to navigate on the village map and to explore the main areas of this map (e.g., temple, catacombs). This real time navigation process is implemented in the three steps: Hand detection, trajectory extraction, and sketching.

7.1. Hand Detection

Detection and tracking was a non trivial step because occlusions can occur due to overlap of each hand on the other or of the other skin colored regions (e.g., face and harms). To solve this problem and to make the detection easier a blue glove was worn from the player. In this way, we could deal with detection and tracking of hand exploiting techniques based on color blobs.

The colored region is detected via the histogram approach as proposed in [14]. Double thresholding is used to ensure connectivity, and to avoid spikes in the binary image. The scheme is composed of training the histogram and threshold values for future use. To cancel the noise, we selected the largest connected component of the detected regions into consideration. Thus we had only one component identified as hand.

7.2. Hand Tracking and Trajectory extraction

The analysis of hand motion is performed by tracking the center of mass (CoM) and calculating the velocity of each segmented hand. However, these hand trajectories are noisy due to the noise introduced at the segmentation step. Thus, we use Kalman filters to smooth the obtained trajectories. The initialization of the Kalman Filter is done when the hand is first detected in the video. At each frame, Kalman filter time update equations are calculated to predict the new hand position. The hand position found by the hand segmentation is used as measurements to correct the Kalman Filter parameters. Posterior states of each Kalman filter is defined as feature vectors for x, y coordinates of CoM and velocity. The hand can be lost due to occlusion or bad lighting in some frames. In that case, Kalman Filter prediction is directly used without correcting the Kalman Filter parameters. The hand is assumed to be out of the camera view if no hand can be detected for some number of (i.e. six) consecutive frames.

7.3. Sketching on the map

The extracted trajectory is then superimposed to the map, so that player can sketch directly the path on the map during the whole game.

Hand gestures performed by player were encoded with respect to the elongation of the hand. Positions of the hand were used as drawing controller, when player puts the hand in the vertical position with respect to the ground, drawing is enabled (Fig. 7a) and s/he can start to sketch trajectory on the map. When the hand is moved to the horizontal position with respect to the ground, the drawing is disabled (Fig. 7b). If the user moves her/his hand to the top left corner of the map, the drawing is deleted and the user may start from the beginning.

(a) (b)

Figure 7: *The detected binary hand. The elongation of the hand sets whether the drawing is (a) on or (b) off.*

The predefined locations on the map are used as the start and stop locations of the path. The drawing only starts when the user is around the starting position and the drawing ends when the path reaches to the stopping position (Fig. 8).

8. SIGN LANGUAGE RECOGNITION

Figure 9 depicts the steps in sign recognition. The first step in hand gesture recognition is to detect and track both hands. This is a complex task because the hands may occlude each other and also come in front of other skin colored regions, such as the arms and the face. To make the detection problem easier, we have used colored gloves worn on the hand (see Fig. 10). Once the hands are detected, a complete hand gesture recognition system

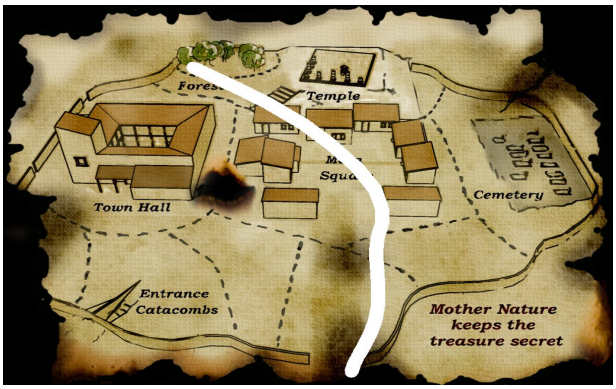


Figure 8: *The sketched trajectory on the map.*

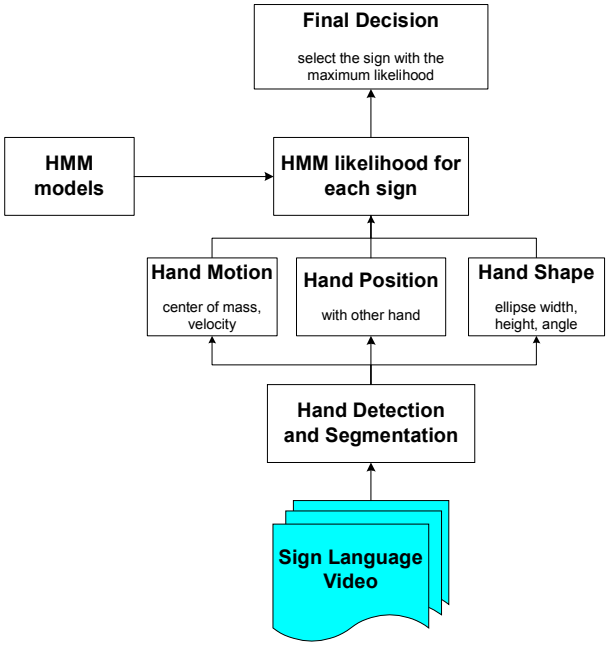


Figure 9: *Sign language recognition system block diagram.*

must be able to extract the hand shape, and the hand motion. We have extracted simple hand shape features and combined them with hand motion and position information to obtain a combined feature vector [15].

Our sign database consists of four ASL signs for directions: *north, south, east, and west*. For each sign, we recorded 15 repetitions from two subjects. The video resolution is 640*480 pixels and the frame rate is 25 fps. A left-to-right continuous HMM model with no state skips is trained for each sign in the database. For the final decision, likelihoods of HMM for each sign class are calculated and the sign class with the maximum likelihood is selected as the base decision.

9. APPLICATION SCENARIO

The aforementioned technologies where integrated in order to create an entertainment scenario. The scenario consists of seven steps. In each step one of the users has to perform one or more actions in order to pass successfully to the next step. The storyboard is about an ancient city that is under attack and citizens of the city try finding the designs in order to create high technology

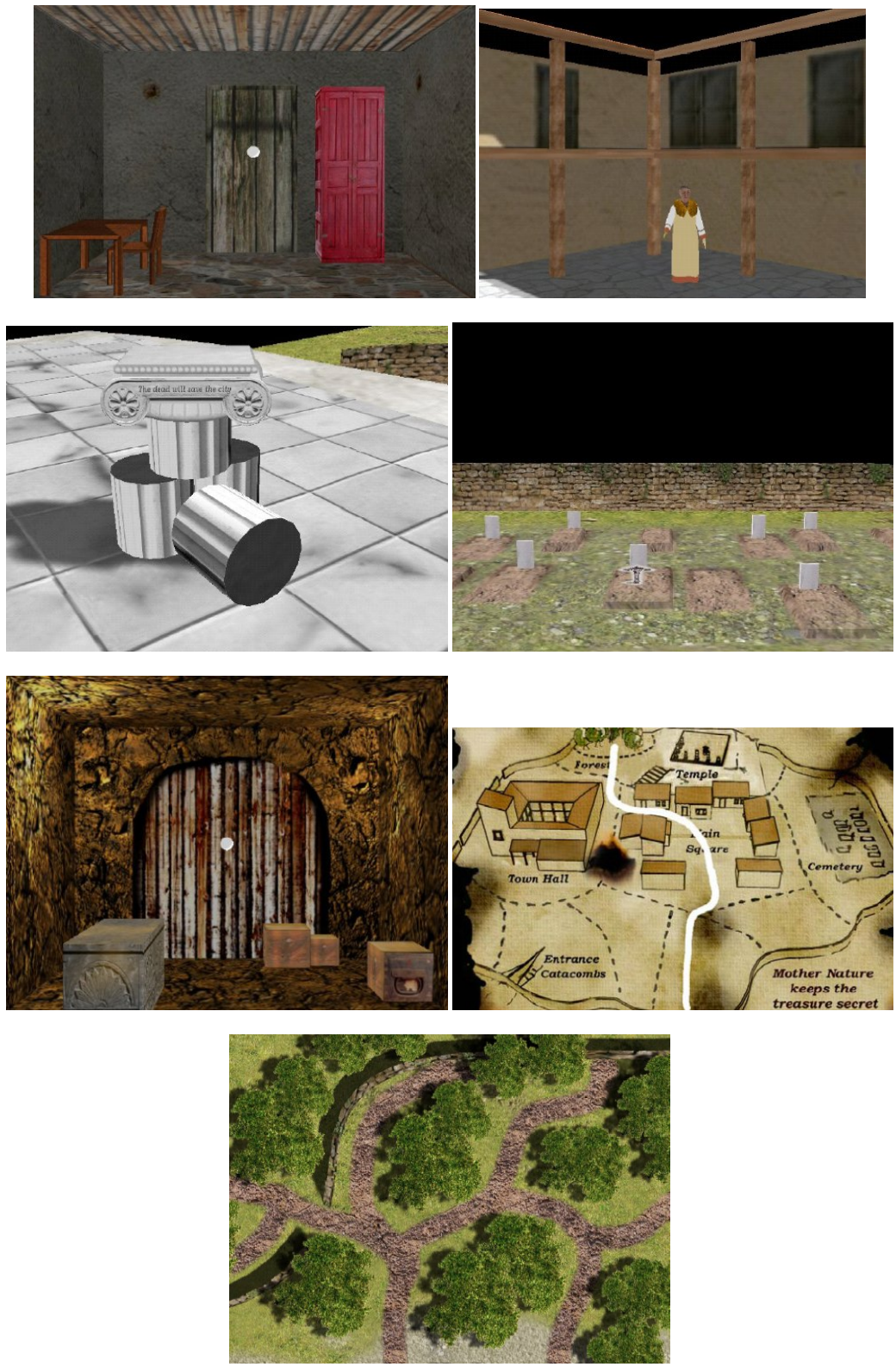


Figure 11: *The seven steps of the virtual game.*

Figure 10: *The user wearing colored gloves.*

war machines.

In the first step, the blind user receives an audio message and is instructed to "find a red closet". Subsequently, the blind user explores the village using the haptic device.

In the second step, the deaf-and-mute person receives the audio message which is converted to text using the speech recognition tool and then to sign language using the sign synthesis tool. Finally, the user receives the message as a gesture through an avatar, as depicted in Fig. 12. This message guides the deaf-and-mute user to the town hall, where the mayor provides the audio message "Go to the temple ruins".

The third step involves the blind user, who hears the message said by the mayor and goes to the temple ruins. In the temple ruins the blind user has to search for an object that has an inscription written on it. One of the columns has an inscription written on it that states, "The dead will save the city". The blind user is informed by an audio message whenever he finds this column and the message is sent to the deaf-mute user's terminal.

The fourth step involves again the deaf and mute user. The user receives the written text in sign language form. The text modality is translated to sign language symbols using the sign synthesis tool. Then the deaf and mute user has to understand the meaning of the inscription "The dead will save the city" and go to the cemetery using the mouse where he/she finds a key with the word "Catacombs" written on it.

In the fifth step, the text-to-speech (TTS) tool is employed to transform the instructions written on the key ("CATACOMBS") to an audio signal that can be perceived by the blind user. The user has to search for the catacombs enter in them and find the box that contains a map (Fig. 11). The map is then sent to the next level.

In the sixth step, the deaf user receives the map, and has to draw the route to the area where the treasure is hidden. The route is drawn on the map and the map is converted to a grooved line map, which is send to for the last level to the blind user.

In the seventh step, the blind user receives the grooved line map and has to find and follow the way to the forest where the treasure is hidden. Although the map is presented again as a 2D image the blind user can feel the 3D grooved map and follow the route to the forest. The 2D image and the 3D map are registered and this allows us to visualize the route that the blind user actually follows on the 2D image. The blind user is asked to press the key of the PHANToM device while he believes that the PHANTOM cursor lies in the path. Finally, after finding the forest he obtains a new grooved line map where the blind user has to search for the final location of the treasure. After searching in the forest streets the blind user should find the treasure.

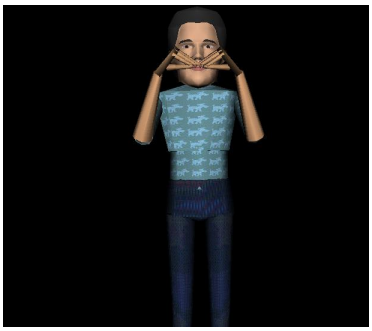


Figure 12: *Sign language synthesis using an avatar.*

10. CONCLUSIONS

In this paper, a novel system for the communication between disabled used and their effective interaction with the computer was presented based on multimodal user interfaces. The main objective was to address the problem caused by the fact that impaired individuals, in general, do not have access to the same modalities. Therefore, the transmitted signals were translated into a perceivable form. The critical point for the automatic translation of information is the accurate recognition of the transmitted content and the effective transformation into another form. Thus, an audio-visual speech recognition system was employed to recognize phonetic commands from the blind user. The translation of the commands for the deaf-mute was performed using a sign synthesis module which produces an animation with an avatar. On the other hand, the deaf-mute user interacts using sign language and gestures. The system incorporates a module which is capable of recognizing user gestures and translate them using text-to-speech applications. As an application scenario, the aforementioned technologies are integrated in a collaborative treasure hunting game which requires the interaction of the users in each level. Future work will focus on the extension of the developed modules in order to support larger vocabularies and enable more natural communication of the users. Furthermore, the structure of the employed modalities should be studied more to reveal their inter- dependencies and exploit their complementary nature more effectively.

11. ACKNOWLEDGEMENTS

This work was supported by the EU funded SIMILAR Network of Excellence.

12. REFERENCES

[1] "W3C Workshop on Multimodal Interaction", July 2004. http://www.w3.org/2004/02/mmi-workshop-cfp.html. 27

[2] I. Marsic, A. Medl, and J. Flanagan, "Natural communication with information systems", *Proc. of the IEEE*, vol. 88, pp. 1354–1366, August 2000. 27

[3] J. Lumsden and S. A. Brewster, "A paradigm shift: Alternative interaction techniques for use with mobile and wearable devices", in *Proc. 13th Annual IBM Centers for Advanced Studies Conference (CASCON 2003)*, (Toronto, Canada), pp. 97–100, 2003. 27

[4] T. V. Raman, "Multimodal Interaction Design Principles For Multimodal Interaction", in *Proc. of Computer Human Interaction (CHI 2003)*, (Fort Lauderdale, USA), pp. 5–10, 2003. 27

[5] C. Colwell, H. Petrie, D. Kornbrot, A. Hardwick, and S. Furner, "Haptic Virtual Reality for Blind Computer Users", in *Proc. of Annual ACM Conference on Assistive Technologies (ASSETS 1998)*, pp. 92–99, 1998. 27

[6] C. Sjostrom, "Touch Access for People With Disabilities". Licentiate Thesis, CERTEC Lund University, Sweden, 1999. 27

[7] V. Scoy, I. Kawai, S. Darrah, and F. Rash, "Haptic Display of Mathematical Functions for Teaching Mathematics to Students with Vision Disabilities", in *Haptic Human-Computer Interaction Workshop*, 2000. 27

[8] K. Moustakas, G. Nikolakis, D. Tzovaras, B. Deville, I. Marras, and J. Pavlek, "Multimodal tools and interafces for the intercommunication between visually impaired and deaf-and-mute people", in *Proc. of eNTERFACE 2006*, (Dubrovnik, Croatia), July 2006. 28

[9] S. Tamura, K. Iwano, and S. Furui, "A Stream-Weight Optimization Method for Multi-Stream HMMS Based on Likelihood Value Normalization", *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005 (ICASSP'05)*, vol. 1, pp. 469–472, 2005. 29

[10] E. Erzin, Y. Yemez, and A. Tekalp, "Multimodal speaker identification using an adaptive classifier cascade based on modality reliability", *IEEE Transactions on Multimedia*, vol. 7, no. 5, pp. 840–852, 2005. 29

[11] S. Yound *et al.*, *The HTK Book, HTK Version 3.4*. Cambridge University Engineering Department, 2006. 29, 30

[12] L. Rabiner and B. Juang, *Fundamentals of Speech Recognition*. Englewood Cliffs, USA: New Jersey: Prentice-Hall, 1993. 30

[13] A. Nefian, L. Liang, X. Pi, X. Liu, and K. Murphy, "Dynamic Bayesian Networks for Audio-Visual Speech Recognition", *EURASIP Journal on Applied Signal Processing*, vol. 2002, no. 11, pp. 1274–1288, 2002. 30

[14] S. Jayaram, S. Schmugge, M. C. Shin, and L. V. Tsap, "Effect of Colorspace Transformation, the Illuminance Component, and Color Modeling on Skin Detection", in *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2004. 31

[15] O. Aran and L. Akarun, "Recognizing two handed gestures with generative, discriminative and ensemble methods via Fisher kernels", in *International Workshop on Multimedia Content Representation, Classification and Security, (MRCS '06)*, (Istanbul, Turkey), September 2006. 32

13. BIOGRAPHIES

Savvas Argyropoulos (S'04) received the Diploma degree in electrical and computer engineering from the Aristotle University of Thessaloniki (AUTH), Thessaloniki, Hellas, in 2004, where he is currently pursuing the Ph.D. degree. He holds teaching and research assistantship positions at AUTH. He is also a graduate Research Associate with the Informatics and Telematics Institute, Centre for Research and Technology Hellas. Since 2005, he has participated in several research projects funded by the EC. His research interests include distributed source coding, video coding/transmission, multimodal signal processing, and biometric recognition.
Email: savvas@ieee.org

Konstantinos Moustakas received the Diploma degree in electrical and computer engineering and the PhD degree in virtual reality from Aristotle University of Thessaloniki, Thessaloniki, Greece, in 2003 and 2007 respectively. Currently, he is a postdoctoral research fellow in the Informatics and Telematics Institute, Centre for Research and Technology Hellas, Thessaloniki. His main research interests include virtual reality, collision detection, haptics, deformable object modeling and simulation, 3D content-based search, computer vision, and stereoscopic image processing. During the last three years, he has been the (co)author of more than 40 papers in refereed journals, edited books, and international conferences. Dr. Moustakas serves as a regular reviewer for various international scientific journals. He has also been involved in many projects funded by the EC and the Greek secretariat of Research and Technology. He is a member of the IEEE and the Technical Chamber of Greece.
Email: moustak@iti.gr

Alexey A. Karpov received the M.S. Diploma from St. Petersburg State University of Airspace Instrumentation and Ph.D. degree in computer science from St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences (SPIIRAS), in 2002 and 2007, respectively. Currently he is a researcher of Speech Informatics Group of SPIIRAS. His main research interests include automatic speech and speaker recognition, multimodal interfaces based on speech and gestures, computer vision techniques. He has been the (co)author of more than 50 papers in refereed journals and international conferences. He has also been involved in SIMILAR Network of Excellence funded by the EC as well as several research projects funded by EU INTAS association and Russian scientific foundations. He is a member of organizing committee of International Conferences "Speech and Computer" SPECOM.
Email: karpov@iias.spb.su

Oya Aran received the BS and MS degrees in Computer Engineering from Boğaziçi University, İstanbul, Turkey in 2000 and 2002, respectively. She is currently a PhD candidate at Boğaziçi University working on dynamic hand gesture and sign language recognition. Her research interests include computer vision, pattern recognition and machine learning.
Email: aranoya@boun.edu.tr

Dimitrios Tzovaras received the Diploma degree in electrical engineering and the Ph.D. degree in 2-D and 3-D image compression from Aristotle University of Thessaloniki, Thessaloniki, Greece, in 1992 and 1997, respectively. He is currently a Senior Researcher Grade B in the Informatics and Telematics Institute of the Centre for Research and Technology Hellas in Thessaloniki, Greece. Prior to his current position, he was a Senior Researcher on 3-D imaging at the Aristotle University of Thessaloniki. His main research interests include virtual reality, assistive technologies, 3-D data processing, haptics and semantics in virtual environments. His involvement with those research areas has led to the co-authoring of more than 50 papers in refereed journals and more than 150 publications in international conferences. He has served as a regular reviewer for a number of international journals and conferences. Since 1992, he has been involved in more than 100 projects in Greece, funded by the EC and the Greek Ministry of Research and Technology. Dr. Tzovaras is a member of the Technical Chamber of Greece.
Email: dimitrios.tzovaras@iti.gr

Thanos Tsakiris received the BSc in Computer Science from the Aristotle University of Thessaloniki in 2000 and the MSc in Computer Games Technology from the University of Abertay Dundee in 2001. He is currently working in ITI/CERTH as a research associate in the fields of 3D Graphics,VR and HCI.
Email: atsakir@iti.gr

Giovanna Varni was born in Genoa (Italy), in 1979. She received summa cum laude her master's degree in Biomedical Engineering at the University of Genoa in 2005. She is currently PhD Student at InfoMus Lab (DIST, University of Genoa) where she is working on multimodal feedback and multimodal streams data analysis.
Email: giovanna@infomus.dist.unige.it

Byungjun Kwon was born in Seoul, Korea. He received his bachelor's degree in French language and literature from Seoul National University (1996) and master's in Art- Science from Royal Conservatory, the Hague, Netherlands (2008). He started his musical career in early 90's as a singer/songwriter and has released 7 albums ranging from alternative rock to minimal house. He creates music for records, sound tracks, fashion collections, contemporary dance, theatre plays and interdisciplinary events. Recent works and performances have been presented in many international venues. Now he lives and works in Amsterdam.
Email: byungjun@gmail.com

SPEECH AND SLIDING TEXT AIDED SIGN RETRIEVAL FROM HEARING IMPAIRED SIGN NEWS VIDEOS

Oya Aran [1], *Ismail Ari* [1], *Pavel Campr* [2], *Erinç Dikici* [3], *Marek Hruz* [2], *Deniz Kahramaner* [4], *Siddika Parlak* [3], *Lale Akarun* [1], *Murat Saraçlar* [3]

[1] PILAB, Boğaziçi University, İstanbul, Turkey
[2] University of West Bohemia, Pilsen, Czech Republic
[3] BUSIM, Boğaziçi University, İstanbul, Turkey
[4] Robert College, İstanbul, Turkey

ABSTRACT

The objective of this study is to automatically extract annotated sign data from the broadcast news recordings for the hearing impaired. These recordings present an excellent source for automatically generating annotated data: In news for the hearing impaired, the speaker also signs with the hands as she talks. On top of this, there is also corresponding sliding text superimposed on the video. The video of the signer can be segmented via the help of either the speech or both the speech and the text, generating segmented, and annotated sign videos. We aim to use this application as a sign dictionary where the users enter a word as text and retrieve sign videos of the related sign with several examples. This application can also be used to automatically create annotated sign databases that can be used for training recognizers.

KEYWORDS

Speech recognition – Sliding text recognition – Sign language analysis – Sequence clustering – Hand tracking

1. INTRODUCTION

This project aims to exploit TRT news for the hearing impaired programs in order to generate usable data for sign language education. The news video consists of three major information sources: sliding text, speech and signs. Fig. 1 shows an example frame from the recordings.

Figure 1: *An example frame from the news recordings. The three information sources are the speech, sliding text, signs.*

The three sources in the video convey the same information via different modalities. The news presenter signs the words as she talks. It is important to note that sign languages have their own grammars and word orderings. Thus, it is not necessary to have the same word ordering in a Turkish spoken sentence and in a Turkish sign sentence [1]. Thus, the signing in these news videos is not Turkish sign language (Turk Isaret Dili, TID) but Signed Turkish: the sign of each word is from TID but their ordering would have been different in a proper TID sentence. In addition to the speech and sign information, a corresponding sliding text is superimposed on the video. Our methodology is to process the video to extract the information content in the sliding text and speech components and to use either the speech alone or both the speech and the text to generate segmented and annotated sign videos. The main goal is to use this annotation to form a sign dictionary. Once the annotation is completed, unsupervised techniques are employed to check consistency among the retrieved signs, using a clustering of the signs.

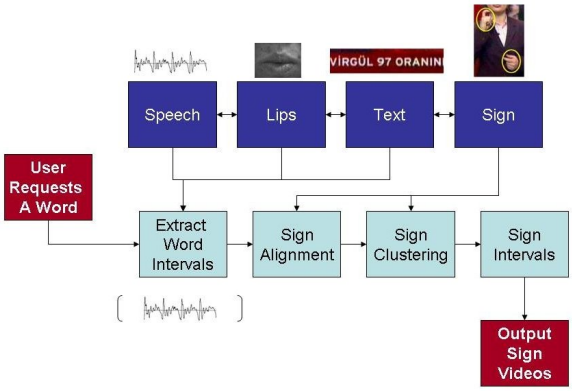


Figure 2: *Modalities and the system flow.*

The system flow is illustrated in Fig. 2. The application receives the text input of the user and attempts to find the word in the news videos by using the speech. At this step the application returns several intervals from different videos that contain the entered word. Then, sliding text information may optionally be used to control and correct the result of the retrieval. This is done by searching for the word in the sliding text modality during each retrieved interval. If the word can also be retrieved by the sliding text modality, the interval is assumed to be correct. The sign intervals will then be extracted by analyzing the correlation of the signs with the speech. Sign clustering is necessary for two reasons. First, there can be false alarms of the retrieval corresponding to some unrelated signs and second, there are homophonic words that have the same phonology but different meanings thus possibly different signs.

2. SPOKEN TERM DETECTION

Spoken utterance retrieval part of this project will be used as a tool to automatically segment the signs in a broadcast news video for the disabled and to display the desired sign to the user after the alignment. The general system diagram is given in Fig. 3 where the input to the search part is the query and the output is the occurrence time, duration and the program name in which the query appears in.

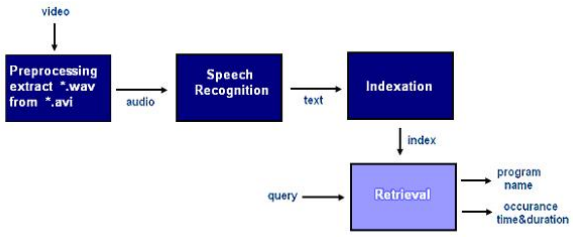


Figure 3: *Block diagram of the spoken term detection system.*

Three main modules of the system are speech recognition, indexing and retrieval which will be explained in detail. Currently we have 127 broadcast news videos, each with a duration of 10 minutes and sampled at 32 kHz. As a preprocessing operation, the audio information is extracted and the sampling rate is reduced to 16 kHz. The speech recognizer converts this audio data into textual information (in terms of weighted finite state automata). Indexation of the text is done via weighted finite state transducers [2]. The index is built in such a way that when it is composed with the query, the output is the search hits with program name, occurrence time and duration as well as the expected count. This type of indexation introduces several benefits as will be explained later.

2.1. Speech recognition

The speech recognizer takes a collection of audio files and converts them into text. Before recognition, audio data is segmented based on the energy constraint. Since the background does not include music or noise it was adequate to identify the speech and non-speech portions only. The method explained in [3] is applied and some further arrangements are made on the output. These post modifications were mainly about setting a minimum duration limit on segments and merging them if they are smaller.

An HMM-based large vocabulary continuous speech recognition (LVCSR) system is used for recognition. The feature vector consists of 12 MFCC components, the energy component as well as delta and delta-delta components. The acoustic models consist of decision tree clustered triphones and the output distributions are GMMs. The acoustic model used in this project was previously trained on approximately 100 hours of broadcast news data. The language models are pruned back off trigram models which are based on words. The recognition networks are represented as weighted finite state machines (FSMs). The output of the ASR system is also represented as an FSM and may be in the form of a best hypothesis string or a lattice of alternate hypotheses. To illustrate this, the lattice output of a recognized utterance is shown in Fig. 4. [4].

The labels on the arcs are the word hypotheses (or subwords such as morphemes) and the values next to the labels are the probabilities of each arc. It is obvious that there is more than one path from the initial state to the final. However, the most probable path is "iyi gUnler" ("good afternoon"), which is the correct transcription.

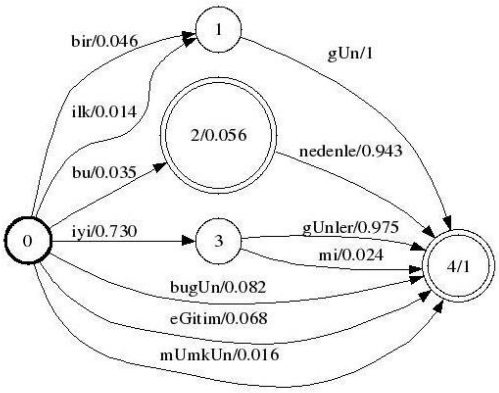


Figure 4: *An example lattice output.It belongs to the utterance "iyi gUnler".*

For the lattice output, more than one hypothesis is returned with corresponding probabilities. Indexation estimates the expected count using path probabilities. By setting a threshold on the expected count, different precision-recall points can be obtained which results in a curve. On the other hand, the one-best hypothesis is represented with only one point. Having a curve allows choosing an operating point by setting the threshold. Use of a higher threshold improves the precision but recall falls. Conversely, a lower threshold value causes small expected counts to be retrieved. This increases recall but decreases precision.

The opportunity of choosing the operating point is crucial. Depending on the application, it may be desirable to retrieve all the documents or only the most probable ones. For our case, it is more convenient to operate at the point where precision is high.

Although this introduces some redundancy, using lattice output of the ASR, improves system performance as shown in the experiments. This improvement is expected to be much higher for noisy data.

The HTK tool[5] is used to produce feature vectors and AT&T's FSM and DCD tools[6] are used for recognition. Currently, the word error rate of the speech recognizer is approximately 20%. Using a morpheme based-model instead of the word-based one reduces the WER and will be left as a future work.

2.2. Indexation

The output of the speech recognizer is a weighted automaton where each speech utterance of the dataset is represented as an FSM and they are concatenated. Since input to indexation is a finite state machine, it is advantageous to represent the indexer as a transducer FSM. Another advantage of this approach is that it simplifies dealing with arc weights or path probabilities in the lattice. Since the input to indexation is uncertain, it is important to keep log-likelihoods. This information should be passed to the search module by the indexer for ranking purposes. Since FSMs are compact representations of alternative hypotheses with varying probabilities, designing the indexer as an FSM is favorable.

The problem is to build an index which accepts any substring of the utterance transcriptions. Thus the index can be represented by a weighted finite-state transducer mapping each factor of each utterance to the indices of the automata. This in-

formation comes from the ASR output and is defined as:

$$T(x,i) = -\log(E_{P_{|i|}}[C_{[x]}]) \quad (1)$$

where x is the substring, i is the automaton index, P is the probability coming from ASR and C_x denotes the number of occurrences of substring x in the utterance.

The factor selection is done by creating a new initial and final state to the automata such that each intermediate state is connected to both the initial and the final state. Thus it is possible to go from the initial state to any of the terms (phones or words) and to reach the final state from there; this is accepted by the transducer. After the new states and connections are added, the automata are optimized with epsilon removal, determination and minimization. The details of this process and the further theory of weighted finite state automata are explained in [2].

A very powerful feature of this indexing method is its speed. The resulting index also does not bother with the expansion of the database. The search time for a query is linear in the size of the query and the number of places it appears in the database. Search duration is currently 1-1.8 seconds depending on the frequency of the query and expanding the database from 20 videos to 120 videos does not make any difference.

2.3. Retrieval

The top three blocks in 3, namely, preprocessing, recognition and indexation are performed only once. When the query is entered, only the retrieval module in the diagram is activated.

First the query is converted into an FSM and then composed with the index FSM. After the optimization, the list of all indices where the query appears and the corresponding log-likelihoods are acquired. Since the probabilities are known, it is possible to select the most probable outputs and rank the documents [2].

Now we have the utterance indices; however this includes not only the query but other words. To clarify each word's starting time and duration we apply forced alignment. The final output is the program name, starting time and duration of the query in seconds.

For the client-server interaction, a server application using sockets is written in PERL. Client side will be explained later. When the server gets the query from the client, the search operation is initiated and the standard output is sent back to the client. If the user asks for a word which is not in the vocabulary, an error message is displayed and no search is performed.

2.4. Experiments and results

2.4.1. Evaluation Metrics

Speech recognition performance is evaluated by the WER (word error rate) metric which was measured to be around 20% in our experiments. For evaluation of the retrieval system precision-recall values and F-measures are used.

Let the reference transcriptions include R(q) occurrences of the search term q. A(q) is the number of retrieved documents and C(q) is the number of documents which are related to the query from the retrieved ones.

Then

$$Precision(q) = \frac{C(q)}{A(q)} \quad (2)$$

$$Recall(q) = \frac{C(q)}{R(q)} \quad (3)$$

and the F-measure is:

$$F(q) = \frac{2 * Precision(q) * Recall(q)}{Precision(q) + Recall(q)} \quad (4)$$

Precision and recall values for each word in the query set will be determined and averaged. Out-of-vocabulary words will be discarded in the precision averaging. However their recall is assumed to be zero for recall averaging. This is also the case for words which exist in the reference but could not be retrieved. The reason behind this assumption is that retrieving a wrong document and retrieving nothing cannot be judged as if they are the same.

2.4.2. Corpora

Although the whole collection consists of 127 videos, we used only 15 for the evaluation.Since we have the reference manual transcripts only for these files. The evaluation task begins with the forced alignment of the manual transcriptions which will be used as the reference. Start times and durations of each word - and the silence - are identified and kept in the format (*.ctm), shown in Fig 5, where the columns represents program name, channel (studio, telephone etc.), start time(in sec.), duration(in sec.) and spoken word (or silence) respectively.

```
2007-06-29-14-40 1 10.06 0.08 Z1
2007-06-29-14-40 1 10.14 0.23 iyi
2007-06-29-14-40 1 10.37 0.57 gUnler
2007-06-29-14-40 1 10.94 1.11 Z1
```

Figure 5: **.ctm file format. It is constructed after the forced alignment of manual transcriptions.*

A query set is created from the reference files by taking each word uniquely. Each item in the query set is searched in the database and search results are kept in a file with the format, shown in Fig. 6. In this format, columns represent the program file name in which the query is claimed to occur, start time (in ms.), duration (in ms.) and relevance respectively. After these files are obtained, precision and recall values are calculated and averaged.

```
2007-06-29-14-40.avi, 255477, 480, 1
2007-06-29-14-40.avi, 271494, 460, 1
2007-06-29-17-40.avi, 530545, 500, 0.0118179
```

Figure 6: *Search result file format. For each query, this file is created and compared with the reference ctm file.*

The retrieval result is checked with the reference ctm file with a margin time. If the beginning and end times are found to agree with margin seconds or less from the correct, search is assumed to be successful.

2.4.3. Results

The index is created for the lattice and one-best output of the recognizer. Precision-recall (no limit is set on the number of retrieved documents) and precision-recall at 10 (maximum number of retrieved documents is limited to 10) graphs are depicted as in Fig. 7 and Fig. 8.

It is obvious from the plots that, use of lattices performs better than one-best. This is also the case for precision and recall at 10. The arrows point at the position where the maximum F-measure is achieved for lattice. Comparison of F-measures of lattice and one-best output of ASR are given in 1. Use of lattices introduces 1-1.5 % of improvement. Since the broadcast news corpora are fairly noiseless, the achievement may seem minor. However for noisy data the difference is much higher [4].

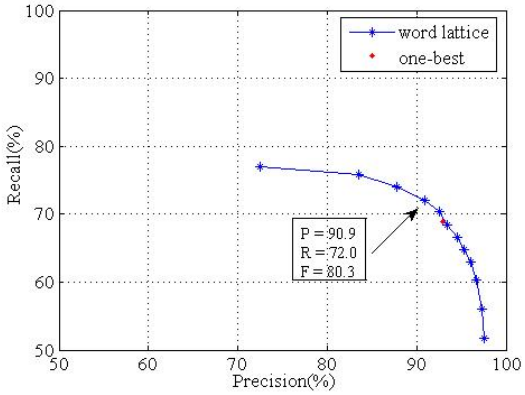


Figure 7: *Precision-Recall for word-lattice and one-best hypotheses when no limit is set on maximum number of retrieved documents.*

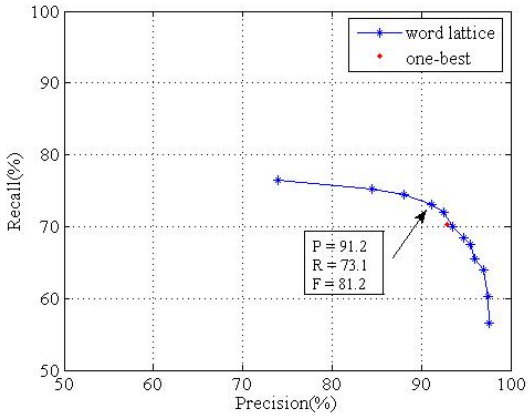


Figure 8: *Precision-Recall for word-lattice and one-best hypotheses when maximum number of retrieved documents is set to 10 (precision at 10).*

3. SLIDING TEXT RECOGNITION

3.1. Sliding Text Properties

The news videos are accompanied by a sliding text band, which includes simultaneous transcription of what is being said. It is placed at the bottom of the screen and contains characters with a specific font, shown with white pixels over a solid background. Speed of the sliding text is approximately constant throughout the whole video sequence (usually 4 pixels/frame), which allows each character to appear on the screen for at least 2.5 seconds. An example of a frame with sliding text is shown in Fig. 9.

3.2. Baseline Method

The method we propose to obtain sliding text information consists of three parts: Extraction of the text line, character recognition and temporal alignment.

3.2.1. Text Line Extraction

Size and position of the sliding text band does not change throughout the video. Therefore, it is found at the first frame and used in the rest of the operations. To find the position of the text, first, we convert the RGB image into a binary image, using grayscale

Table 1: *Comparison of lattice and one-best ASR outputs on maximum F-measure performance.*

	Max-F(%)	Max-F@10(%)
Lattice	80.32	81.16
One-best	79.05	80.06

Figure 9: *Frame snapshot of the broadcast news video.*

quantization and thresholding with the Otsu method [7]. Then we calculate horizontal projection histogram of the binary image, i.e., the number of white pixels for each row. The text band appears as a peak on this representation, separated from the rest of the image. We apply a similar technique over the cropped text band, this time on the vertical projection direction, to eliminate the program logo. The sliding text is bounded by the remaining box, whose coordinates are defined as the text line position. Fig. 10 shows an example of binary image with its horizontal and vertical projection histograms.

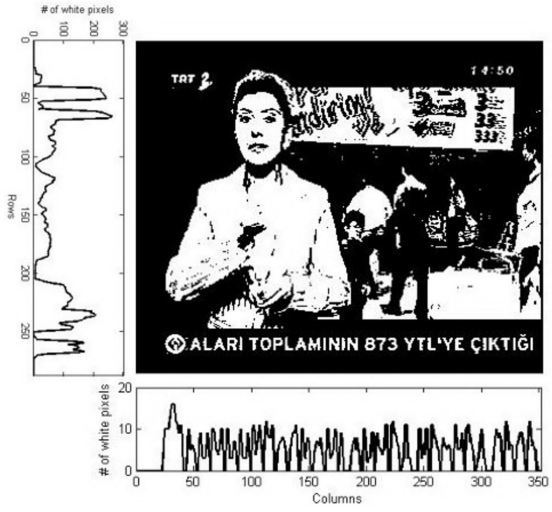


Figure 10: *Binary image with horizontal and vertical projection histograms.*

Since there is redundant information in successive frames, we do not extract text information from every frame. Experiments have shown that updating text transcription once in every

10 frames is optimal for achieving sufficient recognition accuracy. We call these the "sample frames". The other frames in between are used for noise removal and smoothing.

Noise in binary images stems mostly from quantization operations in color scale conversion. Considering the low resolution of our images, it may cause two characters, or two distinct parts of a single character to be combined, which complicates the segmentation of text into characters. We apply morphological opening with a 2x2 structuring element to remove such effects of noise. To further smooth the appearance of characters, we horizontally align binary text images of the frames between two samples, and for each pixel position, decide on a 0 or 1, by voting.

Once the text image is obtained, vertical projection histogram is calculated again to find the start and end positions of every character. Our algorithm assumes that two consecutive characters are perfectly separated by at least one black pixel column. The threshold of an inter-word space is determined adaptively, by searching for outliers in character spacing. For proper alignment, only complete words are taken for transcription.

Each character is individually cropped from the text figure and saved, along with its start and end horizontal pixel positions.

3.2.2. *Character Recognition*

For character recognition, we implement the template matching method. Each binary character image is compared to each template, pixel by pixel. The total number of matching pixels are divided by the size of the character image and used as a similarity score. Matched characters are stored as a string. Fig. 11 depicts a sample text band image, its transcribed text and corresponding pixel positions, respectively.

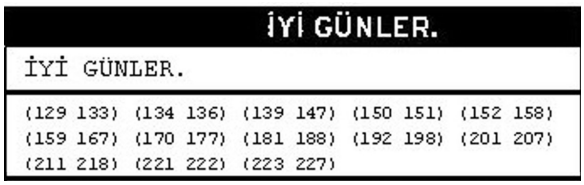


Figure 11: *Sample image and transcription.*

3.2.3. *Temporal Alignment*

The sliding text is interpreted as a continuous band throughout the video. Since we process only selected sample frames, the calculated positions of each character should be aligned in space (and therefore in time) with their positions from the previous sample frame. This is done using frame shift and pixel shift values. For instance, a character which appears in positions 180-189 in the first sample and the one in 140-149 of the second refer to the same character, since we investigate each 10^{th} frame with 4 pixels of shift per frame. Small changes in these values (mainly due to noise) are compensated using shift-alignment comparison checks. Therefore, we obtain a unique pixel position (start-end) pair for each character seen on the text band.

We determine the final transcript as follows: For each frame, we look for the occurrence of a specific start-end position pair, and note the corresponding character to a character candidate list (for different frames, these characters may not be the same, due to recognition errors). The final decision is made by majority voting; the character that is seen the most in the list is assigned to that position pair.

3.3. Baseline Performance

We compare the transcribed text with the ground truth data and use character recognition and word recognition rates as performance criteria. Even if only one character of a word is misrecognized, we label this word as erroneous.

Applying the baseline method on a news video dataset of 40 minutes (around 3500 words), we achieved 94% character recognition accuracy, and 70% word recognition accuracy.

3.4. Discussions

One of the most important challenges of character recognition was the combined effect of low resolution and noise. We work with frames of 352x288 resolution, therefore, each character covers barely an area of 10-14 pixels in height and 2-10 pixels in width. In such a small area, any noise pixel distorts the image considerably, thus making it harder to achieve a reasonable score by template comparison.

Noise cancellation techniques created another disadvantage since they removed distinctive parts of some Turkish characters, such as erasing dots (İ, Ö, Ü), or pruning hooks (Ç, Ş). Fig. 12 shows examples of such characters, which, after noise cancellation operations, look very much the same.

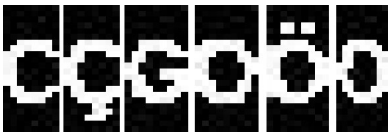


Figure 12: *Highly confused characters.*

3.5. Improvements over the Baseline

We made two major improvements over the baseline system, to improve recognition accuracy. The first one is to use Jaccard's distance for template match score. Jaccard uses pixel comparison with a slightly different formulation: Let n_{ij} be the total number of pixel positions, where binary template pixel has value i and character pixel has value j. Then, the comparison score is formulated as [8]:

$$d_j = \frac{n_{11}}{n_{11} + n_{10} + n_{01}} \tag{5}$$

Using Jaccard's distance for template match scoring resulted in 96.7% accuracy for character recognition, and 80.0% for word recognition accuracy.

The highest rate in the confusion matrix belongs to discrimination of the letter "O" and the number "0". To distinguish these, the knowledge that a zero should be preceded or succeeded by another number, is used as a postprocessing operation. Adding this correction resulted in 98.5% character accuracy and 90% word recognition accuracy, respectively. The confusion rates are listed in Table 2.

4. SIGN ANALYSIS

4.1. Skin color detection

4.1.1. *Skin color detection using a probabilistic model*

Skin color is widely used to aid segmentation in finding parts of the human body [9]. We learn skin colors from a training set and then create a model of them using a Gaussian Mixture Model (GMM). This is not a universal skin color model; but rather, a model of skin colors contained in our training set.

Table 2: *Character confusion rates.*

Character (Original)	Character (Recognized)	Confusion Rate (%)
Ç	C	8.33
Ğ	G	1.49
Ğ	Ö	2.99
H	M	2.94
I	1	0.85
N	M	0.34
Ö	O	9.68
Ş	S	2.73
Ü	U	2.47
0	O	36.36
2	Z	7.14

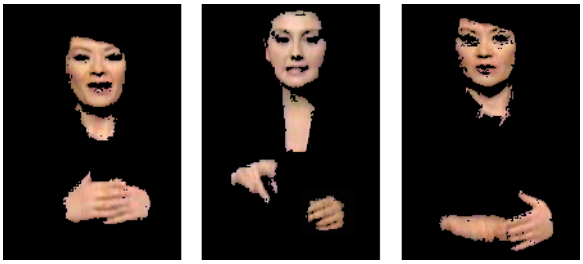


Figure 13: *Examples from the training data*

Training of GMM. We prepared a set of training data by extracting images from our input video sequences and manually selecting the skin colored pixels. We used images of different speakers under different lighting conditions. In total, we processed six video segments. Some example training images are shown in Fig. 13.

For color representation we use the RGB color space. The main reason is that this color space is native for computer vision and therefore does not need any conversion to other color space. The collected data are processed by the Expectation Maximization (EM) algorithm to train the GMM. After inspecting the spatial parameters of the data we decided to use a five Gaussian mixtures model. The resulting model can be observed in Fig. 14.

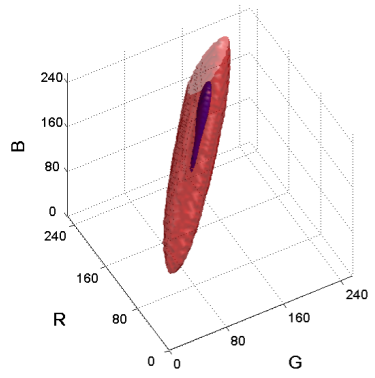


Figure 14: *General look up table. Levels of likelihood can be seen in different colors. Lower level likelihood of 128, higher level likelihood of 220.*

Using GMM for image segmentation. The straight forward way of using the GMM for segmentation is to compute the probability of belonging to a skin segment for every pixel in the image. One can then use a threshold to decide whether the pixel color is skin or not. But this computation would take a long time, provided the information we have is the mean, variance and gain of each Gaussian in the mixture. We have precomputed the likelihoods and used table look-up.

4.1.2. Skin color detection using look up table

Look up table model description. We decided to create a look-up table to store the likelihood that a color corresponds to a skin. In this case the values of the look-up table are computed from the probability density function given by GMM. The range of likelihoods is from 0 to 255. A likelihood of 128 and more means that the particular color belongs to a skin segment. With this procedure we obtain a 256x256x256 look-up table containing the likelihood of all the colors from RGB color space.

The segmentation is straightforward. We compare the color of each pixel with the value in the look-up table. According to the likelihood of the color, we decide whether it belongs to skin segment or not. For better performance, we blur the resulting likelihood image. That way, the segments with low probability disappear and the low probability regions near the high probability regions are strengthened. For each frame, we create a mask of skin color segments by thresholding the likelihood image (See Fig. 15).

Figure 15: *The process of image segmentation. From left to right: original image, probability of skin color blurred for better performance, binary image as a result of thresholding, original image with applied mask.)*

The look up table is created using color information from more than one speaker. Generally we want to segment a video sequence with just one speaker. This means that with the look up table we will obtain some undesirable pixels identified as skin color pixels which do not belong to the particular speaker. Thus, we need to adapt the look up table for each video. Since manual data editing is not possible at this stage, a quick automatic method is needed.

Training of the look up table. We refer to the general look up table modeled from several speakers as the background look-up table and describe how we adapt it to the current speaker. First we need to extract the skin color pixels from the current speaker. Using the OpenCV's Haar classifier [10] we detect the face of the speaker and store it in a separate image. This image is then segmented using the background look up table and a new look up table is created. We process all the pixels of the segmented image and store the pixels' color in the proper positions in the look up table. Information from one image is of course not sufficient and there are big gaps in the look up table. To solve this problem, we use a convolution with a Gaussian kernel to smooth the look-up table. To save time, we process color components separately. At the end of this step, we obtain a speaker dependent look up table.

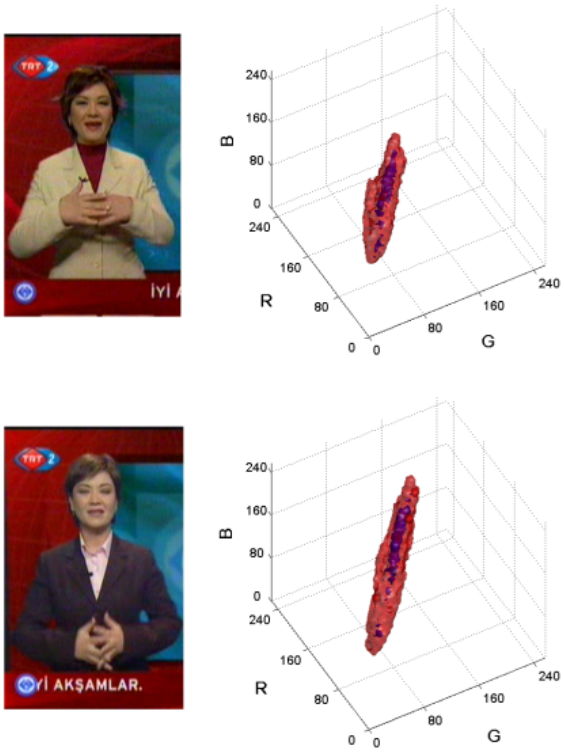


Figure 16: *Examples of different lighting conditions resulting into different skin color of the speaker and the corresponding adapted look up table.)*

4.1.3. Adaptation of general skin color model for a given video sequence

Now we need to adapt our background look up table to the new speaker dependent look up table. We use weighted averaging of the background and the speaker dependent look up tables. The speaker is given more weight in the weighted averaging. This way we eliminate improbable colors from the background look up table thus improving its effectiveness. Some examples of adapted look up tables can be seen in Fig. 16.

4.2. Hand and face tracking

The movement of the face and the hands is an important feature of sign language. The trajectory of the hands gives us a good idea about the performed sign. The position of the head is used for creating a local system of coordinates and normalizing the hand position.

There are many algorithms that are able to detect a face with very good results. We use OpenCV's Haar feature based face detector [10]. From the resulting bounding rectangle we calculate an ellipse around the face. The center of the ellipse is assumed to be the center of mass of the head. We use this point as the head's tracked position.

The input of our tracking algorithm is a segmented image containing only blobs of skin colored objects. As a result of skin color detection, we have an image that contains only skin colored regions. To retrieve the separate blobs from the image we use cvBlobsLib [11]. There can be some blobs which appear in our segmented image but they do not belong to the signer. That means the blob is neither the signer's head nor the signer's hand. We want to eliminate these blobs.

4.2.1. Blob filtering

First we eliminate the blobs with small area. Next we eliminate blobs which belong to the background. The second step is to take into account only the three biggest blobs in the image. They should belong to the head and the hands of the signer as the other false blobs are a result of noise. Sometimes when an occlusion occurs there can be fewer than three blobs belonging to the signer. We need to take this into account. Third, we eliminate all the blobs that are too far away from the previous position of already identified blobs. As a threshold, we use a multiple of the signer's head width. This solves the problem when a false blob appears when a hand and the head are in occlusion.

Generally the biggest challenge in hand tracking is when hands occlude each other, or the head. In this case, two or three objects form a single blob. This situation can be easily detected, but this information alone is not enough to resolve the occlusion. We need to somehow know which objects of interest are in occlusion or whether they are so close to each other that they form a single blob. For this purpose, we try to predict the occlusion. When two objects occlude in the next frame, we assume that they are the predicted ones.

4.2.2. Occlusion prediction

We apply an occlusion prediction algorithm as a first step in occlusion solving. We need to predict whether there will be an occlusion and among which blobs will the occlusion be. For this purpose, we use a simple strategy that predicts the new position, p_{t+1}, of a blob from its velocity and acceleration. The velocity and acceleration is calculated by the first and second derivatives of the position at the previous frames.

$$v_t = p_{t-1} - p_t \quad (6)$$

$$a_t = v_{t-1} - v_t \quad (7)$$

$$p_{t+1} = p_t + v_t + a_t \quad (8)$$

The size of the blob is predicted with the same strategy. With the predicted positions and sizes of the blobs, we check whether these blobs intersect. If there is an intersection, we identify the intersecting blobs and predict that there will be an occlusion between those blobs.

4.2.3. Occlusion solving

We decided to solve the occlusions by trying to find out which part of the blob belongs to one of the objects and divide the blob into two (or three in the case that both hands and head are in occlusion). We divide the blob by drawing a black line or ellipse at the location we think is the best. Thus we always obtain three blobs which can then be tracked as will be shown later. Let us describe the particular cases of occlusion.

Two hands occlude. In this case we separate the blob with a line. We find the bounding ellipse of the blob of occluded hands. The minor axis of the ellipse is computed and a black line is drawn along this axis. For better results we draw the line several pixels wide as can be seen on Fig. 17.

One hand occludes with the head. Usually the blob of the head is much bigger than the blob of the hand. Therefore a division of the blob along the minor axis of the bounding ellipse would have an unwanted effect. We use a template matching method instead [12]. In every frame when the hand is visible we collect its template. The template is a gray scale image defined by the hand's bounding box. When the occlusion is detected a region around the previous position of the hand is defined. We

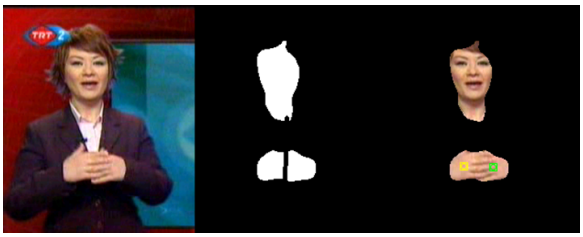


Figure 17: *An example of hand occlusion. From left original image, image with the minor axis drawn, result of tracking.*

calculate the correlation between the hand template and the segmented image. We use a squared difference correlation method to calculate the correlation,

$$R(x,y) = \sum_{x'} \sum_{y'} (T(x',y') - I(x+x', y+y'))^2 \qquad (9)$$

where T is the template, I is the image we search in, x and y are the coordinates in the image, x' and y' are the coordinates in the template. After that we find the minimum in the result of the correlation but we limit the search only to the region around the last position. We draw a black ellipse at the estimated position of the hand as in Fig. 18. The parameters of the ellipse are taken from the bounding ellipse of the hand. As a last step we collect a new template for the hand. Previously the last step was omitted and the last known template of the hand was used. But often the shape of the hand changes in front of the head, so it is necessary to collect the template all the time be able to track the hand properly. The remaining parts of the head in the template image do not have a significant impact on the result of the correlation.

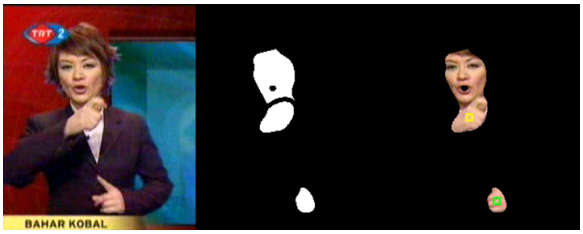


Figure 18: *An example of hand and head occlusion. From left original image, image with the ellipse drawn, result of tracking.*

Both hands occlude with the head. In this case the template matching method proved to be successful. We just need to apply it to both hands.

4.2.4. Tracking of hands

For the tracking of the hands we are able to apply several rules. First, we need to decide which blob belongs to what part of the signer's body. In principle, we assign the blobs in the current frame to the hands and the head by comparing their previous positions and velocities. We assume that the biggest blob closest to the previous head position belongs to the signer's head. The other two blobs belong to the hands.

After examining the signs we found out that the right hand of the signer is most of the time in the left part of the image and the left hand is in the right side of the image. We use this assumption when there is no hand in the previous frame: for the current frame, we set the blob whose center is more left than the other (whose x coordinate is smaller) as the right hand and vice versa.

4.3. Feature extraction

Several features have been extracted from processed video sequences [13] for further use, for example for sign recognition, clustering or alignment. These features can be separated into three groups according to their character.

4.3.1. Tracking features

In the previous section the method for head and hands tracking was introduced. The output of this algorithm is the position and a bounding ellipse of the hands and the head during the whole video sequence. The position of the ellipse in time forms the trajectory and the shape of the ellipse gives us some information about the hand or head orientation. The features extracted from the ellipse are its width, height and angle between ellipse's major axis and x-axis of the image's coordinate system. The tracking algorithm provides five features per object of interest, 15 features in total.

Gaussian smoothing of measured data in time is applied to reduce the noise. The width of the Gaussian kernel is five, i.e. we calculate the actual smoothed value from two previous, actual and two following measured values.

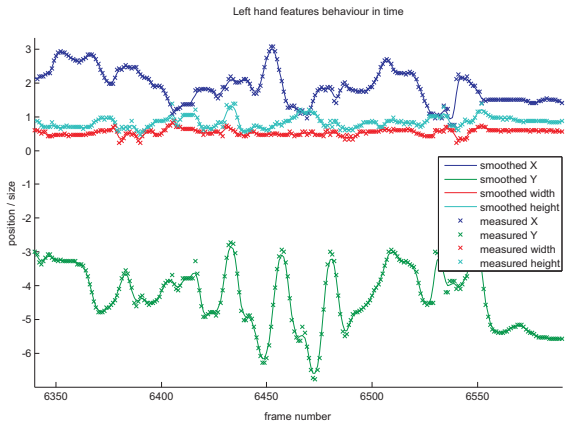


Figure 19: *Smoothing features - example on 10 seconds sequence of four left hand features (x and y position, bounding ellipse width and height).*

We then normalize all the features such that they are speaker independent and invariant to the source resolution. We define a new coordinate system such that the average position of the head center is the new origin and scale with respect to the average width of the head ellipse. Thus the normalization consists of translation to thehead center and scaling.

The coordinate transformations are calculated for all 15 features, which can be used for dynamical analysis of movements. We then calculate differences from future and previous frames and include this in our feature vector:

$$\hat{x}'(t) = \frac{1}{2}(x(t) - x(t-1)) + \frac{1}{2}(x(t+1) - x(t)) \qquad (10)$$

$$= \frac{1}{2}x(t-1) + \frac{1}{2}x(t+1) \qquad (11)$$

In total, 30 features from tracking are provided, 15 smoothed features obtained by the tracking algorithm and their 15 differences.

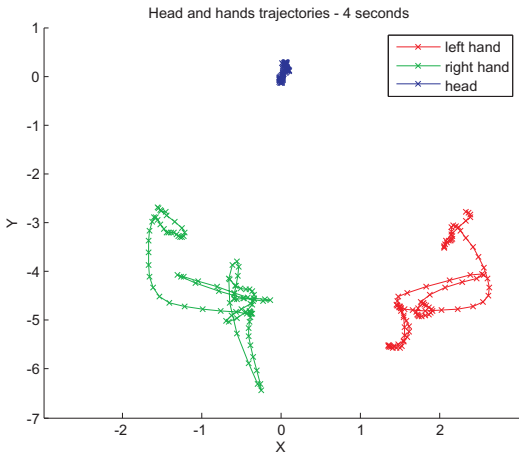


Figure 20: *Trajectories of head and hands in normalized coordinate system.*

4.3.2. DCT features for hands

The hand shape tells us a lot about the performed sign. For some signs it is the only distinguishable feature as there can be very similar signs in trajectory or the signs can be static. In every frame we take a discrete cosine transformation (DCT) of the grayscale template image for each hand. Most of the information in DCT is distributed in the lower frequencies. We extract the upper left triangular part of the DCT matrix which corresponds to those low frequencies. We use 54 features per hand (triangle matrix with width 10). We do not use DC value since this corresponds to the average gray scale value.

Figure 21: *left: reconstructed hand from 54 DCT parameters (by inverse DCT), right: original image.*

The DCT parameters are calculated from a grayscale image where the light areas correspond to the skin colors and the dark areas to the non-skin colors. When a hand is in occlusion with another hand or with the head, then other skin color areas are present in the source image and those areas are included in the DCT parameters. That way a particular hand shape without and with occlusion has different DCT parameters. Only the same hand shape with the same occlusion has the same DCT parameters. This can be either an advantage or a disadvantage depending on the further use.

4.3.3. DCT features for the whole image

At last the DCT parameters of the whole image were calculated. The DCT was calculated from the original image with the skin color mask applied and converted to gray scale. Again the upper left triangular part of the resulting DCT matrix was extracted and the DC parameter was omitted. We have experimented to find that the optimal number of the DCT parameters is 104 (triangular matrix with 14 rows and columns). This number of parameters contains sufficient information about skin color distribution in the image, and does not contain speaker dependent features (such as a specific shape of the head). This information about the specific shapes is stored in higher frequencies, which are cropped from our DCT parameter matrix. We keep only information from the lower frequencies, where the general distribution of skin color in the image is stored.

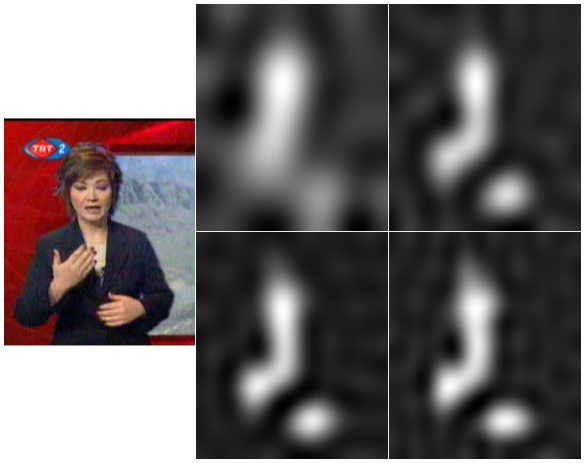


Figure 22: *a) original image, reconstructed images by inverse DCT from b) 35, c) 77, d) 104, e) 135 DCT parameters.*

4.3.4. Summary of extracted features

We have extracted a total of 242 features. Some of them have more information content than others. We extracted all of them in order to have more features for following experiments. The features are summarized in Table 3.

Table 3: *Confusion rates.*

	Number of features	Description
Tracking features	30	head and hands: x, y position width, height, angle of bounding ellipse + derivations
DCT features - hands	2 x 54	DCT parameters for left and right hand shape
DCT features - whole image	104	DCT parameters for skin color distribution in image

4.4. Clustering

Our goal is to cluster two or more isolated signs. If the signs are considered to be the same they should be added to the same cluster. The purpose of clustering is to define a distance (or similarity) measure for our signs. We have to consider that we don't know the exact borders of the sign. The sign news, which contains continuous sign speech, was split into isolated signs by a speech recognizer (see Fig. 23). In the case the pronounced word and the performed sign are shifted, the recognized borders will not fit the sign exactly. We have examined that the spoken word usually precedes the corresponding sign. The starting border of the sign has to be moved backwards in order to compensate the delay between speech and signing. A typical delay of starting instant is about 0.2 seconds backwards. In some cases,

the sign was delayed against the speech, so it's suitable to move the ending instant of the sign forward. To keep our sequence as short as possible, we shifted the ending instant about 0.1 second forward. If we increase the shift of the beginning or the ending border, we increase the probability that the whole sign is present in the selected time interval, at the expense of including parts of previous and next signs into the interval.

Figure 23: *Timeline with marked borders from speech recognizer.*

Now we take two signs from the news whose accompanying speech was recognized as the same word. We want to find out whether those two signs are the same or are different (in case of homonyms). After we extend the borders of those signs (as described above), we suppose that those intervals contain our examined sign and can contain ending part of the previous sign and beginning part of the next sign. The goal is to calculate the similarity of these two signs and to determine if they contain same sign.

Our first experiment was calculating distances between a manually selected short sequence which contains one sign and other same length sequences which were extracted from a longer interval. See Fig. 24, where we calculated distances between a 0.6 second long sequence and the same length sequences extracted from 60 seconds long interval (i.e. we have extracted 0.6 second long sequence starting at frame 1, than from frame 2 and so on). This way we have extracted nearly 1500 sequences and calculated their distances to our selected sequence. This experiment has shown that our distance calculation has high distances for different signs and low distances for similar signs. This was manually evaluated by comparing video sequences containing compared intervals. One may observe that in frame 500, the distance is zero, in fact this is the frame from which we have taken our first, manually selected sequence and then we have compared two same sequences.

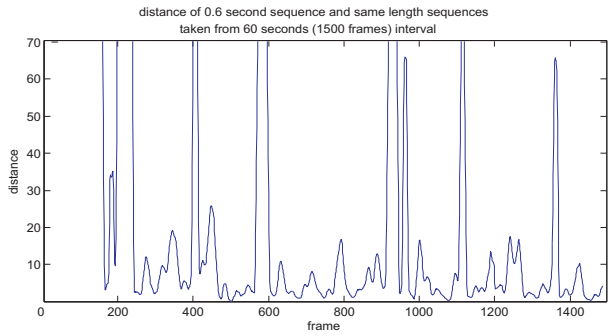


Figure 24: *Experiment: distance calculation between 0.6 second long sequence and same length sequences taken from 60 seconds long interval.*

The distance of the two same length sequences is calculated from tracking features and their derivations. We have not experimented with the hand DCT features, so when the distance is calculated as low, it means that those signs are similar in hand and head position and speed, but can differ in the hand shape.

The distance is calculated in the following way:

1. the difference between corresponding features of two signs is calculated for each frame
2. these differences are squared
3. resulting squared differences are summed over all frames and all those sums for each feature are summed together (we consider same weight for all features)
4. calculated distance is normalized by multiplication with factor 1/*length of sequence*

This distance calculation is very fast, but the problem is when compared sequences contain not only our examined sign, but parts of other signs, then the distance increases. Next problem is time warping. If two same signs are performed with different speed, then the distance increases too.

The solution for both can be usage of a multidimensional DTW (dynamic time warping) algorithm. It's obvious it can handle time warped signals, but we suppose it could solve problem with inclusion of other signs in examined intervals as well: if we calculate the lowest cost path in DTW matrix, than the part where the cost grows slowly correspond to comparing two same signs, otherwise when the cost grows fast two different signs are compared.

Another possible solution is using HMMs (Hidden Markov Models), where one sign is represented by one HMM and each HMM has some additional hidden states at the beginning and at the end, which correspond to the "noise" before and after given sign. Parts of neighboring signs represent this noise. When we cluster more than two signs, we use the following method:

1. Calculate pair wise distances between all compared signs, store those distances in upper triangular matrix
2. Group two most similar signs together and recalculate the distance matrix (it will have one less row and column)
3. Repeat step 2. until all signs are in one group
4. Mark the highest difference between two distances, at which two signs were grouped together in following steps, as distance up to which the signs are in the same cluster (see Fig. 25)

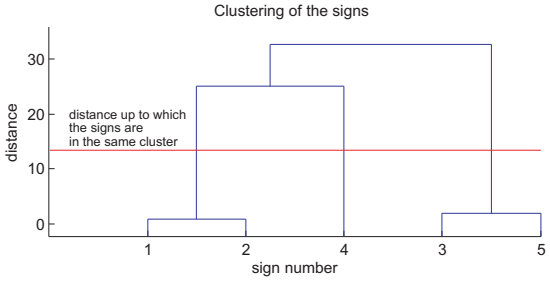


Figure 25: *Dendogram - grouping signs together at different distances.*

Clustering was implemented in Matlab as standalone and server applications. The server application receives a list of signs from a client, calculates their distances, clusters those signs and sends the cluster information for each sign back to the client.

5. APPLICATION AND GUI

5.1. System Design and Application Details

The user interface and the core search engine are separated and the communication between is being done via using TCP/IP

socket connections. This design is also expected to be helpful in the future since a web application is planned to be built using this service. This design can be seen in the Fig. 26.

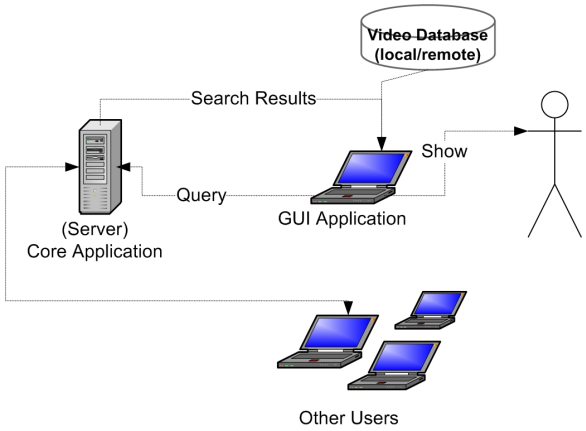


Figure 26: *System Structure.*

In Fig. 27, the screenshot of the user interface is shown. There are mainly five sections in it. The first is the "Search" section where the user inputs a word or some phrases using the letters in Turkish alphabet and sends the query. The application communicates with the engine on the server side using a TCP/IP socket connection, retrieves data and processes it to show the results in the "Search Results" section. The results being done in the same news file is grouped together and shown with date and time information in the tree structure so that the user can also use the application to scan the news archives. A "recent searches" menu is added to the search box aiming to cache the searches and increase the service time. But the user can clear the search history to retrieve the results from the server again for the recent searches. The relevance of the returned results (with respect to the speech) is shown using stars (0 to 10) in the tree structure to inform the user about the reliability of the found results. Moreover, the sign clusters is shown in paranthesis, next to each result.

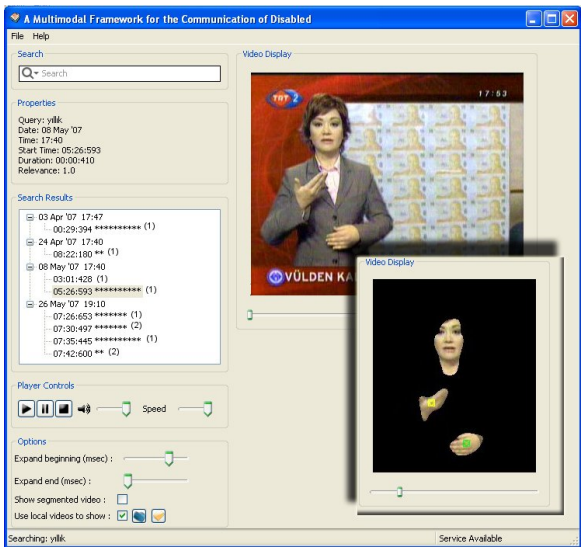


Figure 27: *Screenshot of the User Interface.*

When the user selects a result, it is loaded and played in the "Video Display" section. The original news video or the segmented news video is shown in this section according to the user's "Show segmented video" selection. The segmented video is added separately to figure above to indicate this. In addition to video display, the "Properties" section also informs the user about the date and time of the news, starting time, duration and the relevance of the result. "Player Controls" and "Options" enable the user to expand the duration to left/right or play with the volume/speed of the video to analyze the sign in detail. Apart from using local videos to show in the display, one can uncheck "Use local videos to show" and use the videos on the web. But the speed of loading video files from the web is not satisfactory since the video files are very large.

5.2. Used Tools

The ease and power of `python` language with the `wxPython`'s GUI bindings gave the most help in the user interface creation. `wxFormBuilder` enabled us to design the UI file separately. In addition, `py2exe` is used to create executables from the python code and finally `nsis` is used to create a standalone program setup from separate files and folders.

6. CONCLUSIONS

We have developed a Turkish sign dictionary that can be used as tutoring videos for novice signers. The dictionary is easy to extend by adding more videos and provides a large vocabulary dictionary with the corpus of the broadcast news videos. The application is accessible as a standalone application and will soon be accessible from the Internet.

7. ACKNOWLEDGEMENTS

This work is developed during the eNTERFACE'07 Summer Workshop on Multimodal Interfaces, İstanbul, Turkey and supported by European 6^{th} FP SIMILAR Network of Excellence.

8. REFERENCES

[1] U. Zeshan, "Aspects of Türk Isaret Dili (Turkish Sign Language)", *Sign Language & Linguistics*, vol. 6, no. 1, pp. 43–75, 2003. 37

[2] C. Allauzen, M. Mohri, and M. Saraçlar, "General Indexation of Weighted Automata- Application to Spoken Utterance Retrieval", in *HLTNAACL*, 2004. 38, 39

[3] L. R. Rabiner and M. Sambur, "An Algorithm for Determining the Endpoints of Isolated Utterances", *Bell System Technical Journal*, vol. 54, no. 2, pp. 297–315, 1975. 38

[4] M. Saraçlar and R. Sproat, "Lattice-based Search for Spoken Utterance Retrieval", in *HLTNAACL*, 2004. 38, 39

[5] "HTK Speech Recognition Toolkit". `http://htk.eng.cam.ac.uk`. 38

[6] "AT&T FSM & DCD tools". `http://www.research.att.com`. 38

[7] N. Otsu, "A threshold selection method from gray-level histograms", *IEEE Trans. Systems, Man, and Cybernetics*, vol. 9, no. 1, pp. 62–66, 1979. 40

[8] J. D. Tubbs, "A note on binary template matching", *Pattern Recognition*, vol. 22, no. 4, pp. 359–365, 1989. 41

[9] V. Vezhnevets, V. Sazonov, and A. Andreeva, "A Survey on Pixel-based Skin Color Detection Techniques", in *Graphicon*, pp. 85–92, 2003. 41

[10] "Intel Open Source Computer Vision Library". http://opencvlibrary.sourceforge.net/. 42, 43

[11] "OpenCV Blob Extraction Library". http://opencvlibrary.sourceforge.net/cvBlobsLib. 43

[12] N. Tanibata, N. Shimada, and Y. Shirai, "Extraction of Hand Features for Recognition of Sign Language Words", in *In International Conference on Vision Interface*, pp. 391–398, 2002. 43

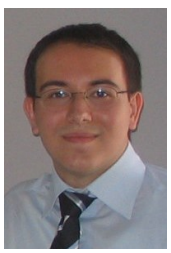
[13] S. Ong and S. Ranganath, "Automatic Sign Language Analysis: A Survey and the Future beyond Lexical Meaning", *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 27, no. 6, pp. 873–891, 2005. 44

9. BIOGRAPHIES

Oya Aran received the BS and MS degrees in Computer Engineering from Boğaziçi University, İstanbul, Turkey in 2000 and 2002,respectively. She is currently a PhD candidate at Boğaziçi University working on dynamic hand gesture and sign language recognition. Her research interests include computer vision, pattern recognition and machine learning.
Email: aranoya@boun.edu.tr

Ismail Ari received his BS degree in computer engineering from Boğaziçi University in 2006. He is currently an MS student in the same department progressing a thesis entitled "Facial Feature Tracking and Recognition for Sign Language" under the supervision of Prof. Lale Akarun. His research interests are computer vision, image processing and computer graphics.
Email: ismailar@boun.edu.tr

Pavel Campr was born in Zatec, Czech Republic. He attended Faculty of Applied Sciences, University of West Bohemia in Pilsen from 2000-2005, where he received master's degree in cybernetics with honors. As a Ph.D. candidate at the Department of Cybernetics, University of West Bohemia in Pilsen, his research interests include sign language and gesture recognition, image and signal processing.
Email: campr@kky.zcu.cz



Erinç Dikici received his B.S. degree in Telecommunications Engineering from İstanbul Technical University (ITU), Turkey, in 2006. He is currently an M.S. student in Electrical and Electronic Engineering Department at Boğaziçi University. His research interests include image and audio/speech processing, specifically, speaker clustering, music transcription, and musical instrument synthesis.
Email: erinc.dikici@boun.edu.tr

Marek Hruz was born in Martin, Slovakia. He attended Faculty of Applied Sciences, University of West Bohemia in Pilsen, Czech Republic from 2001-2006, where he received master's degree in cybernetics. As a Ph.D. candidate at the Department of Cybernetics, University of West Bohemia in Pilsen, his research interests include sign language and gesture recognition, image and signal processing.
Email: mhruz@kky.zcu.cz

Deniz Kahramaner was born in İstanbul, Turkey. He is currently attending Robert College in İstanbul. His research interests include sign language and gesture recognition.
Email: dennizk@gmail.com

Siddika Parlak received her B.S. degree in the Electrical and Electronics Engineering Department of Boğaziçi University in 2006. She is currently an M.S. student at the same department and research asistant in BUSIM (Boğaziçi University Signal and Image Processing) Laboratory. Her main interests are in the field of Speech and Language Processing. She is pursuing her M.S. degree under the supervision of Prof. Murat Saraçlar.
Email: siddika.parlak@boun.edu.tr

Lale Akarun received the B.S. and M.S. degrees in electrical engineering from Boğaziçi University, İstanbul, Turkey, in 1984 and 1986, respectively, and the Ph.D. degree from Polytechnic University, Brooklyn, NY, in 1992. From 1993 to 1995, she was Assistant Professor of electrical engineering at Boğaziçi University, where she is now Professor of computer engineering. Her current research interests are in image processing, computer vision, and computer graphics.
Email: akarun@boun.edu.tr

Murat Saraçlar received his B.S. degree from Bilkent University, Ankara, Turkey in 1994. He earned both his M.S.E. and Ph.D. degrees from Johns Hopkins University, Baltimore, MD, USA in 1997 and 2001 respectively. He worked on automatic speech recognition for multimedia analysis systems from 2000 to 2005 at the AT&T Labs Research. In 2005, he joined the Department of Electrical and Electronic Engineering at Boğaziçi University as an assistant professor. His main research interests include all aspects of speech recognition, its applications, as well as related fields such as speech and language processing, human-computer interaction and machine learning. He authored and co-authored more than two dozen papers in refereed journals and conference proceedings. He has filed four patents, both internationally and in the US. He has served as a reviewer and program committee member for various speech and language processing conferences and all the major speech processing journals. Dr. Saraclar is currently a member of ISCA and IEEE, serving as an elected member of the IEEE Signal Processing Society Speech and Language Technical Committee (2007-2009).
Email: murat.saraclar@boun.edu.tr

MULTIMODAL SPEAKER IDENTITY CONVERSION - CONTINUED -

Zeynep Inanoglu [1], *Matthieu Jottrand* [2], *Maria Markaki* [3], *Kristina Stanković* [4], *Aurélie Zara* [5], *Levent Arslan* [6], *Thierry Dutoit* [2], *Igor S. Panžić* [4], *Murat Saraçlar* [6], *Yannis Stylianou* [3]

[1] University of Cambridge, United Kingdom
[2] Faculté Polytechnique de Mons, Belgium
[3] University of Crete, Greece
[4] University of Zagreb, Croatia
[5] France Télécom R&D, France
[6] Boğaziçi University, Turkey

ABSTRACT

Being able to convert a given the speech and facial movements of a given source speaker into those of another (identified) target speaker, is a challenging problem. In this paper we build on the experience gained in a previous eNTERFACE workshop to produce a working, although still very imperfect, identity conversion system. The conversion system we develop is based on the late fusion of two independently obtained conversion results: voice conversion and facial movement conversion.

In an attempt to perform parallel conversion of the glottal source and excitation tract features of speech, we examine the usability of the ARX-LF source-filter model of speech. Given its high sensitivity to parameter modification, we then use the code-book based STASC model.

For face conversion, we first build 3D facial models of the source and target speakers, using the MPEG-4 standard. Facial movements are then tracked using the Active Appearance Model approach, and facial movement mapping is obtained by imposing source FAPs on the 3D model of the target, and using the target FAPUs to interpret the source FAPs.

KEYWORDS

Voice conversion – Speech-to-speech conversion – Speaker mapping – Face tracking – Avatar control

1. INTRODUCTION

This eNTERFACE'07 project is a continuation of a project started during eNTERFACE'06 in Dubrovnik [1], in which we aimed at converting a given *source speaker* speech and facial movements into those of another (identified) *target speaker*. Such a conversion is typically based on some (separate) parametric models of the speech and facial movements for both speakers (Fig. 1). Two streams of (time-varying) parameters (one for the speech model, one for the face model) are first estimated from an audio-video file of the source speaker; some of these parameters are modified using *mapping functions*; the modified parameter streams are finally converted into an audio-video file which should hopefully be identified as originating from the target speaker.

The final quality of the conversion therefore depends on the quality of the rendering obtained by the parametric models and on the efficiency of the mapping functions, which both result from design choices.

Rendering quality can easily be estimated by copy-synthesis experiments: one takes an audio-video file as input, estimates parameters and performs rendering without modifying the parameters. Errors can be due to *modeling errors* (the model is not able of capturing all the details in the data) and/or to *estimation errors* (the model, when used for rendering, is capable of producing perfect copysynthesis if it is fed with some optimal parameter stream, but the parameter estimation algorithm cannot find the best parameter values). This leads to a classical modeling tradeoff: if too simple, a model lends itself to modeling errors; if too complex, it minimizes modeling errors but opens the doors to estimation errors.

Mapping should produce a sensation of identity conversion while not degrading too much the rendering quality obtained with copy synthesis. Here again, a tradeoff usually has to be made: while applying smoothed mapping preserves copy-synthesis quality, it only partially produces identity conversion; conversely, applying hard mapping modifies the impression of identity but often significantly degrades quality [2].

In addition to being dependent on the models and mapping methods they use, speaker conversion systems are characterized by the type of data they are based on. Mapping functions are usually trained from *aligned* data between source and speaker, although a new trend is to design mapping functions from source and target speakers not uttering the same sentences.

Conversion approaches also differ by the assumptions they make on the size of the available data from the source and speaker, for training the mapping functions. Being able to train an efficient mapping function from limited data is more challenging (and often closer to real applications).

In this project, it is assumed that a large amount of aligned speech data can be recorded from both the source and target. As a matter of fact, even in such advantageous conditions, the state-of-the-art in voice conversion has not yet reached a level which would make it a widely usable tool for commercial applications. In contrast, we assume that only a photo of the target speaker is available. A typical application of this project is therefore that of a human actor controlling the speech and facial movements of a 3D character whose face and voice is well-known to the audience, and from whom a large amount of speech data is available.

The paper is organized as follows. Section 2 summarizes the speech model and mapping function we tested in [1] and examines the new choices made in this year's project. In Section 3, we summarize the face model and mapping function (which has not changed from [1]). Experiments using these models and mappings (using the eNTERFACE06_ARCTIC database created last year [1]) are detailed in Section 4, followed by an analysis of our results, and perspectives, in Section 5.

[1] http://www.enterface.net/enterface06/docs/results/databases/eNTERFACE06_ARCTIC.rar

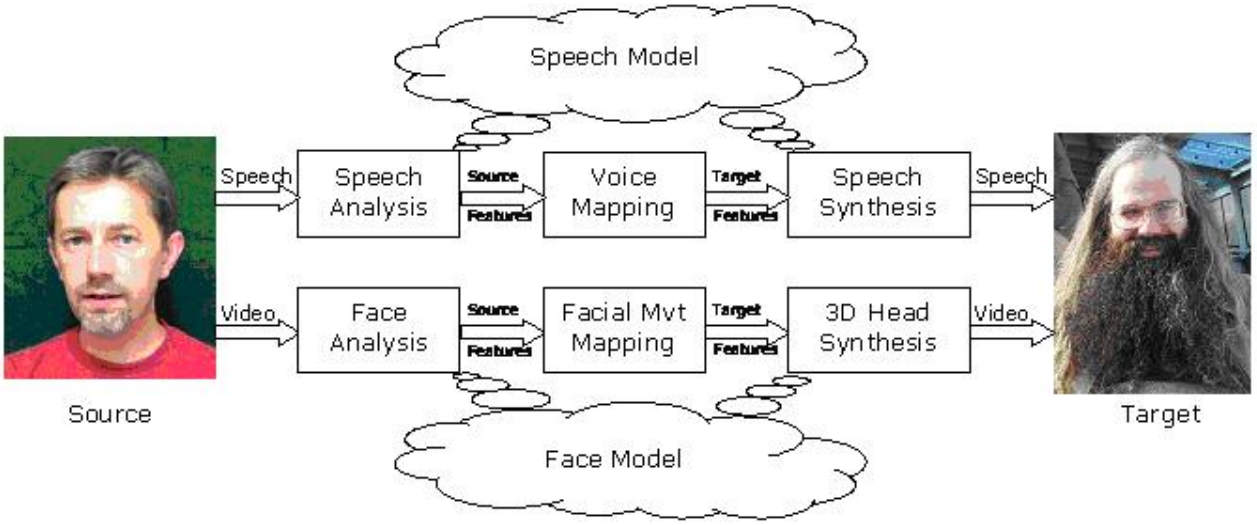


Figure 1: *Principles of speaker identity conversion, using source speech and facial movements, mapping them to target speech and face, and producing target-like speech and facial movement.*

2. SPEECH ANALYSIS, MAPPING, AND SYNTHESIS FOR VOICE CONVERSION

In [1], a number of choices had been made relatively to the modeling and mapping trade-offs mentioned in Section 1.

Speech was modeled using Residue Excited Linear Prediction (RELP), which has the advantage of providing transparent copy synthesis, but consequently provided few means of mapping the glottal source signals from source to target speaker. A mapping function was then applied, frame-by-frame, on the vocal tract filter of the source,

$1/A_s(z)$, based on Gaussian Mixture Models (GMM) of the source and target Mel-Frequency Cepstrum Coefficient (MFCC) distributions [3] (Fig. 2). This produced an estimate of the vocal tract filter of the target, $1/\tilde{A}_t(z)$. The original part of this speech conversion system resided in an additional mapping step. In order to increase the acoustic similarity between converted speech and target speech, we used a large database of target speech, and applied a units selection principle, similar to that used in unit selection for text-to-speech synthesis: we searched in the target database for a sequence of real target vocal tract filters $\{\ 1/A_t(z)\ \}$ whose distance to the sequence of mapped filters $\{\ 1/\tilde{A}_t(z)\ \}$ was minimized. The search for optimal target sequences was based on dynamic programming, for additionally optimizing the length of the real target filter sequences used (in order to avoid discontinuities when switching from one sequence to another). Converted speech was finally obtained by filtering some excitation signal with the sequence of real target vocal tract filters $\{\ 1/A_t(z)\ \}$.

One of the main conclusions of [1] was that, if the *target speaker LP residual* was filtered by the sequence of mapped vocal tract filters $\{\ 1/A_t(z)\ \}$ the converted speech sounded like "processed" target speech, and was therefore clearly identifiable as originating from the target speaker (but its quality was somehow discontinuous)[2]. In contrast, when the *source speaker LP residual* was used to drive the sequence of mapped vocal tract filters (as in Fig. 2), a lot of the source speaker identity was retained in the converted speech.

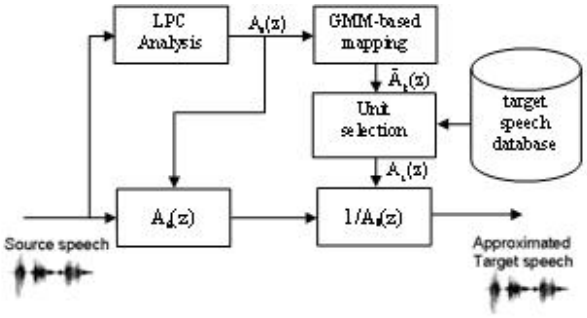


Figure 2: *Voice conversion in [1].*

This led us to focus initially this year on source/tract separation, with separate mappings for source and tract parameters. The main initial idea was to use a recently developed source-tract model, the so-called AutoRegressive eXogeneous Lijlencrants-Fant model (ARX-LF) [4]. We also tested another mapping function (other than the one used in [1]), called STASC [2] and based on a simpler algorithm than [3] while still producing efficient vocal tract mapping.

2.1. ARX-LF

The source-filter model of speech production hypothesizes that a speech signal is formed by passing an excitation through a linear filter. In physiological terms, the excitation refers to the signal produced at the glottis and the filter represents the resonances caused by the cavities in the vocal tract. Linear prediction analysis is a basic source-filter model which assumes a periodic impulse train as the source signal for voiced sounds and white noise for unvoiced speech. Most voice conversion frameworks assume an LPC based analysis-synthesis approach, where only the LPC-based parameters are converted and excitations are left untouched. More elaborate parametric models of the excitation do exist and are interesting in terms of joint vocal tract and source conversion. In our experiments we have chosen to evaluate the LF model which models the voiced excitation by

[2]Notice that, since the length of the source and target files were generally different, using target speaker excitation with modified source speaker vocal tract parameters implied to perform some alignment of the source and speaker files. This was achieved by applying Dynamic Time Warping (DTW) between source and target utterances.

approximating the glottal flow derivative (GFD) with three parameters. Figure 3 illustrates a GFD waveform generated by the LF model. The three parameters of interest are open quotient, asymmetry coefficient and closed phase quotient.

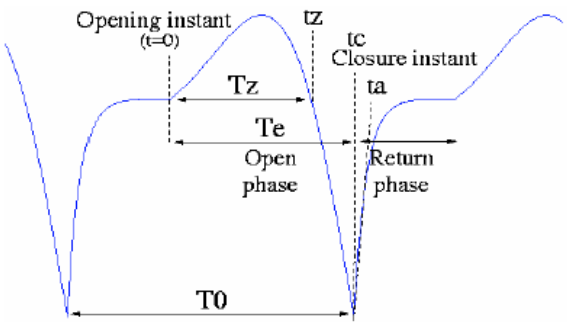


Figure 3: *Glottal flow derivative produced by the LF model.*

With the adoption of LF glottal pulse, the speech signal can then be expressed by an ARX (auto regressive exogenous) process [4]:

$$s(n) = \sum_{k=1}^{p} a(k)s(n-k) + u(n) + e(n) \qquad (1)$$

where u is the LF waveform and e is the residual noise and $a(k)$ are the coefficients of the p^{th} order filter representing the vocal tract. Once an LF waveform is found for a given speech frame, deriving the filter coefficients is a trivial task. For ARX-LF analysis we have used an implementation based on the work of A. Moinet and N. D'Alessandro. Our implementation does not incorporate a residual noise factor, therefore $e(n)$ is always 0 when synthesizing speech with our implementation of the ARX-LF model. The steps of the ARX-LF analysis can be summarized as follows:

1. Extract pitch at regular intervals (5ms) from the wav file.
2. Find the point of initial glottal closure instant (GCI) in each voiced segment in the utterance.
3. Use the first GCI as the anchor point to determine the remaining glottal closure instants in each voiced segment.
4. For each pitch period, search an LF derivative waveform codebook for the waveform which minimizes the error between actual speech frame and the ARX-LF synthesized frame.
5. Obtain the parameter set which produced this waveform (this is stored in the LF derivative waveform codebook, along with each waveform).
6. Given the LF parameter for each frame, determine the filter coefficients.

2.2. STASC

Speaker Transformation Algorithm using Segmental Codebooks (STASC) converts the voice of a source speaker to that of a target speaker maintaining high speech quality [2]. Figures 4 and 5 schematically depict training and conversion stages of STASC which are briefly described below.

2.2.1. Training

First, alignment between the same sentences from source and target speaker is automatically performed by aligning Sentence HMMs for the (source, target) utterance pairs [2]. An advantage of this model is that it doesn't require any knowledge of the text or the language spoken. In every frame of source and target speech, acoustic feature vectors are extracted, including MFCCs, logarithm of energy and voicing probability - as well as their delta coefficients (18 features in total). For each source speaker utterance, the segmental k-means algorithm initializes an HMM whereas Baum-Welch algorithm is employed for training. Both source and target speaker utterances are force-aligned with this HMM using the Viterbi algorithm. A new state is added to the HMM topology every 40 ms of source speaker utterance. The mapping between the acoustic parameters of source and target speakers can subsequently be obtained based on this aligned data. For each HMM state, line spectral frequencies (LSF), fundamental frequencies (F_0), durations, energy and excitation parameters are computed. Their mean values over the corresponding source and target HMM states are stored in the source and target codebooks, respectively.

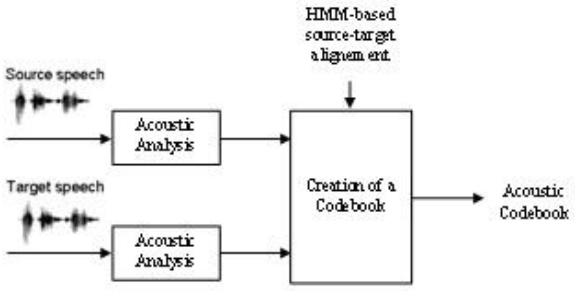


Figure 4: *Training stage flowchart of STASC.*

2.2.2. Conversion

Vocal tract and excitation characteristics are independently modified. Line spectral frequencies (LSF) are selected to represent vocal tract characteristics of each speaker since they are closely related to formant frequencies and, moreover, they can be reliably estimated. After pitchsynchronous linear prediction (LP) analysis of source speaker utterance, LP parameters are converted to LSFs.

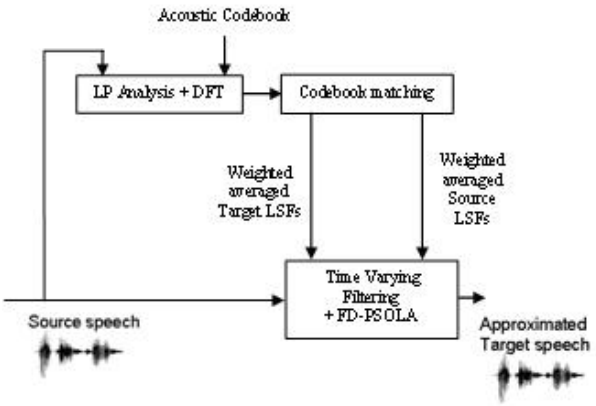


Figure 5: *Conversion stage flowchart of STASC (after [2]).*

The (weighted) distance d_m between the LSF vector u of the input source frame and the m^{th} source codebook LSF vector

is given by the following equations:

$$d_m = \sum_{n=1}^{p} k_n |u_n - C^s_{mn}| \text{ for } m = 1, \cdots, M \quad (2)$$

$$k_n = \frac{1}{\text{argmin}(|u_n - u_{n-1}|, |u_n - u_{n+1}|)}, n = 1, \cdots, P \quad (3)$$

where m is the codebook entry index, M is the codebook size, n is the index of LSF vector entries, P is the dimension of LSF vectors (order of LP analysis), u_n the n^{th} entry of the LSF vector for the input source frame, C^s_{mn} is the n^{th} entry of the m^{th} source codebook LSF vector, and k_n are the LSF weights.

STASC further estimates the vocal tract characteristics of the target speaker, i.e. the n^{th} entry of the estimated target LSF vector $\bar{y}_n$:

$$\bar{y}_n = \sum_{m=1}^{M} \upsilon^m C^t_{mn} \text{ for } n = 1, \cdots, P \quad (4)$$

where C^t_{mn} is the n^{th} entry of the m^{th} target codebook LSF vector, υ^m are the normalized codebook weights, and $\bar{a}W$ is used to show that W is obtained through weighted average of codebook entries. The target LSF vector $\bar{y}_n$ is converted into target LP coefficients in order to obtain target vocal tract spectrum $\hat{H}^t(\phi)$ (where ϕ is the angular frequency in radians. The frequency response $H^{VT}(\phi)$ of the time-varying vocal tract filter for the current frame is then given by:

$$H^{VT}(\phi) = \frac{\hat{H}^t(\phi)}{H^s(\phi)} \text{ or } H^{VT}(\phi) = \frac{\hat{H}^t(\phi)}{\bar{H}^s(\phi)} \quad (5)$$

where the source vocal tract spectrum can be estimated using either the original or "estimated" LP coefficients (see [2] for details).

The frequency-domain pitch synchronous overlap-add algorithm (FD-PSOLA) is finally used for adapting the pitch of the source to the pitch range of the target.

3. FACIAL EXPRESSION ANALYSIS, MAPPING AND SYNTHESIS FOR FACE ANIMATION

Starting from video samples (face and shoulders) of the source speaker, and a photograph of the target speaker, we want to produce video samples of the target speaker acting and speaking as the source speaker.

Solving such a problem implies the following steps: (1) analyze the facial movements of the source speaker, using some parametric 3D facial model; (2) estimate the parameters of the same 3D model for the target speaker (3) normalize the facial movements of the source speaker relatively to some biometrical features of his face (4) animate the target 3D model by applying the same relative face movements as those measured on the source speaker.

While we had only considered steps 1, 3, and 4 in our previous work [1], by using an avatar that was already available (without having to build an avatar corresponding to an identified target speaker), this year's project browsed all 4 steps.

3.1. Face modelling

The parametric 3D facial model used in our project is the one defined in the MPEG-4 standard [5]. It is composed of Facial Definition Parameters (FDPs), which define the 3D position of a number of reference points on the face, Facial Animation Parameters (FAPs), which define frame-byframe define movements of those reference points [3].

3.2. Facial movement mapping

Face movement mapping is made easy by the fact that FAPs are actually not sent as absolute movements, but rather as movements normalized by the so-called Facial Animation Parameter Units (FAPUs), which can be obtained from the FDPs (they are basically some important biometric features, such as interocular distance, mouth width, etc.). Applying measured face movements from source face to target face is thus as simple as imposing source FAPs on the 3D model of the target, and using the target FAPUs to interpret the source FAPs.

3.3. Face analysis

In [1], the analysis part mainly consisted in three tasks: detecting and tracking the face region, extracting facial features such as lips, nose, eyes and brows, and tracking these features throughout the source video.

The first part of analysis was done by computing an ellipse outlining the face. The ellipse parameters were estimated from an optical flow [6]. The centre of the ellipse, approximating the centre of the head, was used to track global head movements by assuming that the head centre moves around a sphere situated on the top of spinal cord. By projecting the displacement of the head onto this sphere, the angles of head rotation (pitch, yaw and roll angles) were approximately estimated.

The next task was to define and track the useful feature points for face components, which are lips, eyes, and eyebrows in this scenario. For this purpose, Active Appearance Model (AAM) approach [7] was used as a means of modeling and tracking the face components. Since AAMs require model training phase before they are used to process entire sequence of frames of a given video, a set of frames which cover a wide range of different facial expressions was selected and used to train the AAM. This training step requires manual labeling of feature points located around the desired face components over all frames in the training set. After the model was created by using the set of manually labeled points, feature tracking for face components could be performed easily. From the position of this feature points and global head movements, facial animations parameters were computed.

In this year's project, we focused on improving the global head movement tracking. In [1], global head movements and facial feature tracking were independent tasks. A way to improve the approximation of head rotation angles is to exploit features tracked from the AAM, especially rigid features such as eyes corners or point located at the beginning of the nose, between nostrils.

As we don't know the depth of these points it is not possible to directly compute rotation angles. In [8], the 2D image coordinates of feature points are mapped to 3D by assuming the projection is approximately scaled orthographic. Another solution is to compute an orthogonal projection of the centre of the head on the plane defined by the three points and compute relations between them. When the face is in a frontal position, the centre and its projection are one and the same in 2D.

Consequently, to compute relations between the eyes corners, the bottom of the nose and the head centre, we used a frontal head pose picture of the source. As the 3D rotation

[3] Notice that how the movement of each reference point influences the final face rendering is not defined by MPEG4. Each face rendering software does it its own way.

change angles and distances in 2D, we used ratios between features points to find points necessary to localize the centre.

Let *RE* be the corner of the right eye, *LE* the corner of the left eye, *N* the bottom of the nose and *C* the center of the head (Fig. 6). Let *RELE* be the straight line defined by *RE* and *LE*, *NRE* the straight line defined by *RE* and *N*, *REC* the straight line defined by *RE* and *C*, and *LEC* the straight line defined by *LE* and *C*. Let $I_{rec-nre}$ be the intersection between *REC* and *NRE*, and $I_{lec-nre}$ be the intersection between *LEC* and *NRE*. The following ratios are then computed:

$$R_1 = \frac{d(I_{rec-nre}, re)}{d(re, n)}, R_2 = \frac{d(I_{lec-nre}, le)}{d(le, n)} \quad (6)$$

where d is the Euclidian distance. For each frame, R_1 and R_2 are used to find the position of each intersection points on the corresponding segments and from the location of these points, the 2D coordinates of the head centre projection are computed.

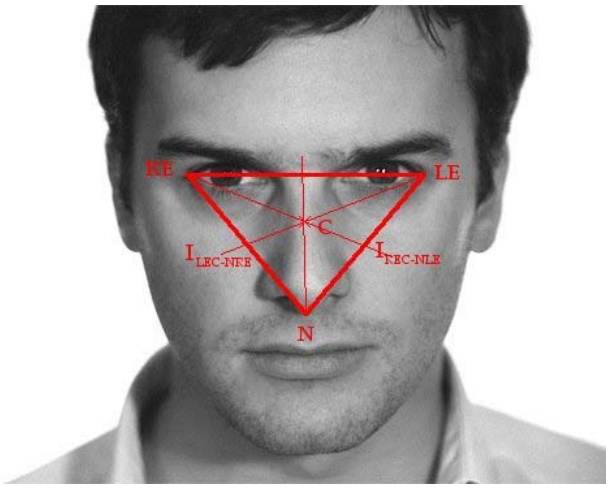


Figure 6: *Relations between the three feature points and the center of the head.*

To compute the depth of the centre and the rotation angles, we used the same method as in [1].

4. EXPERIMENTS AND RESULTS

In order to design the application depicted in Figure 1, we needed to choose a source and target speaker, make sure we could have access to a large amount of speech from the target (for the speech synthesis module), of a reasonable amount of aligned speech data for source *and* target (for the voice mapping module), and of some test speech and video data from the source (in order to test the complete system). The eNTERFACE06_ARCTIC database meets these requirements for the source speaker. It is composed of 199 sentences, spoken by one male speaker, and uniformly sampled from the CMU_ARCTIC database [2]. For each sentence, an .txt, a .avi, and a .wav file are available. The .avi file contains images with 320x240 pixels (Fig. 3), 30 frames per second, of the speaker pronouncing the sentence (Fs=44100 Hz). The .wav file contains the same sound recording as in the .avi file, but resampled to 16 kHz.

In the next paragraphs we expose the results we have obtained with the ARX-LF model and the STASC algorithm; we also report on an improvement we have made to the face analysis module of [1], and we show results related to modeling the target face, and to animating it.

4.1. ARX-LF

Copy synthesis experiments were carried out using the ARX-LF framework (Fig. 7). Here we discuss an informal evaluation of the stimuli.

4.1.1. Copy-synthesis with fixed excitation

In this very simple copy synthesis experiment, a random LF parameter set was chosen from the codebook and the same parameter set was copied to produce all the voiced frames of an input utterance. The purpose of this experiment was twofold: to assess the quality of analysis-resynthesis when the source parameters are modified independently from the filter coefficients and to make an informal evaluation of how much difference in voice quality is perceived when different LF parameter sets are copied throughout an utterance.

The quality of the resynthesis with fixed excitation was acceptable and did not result in many artifacts. There were also perceived differences in voice quality between some parameter sets of the codebook, particularly regarding the breathiness and brightness of the voice. However, there were also many codebook entries which produced no perceptual difference when copied throughout an utterance. This led us to become more skeptical about the potential contribution of LF parameters to speaker identity.

4.1.2. Transplantation of target speaker features to source speaker

Since the results of last year's project indicated the existence of a lot of speaker specific information in the residual, our goal was to code some of that information with the LF parameters using the proposed framework. Therefore a natural experiment to conduct was to align the source and target frames in a parallel utterance and copy the relevant parameters of the target onto the source.

We have compared versions of utterances where only the filter coefficients were copied over, versus ones where both filter coefficients and LF parameters were copied over. We found that for our speaker conversion task, the speaker identity was still mostly coded in the filter coefficients and copying the LF parameters made no perceptual difference in terms of speaker identity. In fact, changing only the LF parameters of the source frames to match those of the target resulted in a stimulus which sounded very much like the source. This again clearly questions the correlation between of LF parameters and speaker identity.

In addition, we have tried copy-synthesis experiments on an emotion conversion task using an expressive speech corpus from a female speaker. For neutral to angry copysynthesis on stimuli, we also found that most of the harshness of anger was coded in the filter parameters rather than the source. On the other hand, going from neutral to sad speech, there was a positive contribution of the LF parameters to perception of sadness.

There are possible reasons why the ARX-LF framework may not have helped as well as we initially hope. We list them here, for future work:

- Glottal waveforms do not contain as much voice quality information as expected.
- It is possible that the current parametric framework is not adapted to model important voice quality information such as spectral tilt. (A suggestion here is to apply pre-emphasis to the speech signal during analysis so as to flatten the vocal tract spectrum and force the modeling of spectral tilt in the LF parameters.)

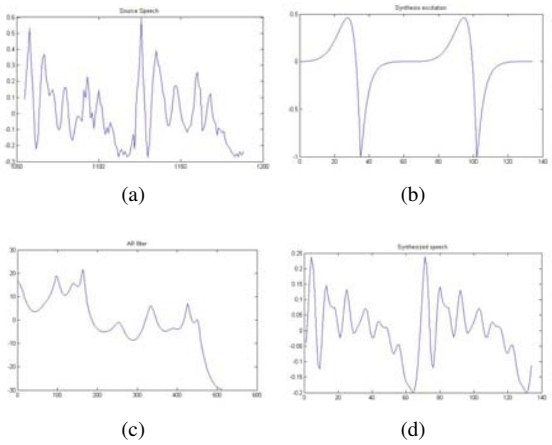


Figure 7: *An example of ARX-LF analysis-synthesis: (a) original speech frame; (b) Estimated Glottal derivative waveform; (c) Frequency response of the estimated vocal tract filter; (d) resulting synthetic speech frame.*

- Adding a component modeling residual noise may help. Probably this component is not even *really* noise-like.
- A better automatic GCI detection algorithm is crucial for this implementation but for the experiments performed here, the GCIs were manually corrected for best analysis/resynthesis results.
- The codebook of LF parameters may be expanded to contain more extreme parameter values and with higher resolution.

4.2. STASC

We reimplemented the STASC voice conversion system described previously. Although the implemented system is very similar to the one described in [2], some differences exist. We had at our disposal a sentence HMM alignment that has been produced as described in [2][4]. Instead of using the average of HMM state features as elements of the mapping codebook, we converted the alignment into a frame-to-frame alignment, leading to a frame-based mapping codebook of LSF parameters, *F0* and energy. We then cleaned this codebook by using the spectral distance confidence measure, the *F0* distance confidence measure and the energy distance confidence measure described in [2]. The original codebook counts 81870 pairs of frames while the cleaned codebook contains 69076 pairs. The frame length is 30 ms, with a shift of 10 ms. The second difference occurs in the conversion step, where we do not compute AR coefficients on pitch synchronous frames, but also on 30 ms length frames, shifted every 10 ms. Two types of tests have been made, referenced below as *method a* and *method b*.

Method *a* is depicted in Fig. 8. LSF coefficients are computed for each frame of the source utterance. The N nearest codebook entries are selected, and a weighted sum of corresponding target LSF coefficients is computed. LSF coefficients of the source and of the target allow, after conversion into AR coefficients, to build the corrective filter as the division of the vocal tract frequency response of the target by that of the source. Each source frame is then filtered by the corrective filter. Speech is resynthesized by overlap-adding the resulting voice-converted frames.

[4]They have been added to the in the eNTERFACE06_ARCTIC archive.

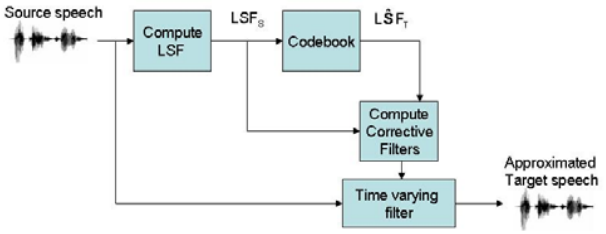


Figure 8: *Conversion method a.*

In order to better check the efficiency of the vocal tract conversion, another synthesis method has been implemented. This method cannot be seen as a genuine conversion system, as it uses the target utterances, but its results give an idea on how well the vocal tract (alone) is transformed from one speaker to another. It uses the target LP residual, and filters it with LSF codebook outputs converted into AR coefficients. The system is represented in Fig. 9.

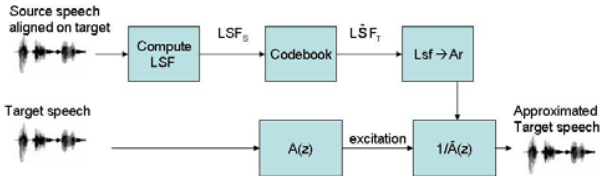


Figure 9: *Conversion method b.*

In theory, if the codebook-based mapping of the valv tract from source to target was perfect, the approximated target speech should be identical to the target speech.

Notice that, as we use the source frames to find the approximated LSF coefficients of target frames and then use the target wavefile for synthesis, source frames and target frames need to be aligned. This alignment is done by dynamic time warping, using the implementation provided by Dan Ellis [5].

Method *a*, the real conversion test, leads to a very good quality, but the identity change is incomplete: it moves from source to something between source and target voices, but still closer to the source. A pitch modification of 10 percent on the converted speech improves a bit the similarity with the target speaker. This pitch modification was done using Praat[6] software.

Method *b* lead to a much better similarity with target speaker (but again, we are using the target excitation, so tests *a* and *b* cannot be compared) although the quality of the converted speech is degraded. Actually, the quality and similarity of the converted voice is somehow similar to the ones obtained from eNTERFACE06 project 4 [1], but the algorithm used this year is very much simpler.

4.3. Face modeling

One of the tools we used for face modelling was the PhotoFit software. This software produces a 3D head model of a person from the person's photograph, as shown in figures 10 and 11. It also needs an FDP file that corresponds to this photograph. FDP file can be created using the visage|annotator software. It displays input 2D picture and automatically detects and marks the feature points in the face. After that some points can be manually corrected. Positions of these points correspond to the positions of the MPEG-4 facial Feature Points (FPs) [9].

[5]http://labrosa.ee.columbia.edu/matlab/dtw/
[6]http://www.fon.hum.uva.nl/praat/

Figure 10: *Input photograph of the source speaker and generated 3D head model by PhotoFit.*

PhotoFit basically takes a generic 3D head model and deforms it in such a way that labelled feature points fit to their positions taken from the FDP file. Other points of the model are also moved so that they form a smooth surface of the face with labelled feature points.

Generation of the facial skin textures from uncalibrated input photographs as well as the creation of individual textures for facial components such as eyes or teeth is addressed in [10]. Apart from an initial feature point selection for the skin texturing, these methods work fully automatically without any user interaction. The resulting textures show a high quality and are suitable for both photo-realistic and real-time facial animation.

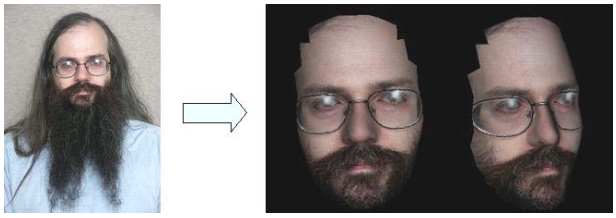


Figure 11: *Input photograph of the target speaker and generated 3D head model by PhotoFit.*

The 3D input mesh is parameterized over the 2D domain (0,1) x (0,1) (part of $\mathbb{R}^2$) in order to obtain a single texture map for the whole mesh. In [10], the face mesh is topologically equivalent to a part of a plane, since it has a boundary around the neck and does not contain any handles. The face mesh can be "flatten" to a part of a plane that is bounded by its boundary curve around the neck. PhotoFit uses described methods to create facial skin textures, but it parameterizes a mesh with a cube instead of a disk.

It is important to mention that the 2D input picture should be a frontal photograph of the person, and should contain the person's face and shoulders. Also, the face should be in the neutral position according to the MPEG-4 FBA specification. If the person on the picture has an expression it will keep that expression through the whole animation, i.e. if the person is smiling generated model will always be smiling.

4.4. Face analysis

In [1], each frame in the training set had been manually labeled with 72 landmarks by using the AAM-API software [8] (freely available for non-commercial use such as research and education). In order to obtain an accurate mapping from the 2D pixel positions to the MPEG-4 parameters, the annotation of the images in the training set should closely match the MPEG-4 FDPs.

The global head movements and the feature points tracking were done on 10 videos. To compute the head ellipse, OpenCv Lib (available on the Intel website) has been used. The calculated values for animation has been smooth, since the measurements in the tracking process are noisy and small scale differences in the parameters for the simulation process may have large effects in the resulting animation. A Kalman Filter was used for this purpose with a state model consisting of positions and velocities of all the key feature points on the face. Fig. 12 shows the results of the global head movements tracking. On the whole, we have noticed an improvement in the head center tracking. The new method is also more robust and has allowed to eradicate big errors due to the use of the optical flow (Fig. 13).

4.5. Face synthesis

We created 10 animations using XFace interface which is an MPEG-4 based open source toolkit for 3D facial animation, developed by Balci [11]. This is a straightforward process: once the FAP values have been computed, the XFace editor (or any other suitable player) can directly synthesize them (given that the parameters file format is correct). Fig. 14 shows the results of the generation of an animation with Xface. The animation is not only depending of the computation of the FAPS, and consequently of the quality of the tracking, but also of the 3D model.

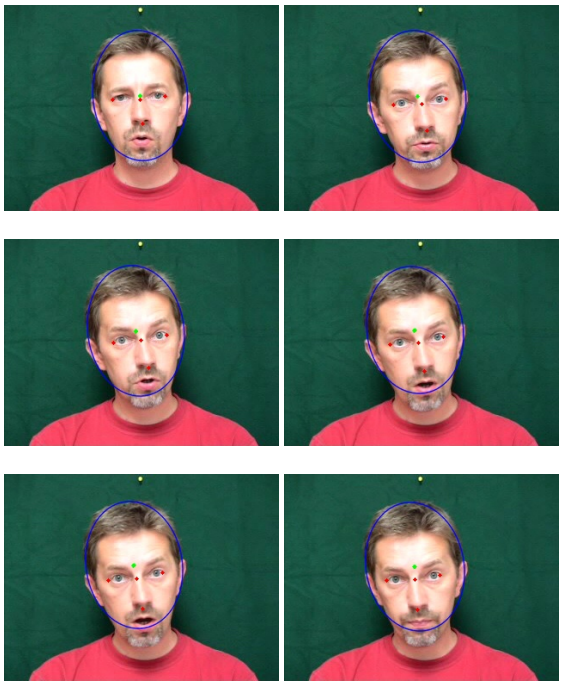


Figure 12: *Results of the global head movements tracking .The ellipse and the green point are the results of last year, red points is the new tracking method.*

Figure 13: *Errors occurring with the ellipse computed from the optical flow.*

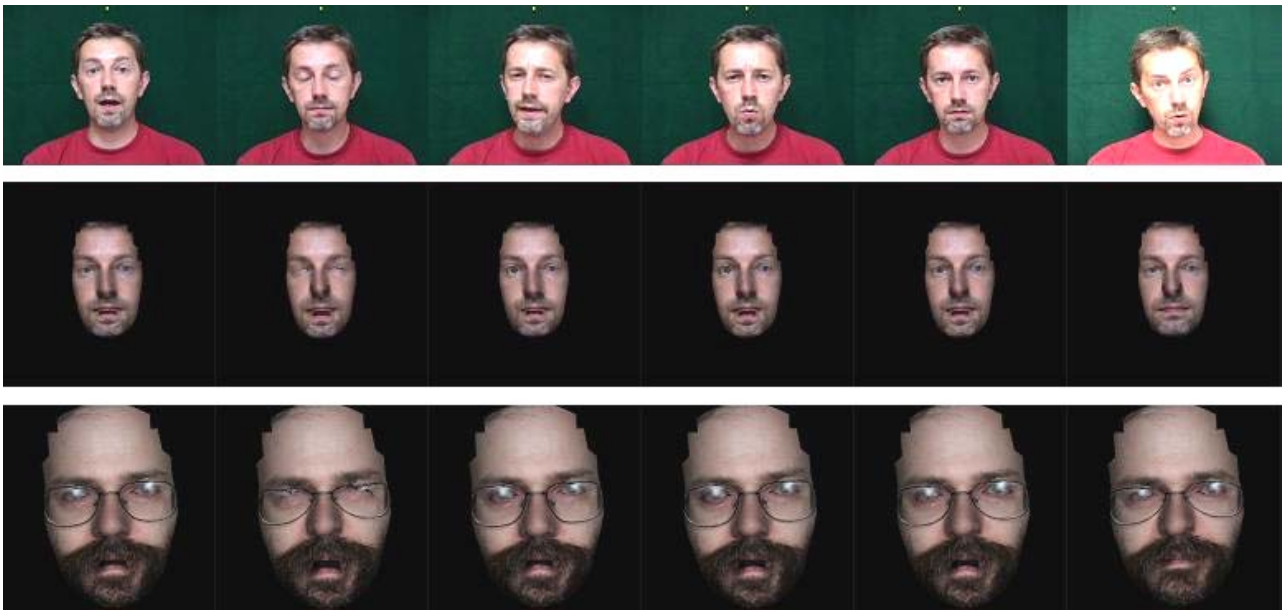


Figure 15: *(Top) Frames extracted from a video of the source speaker; (Center) Frames obtained by animating a 3D model of the source; (Bottom) Frames obtained by animating a 3D model of the target.*

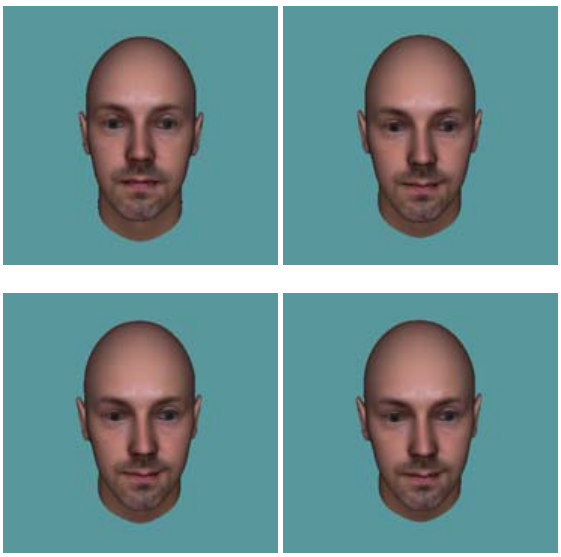


Figure 14: *Animation created with XFace.*

4.6. Face synthesis

PhotoFit uses commercial software visage|SDK [7] for animation and rendering of the generated model. visage|SDK is a Software Development Kit for MPEG-4 character animation. It includes the following main capabilities:

- Animating virtual characters using MPEG-4 Face and Body Animation Parameters (FAPs and BAPs).
- Real-time or off-line character animation driven by SAPI-5 speech synthesis, with on-the-fly rendering of the virtual character and MPEG-4 FBA bitstream output.
- Real-time or off-line lip sync from audio file or microphone, with on-the-fly rendering and MPEG- 4 FBA output.
- Interfaces for plugging-in own interactive or offline animation sources and controls.
- Coding, decoding, merging and other operations on MPEG-4 FBA (Face and Body Animation) bitstreams.

The analysis tool built during eNTERFACE'06 produces ASCII FAP files, and visage|SDK reads binary FBA files. To connect these two tools we had to write code that reads FAP files and calls visage|SDK functions for applying read values of FAPs to the generated model. Animated models of source and target speaker models are shown in figure 15.

5. CONCLUSIONS

In this paper we describe a multimodal speaker conversion system, based on the simultaneous use of facial animation detection, 3D talking face synthesis, and voice conversion. We first try to take advantage of a recently developed source-filter estimation algorithm, namely the ARX-LF model, to perform parallel conversion of voice source parameters and of vocal tract parameters. Copy synthesis using ARX-LF gives acceptable results (although the resulting quality is very sensitive to GCI detection stability), but transplanting target parameters into a source utterance leads to very irregular speech quality.

We then test L. Arslan's STASC algorithm, in a simplified implementation. The results are much more stable, although the ID conversion is still incomplete.

Face conversion is based, as initiated in [1], on the MPEG4 FPSs, FAPS, and FAPUs. After modeling the source and the target speaker faces with the PhotoFit software, we drive the speaker 3D face model using the FAPs of the source. We have also improved the face tracking algorithm, by computing the global head position from the positions of the eyes and nose rather than by simplifying the face shape into an ellipsis.

The results we have obtained so far are more complete then those obtained in [1], although there is still much room for improvement.

The face tracking algorithm still provides only an approximation of the source speaker movements. The face rendering

systems we have tested do not prevent the synthetic face from performing impossible facial movements. Last but not least, the voice conversion algorithms still provides better ID conversion if we keep the target LP residual untouched, which seems to shows that the LP residual still contains some of the speaker identity (although the opinion of all team members did not converge on this last conclusion).

6. ACKNOWLEDGEMENTS

We are grateful to Prof. F. Marques, who collaborated to the initial steps in the project design, but could finally not attend eNTERFACE'07.

We thank A. Moinet and N. D'Alessandro for providing access to their ARX-LF Matlab code.

We also want to thank Alan Black, who provided IDphotos and accepted that we make a demo avatar of him.

7. REFERENCES

[1] T. Dutoit, A. Holzapfel, M. Jottrand, F. Marquès, A. Moinet, F. Ofli, and Y. Stylianou, "Multimodal Speaker Conversion - his master's voice... and face -", in *Proc. eNTERFACE'06 workshop*, (Dubrovnik), pp. 34–45, 2006. 51, 52, 54, 55, 56, 57, 58

[2] O. Turk and L. Arslan, "Robust Processing Techniques for Voice Conversion", *Computer Speech and Language*, vol. 20, pp. 441–467, 2006. 51, 52, 53, 54, 55, 56

[3] Y. Stylianou, O. Cappé, and E. Moulines, "Continuous probabilistic transform for voice conversion", *IEEE Trans. on Speech and Audio Processing*, vol. 6, no. 2, pp. 131–142, 1998. 52

[4] D. Vincent, O. Rosec, and T. Chonavel, "A New Method for Speech Synthesis and Transformation Based On an ARX-LF Source-Filter Decomposition and HNM Modeling", in *Proc. ICASSP*, pp. 525–528, 2007. 52, 53

[5] I. S. Pandžić and R. Forchheimer, *MPEG-4 Facial Animation - The standard, implementations and applications*. John Wiley & Sons, 2002. 54

[6] M. E. Sargin, F. Ofli, Y. Yasinnik, O. Aran, A. Karpov, S. Wilson, E. Erzin, Y. Yemez, and A. M. Tekalp, "Combined gesture speech analysis and synthesis", in *Proc. of the eNTERFACE'05 Workshop*, (Mons, Belgium), 2005. 54

[7] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models", in *Proc. European Conf. on Computer Vision*, vol. 2, pp. 484–498, 1990. 54, 58

[8] H. Tao, R. Lopez, and T. Huang, "Tracking Facial Features using probabilistic networks", in *Proceedings of the 3rd International Conference on Face and Gesture Recognition*, pp. 166–171, 1998. 54, 57

[9] "Visage|SDK". Visage Technologies. http://www.visagetechnologies.com/products_sdk.html. 56

[10] M. Tarini, H. Yamauchi, J. Haber, and H.-P. Seidel, "Texturing Faces", in *Proc. of Graphics Interface*, (Calgary, Canada), pp. 89–98, 2002. 57

[11] K. Balcı, "Xface: Mpeg-4 based open source toolkit for 3D facial animation", in *Proc. Advance Visual Interfaces*, pp. 399–402, 2004. 57

8. BIOGRAPHIES

Zeynep Inanoglu graduated with an Electrical Engineering degree from Harvard University in 1999. Between 1999 and 2002 she worked as a software engineer and product manager at Angel.com, focusing on agile development and dialog systems. Subsequently she received her masters degree in Speech and Language Processing at the University of Cambridge and is currently in her final year of her PhD in Cambridge, focusing on emotional speech synthesis and prosody modeling.
Email: zeynepgates@scholar.org

Matthieu Jottrand holds an Electrical Engineering degree from the Faculté Polytechnique de Mons since June 2005. He did his master's thesis in the Image Coding Group of Linköping Institute of Technology. Matthieu is a researcher fromTCTS lab (FPMs) since September 2005. He is currently working in the field of ASR for the IRMA project (development of a multimodal interface to search into indexed audiovisual documents) and just started a PhD thesis under the supervision of Thierry Dutoit.
Email: matthieu.jottrand@fpms.ac.be

Maria Markaki holds a degree and a MSc in Physics from Athens University, and a MSc in Computer Science from University of Crete. From 1995 until 2001 she was a researcher in the Institute of Applied and Computational Mathematics (IACM - FORTH). She is currently working in the field of audio indexing for the SIMILAR project and pursues a PhD thesis under the supervision of Yannis Stylianou.
Email: mmarkaki@csd.uoc.gr

Kristina Stanković just finished her undergraduate studies at Faculty of Electrical Engineering and Computing, the University of Zagreb, Croatia, and she will also start her PhD studies there. She is a member of Human-Oriented Technologies Laboratory (HOTlab). Her main interest is in the field of face and body animation, and its application with Internet and mobile technologies.
Email: kristina.stankovic@fer.hr

Aurélie Zara received her Master Degree in computer sciences from Université Paris XI, Orsay, France, since 2006. She did her master thesis on the modelisation of multimodal emotional interaction between users and virtual agents at LIMSI-CNRS. She is currently doing a Phd thesis at France Télécom Orange Labs. Her research is on the impact of multimodal expressions of emotions of a virtual agent interacting with an human.
Email: aurelie.zara@orange-ftgroup.com

Levent Arslan has graduated from Boğaziçi University, Turkey in 1991. He received his M.S. and Ph.D. degrees from Duke University, USA in 1993 and 1996 respectively. He has worked at Texas Instruments and Entropic Research until 1998. Since then he has been teaching at Boğaziçi University. His main research interests are Voice conversion, speech recognition, speech synthesis, and Speech enhancement. He has published 13 journal papers and 70 conference papers in speech processing. He holds 9 international patents.
Email: arslanle@boun.edu.tr

Thierry Dutoit graduated as an electrical engineer and Ph.D. in 1988 and 1993 from the Faculté Polytechnique de Mons, Belgium, where he is now a professor. He spent 16 months as a consultant for AT&T Labs Research in Murray Hill and Florham Park, NJ, from July, 1996 to September, 1998. He is the author of two books on speech processing and text-to-speech synthesis, of a forthcoming book on signal processing, and the coordinator of the MBROLA project for free multilingual speech synthesis. T. Dutoit was an Associate Editor of the IEEE Transactions on Speech and Audio Processing (2004-2006) and is a member of the INTERSPEECH'07 organization committee. He was the initiator of eNTERFACE workshops and the organizer of eNTERFACE'05.
Email: thierry.dutoit@fpms.ac.be

Igor S. Panžić is an Associate Professor at the Department of Telecommunications, Faculty of Electrical Engineering and Computing, University of Zagreb, Croatia. His main research interests are in the field of computer graphics and virtual environments, with particular focus on facial animation, embodied conversational agents, and their applications in networked and mobile environments. Igor also worked on networked collaborative virtual environments, computer generated film production and parallel computing. He published four books and around 60 papers on these topics. Formerly he worked MIRALab, University of Geneva, where he finished his PhD in 1998. He was a visiting scientist at AT&T Labs, at the University of Linkping and at Kyoto University. Igor was one of the key contributors to the Facial Animation specification in the MPEG-4 International Standard for which he received an ISO Certificate of Appreciation in 2000.
Email: igor.pandzic@fer.hr

Murat Saraçlar received his B.S. degree from Bilkent University, Ankara, Turkey in 1994. He earned both his M.S.E. and Ph.D. degrees from Johns Hopkins University, Baltimore, MD, USA in 1997 and 2001 respectively. He worked on automatic speech recognition for multimedia analysis systems from 2000 to 2005 at the AT&T Labs-Research. In 2005, he joined the Department of Electrical and Electronic Engineering at Boğaziçi University as an assistant professor. His main research interests include all aspects of speech recognition, its applications, as well as related fields such as speech and language processing, human-computer interaction and machine learning. He authored and co-authored more than two dozen papers in refereed journals and conference proceedings. He has filed four patents, both internationally and in the US. He has served as a reviewer and program committee member for various speech and language processing conferences and all the major speech processing journals. Dr. Saralar is currently a member of ISCA and IEEE, serving as an elected member of the IEEE Signal Processing Society Speech and Language Technical Committee (2007-2009).
Email: murat.saraclar@boun.edu.tr

Yannis Stylianou is Associate Professor at University of Crete, Department of Computer Science. He received the Diploma of Electrical Engineering from NTUA, Athens, in 1991 and the M.Sc. and Ph.D. degrees in Signal Processing from ENST, Paris, France in 1992 and 1996, respectively. From 1996 until 2001 he was with AT&T Labs Research (Murray Hill and Florham Park, NJ, USA) as a Senior Technical Staff Member. In 2001 he joined Bell-Labs Lucent Technologies, in Murray Hill, NJ, USA. Since 2002 he is with the Computer Science Department at the University of Crete. He was Associate Editor for the IEEE Signal Processing Letters from 1999 until 2002. He is Associate Editor of the EURASIP Journal on Speech, Audio and Music Processing. He served on the Management Committee for the COST Action 277: "Nonlinear Speech Processing" and he is one of the two proponents for a new COST Action on Voice Quality Assessment.
Email: styliano@ics.forth.gr

AUDIO-DRIVEN HUMAN BODY MOTION ANALYSIS AND SYNTHESIS

Ferda Ofli [1], *Cristian Canton-Ferrer* [2], *Yasemin Demir* [1], *Koray Balcı* [3], *Joëlle Tilmanne* [4], *Elif Bozkurt* [5], *İdil Kızoğlu* [3], *Yücel Yemez* [1], *Engin Erzin* [1], *A. Murat Tekalp* [1], *Lale Akarun* [3], *A. Tanju Erdem* [5]

[1] Multimedia, Vision and Graphics Laboratory, Koç University, İstanbul, Turkey
[2] Image and Video Processing Group, Technical University of Catalonia, Barcelona, Spain
[3] Multimedia Group, Boğaziçi University, İstanbul, Turkey
[4] TCTS Lab, Faculté Polytechnique de Mons, Belgium
[5] Momentum Digital Media Technologies, İstanbul, Turkey

ABSTRACT

This project is on multicamera audio-driven human body motion analysis towards automatic and realistic audio-driven avatar synthesis. We address this problem in the context of a dance performance, where the gestures or the movements of a human actor are mainly driven by a musical piece. We analyze the relations between the audio (music) and the body movements on a training video sequence acquired during the performance of a dancer. The joint analysis provides us with a correlation model that is used to animate a dancing avatar when driven with any musical piece of the same genre.

KEYWORDS

Body motion analysis – Dance figures – Audio-driven synthesis

1. INTRODUCTION

There exists little research work reported on the problem of audio-driven human body motion analysis and synthesis. The most relevant literature is on speech-driven lip animation [1]. Since lip movement is physiologically tightly coupled with acoustic speech, it is relatively an easy task to find a mapping between the phonemes of speech and the visemes of lip movement. Many schemes exist to find such audio-to-visual mappings among which the HMM (Hidden Markov Model)-based techniques are the most common as they yield smooth animations exploiting temporal dynamics of speech. Some of these works also incorporates synthesis of facial expressions along with the lip movements to make animated faces look more natural [2, 3, 4].

Humans body motion can have many purposes: To go from one place to another, humans walk, or run. Walking is perhaps the most thoroughly studied form of body motions. Upper body motions, such as hand gestures can also have many aims: communicative, deictic or conversational. Sign language relies on hand gestures as well as upper body motions and facial expressions to convey a whole language [5]. On the other hand, some body motions express emotions. Dancing is a special type of body motion that has some predefined structure; as well as emotional aspects. Analysis of gestures in dance with the purpose of uncovering the conveyed emotions has been undertaken in recent researches [6, 7].

There are several challenges involved in audio-driven human body motion analysis and synthesis: First, there does not exist a well-established set of elementary audio and motion patterns, unlike phonemes and visemes in speech articulation. Second, body motion patterns are person dependent and open to interpretation, and may exhibit variations in time even for the same person. Third, audio and body motion are not physiologically coupled and the synchronicity in between may exhibit variations. Moreover, motion patterns may span time intervals of different length with respect to its audio counterparts. The recent works [8, 9] address the challenges similar to those mentioned above in the context of facial expression analysis and prosody-driven facial expression synthesis, using a multi-stream parallel HMM structure to find the jointly recurring gesture-prosody patterns and the corresponding audio-to-visual mapping.

In this work, our aim is to learn the predefined structure of a given dance. We analyze the relations between the music and the body movements on a training video sequence acquired during the performance of a dancer. We track the movements of the dancer using marker-based and markerless methods. We train an HMM with the basic pieces from the given genre. Then, given a music piece, we classify its genre and use the corresponding HMM for synthesis. We use a body animation software developed in this project to animate an avatar with the motion parameters produced by the HMM. We analyze the audio to extract parameters about the speed of the music and adapt the animation accordingly.

2. SYSTEM OVERVIEW

Our audio-driven body animation system is composed of multimodal analysis and synthesis blocks. In the analysis, we observe the recurring body motion patterns in the video and segment the video into partitions of meaningful dance figures. Then, we try to learn the recurring body motion patterns by training HMMs over these segments. In the mean time, we try to learn the audio by looking at its beat frequency within each time window we have for the video. We expect to have similar beat frequency values in the time windows that correspond to the same dance figure. However, beat frequency may vary throughout a single musical piece, or among different musical pieces. This variation is smaller in the former case whereas it is expected to be larger in the latter case. Therefore, the variation in beat frequency is used to determine the duration of dance figures as well as to specify the genre of the given musical piece.

In the synthesis, given a musical piece of one of the types we learn in the analysis part, we first classify the audio into the correct audio class. Then extracting the beat information of the given audio signal, we decide on which dance figure is going to be generated and how much time that the expected dance figure is going to occupy. After we obtain the outline of the dance

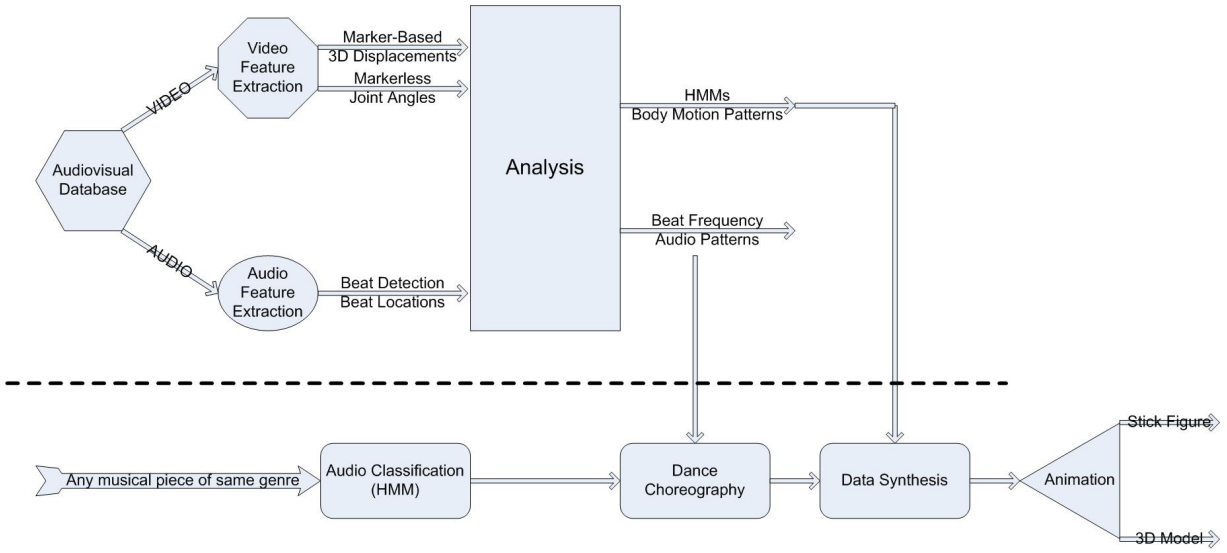


Figure 1: *Block diagram of the analysis-synthesis system.*

figure synthesis task, we start generating the dance figures using the corresponding HMMs for this specific set of dance figures. Fig. 1 shows the overall analysis-synthesis system.

3. VIDEO FEATURES EXTRACTION

Body motion capture and feature extraction involves automated capture of body motion from multiview video recorded by a multicamera system. We will employ two different methods for body motion tracking in parallel. One method will be based on 3D tracking of the markers attached to the person's body in the scene. The other method will be based on 3D reconstruction of background segmented images of the scene. We will make use of the multistereo correspondence information from multiple cameras to obtain 3D motion information in both methods. This task will provide us with a set of features of joint angles over time that expresses the alignment of the body parts of the dancer in the scene.

3.1. Marker-based Motion Capture

The motion capture process involves tracking a number of markers attached to the dancer's body as observed from multiple cameras and extraction of the corresponding motion features. Fig. 2 demonstrates our setting for this scenario. Markers in each video frame are tracked making use of their chrominance information. The 3D position of each marker at each frame is then determined via triangulation based on the observed projections of the markers on each camera's image plane.

3.1.1. Initialization

Markers on the subject are manually labeled in the first frame for all camera views. We change the color space from RGB to YCrCb which gives flexibility over intensity variations in the frames of a video as well as among the videos captured by cameras at different views. We assume that the distributions of Cr and Cb channel intensity values belonging to marker regions are Gaussian. Thus, we calculate the mean, μ, and the covariance, $\mathbf{\Sigma}$, over each marker region (a pixel neighborhood around the labeled point), where $\mu = [\mu_{Cr}, \mu_{Cb}]^T$ and $\mathbf{\Sigma} = (\mathbf{c}-\mu)(\mathbf{c}-\mu)^T$, $\mathbf{c}$ being $[c_{Cr}, c_{Cb}]^T$.

Let M be the number of markers on the subject and $\mathbf{W}$ be the set of search windows, where $\mathbf{W} = [\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_M]$ such that each window $\mathbf{w}_m$ is centered around the location, $[x_m, y_m]^T$, of the corresponding marker. The set $\mathbf{W}$ is used to track markers over frames. Thus the center of each search window, $\mathbf{w}_m$, is initialized as the point manually labeled in the first frame and specifies the current position of the marker.

3.1.2. Tracking

To track the marker positions through the incoming frames, we use the Mahalanobis distance from $\mathbf{c}$ to $(\mu, \mathbf{\Sigma})$ where $\mathbf{c}$ is a vector containing Cr and Cb channel intensity values $[c_{Cr}, c_{Cb}]^T$ of a point $\mathbf{x}_n \in \mathbf{w}_m$. Let $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N]$ be the set of candidate pixels for which the chrominance distance is less than a certain threshold. If the number of these candidate pixels, N, is larger than a predefined value, then we label that marker as visible in the current camera view and update its position as the mean of the points in $\mathbf{X}$ for the current camera view. The same process is repeated for all marker points in all camera views. Hence, we have the visibility information of each marker from each camera, and for those that are visible, we have the list of 2D positions of the markers on that specific camera image plane.

Once we scan the current scene from all cameras and obtain the visibility information for all markers, we start calculating the 3D positions of the markers by back-projecting the set of 2D points which are visible in respective cameras, using triangulation method. Theoretically, it is sufficient to see a marker at least from two cameras to be able to compute its position in 3D world. If a marker is not visible at least from two cameras, then its current 3D position is estimated from the information in the previous frame.

The 3D positions of markers are tracked over frames by Kalman filtering where the filter states correspond to 3D position and velocity of each marker. The list of 3D points obtained by back-projection of visible 2D points in respective camera image planes constitute the observations for this filter. This filtering operation has two purposes:

- to smooth out the measurements for marker locations in the current frame,
- to estimate the location of each marker in the next frame and to update the positioning of each search window,

Figure 2: *An example scene.*

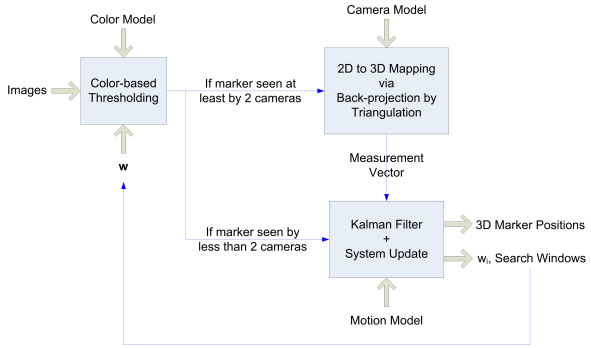


Figure 3: *Block diagram of the proposed tracking system.*

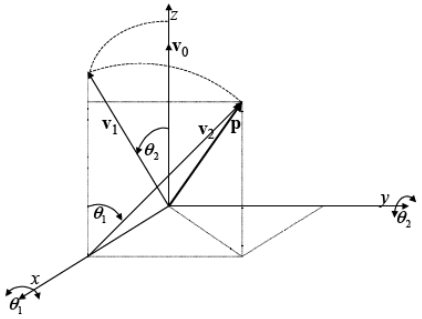


Figure 4: *The vector that connects neck joint and head centroid.*

$\mathbf{w}_m$, on the corresponding image plane accordingly.

Fig. 3 summarizes the overall system. Having updated the list of 3D marker positions for the current frame and estimated the location of the search windows for the next frame, we move on to the next frame and search the marker positions within the new search windows. This algorithm is repeated for the whole video.

3D positions of each marker is used to extract the euler angles for each joint by using inverse kinematic chain structure. The derivation of the euler angles that belong to neck is given exactly here and the other derivations results are given in Fig. 5 which are calculated in an analogous way with the given neck angles. In Fig. 4, the derivation of the angles in neck are delineated. The given vector **p** demonstrates the vector from neck joint to head centroid in the torso-centered coordinate system, these are defined as:

$$\vec{p} = R_0^{-1} \times (\vec{p}_{neck} - \vec{p}_{headCen}). \quad (1)$$

$\vec{R}_0$ is the rotation matrix of torso $\vec{v}_0$ is the vector along z axis, $\vec{v}_1$ is the projection on x-z plane and $\vec{v}_2$ is the projection on y-z plane.

We can write the vectors as:

$$\vec{v}_0 = \begin{bmatrix} 0 & 0 & \sqrt{x^2+y^2+z^2} \end{bmatrix}^T \quad (2)$$

,

$$\vec{v}_1 = \begin{bmatrix} x & 0 & \sqrt{y^2+z^2} \end{bmatrix}^T \quad (3)$$

,

$$\vec{v}_2 = \begin{bmatrix} 0 & y & z \end{bmatrix}^T \quad (4)$$

and with these given vectors we can calculate the angles for neck as given below:

$$\Theta 1 = -\arccos((e_3 v_2) \div (\|v_2\|))sign(y) \quad (5)$$

$$\Theta 2 = \arccos((e_3 v_1) \div (\|v_1\|))sign(x) \quad (6)$$

3.2. Markerless Motion Capture

Retrieving the body configuration in terms of its defining parameters, i.e. joint angles, from unlabeled video data presents a number of challenges. The main advantage of this technique is that no intrusive markers are required. However, the precision of the output may no be as accurate as the one obtained from marker based techniques since the input data is corrupted by noise.

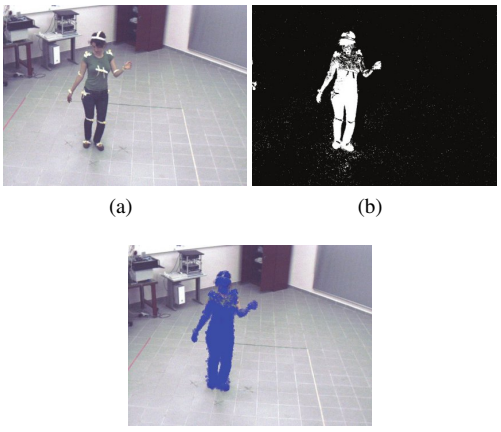
3.2.1. 3D Data generation

For a given frame in the video sequence, a set of N images are obtained from the N cameras. Each camera is modeled using a pinhole camera model based on perspective projection with camera calibration information available. Foreground regions from input images are obtained using a segmentation algorithm based on Stauffer-Grimson's background learning and substraction technique [10] as shown in Fig. 6b.

Angle	Formula	Vector $\mathbf{v}$	Vector $\mathbf{p} = \begin{bmatrix} x & y & z \end{bmatrix}^T$
Neck: θ_1	$\theta_1 = -\arccos(\mathbf{e}_3\mathbf{v})\text{sgn}(y)$	$\mathbf{v} = \frac{1}{\sqrt{y^2+z^2}} \begin{bmatrix} 0 & y & z \end{bmatrix}^T$	$\mathbf{p} = \mathbf{R}_0^T(\mathbf{p}_1 - \mathbf{p}_{10})$
Neck: θ_2	$\theta_2 = \arccos(\mathbf{e}_3\mathbf{v})\text{sgn}(x)$	$\mathbf{v} = \frac{1}{\sqrt{x^2+y^2+z^2}} \begin{bmatrix} x & 0 & \sqrt{y^2+z^2} \end{bmatrix}^T$	$\mathbf{p} = \mathbf{R}_0^T(\mathbf{p}_1 - \mathbf{p}_{10})$
Left shoulder: θ_3	$\theta_3 = \arccos(\mathbf{e}_2\mathbf{v})\text{sgn}(z)$	$\mathbf{v} = \frac{1}{\sqrt{y^2+z^2}} \begin{bmatrix} 0 & y & z \end{bmatrix}^T$	$\mathbf{p} = \mathbf{R}_0^T(\mathbf{p}_2 - \mathbf{p}_{11})$
Right shoulder: θ_4	$\theta_4 = \arccos(\mathbf{e}_2\mathbf{v})\text{sgn}(z)$	$\mathbf{v} = \frac{1}{\sqrt{y^2+z^2}} \begin{bmatrix} 0 & y & z \end{bmatrix}^T$	$\mathbf{p} = \mathbf{R}_0^T(\mathbf{p}_3 - \mathbf{p}_{12})$
Left shoulder: θ_5	$\theta_5 = -\arccos(\mathbf{e}_2\mathbf{v})\text{sgn}(x)$	$\mathbf{v} = \frac{1}{\sqrt{x^2+y^2+z^2}} \begin{bmatrix} x & \sqrt{y^2+z^2} & 0 \end{bmatrix}^T$	$\mathbf{p} = \mathbf{R}_0^T(\mathbf{p}_2 - \mathbf{p}_{11})$
Right shoulder: θ_6	$\theta_6 = \arccos(\mathbf{e}_2\mathbf{v})\text{sgn}(x)$	$\mathbf{v} = \frac{1}{\sqrt{x^2+y^2+z^2}} \begin{bmatrix} x & \sqrt{y^2+z^2} & 0 \end{bmatrix}^T$	$\mathbf{p} = \mathbf{R}_0^T(\mathbf{p}_3 - \mathbf{p}_{12})$
Left shoulder: θ_7	$\theta_7 = -\arccos(\mathbf{R}_3\mathbf{R}_5\mathbf{e}_3(\mathbf{v}_2 \times \mathbf{v}_1))\cdot$ $\text{sgn}(\mathbf{v}_1(\mathbf{v}_2 \times (\mathbf{v}_2 \times \mathbf{v}_1)))$	$\mathbf{v}_1 = \mathbf{R}_0^T(\mathbf{p}_2 - \mathbf{p}_{15})$ $\mathbf{v}_2 = \mathbf{R}_0^T(\mathbf{p}_6 - \mathbf{p}_{15})$	
Right shoulder: θ_8	$\theta_8 = -\arccos(\mathbf{R}_4\mathbf{R}_6\mathbf{e}_3(\mathbf{v}_1 \times \mathbf{v}_2))\cdot$ $\text{sgn}((\mathbf{v}_2 \times (\mathbf{v}_1 \times \mathbf{v}_2))\mathbf{v}_1)$	$\mathbf{v}_1 = \mathbf{R}_0^T(\mathbf{p}_3 - \mathbf{p}_{16})$ $\mathbf{v}_2 = \mathbf{R}_0^T(\mathbf{p}_7 - \mathbf{p}_{16})$	
Left elbow: θ_9	$\theta_9 = \pi - \arccos(\mathbf{v}_1\mathbf{v}_2)$	$\mathbf{v}_1 = \mathbf{R}_0^T(\mathbf{p}_2 - \mathbf{p}_{15})$ $\mathbf{v}_2 = \mathbf{R}_0^T(\mathbf{p}_6 - \mathbf{p}_{15})$	
Right elbow: θ_{10}	$\theta_{10} = -\pi + \arccos(\mathbf{v}_1\mathbf{v}_2)$	$\mathbf{v}_1 = \mathbf{R}_0^T(\mathbf{p}_3 - \mathbf{p}_{16})$ $\mathbf{v}_2 = \mathbf{R}_0^T(\mathbf{p}_7 - \mathbf{p}_{16})$	
Left hip: θ_{11}	$\theta_{11} = \arccos(-\mathbf{e}_3\mathbf{v})\text{sgn}(y)$	$\mathbf{v} = \frac{1}{\sqrt{y^2+z^2}} \begin{bmatrix} 0 & y & z \end{bmatrix}^T$	$\mathbf{p} = \mathbf{R}_0^T(\mathbf{p}_4 - \mathbf{p}_{13})$
Right hip: θ_{12}	$\theta_{12} = \arccos(-\mathbf{e}_3\mathbf{v})\text{sgn}(y)$	$\mathbf{v} = \frac{1}{\sqrt{y^2+z^2}} \begin{bmatrix} 0 & y & z \end{bmatrix}^T$	$\mathbf{p} = \mathbf{R}_0^T(\mathbf{p}_5 - \mathbf{p}_{14})$
Left hip: θ_{13}	$\theta_{13} = -\arccos(-\mathbf{e}_3\mathbf{v})\text{sgn}(x)$	$\mathbf{v} = \frac{1}{\sqrt{x^2+y^2+z^2}} \begin{bmatrix} x & 0 & \sqrt{y^2+z^2} \end{bmatrix}^T$	$\mathbf{p} = \mathbf{R}_0^T(\mathbf{p}_4 - \mathbf{p}_{13})$
Right hip: θ_{14}	$\theta_{14} = -\arccos(-\mathbf{e}_3\mathbf{v})\text{sgn}(x)$	$\mathbf{v} = \frac{1}{\sqrt{x^2+y^2+z^2}} \begin{bmatrix} x & 0 & \sqrt{y^2+z^2} \end{bmatrix}^T$	$\mathbf{p} = \mathbf{R}_0^T(\mathbf{p}_5 - \mathbf{p}_{14})$
Left knee: θ_{15}	$\theta_{15} = \pi - \arccos(\mathbf{v}_1\mathbf{v}_2)$	$\mathbf{v}_1 = \mathbf{R}_0^T(\mathbf{p}_8 - \mathbf{p}_{19})$ $\mathbf{v}_2 = \mathbf{R}_0^T(\mathbf{p}_4 - \mathbf{p}_{19})$	
Right knee: θ_{16}	$\theta_{16} = \pi - \arccos(\mathbf{v}_1\mathbf{v}_2)$	$\mathbf{v}_1 = \mathbf{R}_0^T(\mathbf{p}_9 - \mathbf{p}_{20})$ $\mathbf{v}_2 = \mathbf{R}_0^T(\mathbf{p}_5 - \mathbf{p}_{20})$	

Figure 5: *The formulas for calculation of joint euler angles.*

Redundancy among cameras is exploited by means of a Shape-from-Silhouette (SfS) technique [11]. This process generates a discrete occupancy representation of the 3D space (voxels). Each voxel is labelled as foreground or background by checking the spatial consistency of its projection on of the N segmented silhouettes. The data obtained with this 3D reconstruction is corrupted by spurious voxels introduced due to wrong segmentation, camera calibration inaccuracies, etc. A connectivity filter is introduced in order to remove these voxels by checking its connectivity consistency with its spatial neighbors. An example of the output of the whole 3D processing module is depicted in Fig. 6c. For the research presented whithin this paper, it is assumed that only one person is present in the scene. Let us refer to the obtained voxel data as $\mathcal{V}$.



(a) (b)

(c)

Figure 6: *3D data generation. In (a), one of the original N images. In (b), the foreground/background binary segmentation and, in (c), the projection of the voxels defining the input 3D data.*

3.2.2. *Human Body Model*

In order to analyze the incoming data $\mathcal{V}$, an articulated body model will be used. This body model allows exploiting the underlying antropomorphic structure of the data; let us refer to this model as $\mathcal{H}$. Model based analysis of humans as been already addressed in the literature in [12, 13]. The employed model is formed by a set of joints and links representing the limbs, head and torso of the human body and a given number of degrees of freedom (DoF) are assigned to each articulation (joint). Particularly, our model has 22 DoF to properly capture the possible movements of the body: position of the center of the torso (3 DoF), rotation of the torso (3 DoF), rotation of the neck (2 DoF), rotation of the shoulders (3+3 DoF), rotation of the elbows (1+1 DoF), rotation of the hips (2+2 DoF) and rotation of the ankles (1+1 DoF). An example of this body model is depicted in Fig. 7.

The equations driving the behavior of the joints rely on kinematic chains formulated by means of exponential maps [14, 15].

3.2.3. *Human Body Tracking*

Particle Filtering (PF) [16] algorithms are sequential Monte Carlo methods based on point mass (or "particle") representations of

probability densities. These techniques are employed to tackle estimation and tracking problems where the variables involved do not hold Gaussianity uncertainty models and linear dynamics. In the current scenario, the hidden state to be estimated, that is the set of 22 DoF of the human body model, falls in the aforementioned conditions hence particle filtering techniques may efficiently retrieve this state vector.

Two major issues must be addressed when employing PF: likelihood evaluation and propagation model. The first one, establishes the observation model, that is how a given configuration of a the body matches the incoming data. The propagation model is adopted to add a drift to the angles of the particles in order to progressively sample the state space in the following iterations [16]. For complex PF problems involving a high dimensional state space such as in this articulated human body tracking task [13], an underlying motion pattern is employed in order to efficiently sample the state space thus reducing the number of particles required. This motion pattern is represented by the kinematical constrains and physical limits of the joints of the human body.

Likelihood evaluation being a crucial step is described as follows. For any particle, a volumetric representation of the human body using hyper-ellipsoids is generated, $\tilde{\mathcal{H}}$ (see Fig. 7 for an example). Within this representation, every limb of the generated model is denoted as $\mathcal{L}_k$. Likelihood

$$p(\tilde{\mathcal{H}}^j|\mathcal{V}) = \prod_{k=0}^{K} \mathcal{L}_k \cap \mathcal{V}, \tag{7}$$

where $\cap$ operator denotes the volumetric intersection between the limb of the model $\mathcal{L}_k$ and the incoming data $\mathcal{V}$. Individual likelihoods of each limb are assumed to be independent in order to generate the global human body likelihood function. Current research involves employing more informative measures including color information.

Figure 7: *Articulated human body model with 22 DoF and hyper-ellipsoids to represent the limbs.*

4. AUDIO FEATURES EXTRACTION

An appropriate set of features will be extracted from the audio signal that is synchronized with the body motion parameters. The mel frequency cepstral coefficients (MFCC) along with additional prosodic features can be considered as audio features. Audio feature extraction will be performed using the well known HTK Tool.

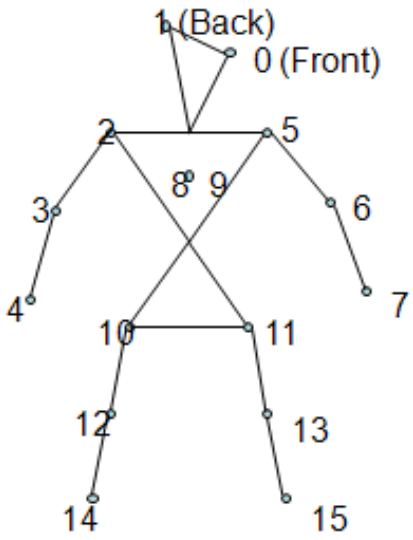


Figure 8: *Markers positions (10 to 15 for lower body, 2 to 7 for upper body).*

5. ANALYSIS

The feature sets resulting from body motion and audio will jointly be analyzed to model the correlation between audio patterns and body motion patterns. For this purpose, we plan to use a two-step HMM-based unsupervised analysis framework as proposed in [5]. At the first step, the audio and motion features will separately be analyzed by a parallel HMM structure to learn and model the elementary patterns for a particular performer. A multi-stream parallel HMM structure will then be employed to find the jointly recurring audio-motion patterns and the corresponding audio-to-visual mapping. All the simulations at this second step will be implemented by using the HTK Toolkit.

The body motion synthesis system will take an audio signal as an input and produce a sequence of body motion features, which are correlated with the input audio. The synthesis will be based on the HMM-based audio-body motion correlation model derived from the multimodal analysis. The synthesized body motion will then be animated on an avatar.

5.1. Video Analysis

Human body motion analysis will be tackled through HMMs. Dance motion can be addressed by analyzing patterns that are repeated sequentially by the dancer and a set of HMMs is trained separately for each dance figure. Data employed to train the HMMs are the normalized 3D positions of some landmarks defined on the body (typically the body joints) and tracked along time by means of the two vision based analysis systems. For each figure, two sub-HMMs are defined to better capture the dynamics behavior of the upper and lower part of the body. The HMM modelling the upper part of the body addresses the arms movement (described by the (x, y, z) positions of the six landmarks placed in shoulders, elbows and wrists) while the other HMMs accounts for the legs (described by the (x, y, z) position for the six landmarks placed in hips, knees and ankles) (Fig 8).

To start evaluating the performance of the system presented in this report, a simple HMM is adopted. Typically, dance figures always contain a very concrete sequence of movements hence a left-right HMM structure is employed (Fig 9). Each of the parameters is represented by a single Gaussian function

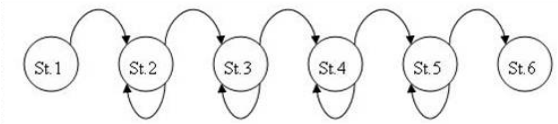


Figure 9: *Simple left-right HMM structure.*

and one full covariance matrix is computed for each state. This rather simple scheme leads to satisfactory results hence no further complexity is added to the simple. All computation as been done by means of the "Hidden Markov Model Toolkit" (HTK) package developed at the Cambridge University. This packages allowed us to efficiently model the HMM structures employed within this project.

5.2. Audio Analysis

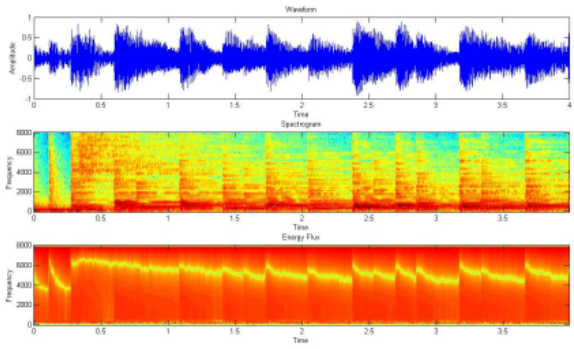


Figure 10: *From top to bottom: time waveform, spectrogram and spectral energy flux of four seconds of Salsa music audio file.*

Both for Salsa and Belly dance music audio files we measure tempo in terms of beats per minute (BPM) using the estimation algorithm suggested in [17]. Tempo estimation can be broken down into three sections: onset detection, periodicity estimation and beat location estimation. Onset detection aims to point out where musical notes begin and tempo is established by periodicity of the detected onsets. It is straight forward to detect beat locations after periodicity estimation.

First, we detect onsets based on the spectral energy flux of the input audio signal that signify the most salient features of the audio file as shown Fig. 10. Onset detection is important, since beat tends to occur at onsets. Algorithm works best for four second analysis window with 50% overlap. The below figure belongs to the first four seconds analysis window of the Salsa music audio file.

Next, we estimate the periodicity of the detected onsets using the autocorrelation method. The distance between the largest peak in the interval from 300 ms to 1 s of the autocorrelated signal and its origin gives the periodicity value. Once the periodicity is determined we can calculate tempo in terms of number of beats per minute, which lies between the values 60 and 200 BPM. For Salsa music audio file we estimate tempo as 185 BPM and for the Belly dance it is 134 BPM.

Furthermore, we estimate beat locations regarding the periodicity value in the previous step. We generate an artificial pulse train with the estimated periodicity and cross - correlate with the onset sequence where maximum value of the cross - correlation gives the starting beat location. Successive beats are expected in every beat period *T*.

Analyzing the results from labeling of the dance figures in the video frames, we conclude that each Salsa figure corresponds to 8 beats in the Salsa music audio file and each Belly dance figure corresponds to 3 beats in the Belly dance music. We also use this information during the synthesis time to determine the beginning and ending frames of a dance figure.

6. SYNTHESIS

Given a audio track, the generated classifier is used to classify the tracks as belly or salsa. The classified track beats is extracted by the method explained in section 4 and the beat periods of each track are used to generate figures that are learnt by HMMs in the motion analysis part. The appropriate HMM that would generate figures with the given audio track is used by the synthesizer. We have one HMM model for salsa which generates the basic figure with the given salsa track. Belly dance sequence is more complicated than the salsa dance sequence. It yields three independent figures with just one beat period. An HMM is trained for this scenario. We will generate a coupled HMM with individual HMM models in each state that correspond to different figures that are recognized during training. The state transitions are determined according to the dancer sequence and the transition probabilities are calculated from the co-occurrence matrices of audio beat numbers and video labels.

6.1. Audio Classification

This part is a simple music genre classification problem. We have two types of music audio files where one is Salsa and the other is Belly dance. We use supervised HMMs and the well - known Mel Frequency Cepstral Frequency coefficients (MFCCs) to discriminate the 16 kHz, 16 bit mono PCM wavefiles. The music audio signals are analyzed over 25 ms Hamming window for each 10 ms frame. Finally, 13 MFCC coefficients together with the accelaration and the delta parameters, adding up to 39 features, form the audio feature vectors. Use of MFCCs as the only audio feature set is sufficient for the classification problem, since we have only two kinds of audio files. For the extraction of parameters and classification steps, we use HTK toolkit.

Using the HMMs generated in the analysis step we first classify the input music audio files as Salsa or Belly dance as depicted in Fig. 11, below. Then, we estimate the beat signal for the detected music audio file following the steps onset detection, periodicity estimation and beat location. Next, we identify the beat segmentation times in the music audio and determine the duration (in terms of frame numbers) of figures to be performed during the animation. Precalculated beats per frame information that we got in the analysis section is used for this purpose. For example, for Salsa, each figure corresponds to a time segment of eight beats, so by multiplying the start and end time of the each segment with the number of frames per second (30 in our case), we simply get the beginning and ending frame numbers for Salsa dance figures.

6.2. Body Motion Parameters Generation

Once we have the list of durations and types of consecutive dance figures in a file, we can use that file to generate the appropriate values for the animation parameters according to the mean and standard deviation values of the corresponding HMM states. This file basically determines how much time each dance figure takes in the sequence. This helps us to allocate exactly the necessary amount of time to perform each dance figure.

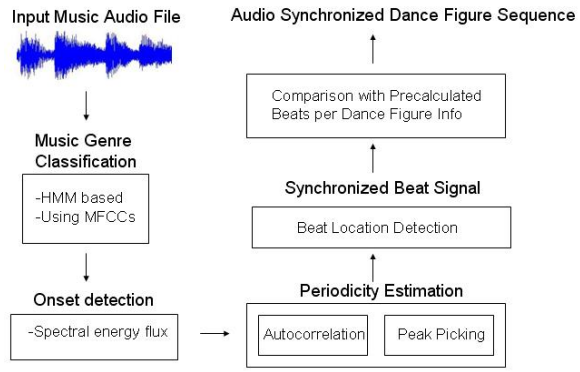


Figure 11: *Audio processing steps in the synthesis part.*

7. VISUALIZATION

For avatar model, we used a free 3D model named Douglas F. Woodward shown in Figure 12 with 9599 vertices and 16155 faces. The model comes with segmented hierarchy, which let us create a kinematic chain of segments in a conventional directed acyclic graph (DAG) structure.

We decided to implement a generic synthetic body representation and animation tool instead of relying on a single model. Our tool, namely Xbody, can open models in 3DS format and display the DAG and submesh info and enables labeling of the segments for animation as can be seen in Figure 12. For rendering, Xbody relies on OpenGL and existing Xface [18] codebase. We implemented an additional forward kinematics pipeline for rendering and animation of DAG.

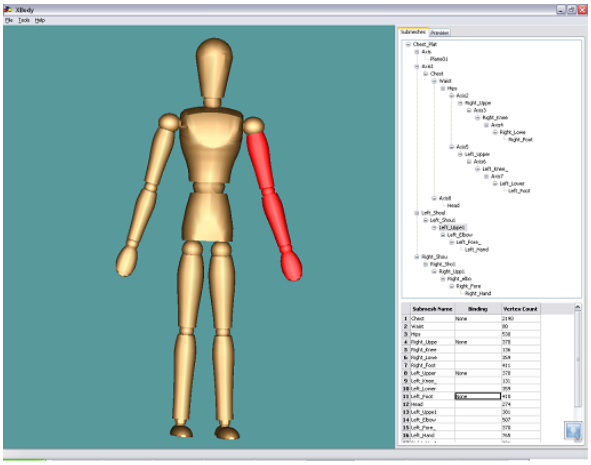


Figure 12: *Xbody DAG view and labelling pane.*

As for animation, the generated data from analysis and synthesis part can be fed to Xbody and animated with the same frame per second of video. The previewing interface of the tool enables us to inspect each frame by entering the frame number and using rotation, zooming in/out and panning the model on the screen. In Figure 13, previewing of frame 180 for markerles tracking analysis result for "Zeybek" dance is shown. The tool can also export the animation as video in avi format.

As its current state, Xbody can be used for better analyzing the results of motion tracking algorithms and HMM based motion generation. In the future, we plan to improve on Xbody and implement full support for MPEG-4 Body Animation by parsing BAP and BDP formats.

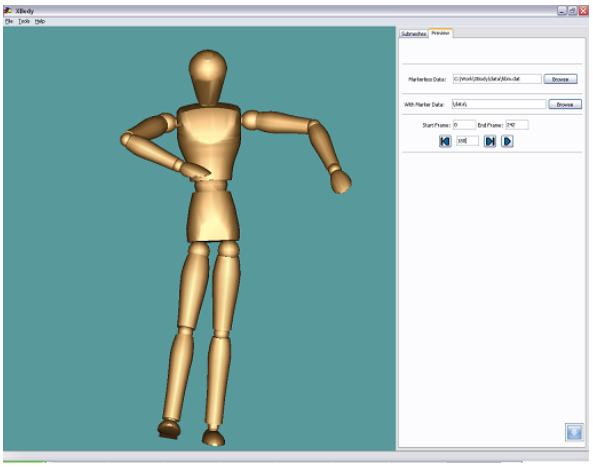


Figure 13: *Xbody preview pane.*

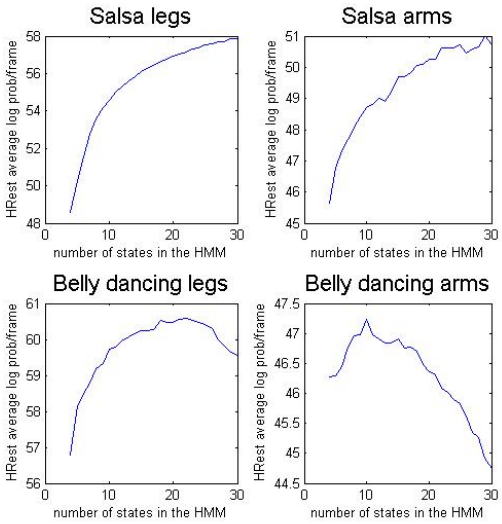


Figure 14: *Evolution of the logarithmic probability per frame for the 4 HMMs types.*

8. EXPERIMENTS AND RESULTS

As we modeled two dance figures (salsa and belly dancing) and that the whole body movement modeling was splitted into two HMMs, we had four HMMs to train.

The training of the HMM was performed using the HTK function HERest. It takes as input the data parameters file in HTK format, a prototype of the HMM containing its structure, and the transcription of the data file. In our case, we trained only one HMM at a time, and the transcription is only a succession of the same HMM name.

At the end of each training iteration, HERest gives an average logarithmic probability per frame. The evolution of this parameter enables us to follow the progression of the learning process and the accuracy of the trained model.

As we had no prior knowledge of the optimal number of states, we trained HMMs with an increasing number of states (from 4 to 30) and compared their average logarithmic probability per frame. The evolution of this parameter for the four types of HMMs we trained is shown in figure 14.

As the arms are hardly moving in the belly dance and as the salsa motion pattern is much more complicated than the belly

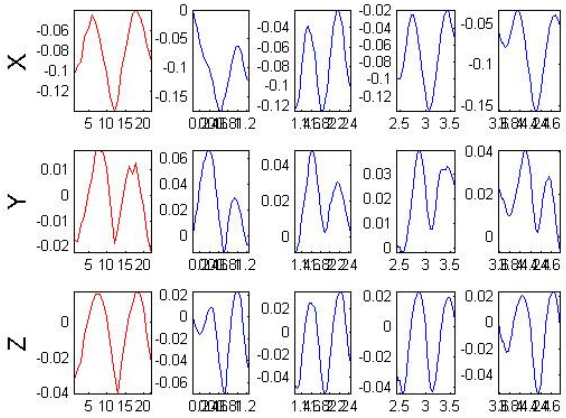


Figure 15: *Comparison of the means of the Gaussian distributions for 22 states of the belly dance HMM (in red) to the evolution of the 3 corresponding parameters during 4 dance figures (in blue) (x,y and z values of the left hip) during one salsa figure.*

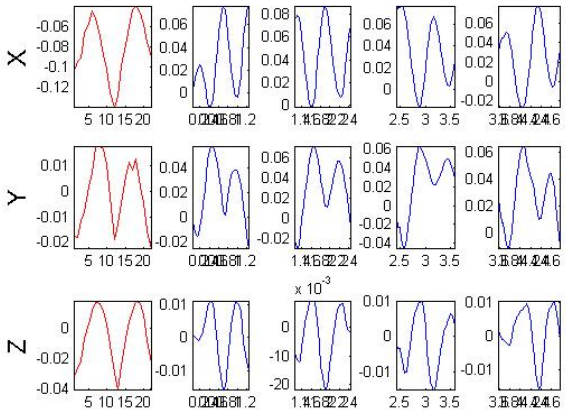


Figure 16: *Comparison of the HMM Gaussian means and the corresponding parameter (left knee) in 4 occurrence of the belly figure where the inter occurrence variability is very high*

dance one, we can see that the number of states required is linked to the complexity of the motion to model. For belly dancing, the optimal number is clear (around 20 for the legs and 10 for the arms), but in salsa there is no decrease in the logarithmic probability parameter before 30 states. For the salsa motion, we decided to keep around 20 states for both legs and arms as saturation began around that number and that the number of states (and thus the complexity of the HMM model) has to be kept reasonable.

In order to verify that the modeling of the data parameters was correct, we compared, for each parameter, the evolution of the mean of its Gaussian distribution across the states to the evolution of the same parameter for a few occurrence of the dance figure in the training dataset. The shape of the evolution can clearly be recognized (Fig. 15), even if some parameters vary highly between two occurrences of the same dance figure in the training set and are thus more difficult to model (Fig. 16).

9. CONCLUSIONS AND FUTURE WORK

In this research work, we first developed an automated human body motion capture system based solely on image processing and computer vision tools using standard digital video cameras. Second we provided a framework for joint analysis of loosely correlated modalities such as motion and audio and demonstrate how this framework can be used for audio-driven motion synthesis.

10. SOFTWARE NOTES

As a result of this project, the following resources are available:

- Dance databases
- Color-marker-based motion tracking software

11. REFERENCES

[1] T. Chen, "Audiovisual speech processing", *IEEE Signal Processing Magazine*, vol. 18, no. 1, pp. 9–21, 2001. 61

[2] C. Bregler, M. Covell, and M. Slaney, "Video Rewrite: driving visual speech with audio", in *SIGGRAPH '97: Proceedings of the 24th annual conference on Computer graphics and interactive techniques*, (New York, NY, USA), pp. 353–360, ACM Press/Addison-Wesley Publishing Co., 1997. 61

[3] M. Brand, "Voice puppetry", in *SIGGRAPH '99: Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, (New York, NY, USA), pp. 21–28, ACM Press/Addison-Wesley Publishing Co., 1999. 61

[4] Y. Li and H.-Y. Shum, "Learning dynamic audio-visual mapping with input-output Hidden Markov models", *IEEE Transactions on Multimedia*, vol. 8, no. 3, pp. 542–549, 2006. 61

[5] O. Aran, I. Ari, A. Benoit, A. H. Carrillo, F. Fanard, P. Campr, L. Akarun, A. Caplier, M. Rombaut, and B. Sankur, "SignTutor: An Interactive Sign Language Tutoring Tool", in *Proceedings of eNTERFACE 2006, The Summer Workshop on Multimodal Interfaces, Dubrovnik, Croatia*, 2006. 61

[6] A. Camurri, B. Mazzarino, M. Ricchetti, R. Timmers, and V. G, *Multimodal Analysis of Expressive Gesture in Music and Dance Performances*, vol. 2915/2004. Heidelberg: Springer Berlin, 2004. 61

[7] F. Ofli, Y. Demir, E. Erzin, Y. Yemez, and A. M. Tekalp, "Multicamera Audio-Visual Analysis of Dance Figures", *IEEE International Conference on Multimedia and Expo, 2007. ICME 2007*, 2007. 61

[8] F. Ofli, E. Erzin, Y. Yemez, and A. M. Tekalp, "Estimation and Analysis of Facial Animation Parameter Patterns", *to appear in IEEE International Conference on Image Processing, 2007. ICIP 2007*, 2007. 61

[9] M. E. Sargin, E. Erzin, Y. Yemez, A. M. Tekalp, A. T. Erdem, C. Erdem, and M. Ozkan, "Prosody-Driven Head-Gesture Animation", *IEEE International Conference on Acoustics, Speech and Signal Processing, 2007. ICASSP 2007*, vol. 2, pp. 677–680, 2007. 61

[10] C. Stauffer and W. Grimson, "Adaptive background mixture models for real-time tracking", in *IEEE Int. Conf. on Computer Vision and Pattern Recognition*, pp. 252–259, 1999. 63

[11] G. Cheung, T. Kanade, J.-Y. Bouguet, and M. Holler, "A real time system for robust 3D voxel reconstruction of human motions", in *IEEE Conf. on Computer Vision and Pattern Recognition*, vol. 2, pp. 714–720, 2000. 64

[12] F. Caillette and T. Howard, "Real-Time Markerless Human Body Tracking with Multi-View 3-D Voxel Reconstruction", in *Proceedings of British Machine Vision Conference (BMVC)*, vol. 2, pp. 597–606, September 2004. 64

[13] J. Deutscher and I. Reid, "Articulated Body Motion Capture by Stochastic Search", *International Journal of Computer Vision*, vol. 61, pp. 185–205, Feb. 2005. 64, 65

[14] M. J. Bregler, C., "Tracking people with twists and exponential maps", in *IEEE Conf. on Computer Vision and Pattern Recognition*, 1998. 64

[15] F. Sebastin-Grassia, "Practical parameterization of rotations using the exponential map", *J. Graph. Tools*, vol. 3, no. 3, pp. 29–48, 1998. 64

[16] M. Arulampalam, S. Maskell, N. Gordon, and T. Clapp, "A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking", *Signal Processing, IEEE Transactions on [see also Acoustics, Speech, and Signal Processing, IEEE Transactions on]*, vol. 50, no. 2, pp. 174–188, 2002. 64, 65

[17] M. Alonso, B. David, and G. Richard, "Tempo and Beat Estimation of Music Signals", in *Proceedings of ISMIR 2004, Barcelona, Spain*, 2004. 66

[18] K. Balci, E. Not, M. Zancanaro, and F. Pianesi, "Xface open source project and smil-agent scripting language for creating and animating embodied conversational agents", in *MULTIMEDIA '07: Proceedings of the 15th international conference on Multimedia*, (New York, NY, USA), pp. 1013–1016, ACM, 2007. 67

12. BIOGRAPHIES

Lale Akarun is a professor in computer engineering of Boğaziçi University, İstanbul. Her research interests are in face recognition and gesture recognition. She is currently involved in FP6 projects Biosecure and SIMILAR. She would like to participate in projects in Biometrics and in Human computer interaction.
Email: akarun@boun.edu.tr

Koray Balcı received his B.S degree in Electrical and Electronics Engineering and M.S. degree in Cognitive Sciences from Middle East Technical University (METU), Ankara, Turkey, in 2000 and 2003. Since 2005, he is a PhD student at Boğazici University, İstanbul, Turkey. His research topic is human body animation. He is also a research consultant in Bruno Kessler Foundation (formerly ITC-irst) in Trento, Italy since 2003. He has participated in European projects PF-STAR and NetCarity and implemented Xface 3D Facial Animation toolkit.
Email: koraybalci@boun.edu.tr

Elif Bozkurt is a researcher at Momentum Digital Media Technologies, Gebze, Kocaeli, Turkey. She received her BS degree in Telecommunications Engineering from Sabancı University, İstanbul, Turkey, in July 2004. Since August 2004 she is with the Momentum Digital Media Technologies. Her research interests are speech synthesis, speech - driven 3D facial animation analysis and synthesis and emotion recognition from speech.
Email: ebozkurt@momentum-dmt.com

Cristian Canton-Ferrer received the Electrical Engineering degree from the Technical University of Catalonia (UPC), Barcelona, Spain, in 2003. He is currently working towards the PhD degree at the Image and Video Processing Group in the UPC. He made his Ms. Thesis at the Signal Processing Institute at the Swiss Federal Institute of Technology (EPFL). He has contributed to European projects such as IST CHIL and NoE SIMILAR and MUSCLE. His research interests focuses in multiview image analysis, gesture recognition, human body motion and gait analysis.
Email: ccanton@gps.tsc.upc.edu

Yasemin Demir received her B.S degree in Telecommunication Engineering from İstanbul Technical University (İTÜ), İstanbul Turkey in 2006. Since 2006, she is a master student at Koç University, İstanbul, Turkey. Her research topic is audio-visual analysis of dance figures.
Email: ydemir@ku.edu.tr

A. Tanju Erdem is the CTO and co-founder of Momentum Digital Media Technologies. He received a B.S. degree in electrical and electronics engineering and a B.S. degree in physics, both in 1986, from Boğaziçi University. He received an M.S. degree in 1988 and a Ph.D. degree in 1990, both in electrical engineering, from University of Rochester. Prior to Momentum, he was with the Research Laboratories of Eastman Kodak Company from 1990 to 1998. He holds 9 U.S. patents in the field of video processing and 3D face animation and has authored and co-authored more than 50 technical publications in these fields.
Email: terdem@momentum-dmt.com

Engin Erzin is an assistant professor in the Electrical and Electronics Engineering and the Computer Engineering Departments of Koç University, İstanbul, Turkey. His research interests include speech signal processing, pattern recognition, and adaptive signal processing. Erzin received a PhD, MS, and BS from the Bilkent University, Ankara, Turkey, all in electrical engineering.
Email: eerzin@ku.edu.tr

İdil Kızoğlu is a B.S. student in Computer Engineering Department at Boğazici University, İstanbul, Turkey.
Email: idilkizoglu@boun.edu.tr

Ferda Ofli received B.Sc. degree in Electrical and Electronics Engineering, and B.Sc. degree in Computer Engineering from Koç University, İstanbul, Turkey in 2005. He is now a M.Sc. student in Electrical and Computer Engineering Department and a member of Multimedia, Vision and Graphics Laboratory at Koç University. He is currently taking part in European projects, SIMILAR NoE and 3DTV NoE. His research interests include image and video processing, specifically, object segmentation and tracking, facial expression analysis, human body modeling, motion capture and gait/gesture analysis.
Email: fofli@ku.edu.tr

A. Murat Tekalp is a professor at Koç University. His research interests are in digital image and video processing. Tekalp received an MS and a PhD in electrical, computer, and systems engineering from Rensselaer Polytechnic Institute. He received a Fulbright Senior Scholarship in 1999 and the Tübitak Science Award in 2004. Tekalp, editor in chief of the EURASIP journal Signal Processing: Image Communication, authored Digital Video Processing (Prentice Hall, 1995). He holds seven US patents and is a Fellow of the IEEE.
Email: mtekalp@ku.edu.tr

Joëlle Tilmanne holds an Electrical Engineering degree from the Faculté Polytechnique de Mons (FPMs, Belgium) since June 2006. She did her master thesis in the field of sleep signals analysis, at Lehigh University (USA). She is pursuing a PhD thesis in the TCTS lab of FPMs since September 2006, in the field of HMM based motion synthesis.
Email: joelle.tilmanne@fpms.ac.be

Yücel Yemez is an assistant professor in the Computer Engineering Department at Koç University. His current research is focused on various fields of computer vision and 3D computer graphics. Yemez received a BS from Middle East Technical University, Ankara, Turkey, and an MS and PhD from Boğaziçi University, İstanbul, Turkey, all in electrical engineering.
Email: yyemez@ku.edu.tr

EVENT RECOGNITION FOR MEANINGFUL HUMAN-COMPUTER INTERACTION IN A SMART ENVIRONMENT

Ramon Morros [1], *Albert Ali Salah* [2], *Ben Schouten* [2], *Carlos Segura Perales* [1], *Jordi Luque Serrano* [1], *Onkar Ambekar* [2], *Ceren Kayalar* [3], *Cem Keskin* [4], *Lale Akarun* [4]

[1] Technical University of Catalonia, Barcelona, Spain
[2] Centrum voor Wiskunde en Informatica, Amsterdam, The Netherlands
[3] Computer Graphics Laboratory (CGLAB), Sabanci University, Turkey
[4] Perceptual Intelligence Laboratory, Department of Computer Engineering, Boğaziçi University, İstanbul, Turkey

ABSTRACT

The aim of this project is to monitor a room for the purposes of analysing the interactions and identities of a small set of individuals. We work with multiple uncalibrated sensors that observe a single environment and generate multimodal data streams. These streams are processed with the help of a generic client-server middleware called SmartFlow. Modules for visual motion detection, visual face tracking, visual face identification, visual opportunistic sensing, audio-based localization, and audio-based identification are implemented and integrated under SmartFlow to work in coordination across different platforms.

KEYWORDS

Pattern recognition – Image sensors – Image motion analysis – Audio systems

1. INTRODUCTION

The localization and recognition of spatio-temporal events are problems of great theoretical and practical interest. Two specific scenarios are nowadays of special interest: home environment and smart rooms. In these scenarios, context awareness is based on technologies like gesture and motion segmentation, unsupervised learning of human actions, determination of the focus of attention or intelligent allocation of computational resources to different modalities. All these technologies pose interesting and difficult research questions, especially when higher levels of semantic processing is taken into account for complex interaction with the users of the environment (See Fig. 1).

It is possible to envision a useful sensor-based system even without an elaborate hierarchical reasoning structure in the home environment, where low-cost sensor equipment connected to a home computer can be used for activity monitoring and in helping the user in various settings. In smart rooms, more sophisticated equipment is usually used, allowing for more robust applications. Furthermore, wearable sensors, sensor-instrumented objects and furniture are also used frequently [4].

This paper inspects some of the problems that arise from the use of non-calibrated, low-cost sensors for observing the room. Through the observation and subsequent processing, we try to determine some basic facts about the persons in the room: identity, spatial position and interactions between people. Using this knowledge, the application can also aim at higher level goals, such as controlling a door.

In the proposed application, a set of cameras and microphones continuously monitor a room. Each person entering the room is identified (by means of a camera aimed at the entrance door and also using the microphones inside the room). Then, the persons are tracked inside the room so that its spatial position is always known. Interactions between people can be detected by using their positions, ID, head orientations and information about if they are speaking or not. Tracking can help determining if a given person is approaching some specific part of the room (i.e. the door) so that a convenient action can be taken. Tracking need not to be continuously and people can be out of sight of camera's and/ or microphone. The challenge is to still have a good knowledge about the whereabouts.

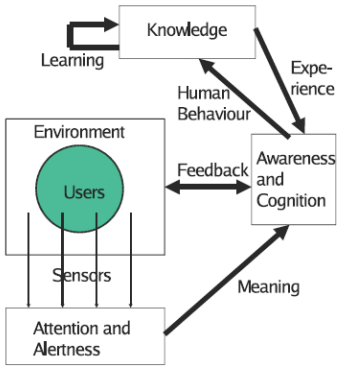


Figure 1: *The conceptual design of a human-centered smart environment (From [31]).*

There are many possible applications of such an approach. Tracking babies, kids, or elderly people for particular events, intrusion detection, gesture or speech based controlling of environmental parameters (e.g. lights, audio volume of the TV set, etc.) can be implemented. The aim of the project is to implement the tools as black-box modules that would allow straightforward application to flexible scenarios. Some possible technologies needed to fulfill these goals are motion detection (visual), person identification (visual, acoustic), tracking (visual, acoustic), speech detection, head pose estimation (visual, acoustic), gesture recognition (visual) and gait analysis.

This paper is organized as follows. In Section 2, we describe the setup of the room, and the low-cost sensors we have employed. We also briefly describe the SmartFlow middleware developed by NIST for cross-platform sensor communication in Section 2.3. Section 3 describes the separate software modules we have implemented and interfaced to SmartFlow. A discussion and our conclusions follow in Section 4.

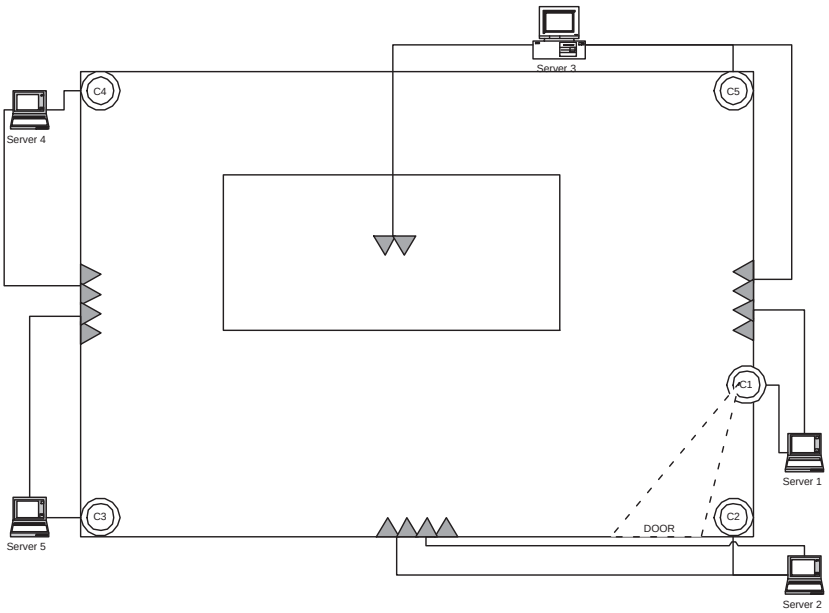


Figure 2: *The design of the room. The ceiling cameras are shown as circles at the corners of the room, and one camera that is roughly placed at eye-level is facing the door. The microphones are placed on the walls and on the table in the centre of the room.*

Figure 3: *Sample recordings from the ceiling cameras. The illumination conditions and quality of images are very different, also because the cameras were of different models.*

2. THE BU SMART ROOM

The BU smart room is established for the whole duration of the Enterface Workshop in Boğaziçi University. Since there is no sensor calibration, the actual installation was fast. We have used five cameras and fourteen microphones. The sensor equipment is low-cost, but five computers are used to drive the sensors in the absence of dedicated cheaper hardware. The sensors are connected to the computers via USB interface, and the computers are connected to an Ethernet switch for communication.

The design of our smart room is given in Fig. 1. There are four ceiling cameras in the corners of the room, and one camera facing the door. The microphones are placed in three groups of four microphones on three walls, plus a group of two microphones in the middle of the room. The room has windows on one side, which change the illumination conditions for each of the cameras. The ceiling illumination is fluorescent, and there are reflections from the ground tiles. The focus and resolutions of the cameras are minimally adjusted.

2.1. Visual Sensors

We have used two Philips 900 NC cameras, two Philips 700 NC cameras, and a Logitech UltraVision camera. The Philips cameras are placed at the ceiling corners, and the Logitech camera faces the door. All cameras have narrow angles of view, and this equipment costs less than 400$. Fig. 3 shows images acquired from these cameras at a given time. As it is evident from these images, the illumination, saturation and resolution properties are not consistent across the cameras.

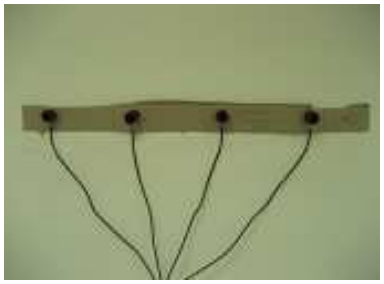


Figure 4: *The home-made low-cost microphone array.*

2.2. Audio Sensors

For the audio equipment, 14 cheap USB microphones were taken, and their protective casings were removed. The microphone arrays are constructed by attaching the stripped microphones to cardboard panels with equal spacing. Fig. 4 shows one such microphone array. The presence of five computers in the room and

the large amount of ambient noise makes the audio system very challenging.

2.3. The SmartFlow Middleware

The SmartFlow system, developed by NIST, is a middleware that allows the transportation of large amounts of data from sensors to recognition algorithms running on distributed, networked nodes [20, 23]. The working installations of SmartFlow is reportedly able to support hundreds of sensors [28].

The SmartFlow provides the user with a way of packaging algorithms into processing blocks, where the data are pictured as a distributed flow among these blocks. Sensor data are captured by clients, cast into a standard format, and a flow for each data stream is initiated. The processing blocks are themselves SmartFlow clients, and they can subscribe to one or more flows to receive the data required for their processing. Each client can output one or more flows for the benefit of other clients.

The design of a working system is realized through a graphical user interface, where clients are depicted as blocks and flows as connections. The user can drag and drop client blocks onto a map, connect the clients via flows, and activate these processing blocks.

There are two versions of SmartFlow. Version 1 is only usable under Unix OS, whereas Version 2 is usable with Windows OS. Our final system contains components that were developed under Linux and Windows, therefore we chose Version 2 as our platform. However, some issues are not completely solved in Version 2, and designs with many components suffer.

The synchronization of the clients is achieved by synchronizing the time for each driving computer, and timestamping the flows. The network time protocol (NTP) is used to synchronize the clients with the server time, and this functionality is provided by SmartFlow. A separate client is used to start the processing clients simultaneously. The video streams are not completely in one-to-one correspondence, as clients sometimes drop frames.

2.4. Data Collection

We have used three different datasets to train the speech/non-speech classes. In addition to the RT05 and RT06 datasets [33] a new dataset was collected for adaptation to the BU smart room acoustics. This dataset contains segments of silence (i.e. ambient noise) and a host of non-silence events. Non-silence events include sounds of chairs moving, keyboard typing, claps, coughs, switching on/off the light, knocking on the door, laughs, and steps. A total of 10 minutes of 14 different microphone streams were recorded, leading to a total of $840,000$ frames at a rate of 10ms. for each of the speech/non-speech classes.

For person localization, we have recorded a long sequence of frames (about eight minutes) from the four ceiling cameras simultaneously. During this sequence, the group members entered the room one by one, and walked on a pre-determined path through the room, visiting each location and stopping briefly on marker areas (centres of floor tiles). While walking, the members were asked to talk continuously to provide ground truth data for the audio based localization system, as well as the speaker identification system.

For gesture recognition, we have collaborated with Enterface Group 12, and a dataset of 90 persons was collected through their efforts. Each session contains a hand gesture part, where four gestures are repeated three times in sequence. These gestures are typically associated with turn-on (clacking the finger), turn off (cutting off the head), volume up (increase indication) and volume down (decrease indication) events. This dataset also contains sequences where the head was moved from right to left during ordinary speech.

3. SOFTWARE MODULES

3.1. Motion Detection Module

The motion detection module attempts to separate the foreground from the background for its operation. Foreground detection using background modeling is a common computer vision task particularly in the field of surveillance [29]. The method is based on detecting moving objects under the assumption that images of a scene without moving object show regular behavior which can be modeled using statistical methods.

In a practical environment like our smart room illumination could change according to the need of the user or it could also change due to gradual sun lighting change. In order to adapt changes we can update the training set by adding new samples. Each pixel observation consist of color measurement. At any time, t, pixel value at pixel i can be written as $X_{i,t} = [R_{i,t}, G_{i,t}, B_{i,t}]$.

The recent history of every pixel within an image is stacked which can be represented as $[X_{i,1}, X_{i,2}, ..., X_{i,t-1}]$ and is modeled as a set of Gaussian distributions. Now the probability of the current observation at a pixel i can be estimated using the model built from previous observations.

$$P(X_{i,t}|X_{i,1}, ..., X_{i,t-1}) = \sum w_{i,t-1} * \eta(X_{i,t}, \mu_{i,t-1}, \sigma^2_{i,t-1}) \quad (1)$$

where η is the Gaussian probability density function. $\mu_{i,t-1}$ and $\sigma^2_{i,t-1}$ are mean and covariance matrix of the Gaussian. $w_{i,t-1}$ is the weight associated with the Gaussian. To make the process on-line, a matching process is carried out; a new pixel is considered to be background if it matches with the current Gaussian component, i.e. if the distance between the pixel and the mean of the Gaussian in question is less than ϵ. In this study we have chosen $\epsilon = 2 * \sigma$. If a current pixel doesn't match the mean of the given distribution, then the parameters of the distribution are updated with a higher weight, otherwise it is updated with a lower weight $w_(i, t)$. The adaptation procedure is as follows:

$$w_{i,t} = (1 - \alpha)w_{i,t-1} + \alpha M_{i,t} \quad (2)$$

where α is learning rate, $\alpha \in [0, 1]$ and $1/\alpha$ determines the speed of the adaptation process. And $M_{i,t} = 1$ if the current pixel matches a model, otherwise it is 0 for rest of the models. In a similar vein μ and σ are updated as follows:

$$\mu_{i,t} = (1 - \lambda)\mu_{i,t-1} + \lambda X_{i,t} \quad (3)$$

$$\sigma^2_{i,t} = (1 - \lambda)\sigma^2_{i,t-1} + \lambda(X_{i,t} - \mu_{i,t})^T(X_{i,t} - \mu_{i,t}) \quad (4)$$

where

$$\lambda = \alpha\eta(X_t|\mu_{i,t-1}, \sigma_{i,t-1}) \quad (5)$$

One significant advantage of this technique is, as the new values are allowed to be part of the model, the old model is not completely discarded. If the new values become prominent overtime, the weighting changes accordingly, and new values tend to have more weight as older values become less important. Thus, if there is a movement of furniture with the room, the background model is updated rather quickly; and the same is true for lighting changes.

Fig. 5 shows a sample frame and the detected moving foreground.

Figure 5: *A frame with motion, and the detected moving foreground.*

3.2. Face Detection Module

Face detection is needed for face identification and opportunistic sensing modules. In this module, the face of each person present in the scene must be detected roughly (i.e. a bounding box around a face will be the output of this module). We use the OpenCV face detection module that relies on the adaboosted cascade of Haar features, i.e. the Viola-Jones algorithm [35]. The client that performs face detection receives a video flow from a client that in its turn directly receives its input from one of the cameras, and outputs a flow that contains the bounding box of the detected face.

3.3. Face Identification Module

The face recognition module is semi-automatic, in that it takes motion tracking and face detection for granted. This module therefore subscribes to the face detection flow that indicates face locations, and to the video flow to analyze the visual input to a camera. A technique for face recognition in smart environments [16, 34] is used. The technique takes advantage of the continuous monitoring of the environment and combines the information of several images to perform the recognition. Models for all individuals in the database are created off-line.

The system works with groups of face images of the same individual. For each test segment, face images of the same individual are gathered into a group. Then, for each group, the system compares these images with the model of the person. We first describe the procedure for combining the information provided by a face recognition algorithm when it is applied to a group of images of the same person in order to, globally, improve the recognition results. Note that this approach is independent of the distance measure adopted by the specific face recognition algorithm.

3.3.1. Combining groups of images

Let $\{x\}_i = \{x_1, x_2, ..., x_P\}$ be a group of P probe images of the same person, and let $\{C\}_j = \{C_1, C_2, ..., C_S\}$ be the different models or classes stored in the (local or global) model database. S is the number of individual models. Each model C_j contains N_j images, $\{y_n\}^j = \{{y_1}^j, {y_2}^j, ..., {y_{N_j}}^j\}$ where N_j may be different for every class. Moreover, let

$$d(x_i, {y_n}^j) : \mathcal{R}^Q x \mathcal{R}^Q \quad \rightarrow \quad \mathcal{R} \tag{6}$$

be a certain decision function that applies to one element of $\{x\}_i$ and one element of $\{y\}_n^j$, where Q is the dimension of x_i and y_n^j. It represents the decision function of any face recognition algorithm. It measures the similarity of a probe image x_i to a test image y_n^j. We fix a decision threshold R_d so that x_i and y_n^j represent the same person if $d(x_i, y_n^j) < R_d$. If, for a given x_i the decision function is applied to every $y_n^j \in C_j$, we can define the δ value of x_i relative to a class C_j, δ_{ij} as

$$\delta_{ij} = \#\{y_n^j \in C_j | d(x_i, y_n^j) < R_d\} \tag{7}$$

That is, δ_{ij} counts the number of times that the face recognition algorithm matches x_i with an element of C_j. With this information, the δ-Table is built, and based on this table we define the following concepts:

- Individual Representation of x_i: It measures the representation of sample x_i by class C_j:

$$R(x_i, C_j) = \frac{\delta_{ij}}{N_j} \tag{8}$$

- Total representation of x_i: It is the sum of the individual representations of x_i through all the classes:

$$R(x_i) = \frac{1}{P} \sum_{j=1}^{S} R(x_i, C_j) = \sum_{j=1}^{S} \frac{\delta_{ij}}{N_j} \tag{9}$$

- Reliability of a sample x_i given a class C_j: It measures the relative representation of sample x_i by class C_j considering that sample x_i could be represented by other classes:

$$\rho(x_i, C_j) = \begin{cases} \frac{R(x_i, C_j)}{R(x_i)} = \frac{\delta_{ij}/N_j}{\sum_{k=1}^{S} \frac{\delta_{ik}}{N_k}} & R(x_i) > 0 \\ 1 & R(x_i) = 0 \end{cases} \leq 1$$

- Representation of C_j: It estimates the relative representation of a group of samples $\{x\}_i$ by a class C_j. Weighting is performed to account for the contribution of the group $\{x\}_i$ to other classes:

$$R(C_j) = \frac{1}{P} \sum_{i=1}^{P} \rho_{ij} \delta_{ij} \tag{10}$$

- Match Likelihood M for class C_j: It relates a class representation and its match probability. If $r = R(C_j)$, then:

$$M(C_j) = \frac{1 - e^{\frac{-r^2}{\sigma^2}}}{1 - e^{\frac{-N_j^2}{\sigma^2}}} \tag{11}$$

  where σ adjusts the range of $R(C_j)$ values.

- Relative Match Likelihood for a class C_j: It relates the M of a class C_j and the maximum M of the other classes:

$$RML(C_j) = \begin{cases} \frac{M(C_j)}{\max_{k \neq j}(M(C_k))} & M(C_j) \geq 0.5 \\ 0 & M(C_j) < 0.5 \end{cases} \tag{12}$$

  This measure determines if the selected class (that with the maximum M) is widely separated from other classes. A minimum value of M is required, to avoid analyzing cases with too low M values.

Relying on the previous concepts, the recognition process is defined, in the identification mode, as follows:

- Compute the δ-Table.
- Compute the match likelihood M for every model.
- Compute the RML of the class with the highest $M(C_j)$.

The group is assigned to the class resulting in a highest RML value In this work, a PCA based approach [14] has been used. This way, the decision function is the Euclidean distance between the projections of x_i and y_n^j on the subspace spanned by the first eigenvectors of the training data covariance matrix:

$$d(x_i, y_n^j) = ||W^T x_i - W^T y_n^j|| \tag{13}$$

where W^T is the projection matrix.

The XM2VTS database [18] has been used as training data for estimating the projection matrix and the first 400 eigenvectors have been preserved. Only frontal faces are used for identification. Note that, in our system, models per each person have been automatically generated, without human intervention. All images for a given individual in the training intervals are candidates to form part of the model. Candidate face bounding boxes are projected on the subspace spanned by the first eigenvectors of the training data covariance matrix W^T. The resulting vector is added to the model only if different enough from the vectors already present in the model.

3.4. Opportunistic Sensing Module

The opportunistic sensing module aims at identifying persons in the room when the face information is not available, or not discriminatory. The primary assumption behind the operation of this module is that the variability in a users appearance for a single camera is relatively low for a single session (this case is termed intra session in [34]), and a user model created on-the-fly can provide us with useful information [32]. We use the following general procedure for this purpose: Whenever the face identification module returns a reliable face identification, cameras with access to the area with the detected face consult the motion detection module, and grab a window from the heart of the motion blob. The pixel intensities within this window are modeled statistically, and this statistical model is then used to produce the likelihood values for every candidate person for which the system stored a mixture model.

The general expression for a *mixture model* is written as

$$p(\boldsymbol{x}) = \sum_{j=1}^{J} p(\boldsymbol{x}|\mathcal{G}_j) P(\mathcal{G}_j) \tag{14}$$

where $\mathcal{G}_j$ stand for the components, $P(\mathcal{G}_j)$ is the prior probability, and $p(\boldsymbol{x}|\mathcal{G}_j)$ is the probability that the data point is generated by component j. In a *mixture of Gaussians* (MoG), the components in Eq. 14 are Gaussian distributions:

$$p(x|\mathcal{G}_j) \sim \mathcal{N}(\mu_j, \boldsymbol{\Sigma}_j) \tag{15}$$

In a MoG, the number of parameters determine the complexity of the model, and the number of training samples required for robust training increases proportionally to the complexity of the model. Introducing more components means that the ensuing mixture will be more complex, and more difficult to train on the one hand, but potentially more powerful in explaining the data, as complex models usually go.

One way of controlling the complexity is to state some assumptions regarding the shape of the covariance matrix. A complete covariance matrix $\boldsymbol{\Sigma}_j$ for a d-dimensional data set has $O(d^2)$ parameters, and specifies a flexible and powerful statistical model. Learning a sample covariance of this shape often creates serious overfitting issues and singularities due to limited training data. Frequently, a diagonal covariance matrix is adopted, which means the covariances between dimensions are discarded, and only the variances are modeled.

To build our statistical models for the users of the smart room, we use a *factor analysis* (FA) approach, which represents a trade-off between model complexity and modeling covariances. In FA, the high dimensional data $\boldsymbol{x}$ are assumed to be generated in a low-dimensional manifold, represented by latent variables $\boldsymbol{z}$. The *factor space* spanned by the latent variables is similar to the principal space in the PCA method, and the relationship is characterized by a *factor loading matrix* Λ, and independent Gaussian noise $\boldsymbol{\epsilon}$:

$$\boldsymbol{x} - \boldsymbol{\mu}_j = \boldsymbol{\Lambda}_j \boldsymbol{z} + \boldsymbol{\epsilon}_j \tag{16}$$

The covariance matrix in the d-dimensional space is then represented by $\boldsymbol{\Sigma}_j = \boldsymbol{\Lambda}_j \boldsymbol{\Lambda}_j^T + \Psi$, where Ψ is a diagonal matrix and $\boldsymbol{\epsilon}_j \sim \mathcal{N}(0, \Psi)$ is the Gaussian noise. We obtain a *mixture of factor analysers* (MoFA) by replacing the Gaussian distribution in Eq. 14 with its FA formulation.

In the opportunistic sensing module, the surface features are modeled with mixture models. A single frame is used to produce a training set for the MoFA model. If the the number of components and the number of factors per component are specified beforehand, the maximum likelihood parameters can be computed with the Expectation-Maximization (EM) algorithm [9]. With no prior information, an incremental algorithm can be used to determine model parameters automatically [26]. The incremental MoFA algorithm (IMoFA) starts with a one-factor, one-component mixture and adds new factors and new components until the likelihood is no longer increased on the validation set.

For component addition, a multivariate kurtosis-based measure is adopted to select components that look least like unimodal Gaussians:

$$\gamma_j = \{b_{2,d}^j - d(d+2)\} \left[\frac{8d(d+2)}{\sum_{t=1}^{N} h_j^t} \right]^{-\frac{1}{2}} \tag{17}$$

$$b_{2,d}^j = \frac{1}{\sum_{l=1}^{N} h_j^l} \sum_{t=1}^{N} h_j^t \left[(\boldsymbol{x}^t - \boldsymbol{\mu}_j)^T \boldsymbol{\Sigma}_j^{-1} (\boldsymbol{x}^t - \boldsymbol{\mu}_j) \right]^2 \tag{18}$$

with $h_j^t \equiv E[\mathcal{G}_j|\boldsymbol{x}^t]$, and the component with greatest γ_j is selected for splitting.

For factor addition the difference between sample covariance and modeled covariance is monitored and the component with the largest difference is selected for factor addition. The new factor is the principal axis of the residual error vectors.

There are currently no accuracy results for the opportunistic sensing module, since the recorded data is not annotated with ground truth. However, visual inspection shows that this model is successful under two conditions:

- The training frame should be discriminative.
- The overall image intensity conditions during operation should be consistent with the training conditions.

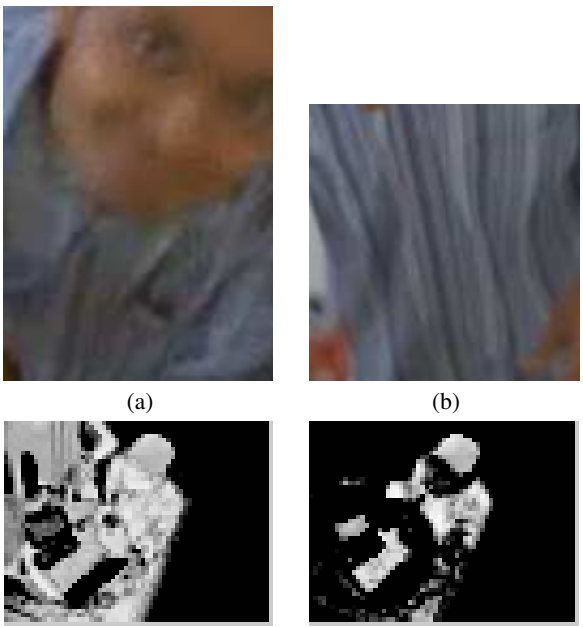
In Fig. 6 three sample training frames are shown. To ensure fast operation, only a portion of these frames is used for training. 5×5 pixels with RGB values are concatenated to form $75-$dimensional vectors, which are then modeled with IMoFA mixtures.

Fig. 9 shows the detection results for each of the three people in the system. Since we use generative models, an increase in the number of subjects does not decrease the localization accuracy of the system, but considering additional candidates brings a computational cost. The client that performs likelihood computation pre-loads the mixture parameters, computes inverse covariance and determinant parameters, and waits for data packets. The real-time operation is ensured by dropping data frames.

(a) (b) (c)

Figure 6: *Training frames for a) Albert, b) Ramon, c) Onkar.*

Fig. 7 shows the effect of a non-discriminative frame. The presence of generic intensities results in a too general mixture, and many data packets produce high likelihoods.



(a) (b)

(c) (d)

Figure 7: *a) Non-discriminative training frame for class Onkar. b) A more discriminative training frame for class Onkar. c) Detection of Onkar results in many false-positives with the non-discriminative training frame. d) The discriminative frame works fine.*

3.5. Speaker Identification Module

The Speaker ID module (SID) segments the audio data, identifying the silences and the speaker changes. Furthermore, the algorithm is able to identify different speakers with the help of prior knowledge. With this purpose, a small database of ten persons was collected in the BU smart room. The speaker models are computed off-line.

The speaker identification system is composed of several modules developed in the SmartFlow framework, and it can be inspected in three groups. The Speech Activity Detection (SAD) is in charge of detecting the speech activity in the room and discriminating it from non-speech events. The Acoustic SEgmentation (ASE) module retrieves speech information from the SAD and provides the Speaker IDentification (SID) stage with homogenous acoustic data from a single identity. All three stages work in parallel, continuously computing the high level information, speech/non-speech segmentation, acoustic change detection and frame-score accumulation for the identity labels, thereby monitoring the room constantly in real time. However, an identity label is only provided in the case that the acoustical information is sufficient, and homogenously collected from a single person, and the speech data is sufficient for discrimination.

Fig. 8 depicts the SmartFlow map employed in the SID implementation. The frequency downsampling stage produces acoustic signals at a rate of 16.000 KHz for the remaining stages, for a total of 14 microphones. The flow connections indicate the feedback and interactions between the three main modules.

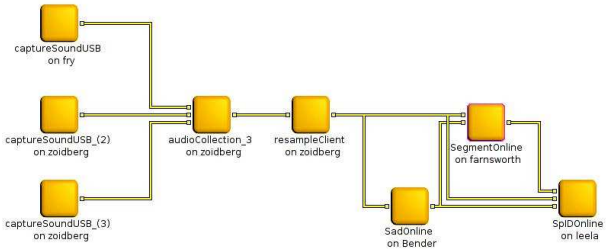


Figure 8: *Smarflow Map implementation of the SID module. The first stages are related to capturing signal modules and downsampling. The SelectChannel box gathers all the input channels coming from different machines and selects one of them based on a simple energy threshold decision.*

3.5.1. Speech Activity Detection

The SAD module used in this work is based on a support vector machine (SVM) classifier [27]. The performance of this system was shown to be good in the 2006 Rich Transcription SAD Evaluations (RT06) [33]. A GMM-based system that ranked among the best systems in the RT06 evaluation was selected as a baseline for performance comparison [33].

For classical audio and speech processing techniques that involve GMMs, the training data are in the order of several hundred thousand examples. However, SVM training usually involves far less training samples, as the benefit of additional samples is marginal, and the training becomes infeasible under too large training sets. Some effective methods should be applied to reduce the amount of data for training without losing accuracy. Therefore, we employ Proximal SVMs (PSVM) in the same way as proposed in [33].

Proximal Support Vector Machine (PSVM) has been recently introduced in [8] as a result of the substitution of the inequality constraint of a classical SVM $y_i(wx_i + b) \geq 1$ by the equality constraint $y_i(wx_i + b) = 1$, where y_i stands for a label of a vector x_i, w is the norm of the separating hyperplane H_0, and b is the scalar bias of the hyperplane H_0. This simple modification significantly changes the nature of the optimization problem. Unlike conventional SVM, PSVM solves a single square system of linear equations and thus it is very fast to train. As a consequence, it turns out that it is possible to obtain an explicit exact solution to the optimization problem [8].

Depending on the speech activity inside the room, we should penalize the errors from the Speech class more than those from the Non-Speech class, in the case of a meeting for example. It is possible to selectively penalize errors for the SVM in the training stage by introducing different costs for the errors in two

classes by introducing different generalization parameters C_- and C_+. This approach will adjust the separating hyperplane H_0, and it will no longer be exactly in the middle of the H_{-1} and H_1 hyperplanes (Fig. 9). It is worth mentioning that favouring a class in the testing stage (after the classifier is trained) is still possible for SVM through the bias b of the separating hyperplane.

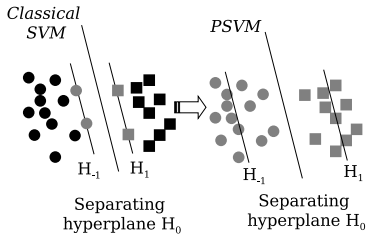


Figure 9: *Proximal Support Vector Machine based SVM.*

3.5.2. Bayesian Information Criterion based Segmentation

Audio segmentation, sometimes referred to as acoustic change detection, consists of exploring an audio file to find acoustically homogeneous segments, detecting any change of speaker, background or channel conditions. It is a pattern recognition problem, since it strives to find the most likely categorization of a sequence of acoustic observations. Audio segmentation becomes useful as a pre-processing step in order to transcribe the speech content in broadcast news and meetings, because regions of different nature can be handled in a different way.

The ALIZE Toolkit was used in this work to implement a XBIC-based segmentation system, making use of the algorithms for the GMM data modeling and the estimation of the parameters. The ALIZE toolkit is a free open tool for speaker recognition from the LIA, Université d'Avignon [1]. The "Bayesian Information Criterion" (BIC) is a popular method that is able to perform speaker segmentation in real time.

Let the set $\Theta = \{\theta_j \in \Re^d \mid j \in 1 \dots N\}$, a sequence of N parameter vectors, the BIC is defined as:

$$BIC_\Theta = L - \alpha P \tag{19}$$

where P is the penalty and α a free design parameter. L is the logarithm of the probability performed by the set of observations Θ over the model λ_i,

$$L = P(\Theta|\lambda_i) = \sum_{k=1}^{N} log(\theta_k|\lambda_i) \tag{20}$$

Let an instant $\theta_j \in \Theta$, it can be defined two partitions of Θ: $\Theta_1 = \{\theta_1 \dots \theta_j\}$ and $\Theta_2 = \{\theta_{j+1} \dots \theta_N\}$ with length N_1 and N_2. In order to take a decision about a change in the speakers, two different hypotheses can be considered:

- H_0: An unique class λ is better at modeling the data from the whole set of observations Θ.
- H_1: Two independent classes λ_1 and λ_2 are better in jointly modeling the before and after of the postulated instance of speaker change θ_j.

The best hypothesis is chosen by evaluating the following expression:

$$\Delta BIC = BIC_{H_0} - BIC_{H_1} = BIC_\Theta - (BIC_{\Theta_1} + BIC_{\Theta_2}) \tag{21}$$

Therefore, and looking at Eqs. 19 and 20 the criterion of change speaker turn at time i can be defined by evaluating the following expression:

$$\Delta BIC(i) = P(\Theta|\lambda) - P(\Theta_1|\lambda_1) - P(\Theta_2|\lambda_2) - \alpha P \tag{22}$$

where the penalty term depends on the number of parameters (complexity) employed to estimate the whole model λ and the two sub-models λ_1,λ_2. This constant is set to a low value, since it only affects the decision threshold around 0.

In overall, a speaker change turn is decided at the time instant i for which $\Delta BIC(i) > 0$, which means that the dual model is now better adapted to the observation data in comparison to the single model. Taking into account the definition of the Rabiner distance, the distance among two Hidden Markov Models is

$$D_{rab} = \frac{D(\lambda_a, \lambda_b) + D(\lambda_a, \lambda_b)}{2} \tag{23}$$

with D as,

$$D(\lambda_a, \lambda_b) = \frac{1}{N_b}\left(\sum_{k=1}^{N_b} log\, p(\theta_k|\lambda_a) - \sum_{k=1}^{N_a} log\, p(\theta_k|\lambda_b)\right) \tag{24}$$

Assuming that the two data segments have the same duration, and sorting the terms of the Eq. 23, we obtain a probabilistic definition of the Rabiner distance,

$$D'_{rab} = \big(P(\Theta_1|\lambda_2) + P(\Theta_2|\lambda_1)\big) - \big(P(\Theta_1|\lambda_1) + P(\Theta_2|\lambda_2)\big) \tag{25}$$

where $P(\Theta_i|\lambda_j)$ is defined in Eq. 20. Eq. 25 is similar to the one presented in the BIC procedure 19, but using a second shared term: $P(\Theta_1|\lambda_1) + P(\Theta_2|\lambda_2)$, which changes as the models adapt to the training set.

The first term of both equations, Eq. 19 and Eq. 25, measure how well the whole audio segment is adapted to a single model or to two models, respectively. In both cases a high probability is obtained if the segment belongs to the same speaker.

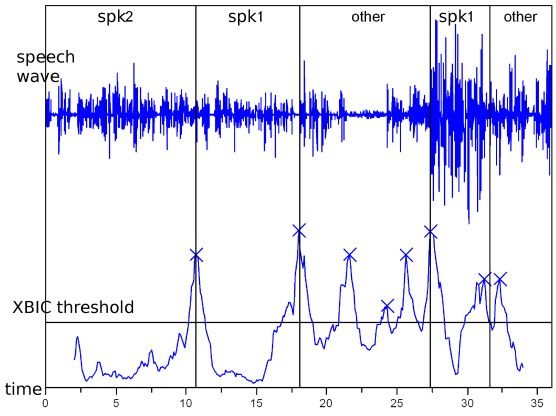


Figure 10: *Segmentation example.*

In the case of the Eq. 25, when two acoustically similar segments are evaluated, the measure oscillates around the 0 value. When a speaker change occurs, the XBIC distance peaks, mostly due to the cross probability term. By this reason the XBIC measure is defined as:

$$XBIC(i) = P(\Theta_1|\lambda_2) + P(\Theta_2|\lambda_1) \tag{26}$$

If a suitable threshold is chosen, the speaker change can be determined at the time instant i as $XBIC(i) < \text{threshold}_{XBIC}$

3.5.3. Speaker Identification

The Person Identification (PID) problem consists of recognizing a particular speaker from a segment of speech spoken by a single speaker. Matched training and testing conditions and far-field data acquisition are assumed, as well as a limited amount of training data and no a priori knowledge about the room environment. The algorithm implemented in this work was tested in the CLEAR'07 PID evaluation campaign, obtaining the best position in mean in all test/train conditions [17].

The algorithm commences by processing each data window by subtracting the mean amplitude, supposing the DC offset is constant throughout the waveform. A Hamming window was applied to each frame and a FFT is computed. The FFT amplitudes are then averaged in 30 overlapped triangular filters, with central frequencies and bandwidths defined according to the Mel scale. Their output represents the spectral magnitude in that filter-bank channel. Instead of the using Discrete Cosine Transform, such as in the Mel-Frequency Cepstral Coefficients (MFCC) procedure [5], the samples are parametrised using the Frequency Filtering (FF):

$$H(z) = z - z^{-1} \tag{27}$$

over the log of the filter-bank energies. Finally, we obtain 30 FF coefficients. A vector of parameters is calculated from each speech segment. The choice of this setting was justified through experimental validation, showing that FF coefficients result in higher accuracies than MFCC coefficients for both speech and speaker recognition tasks, as well as being more robust against noise and computationally more efficient [19]. Additionally, the FF coefficients are uncorrelated, and they have an intuitive interpretation in the frequency domain.

In order to capture the temporal evolution of the parameters, the first and second time derivatives of the features are appended to the basic static feature vector. The so-called Δ and Δ-Δ coefficients are also used in this work. Note that the first coefficient of the FF output, which is proportional to the signal energy, is also employed to compute the model estimation, as well as its velocity and acceleration parameters. Next, for each speaker that the system has to recognize, a model of the probability density function of the parameter vectors is estimated. Gaussian Mixture Models with diagonal covariance matrices are employed, and the number of components is set to 64 for each mixture. The large amount of components assure that the statistical learning is robust, and is further justified by the availability of a very large training set.

The parameters of the models are estimated from speech samples of the speakers using the well-known Baum-Welch algorithm [24]. Given a collection of training vectors, the maximum likelihood model parameters are estimated using the iterative Expectation-Maximisation (EM) algorithm. The sensitivity of EM to cases with few training data is well-known, yet under the present training conditions (with ample data), 10 iterations are demonstrably enough for parameter convergence. This parameter is retained for both training conditions and for all the client models.

In the testing phase of the speaker identification system, a set of parameters $\mathbf{O} = \{\mathbf{o}_i\}$ is computed from the speech signal. Next, the likelihood of obtaining the signal under each client model $\mathbf{O}$ is calculated and the speaker model with the largest posterior probability is chosen,

$$s = \arg\max_j \left\{ L\big(\mathbf{O}|\lambda_j\big) \right\} \tag{28}$$

where s is the recognised speaker, and L is the likelihood function from a linear combination of M unimodal Gaussians of dimension D [3]. Therefore, $L\big(\mathbf{O}|\lambda_j\big)$ is the likelihood that the vector $\mathbf{O}$ has generated by the speaker model λ_j.

3.6. Sound Based Localization Module

Conventional acoustic person localization and tracking systems usually consist of three basic stages. In the first stage, estimation of such information as Time Difference of Arrival or Direction of Arrival is usually computed by the combination of data coming from different microphones. The second stage involves a set of relative delays or directions of arrival estimations to derive the source position that agrees most with the data streams and with the given geometry. In the third (and optional) stage, possible movements of the sources can be tracked according to a motion model. These techniques need several synchronized high-quality microphones. In the BU Smart Room setting we have only access to low-quality, uncalibrated and unsynchronized microphones. As explained in the room setup description, only pairs of microphones within an audio capture device are synchronized, there is no synchronization across different capture devices. Moreover, the channels within a single capture device have a highly correlated noise of the same power level as the recorded voice, thus estimating the position of speakers in this environment is a very difficult and challenging task.

The acoustic localization system used in this project is based on the SRP-PHAT localization method, which is known to perform robustly in most scenarios [6]. The SRP-PHAT algorithm (also known as Global Coherence Field) tackles the task of acoustic localization in a robust and efficient way. In general, the basic operation of localization techniques based on SRP is to search the room space for a maximum in the power of the received sound source signal using a delay-and-sum or a filter-and-sum beamformer. In the simplest case, the output of the delay-and-sum beamformer is the sum of the signals of each microphone with the adequate steering delays for the position that is explored. Concretely, the SRP-PHAT algorithms consists in exploring the 3D space, searching for the maximum of the contribution of the PHAT-weighted cross-correlations between all the microphone pairs. The SRP-PHAT algorithm performs very robustly due the the PHAT weighting, keeping the simplicity of the steered beamformer approach.

Consider a smart-room provided with a set of N microphones from which we choose M microphone pairs. Let $\mathbf{x}$ denote a $\mathbf{R}^3$ position in space. Then the time delay of arrival $TDOA_{i,j}$ of an hypothetic acoustic source located at $\mathbf{x}$ between two microphones i, j with position $\mathbf{m}_i$ and $\mathbf{m}_j$ is:

$$TDOA_{i,j} = \frac{\| \mathbf{x} - \mathbf{m}_i \| - \| \mathbf{x} - \mathbf{m}_j \|}{s}, \tag{29}$$

where s is the speed of sound.

The 3D room space is then quantized into a set of positions with typical separations of 5-10cm. The theoretical TDOA $\tau_{\mathbf{x},i,j}$ from each exploration position to each microphone pair are precalculated and stored.

PHAT-weighted cross-correlations of each microphone pair are estimated for each analysis frame [21]. They can be expressed in terms of the inverse Fourier transform of the estimated cross-power spectral density ($G_{m_1 m_2}(f)$) as follows:

$$R_{m_i m_j}(\tau) = \int_{-\infty}^{\infty} \frac{G_{m_i m_j}(f)}{|G_{m_i m_j}(f)|} e^{j2\pi f \tau} df, \tag{30}$$

The estimated acoustic source location is the position of the quantized space that maximizes the contribution of the cross-correlation of all microphone pairs:

$$\hat{\mathbf{x}} = \underset{\mathbf{x}}{\operatorname{argmax}} \sum_{i,j \in \mathbb{S}} R_{m_i m_j}(\tau_{\mathbf{x},i,j}), \tag{31}$$

where $\mathbb{S}$ is the set of microphone pairs. The sum of the contributions of each microphone pair cross-correlation gives a value of confidence of the estimated position, which can be used in conjunction with a threshold to detect acoustic activity and to filter out noise. In our work, we use a threshold of 0.5 per cluster of 4 microphones. It is important to note that in the case of concurrent speakers or acoustic events, this technique will only provide an estimation for the dominant acoustic source at each iteration.

3.7. Gesture Tracking Module

Hand gestures used in human–computer interaction (HCI) systems are either communicative or manipulative. Manipulative hand gestures are used to act on virtual objects or parameters, whereas communicative gestures are combination of dynamic acts and static symbols that have communication purposes. HCI applications often focus on a subset of these classes of gestures and consider only these classes as meaningful. The hand gesture recognition module we have designed for the smart room application is aimed to use both communicative and manipulative gestures for HCI and to automatically distinguish these in real time. This module is based on the work presented in [11, 12, 13].

The gesture recognition module is the main channel of interaction with between the users and the computers in BU Smart Room, since it allows the users to give commands to the underlying smart system. The gesture recognition system our module is based on, handles the event recognition problem using two primary methods:

- by manipulative gestures, i.e. by allowing direct translation of the hand trajectory and shape into continuous events, such as manipulation of virtual objects, or
- by communicative gestures, i.e. by recognizing certain patterns performed by the user and firing high level discrete events.

In order to make use of both of these methodologies, we have designed a simple extendible gesture language that consists of actions and objects. The structure of each command consists of a combination of one action and one object, the former corresponding to a certain pattern and the latter to a simple hand movement, specifically in this order to prevent false positives. The reasoning behind this is based on the fact that communicative gestures are chosen in a way that they are unlikely to be performed unintentionally, wheras the manipulative ones may correspond to very common hand movements. Since we restrict the objects to follow actions, the intentional manipulative gestures corresponding to the objects can unambiguously be recognized.

Gestures are performed by the flexible hands in 3D space. The motion and change in the shape of the hand have different characteristics. Thus, they are often modeled separately and considered to be different modalities. Both appearance based and 3D model based approaches have been pursued for computer vision based hand shape modeling. 3D models are complex geometrical models using a priori knowledge about hand geometry, while appearance based models use features extracted from the projected intensity image of the hand. In general, most of the computer vision techniques extract the silhouettes of the hands first.

To simplify the hand segmentation problem, either simple backgrounds are assumed, or colored markers are used. Since hands are homogeneously colored, hand tracking can be accomplished using color–based features. However, other skin colored regions such as the face make this task difficult and time consuming. Due to real time execution considerations, markers are used in our system to detect the hand in each image. The choice of color for the marker is not predefined, but is restricted to colors that are significantly different than the skin color.

In the case of smart rooms, simple backgrounds can not be assumed. By employing colored markers, it is possible to robustly detect and track the hands by using simple color based detection algorithms. We have developed two different techniques for marker detection in BU Smart Room, which basically make use of different color systems, namely RGB and HSV. While the RGB–based method uses either fixed or floating ranges to connect neighboring colors for segmentation, the HSV–based method calculates the Hue–Saturation histogram of the marker and uses it as a signature. These methods will be explained in more detail in the following section.

The choice of the hand features extracted from the hand region primarily depends on the target application. Real time HCI applications often prefer to employ simpler models, while establishing a certain amount of variability and descriptive power for specific hand postures. The features used may range from the easily detectable motion of the hand and openness or closeness of the fingers, to the angles of each finger articulation, which might be needed for an advanced sign language recognition system. Since the setup or the quality of the cameras in different smart room environments are likely to be subject to considerable change, the system is assumed to consist of the most basic cameras. Therefore, simple appearance–based methods are preferred for our application.

Appearance–based approaches are not based on prior and explicit knowledge of the hand model. Therefore, model parameters are not directly derived from the 3D spatial description of the hand. Such methods learn to relate the appearance of a given hand shape image to its actual posture.

The most common appearance–based approach is using low level features derived from the images to model hand shapes. These parameters include image moments, image eigenvectors, i.e. "eigenhands", contours and edges, orientation histograms and Hu or Zernike moments. Currently, the training and recognition algorithms rely on Hu moments and the angle of the hand image. The Smartflow environment allows to plug in different feature extractors in a simple manner. Therefore the feature extraction module can easily be replaced with other methods for testing. Yet, the training phase should be repeated with the new feature extraction module.

Dynamic hand gestures are sequential data that involve both temporal and spatial context. Hidden Markov Models (HMMs), which can deal with the temporal invariability of signals are often employed for gesture recognition [15, 25, 36, 37]. In the recognition process, a given trajectory in the model parameter space of the gestures is tested over the set of trained HMMs in the system. The model that gives the highest likelihood is taken to be the gesture class the trajectory belongs to.

HMM is a powerful tool for sequential data modeling and recognition. Yet, it is not directly extendible to multiple synchronous or asynchronous data sequences. HMM variants such as coupled HMMs have been proposed to overcome this problem. Input–Output HMMs, which are proposed by Bengio *et al.* in [2] are neural network–HMM hybrids, which attack the shortcomings of HMMs by representing the local models, i.e. the hidden states, with possibly nonlinear and flexible complex structures such as Multi Layer Perceptrons (MLPs). IOHMMs condition the stochastic architecture of HMMs on an input sequence that has no independence assumptions. The architecture can be conditioned on any function of observable or system parameters. Therefore, the stochastic network of an IOHMM is not homogeneous. Due to their inhomogeneity, IOHMMs are expected to learn long time dependencies better than HMMs.

The gesture recognition module in BU Smart Room is designed to make use of discrete, continuous and input–output HMMs.

In the following section, the marker registration tool and the marker color signatures used in the BU Smart Room system will be explained in detail.

3.7.1. Marker Registration and Detection

The marker registration client developed for Smartflow, namely "ipl marker registrar", aims to learn the parameters of the chosen marker. This module employs two different methodologies:

- RGB–based manual detection method
- HSV–based automatic detection method

The capabilities of the client are accessible in run–time by pressing "h", which displays a help screen. This screen shows how to switch modes, methods and settings. The manual and automatic detection modes make use of different parameters and therefore, they have different user interfaces.

In the manual registration mode, the user needs to click on the marker with the left button to run a connected components algorithm, starting with the pixel clicked. The connection criteria is euclidean distance:

- between the first pixel and the pixel being considered in the fixed range mode (activated by pressing "f"), or
- between the last pixel connected and its neighbors in the floating range mode (activated by pressing "g").

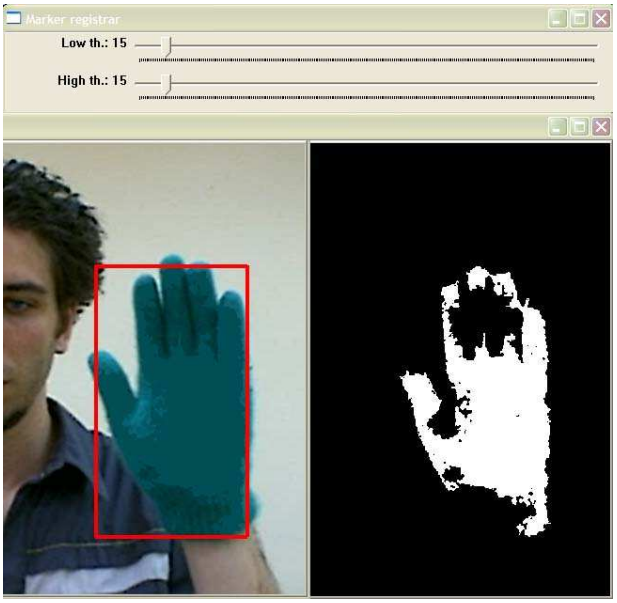


Figure 11: *Effect of low settings for the fixed range mode*

Typically, fixed range mode needs much higher thresholds to be set than the floating mode. Examples of too low, too high and good settings for both of the modes are given in Figures 11, 12, 13, 14, 15 and 16. The settings in the user interface are simply the asymmetric range thresholds for values higher and lower than the original pixel. The best settings can be selected and visually confirmed by the user, which are then saved for further utilization by the hand tracking module.

There are two problems with this approach. First of all, this mode of execution requires human intervention for marker registration, which might cause a problem if the computer is not inside the smart room. Also, RGB color space based tracking

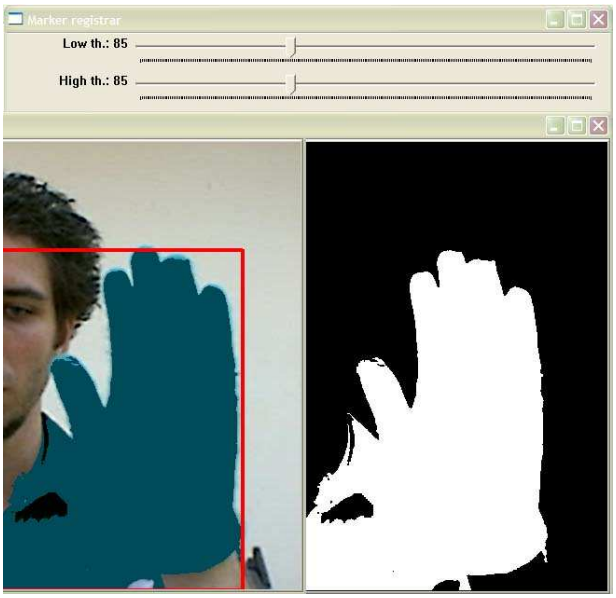


Figure 12: *Effect of high settings for the fixed range mode*

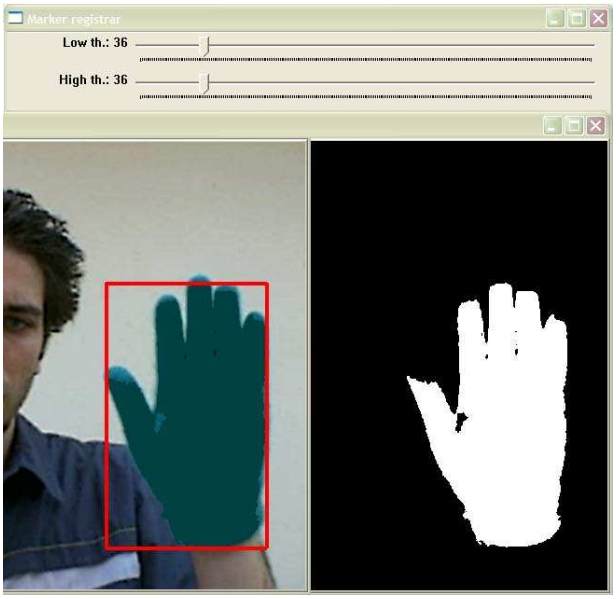


Figure 13: *Good settings for the fixed range mode*

Figure 14: *Effect of low settings for the floating range mode*

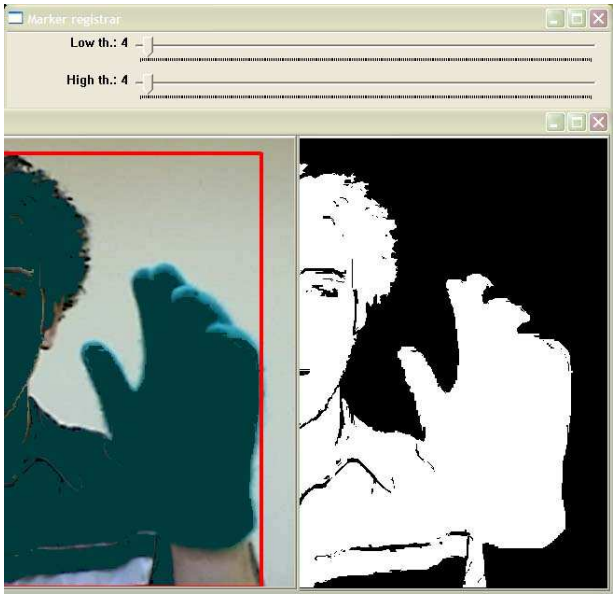


Figure 15: *Effect of high settings for the floating range mode*

is not robust to illumination changes, which would make the registration settings obsolete. Therefore, a second method is implemented, which is an automatic HSV based approach. This method can be used to detect markers without the supervision of a user, and the corresponding tracking method is more robust to illumination changes.

The main difference of the automatic method from its manual counterpart is that a supervisor does not need to click on the marker. Instead, by assuming that the marker has the fastest motion in an image sequence, this module attempts to distinguish the marker from image differences, i.e. motion. It first captures two images separated by a few frames, smoothes them with a Gaussian filter and finds their difference. The difference image is thresholded according to a user setting, and the pixels that are too dark or too bright are also eliminated using two more thresholds, also selectable via user settings. The final image is converted into HSV color space and the 2D Hue–Saturation histogram is calculated. The main consideration in this method is that the pixels corresponding to the marker fall to a single bin. In order to ensure this the number of hue and saturation bins should not be selected too high. Selecting the numbers too low on the other hand, causes the final bin to not only correspond to the marker, but possibly to some other colors, and accordingly, to other objects.

The settings in the automatic mode are as follows:

- RGB threshold - Used to determine the pixels to consider in the difference image
- Dark threshold - Leaves out the pixels that are too dark
- Bright thresold - Leaves out the pixels that are too bright
- # Hue bins - Number of hue bins in the histogram
- # Saturation bins - Number of saturation bins in the histogram

The effects of choosing wrong numbers for the RGB, dark and bright thresholds are straightforward. To clarify the effect of selection of number of bins, two cases are reproduced that correspond to very high and good selection of number of bins for hue and saturation in Figures 17 and 18.

In the case of automatic registration, the marker signature is the histogram bin with the maximum number of pixels. This bin is back–projected in the tracking phase to detect the marker in the incoming frames. The tracking and feature extraction modules are explained in the following section.

3.7.2. Hand Tracking and Gesture Recognition

The hand detection module attempts to track the hand using the algorithm corresponding to the method used during the final registration phase. For the manual case, as explained in the previous section, the low and high thresholds and a reference color are needed to detect the hand. In order to enhance the method, double thresholding is applied. The algorithm first looks for pixels that are close enough to the refence pixel, using a threshold that is $\frac{1}{5}$ of the saved user settings. Starting from such pixels, the algorithm connects the neighboring pixels according to the criteria supplied by the registration module.

In the automatic registration case, the detection module back–projects the hue–saturation histogram calculated by the registration module and employs a region growing algorithm, where each connected pixel falls into the selected bin.

Even though the hue–based algorithm is more robust to illumination changes, both methods may generate noisy results in more challenging conditions. Therefore filtering is a necessity, especially if recognition is to be performed on the data. Due to high framerate, the marker can be assumed move linearly between the frames, and the state change can be described by a

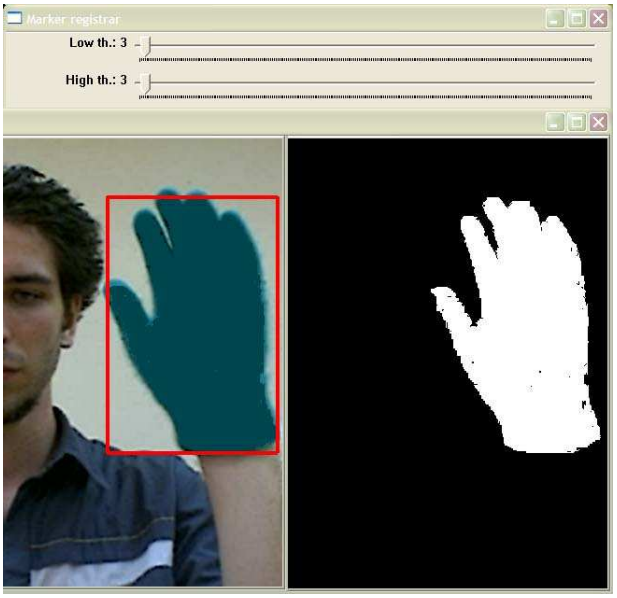


Figure 16: *Good settings for the floating range mode*

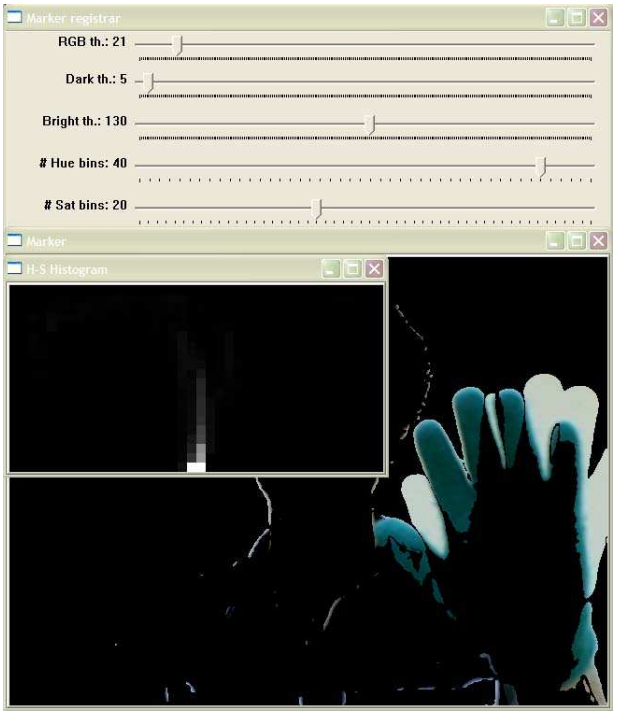


Figure 17: *Effect of too many bins. The majority of pixels fall into two bins.*

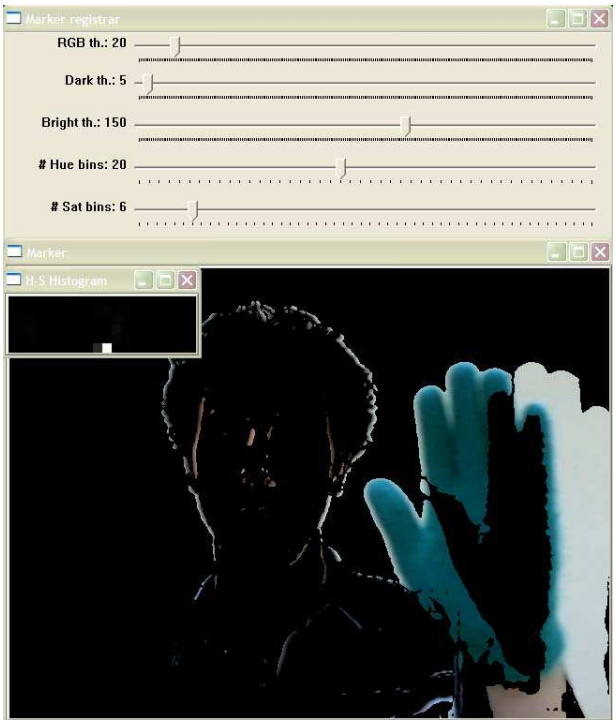


Figure 18: *Example of a good setting for number of bins.*

linear dynamic system. Therefore, a Kalman filter is directly applicable. This filtering method is used in the gesture recognition system recognition module is shown to significantly enhance the accuracy of the results [11, 12, 13].

3.7.3. Feature Extraction and Gesture Recognition

The current design of the gesture recognition module is using Hu moments and the hand image angle as features. Shape features are only necessary when the defined gestures cannot be distinguished otherwise. If each gesture has a unique trajectory, discrete HMMs can be used for the recognition phase. The gestures chosen for the BU Smart Room require shape information to be present. Therefore, discrete HMMs cannot be used unless hand shapes are quantized, which is undesirable. Therefore, our current design produces two data streams; one corresponding to the motion vector of the hand for each frame, and one consisting of the Hu moments and the angles.

In our design, the produced data streams are to be used in a continuous HMM or IOHMM based recognition framework. Both frameworks are implemented and tested for different databases, but are not ported into Smartflow system yet. This is left as a future work. Once implemented, the module will be able to recognize the gestures in Table 1 automatically from continuous streams. The first column in Table 1 corresponds to the possible actions, and the second column corresponds to the objects that can be embedded in commands. Any combination of actions and objects is allowed, as long as actions precede the objects. Using this system, the users will be able to turn on–off or dim the lights, open–close the door gradually or entirely, command a program running on the computer or answer the phone remotely by performing hand gestures.

3.8. User Interface of the BU Smart Room System

To monitor the activity and interactions in the smart room, we implemented a graphical user interface module using OpenGL

Action	Object
Turn On	Door
Turn Off	Light
Turn Up	Phone
Turn Down	Computer

Table 1: *Defined gestures for the BU Smart Room setting.*

[22] and GLUT libraries [10], and integrated it into the SmartFlow system. For visualization purposes, we were inspired by the Marauder's Map in the Harry Potter book series. The basic idea is that the map is covered with tiny ink dots accompanied by minuscule names, indicating every person's location in Hogwarts, the magic academy. In a similar manner, we have created an abstract 2D map of the room and represented people inside the room with color coded feet.

After identifying a person using recognition modules, his state is set "online" in the flow providing the UI module with the necessary data. On the members list of the interface, people who are present in the room are represented with higher opacity and absent ones are represented with decreased opacity. Currently, the members list is fixed. As an improvement to the module, we want to connect it to a database and create an editable room member list. Thus, a flexible UI will be provided by preserving the object-oriented structure.

Sound based localization module feeds the UI module with coordinates relative to the microphone arrays. The data are converted to screen coordinates and an approximate location of the person inside the room is marked on the map grid. As shown in Fig. 19, the map grid represents the tiles in the smart room in a smaller scale. The map elements include the server tables, cameras and simply the entrance of the room, which constitute the obstacles of the room. As a further improvement, we want to visualize specific activities and interactions between multiple people by integrating the audio domain with visual domain. Another visualization method can be implemented on an augmented reality interface. Such an innovative interface can be implemented by annotating the people with virtual text labels containing their name on the real-time video capture [7]. Moreover, interactions can be represented by registering computer graphics objects to the video capture. Thus, the cameras in the room will provide more information than the appearance to an ordinary viewer of the smart room.

4. CONCLUSIONS AND FUTURE WORK

The aim of this project is to monitor a room for the purposes of analyzing the interactions and identities of a small set of individuals. We worked with different modalities from multiple uncalibrated sensors to observe a single environment and generate multimodal data streams. These streams are processed with the help of a generic client-server middleware called SmartFlow and signal processing modules.

Modules for visual motion detection, visual face tracking, visual face identification, visual opportunistic sensing, gesture recognition, audio-based localization, and audio-based identification were implemented and integrated under SmartFlow to work in coordination across two different platforms (Linux and Windows).

A graphical interface was implemented to visualize the functionalities which was inspired by the Marauder's Map in the Harry Potter series. In its current state, it shows an identified person and his or her position in the room, based on audio-visual data.

Figure 19: *First version of the GUI inspired by the Marauder's Map.*

Although SmartFlow is a powerful tool to manage different datastreams, it has its problems, primarily due overload of hardware and limitations of network bandwidth. However in the philosophy of our project, we have attempted to solve these problems not by increasing the available computational resources, or by employing better sensors, but by intelligent allocation of the available resources to different datastreams. This was mainly achieved by allowing different framerates for video. A second objective was to tune this framerate according to various quality measures that indicate the relevance of a given datastream with respect to the identification task. This part was implemented for audio, where the quality is based on the energy of the signal.

The first stage of the project concerned setting up the architecture, collecting the datasets, implementing different modules, and experimenting with them. Since we were working with sensors of much lower cost and quality than related approaches in the literature (see for instance the results of the International Evaluation Workshop on Classification of Events, Activities, and Relationships (CLEAR 2006) [30]), we have identified two subgoals at the start of the project: opportunistic sensing and multimodal fusion. The former aims at identifying persons in the absence of robust (face) recognition. Our opportunistic sensing was implemented to build a statistical model of the pixels representing a person entering the room on the fly, for each of the cameras. Our second aim (fusion of audio-visual information) will be the starting point for future research between the participants after the project. In particular, we have identified the following future research questions:

- In our project we have managed to localize persons into the smart room using audio as well as video streams. In UPC, research has been done on event recognition using audio signals. Different modules have been written for the recognition of speech-non speech identification, hand clapping, footstep recognition, keyboard typing etc. Some of these modules were tested in the BU room. We like to work on visual recognition algorithms that can be used to make this recognition process more robust.
- Another possible research direction is implementing a biologically motivated focus of attention in order to limit the amount of (video) data processing. By allowing quality measures in the recognition process we can ignore some datastreams and/or call upon another modality or

sensor to support the recognition process. In the current version of our system, audio and video streams are used continuously to process the data. In many cases this means processing irrelevant data and overloading the system.

- For better identification and tracking results we will experiment with the fusion of the different modalities, using different methodologies (score level, decision level etc.), to improve robustness. Also, modalities can be used to support other modalities; e.g. the head pose module can improve the results for the face recognition. We will experiment with simple, limited-context discriminative patterns. A typical example is the hair colour to identify persons in the cases that the face is seen from the back, and the face recognition module can not be used.
- Finally, we want to improve our visualization modules. Determining the orientation and focus of attention is of interest in several settings (e.g. museums). If the camera projection matrices could be retrieved accurately, localization data could be represented in a 3D virtual environment in real-time. Diverse types of interaction modules can be integrated to the system. Orientation of people can be tracked by letting people wear objects with inertial orientation sensors or digital compasses, or by tracking the walk trajectory and predicting the next move.

5. ACKNOWLEDGEMENTS

This report, as well as the source code for the software developed during the project, is available online from the eNTERFACE'07 web site: www.cmpe.boun.edu.tr/enterface07.

6. REFERENCES

[1] ALIZE Toolkit, Université d'Avignon: http://www.lia.univ-avignon.fr/heberges/ALIZE/ 77

[2] Bengio, Y., P. Frasconi, "Input/Output HMMs for sequence processing", *IEEE Trans. Neural Networks* vol. 7(5), pp. 1231–1249, 1996. 79

[3] Bimbot, F., J.-F. Bonastre, C. Fredouille, G. Gravier, I. Magrin-Chagnolleau, S. Meignier, T. Merlin, J. Ortega-García, D. Petrovska-Delacrétaz, D.A. Reynolds, "A Tutorial of Text-Independent Speaker Verification," *EURASIP Journal on Applied Signal Processing*, vol.4, pp.430-451, 2004. 78

[4] Cook, D.J., S.K. Das, (eds.) *Smart Environments: Technologies, Protocols, and Applications*, John Wiley & Sons, Inc., 2005. 71

[5] Davis, S. B. and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences", *IEEE Trans. ASSP*, No. 28, pp. 357-366, 1980. 78

[6] DiBiase, J., H. Silverman, M. Brandstein, "Robust localization in reverberant rooms," in M. Brandstein and D. Ward (eds.) *Microphone Arrays*, Springer Verlag, 2001. 78

[7] Feiner, S., MacIntyre, B., Höllerer, T., and Webster, A., "A touring machine: Prototyping 3D mobile augmented reality systems for exploring the urban environment," in *Proc. ISWC'97*. Cambridge, MA, USA, pp.74-81, 1997. 83

[8] Fung, G., and O. Mangasarian, "Proximal Support Vector Machine Classifiers, "*Proc. KDDM*, pp. 77-86, 2001. 76

[9] Ghahramani, Z., and G.E. Hinton, "The EM algorithm for mixtures of factor analyzers," Technical Report CRG-TR-96-1, University of Toronto, (revised), 1997. 75

[10] GLUT: http://www.opengl.org/resources/libraries/glut/spec3/spec3.html 83

[11] Keskin, C., A.N. Erkan and L. Akarun, "Real time hand tracking and 3D gesture recognition for interactive interfaces using HMM", *Proc. ICANN/ICONIP*, İstanbul, Turkey, 2003. 79, 82

[12] Keskin, C., O. Aran and L. Akarun, "Real time gestural interface for generic applications", *European Signal Processing Conference, EUSIPCO Demonstration Session*, Antalya, Turkey, 2005. 79, 82

[13] Keskin, C., K. Balci, O. Aran, B. Sankur and L. Akarun, "A multimodal 3D healthcare communication system", *3DTV Conference*, Kos, Greece, 2007. 79, 82

[14] Kirby, M., L.Sirovich, "Application of the Karhunen-Loeve procedure for the characterization of human faces," *IEEE Trans. PAMI*, vol 12, no. 1, pp. 103-108, Jan. 1990. 75

[15] Lee, H. and J.H Kim, "An HMM–Based threshold model approach for gesture recognition", *IEEE Trans. PAMI*, vol. 21, no. 10, pp. 961-973, 1999. 79

[16] Luque, J., R. Morros, A. Garde, J. Anguita, M. Farrus, D. Macho, F. Marqués, C. Martínez, V. Vilaplana, J. Hernando, "Audio, Video and Multimodal Person Identification in a Smart Room," CLEAR 2006, LNCS, 2007. 74

[17] Luque, J. and J. Hernando, " Robust Speaker Identification for Meetings: UPC CLEAR-07 Meeting Room Evaluation System," to appear in *CLEAR 2007, LNCS*, 2007. 78

[18] Messer, K., J. Matas, J.V. Kittler, J. Luettin and G. Maitre, "XM2VTSDB: The extended M2VTS Database," *Proc. AVBPA*, 1999. 75

[19] Nadeu C., D. Macho, and J. Hernando, "Time and Frequency Filtering of Filter-Bank Energies for Robust Speech Recognition", *Speech Communication*, 34, pp. 93-114, 2001. 78

[20] NIST SmartFlow system: http://www.nist.gov/smartspace/nsfs.html 73

[21] Omologo, M., P. Svaizer, "Use of the crosspower-spectrum phase in acoustic event location," *IEEE Trans. on Speech and Audio Processing*, 1997. 78

[22] OpenGL: http://www.opengl.org/. 83

[23] *Proceedings of the 1998 DARPA/NIST Smart Spaces Workshop*, National Institute of Standards and Technology, pp.3-1 to 3-14, July 1998. 73

[24] Rabiner, L.R. "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition", *Proc. IEEE*, vol.17, no.2, Feb. 1989. 78

[25] Rabiner, L.R. and B. Juang, "An introduction to hidden Markov models", *IEEE ASSP Magazine*, pp. 4–16, Jan. 1996. 79

[26] Salah, A.A., E. Alpaydın, "Incremental Mixtures of Factor Analyzers," in *Proc. ICPR*, vol.1, pp. 276-279, 2004. 75

[27] Schölkopf, B., A. Smola, *Learning with Kernels*, MIT Press, Cambridge, MA, 2002. 76

[28] Stanford, V., J. Garofolo, O. Galibert, M. Michel, and C. Laprun, "The NIST Smart Space and Meeting Room Projects: Signals, Acquisition, Annotation, and Metrics," *Proc. ICCASP*, vol.4, pp.736-739, 2003. 73

[29] Stauffer, C., and Grimson,W., "Adaptive background mixture models for real-time tracking", in *Proc. IEEE CVPR*, 1999. 73

[30] Stiefelhagen, R., and Garofolo, J., (eds.) *Multimodal Technologies for Perception of Humans*, LNCS 4122, Springer Verlag, 2007. 83

[31] Tangelder, J.W.H., Ben A.M. Schouten, Stefan Bonchev, "A Multi-Sensor Architecture for Human-Centered Smart Environments," *Proceedings CAID&CD Conference*, 2005. 71

[32] Tangelder, J.W.H., Ben A.M. Schouten, "Sparse face representations for face recognition in smart environments," *Proc. ICPR*, 2006. 75

[33] Temko, A., D. Macho, C. Nadeu, "Enhanced SVM Training for Robust Speech Activity Detection," *Proc. ICCASP*, 2007. 73, 76

[34] Vilaplana, V., Martínez, C., Cruz, J., Marques, F., "Face recognition using groups of images in smart room scenarios," *Proc. ICIP*, 2006. 74, 75

[35] Viola, P., and M. Jones, "Rapid Object Detection Using a Boosted Cascade of Simple Features," in *Proc. IEEE CVPR*, vol.1, pp.511-518, 2001. 74

[36] Vogler, C. and D. Metaxas, "Adapting hidden Markov models for ASL recognition by using 3D computer vision methods", *In Conference on Systems, Man and Cybernetics*, pp.156–161, 1997. 79

[37] Vogler, C. and D. Metaxas, "ASL recognition based on a coupling between HMMs and 3D motion analysis", *Proc. ICCV*, 1998. 79

7. BIOGRAPHIES

Ramon Morros is an Associate Professor at the Technical University of Catalonia (UPC), Barcelona, Spain, where he teaches in the areas of digital communications, signal processing, image processing and audiovisual networks. He received a degree in Physics from the University of Barcelona (UB) in 1989 and the Ph.D. degree from the UPC in 2004. His research interests are on the fileds of image sequence segmentation and coding, sequence analysis, and visual person identification. He has participated in many public and private research projects in these areas and currently is involved in the CHIL project in the fields of video and multimodal person identification. He is author of conference and journal papers and holds one patent. He has also worked on several companies, either as a staff member or as a free-lance consultant.
Email: morros@gps.tsc.upc.edu

Albert Ali Salah worked as a research assistant in the Perceptual Intelligence Laboratory of Boğaziçi University, where he was part of the team working on machine learning, face recognition and human-computer interaction. After receiving his PhD in Computer Engineering under the supervision of Prof. Lale Akarun in 2007, he joined the Signals and Images research group at Centrum voor Wiskunde en Informatica (CWI) in Amsterdam. With his work on facial feature localization, he received the inaugural EBF European Biometrics Research Award in 2006. His current research interests include sensor networks, ambient intelligence, and spatio-temporal pattern recognition.
Email: a.a.salah@cwi.nl

Ben A.M. Schouten was born in 1953 and graduated from the Rietveld Art Academy in 1983. For many years he worked as a computer artist and his work has been exhibited in many places in the Netherlands and abroad. He received his master's degree in mathematics, specializing in chaos theory, in August 1995, founded Desk.nl in 1996, and he received his PhD in 2001, on content based image retrieval schemes and interfaces that express in an adaptive and intuitive way image similarities according to human perception. His thesis was awarded a bronze medal for Information Design at the New York Arts Festival. Currently he is a researcher at the Centre for Mathematics and Computer Science (CWI), Amsterdam, The Netherlands. His main research interests are human computer interfacing, biometrics, smart environments, image and video understanding and data visualization. Ben Schouten teaches interface & design at the School of Arts and Technology (USAT) in the Netherlands and is the director of the European Biometrics Forum.
Email: bens@cwi.nl

Carlos Segura Perales received the BS and MS degrees in Telecommunication Engineering at the Technical University of Catalonia, and the MS degree at the Technical University of Berlin in 2003. He is currently a research fellow at the Technical University of Catalonia, working on his PhD dissertation entitled "Speaker Localization in Multimodal Smart Environments".
Email: csegura@gps.tsc.upc.edu

Jordi Luque Serrano is a PhD student at Technical University of Catalonia, Spain. He received his Engineering of Telecommunication degree from that university in 2005, and he is finishing his PhD Thesis. His research interests are related to the field of speech processing. Specifically, he has worked on the speaker identification and verification problems, diarization of meetings and automatic speech recognition.
Email: luque@gps.tsc.upc.edu

Onkar Ambekar received his BS in Instrumentation from B.A.M. University, India, in 2000, and his MS on Sensor System Technology from the University of Applied Sciences in Karlsruhe, Germany in 2004. He has worked in Germany and Austria on object detection and tracking for vehicle guidance, and holds a patent in this area. He is currently a PhD student at the CWI institute in Amsterdam.
Email: ambekar@cwi.nl

Ceren Kayalar received her BS degree in computer engineering from Dokuz Eylül University in 2004, and her MS degree from Sabancı University with a thesis entitled "Natural Interaction Framework for Pedestrian Navigation Systems on Mobile Devices". Her research interests are mobile devices and applications, virtual reality, computer graphics, augmented reality, user interfaces, context-aware computing, and human-computer interaction. Her PhD work on an outdoor augmented reality system for cultural heritage at the Sabancı University is sponsored by TUBITAK.
Email: ckayalar@su.sabanciuniv.edu

Cem Keskin was born in 1979. He received a double-major BS degree in physics and computer engineering from Boğaziçi University, in 2003. He received the MS degree in 2006, with a thesis entitled "Vision Based Real-Time Continuous 3D Hand Gesture Recognition Interface For Generic Applications Based On Input-Output Hidden Markov Models", under the supervision of Prof. Lale Akarun. Currently he pursues a PhD degree at the same institution, where he also works as a research assistant.
Email: keskinc@boun.edu.tr

Lale Akarun received the BS and MS degrees in Electrical Engineering from Boğaziçi University, İstanbul, in 1984 and 1986, respectively. She obtained her PhD from Polytechnic University, New York in 1992. Since 1993, she has been working as a faculty member at Boğaziçi University. She became a professor of Computer Engineering in 2001. Her research areas are face recognition, modeling and animation of human activity and gesture analysis. She has worked on the organization committees of IEEE NSIP99, EUSIPCO 2005, and eNTERFACE 2007. She is a senior member of the IEEE.
Email: akarun@boun.edu.tr

3D FACE RECOGNITION PERFORMANCE UNDER ADVERSARIAL CONDITIONS

Arman Savran [1], Oya Çeliktutan [1], Aydın Akyol [2], Jana Trojanová [3], Hamdi Dibeklioğlu [4], Semih Esenlik [1], Nesli Bozkurt [5], Cem Demirkır [1], Erdem Akagündüz [5], Kerem Çalışkan [6], Neşe Alyüz [4], Bülent Sankur [1], İlkay Ulusoy [5], Lale Akarun [4], Tevfik Metin Sezgin [7]

[1] Boğaziçi University, Department of Electrical and Electronic Engineering, İstanbul, Turkey
[2] İstanbul Technical University, Department of Computer Engineering, Turkey
[3] Department of Cybernetics, University of West Bohemia in Pilsen, Czech Republic
[4] Boğaziçi University, Department of Computer Engineering, Turkey
[5] Middle Eastern Technical University, Dept. of Electrical and Electronics Engineering, Turkey
[6] Middle Eastern Technical University, Informatics Institute, Turkey
[7] University of Cambridge, Computer Laboratory, UK

ABSTRACT

We address the question of 3D face recognition and expression understanding under adverse conditions like illumination, pose, and accessories. We therefore conduct a campaign to build a 3D face database including systematic variation of poses, different types of occlusions, and a rich set of expressions. The expressions consist of a judiciously selected subset of Action Units as well as the six basic emotions. The database is designed to enable various research paths from face recognition to facial landmarking and to expression estimation. Preliminary results are presented on the outcome of three different landmarking methods as well as one registration method. As expected, observed non-neutral and non-frontal faces demand new robust algorithms to achieve an acceptable performance.

KEYWORDS

3D face database – Facial expressions – Facial landmarking – Face recognition

1. INTRODUCTION

The history of automated face recognition dates back to the early 1960s. The overwhelming majority of these techniques are based on 2D face data. The early approaches were focused on the geometry of key points (eyes, mouth and nose) and relations between them (length, angles). In the 1990s, the principal component analysis (PCA) based eigenface algorithm was introduced and it became a standard reference in face recognition. Despite the plethora of 2D algorithms, there is not yet a fully reliable identification and verification method that can operate under adverse conditions of illumination, pose, accessories and expression. Two solutions to this impasse consist in i) Exploring new modalities, such as 3D facial data; ii) The use of multibiometrics, that is, the employment of more than one biometric modality and their judicious fusion.

Recently face recognizers using 3D facial data have gained popularity due to their lighting and viewpoint independence. This has also been enabled by the wider availability of 3D range scanners. The 3D face processing can be envisioned as a single modality biometric approach in lieu of the 2D version or in a complementary mode in a multi-biometric scheme. Another goal application of 3D facial data is the understanding of facial expressions in an affective human-computer interface.

Based on the ideas above, two main goals are addressed in this project:

1.1. 3D Recognition of Non-cooperative Persons

Most of the existing methods for facial feature detection and person recognition assume frontal and neutral views only. In these studies data are therefore collected from cooperative subjects, who expose their faces in a still position in front of the scanner, frontal poses, and avoid extreme expressions and any occluding material. This may be uncomfortable to subjects. In fact, the second generation techniques vie for improved quality of life via biometry, hence enrollment and updating should proceed in ways that do not encumber the subject. On the other extreme, a subject, aware of person identification cameras, may try to eschew being recognized by posing awkwardly and worse still, by resorting to occlusions via dangling hair, eyeglasses, facial hair etc. Also, the 3D data may have different translation, rotation or scaling due to the controlled environmental parameters such as the acquisition setup and device properties. Both the natural, uncontrolled behaviour of subjects and the mimics and acrobatics of the eschewer seriously damage the performance of both 2D and 3D recognition algorithms. Furthermore, we believe that 3D capture of face data can mitigate significantly most of these effects. Using a database built specifically for this purpose, we test the performance of 3D face identification algorithms as well as those of automatic landmarking and registration under non-frontal poses, in presence of facial expression, gestures and occlusions.

1.2. Facial expression understanding

Understanding of facial expressions has wide implications ranging from psychological analysis to affective man-machine interfaces. Once the expression is recognized, it can also be neutralized for improved person recognition. Automatic recognition of facial expressions is challenging since the geometry of the face can change rapidly as a result of facial muscle contractions. Expressions are often accompanied by purposeful or involuntary pose variations of the subject's face. 3D representation of the human face contains more information than its 2D appearance. We conjecture that not all shape changes on the facial surfaces due to expressions are reflected in the 2D images. Certain surfaces deformations, bulges and creases, especially in the smooth areas of the face, are hard to track in 2D while they are apparent on 3D data.

Among the facial signal processing problems, we want to address to the above stated goals by providing solutions to the following problems:

1.3. 3D face registration

3D shapes need to be aligned to each other and should be brought into a common coordinate frame before any comparative analysis can be made. The 3D face registration is the process of defining a transformation that will closely align two faces. First of all, the permissible transformations should be set. Rigid transformations allow only for translation, rotation or scaling. Non-rigid transformations go one step further and allow patchwise deformation of facial surfaces within the constraints of estimated landmark points. Secondly, a similarity measure should be determined that favors successful recognition. Different similarity measures are possible for registration techniques such as the point-to-point or point-to-surface distances.

1.4. 3D face landmarking

Facial landmarks are essential for such tasks as face registration, expression understanding and any related processing. In general, registration process is guided by a set of fiducial points called landmarks [1] that are used to define the transform between two surfaces. The registration techniques that are examined in the scope of this project make use of a set of landmark points that are labeled for each face surface to be aligned. Preliminary studies of landmarking of uncontrolled facial expressions [2] indicate that it remains still as an open problem.

1.5. 3D facial expression and face recognition database construction

There are very few publicly available databases of annotated 3D facial expressions. Wang et. al [3] studied recognition of expressions by extracting primitive surface features that describe the shape. They use their database dedicated to expressions BU-3DFE [4]. It includes only emotional expressions (happiness, surprise, fear, sadness, anger, disgust) with four intensity levels from 100 people.

In this project, we aim to study a more comprehensive set of expressions by covering many expressions based on Facial Action Coding System (FACS) [5] in addition to the six basic emotional expressions. There are 44 Action Units (AUs) defined in FACS. AUs are assumed to be building blocks of expressions, and thus they can give broad basis for facial expressions. Also, since each of them is related with activation of distinct set of muscles, they can be assessed quite objectively. We also include various poses and occlusion conditions. We have collected data from 81 subjects, preprocessed and manually landmarked the data. We have defined the metadata standard. We have also performed basic registration and classification tests. This multi-expression and multi-pose database, which will eventually be made public, will be instrumental in advancing the expression analysis research and develop algorithms for 3D face recognition under adverse conditions.

The rest of the report is as follows: Section 2 describes the database content, and the database structure is detailed in Section 3. Section 4 gives the necessary details of the data acquisition setup. Landmarking issues are dealt in two separate sections: Section 5 and 6, respectively, presents manual landmarking and automatic landmarking techniques. The 3D face registration method applied to our database is given in Section 7. Finally, Section 8 shows the performance results of registration, automatic landmarking and recognition. Conclusions are drawn in Section 9.

2. DATABASE CONTENT

In this project, on the one hand we model attempts to invalidate 3D face recognition and any other effort to mislead the system or to induce a fake character. To this effect, we capture 3D face data imitating difficult surveillance conditions and non-cooperating subjects, trying various realistic but effective occlusions and poses. On the other hand, we collect 3D facial data corresponding to various action units and to various emotional expressions.

The database consists of 81 subjects in various poses, expressions and occlusion conditions. The images are captured using Inspeck Mega Capturor II [6], and where each scan is manually labeled for facial landmark points such as nose tip, inner eye corner, etc. Our database has two versions:

- Database1 is designed for both expression understanding and face recognition. Here there are 47 people with 53 different face scans per subject. Each scan is intended to cover one pose and/or one expression type, and most of the subjects have only one neutral face, though some of them have two. Totally there are 34 expressions, 13 poses, four occlusions and one or two neutral faces. In addition, Database1 also incorporates 30 professional actors/actresses out of 47, which hopefully provide more realistic or at least more pronounced expressions.
- Database2 includes 34 subjects with only 10 expressions, 13 poses, four occlusions and four neutral faces and four neutral faces, thus resulting in a total of 31 scans per subject.

The majority of the subjects are aged between 25 and 35. There are 51 men and 30 women in total, and most of the subjects are Caucasian.

This database can be used for automatic facial landmark detection, face detection, face pose estimation, registration, recognition and facial expression recognition purposes. Hence we also labeled facial fiducial points manually, and provide this information with the database.

In the following subsections, the collected facial expressions, head poses and occlusions are explained.

2.1. Facial Expressions

We have considered two types of expressions. In the first set, the expressions are based on AU of the FACS [5]. However, within this project a subset of them, which are more common and easier to enact, is considered. The selected action units are grouped into 19 lower face AUs, five upper face AUs and three AU combinations. It is important to note that, for some subjects properly producing some AUs may not be possible, because they are not able to activate related muscles or they do not know how to control them. Therefore, in the database some expressions are not available for some subjects. Also, unless all the acquired AUs are validated by trained AU coders, the captured AUs can not be verified.

In the second set, facial expressions corresponding to certain emotional expressions are collected. We have considered the following emotions: happiness, surprise, fear, sadness, anger and disgust. It is stated that these expressions are universal among human races [7].

During acquisition of each action unit, subjects were given explications about these expressions and they were given feedback if they did not enact correctly. Also to facilitate the instructions, a video clip showing the correct facial motion for the corresponding action unit is displayed on the monitor [8, 9]. However, in the case of emotional expressions, there were no video

1. Lower Face Action Units
   - Lower Lip Depressor - AU16
   - Lips Part - AU25
   - Jaw Drop - AU26
   - Mouth Stretch - AU27 (*)
   - Lip Corner Puller - AU12 (*)
   - Left Lip Corner Puller - AU12L
   - Right Lip Corner Puller - AU12R
   - Low Intensity Lip Corner Puller - AU12LW
   - Dimpler - AU14
   - Lip Stretcher - AU20
   - Lip Corner Depressor - AU15
   - Chin Raiser - AU17
   - Lip Funneler - AU22
   - Lip Puckerer - AU18
   - Lip Tightener - AU23
   - Lip Presser - AU24
   - Lip Suck - AU28 (*)
   - Upper Lip Raiser - AU10
   - Nose Wrinkler - AU9 (*)
   - Cheek Puff - AU34 (*)
2. Upper Face Action Units
   - Outer Brow Raiser - AU2 (*)
   - Brow Lowerer - AU4 (*)
   - Inner Brow Raiser - AU1
   - Squint - AU44
   - Eyes Closed - AU43 (*)
3. Some Action Unit Combinations
   - Jaw Drop (26) + Low Intensity Lip Corner Puller
   - Lip Funneler (22) + Lips Part (25) (*)
   - Lip Corner Puller (12) + Lip Corner Depressor (15)
4. Emotions
   - Happiness (*)
   - Surprise
   - Fear
   - Sadness
   - Anger
   - Disgust

The expressions marked with (*) are available in both Database1 and Database2, but the others are only found in Database1.

Table 1: *Lower face action units.*

or photo guidelines so that subjects tried to improvise. Only if they were able to enact, they were told to mimic the expression in a recorded video. Moreover, a mirror is placed in front of the subjects in order to let them check themselves.

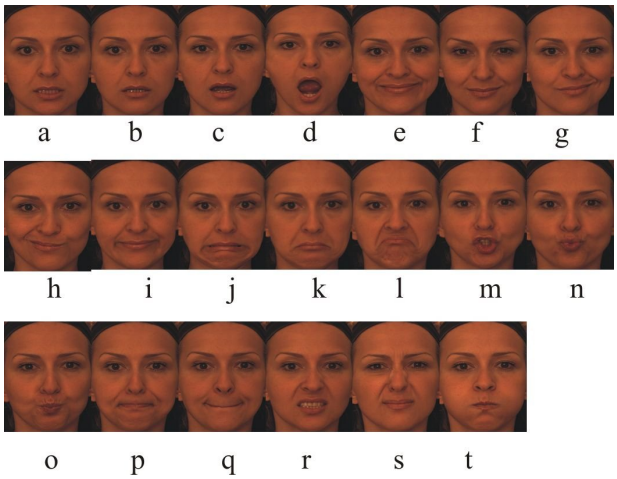


Figure 1: *Lower Face Action Units: lower lip depressor (a), lips part (b), jaw drop (c), mouth stretch (d), lip corner puller (e), low intensity lower lip depressor (f), left lip corner puller (g), right lip corner puller (h), dimpler (i), lip stretcher (j), lip corner depressor (k), chin raiser (l), lip funneler (m), lip puckerer (n), lip tightener (o), lip presser (p), lip suck (q), upper lip raiser (r), nose wrinkler (s), cheek puff (t).*

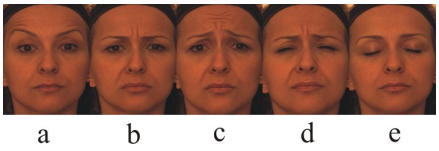


Figure 2: *Upper Face Action Units: outer brow raiser (a), brow lowerer (b), inner brow raiser (c), squint (d), eyes closed (e).*

Figure 3: *Some Action Unit Combinations: lip corner puller + lip corner depressor (a), jaw drop + lip corner puller (b), lip funneler + lips part (c).*

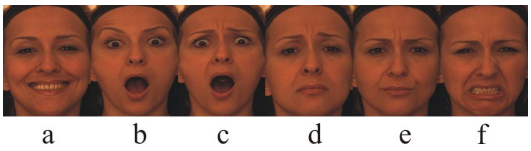


Figure 4: *Emotional expressions: happiness (a), surprise (b), fear (c), sadness (d), angry (e), disgust (f).*

The total of 35 expressions consist of 19 AUs belonging to lower part of the face (Fig. 1), five AUs from the upper face (Fig. 2), three AU combinations (Fig. 3) and six emotional expressions (Fig. 4) as listed in Table 1. In these tables, the expressions marked with (*) are available in both Database1 and Database2, but the others are only to be found in Database1.

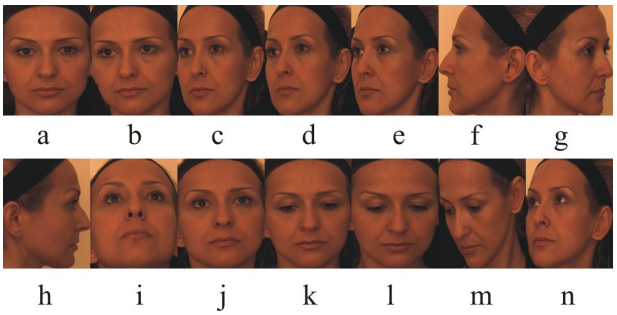


Figure 5: *Poses: neutral pose (a); yaw rotations of +10° (b), +20° (c), +30° (d), +45° (e), +90° (f), -45° (g) and -90° (h) respectively; pitch rotations of strong upwards (i), slight upwards (j), slight downwards (k), strong downwards (l); bottom-right (m) and upper right (n).*

2.2. Head Poses

Various poses of head are acquired for each subject (Fig. 5). There are three types of head which are seven angles of yaw, four angles of pitch, and two cross rotations which incorporate both yaw and pitch. For the yaw rotations, subjects align themselves by rotating the chair on which they sit to align with straps placed on the floor corresponding to various angles. For pitch and cross rotations, we requested the subjects to look at marks placed on the walls by turning their heads only (i.e., no eye rotation). Thus, we can obtain a coarse approximation of rotation angles. The head poses are listed Table 2.

1. Yaw Rotations
   - +10°
   - +20°
   - +30°
   - +45°
   - +90°
   - -45°
   - -90°
2. Pitch Rotations
   - Strong upwards
   - Slight upwards
   - Slight downwards
   - Strong upwards
3. Cross Rotations
   - Yaw and pitch 1 (approximately 20° pitch and 45° yaw)
   - Yaw and pitch 2 (approximately -20° pitch and 45° yaw)

Table 2: *Head poses*

2.3. Occlusions

This database also contains several types of occlusions that are listed in Table 3 and shown in Fig. 6.

For the occlusion of eyes and mouth, subjects choose a natural pose of themselves; for example, as if they were cleaning their eyes or they were surprised by putting their hands over their mouth. Second, for the eyeglasses occlusion, we made so that subjects used different eyeglasses from a pool. Finally, if subjects' hairs were long enough, we have also scanned their faces with hair partly occluding their face.

Figure 6: *Occlusions: eye occlusion (a), mouth occlusion (b), eye glasses (c), hair (d).*

- Occlusion of eye with hand - as natural as possible
- Occlusion of mouth with hand - as natural as possible
- Eye glasses (not sunglasses, normal eyeglasses)
- Hair

Table 3: *Occlusions*

3. THE STRUCTURE OF THE DATABASE

In many image retrieval database applications two approaches are widely used. One is storing any image-like data onto hard disk and the rest of the data to an RDBMS (Relational Database Management System) while the second approach is storing image and its metadata directly on the hard disk. While the first one provides easy querying, it has the disadvantage of using and managing third-party RDBMS software. Also not every RDBMS may work properly in cross-platform systems. In addition it necessitates knowledge of using SQL (Structured Query Language). We suspect that not all researchers are well versed in SQL. For these reasons we decided to use the second alternative, that is, storage of images and their metadata on the harddisk. For this purpose, data storage and mapping structures are designed and a traditional database design used in RDBMSs is considered. In this design we make use of primary key and weak referencing. A primitive referential integrity is used inside the software for preventing redundancy and making relations between tree-like structures. The first version of the design is given in Appendix 12.1.

One can map the entire database and the storage structure based on the above design. We choose to use XML structure in data storage because in addition to having a human readable format XML is also commonly used in non RDBMS applications in the recent years. A record is a row in a table, each column in a record shows us an attribute of the record. Every table defines its attributes and lists its data after declaration of attributes. Content of the subject table exported from Subject.xml is given as an example in Appendix 12.2.

Populating the database manually is always very hard and thus a problematic application. One can easily commit errors during acquisition and processing stages. In order to overcome this problem we decided to write a program using our database structure which helps the acquisition application. Class diagram of the program shows that it is a general time-saving stratagem. Initial class designs involve inheritance of the form structures for easy usage and also involve usage of basic anti-patterns like Singleton for assuring that only single instances of Business Classes exist inside the program for managing data sessions. A schema of the class diagram is found in Fig. 7.

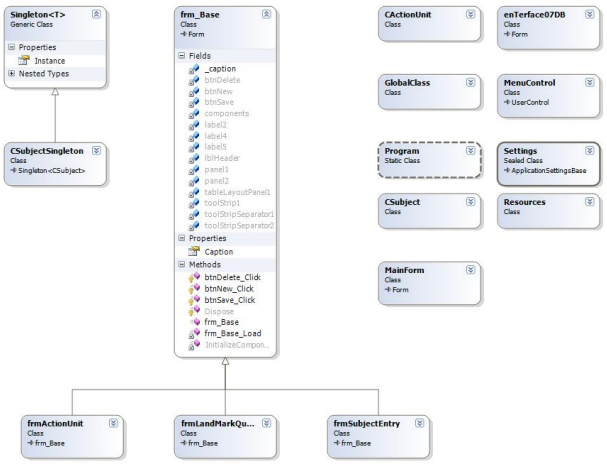


Figure 7: *A schema of the class diagram. Business classes are used for constructing the objects like Subject, Session and Records. Since these objects are used as object construction architecture, each class has its persistency structure, letting it to be written to database in any change. This is a standard mechanism called "Reflection" in "Object Relational"*

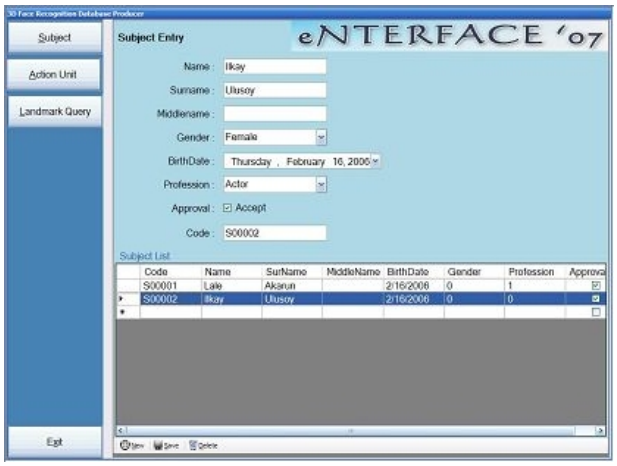


Figure 8: *User can insert new subjects to the database from this screen; editing or deleting previously saved subjects is also done from here.*

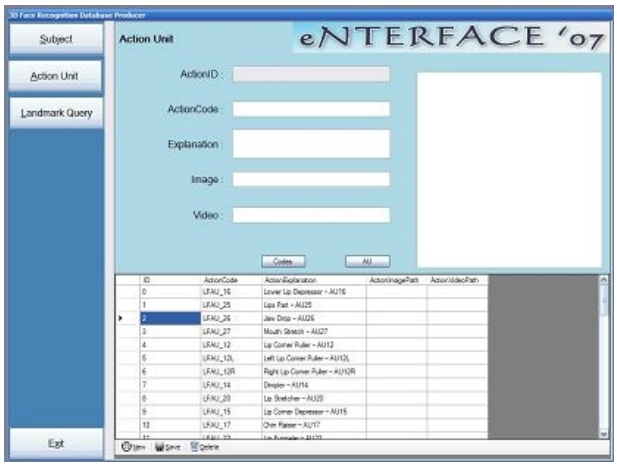


Figure 9: *User can insert the action units used during posing from this screen. Each action unit, its code and explanation has to be inserted. This also lets user to relate his/her action units to be used in their landmarking study to general structure. This part is vital for researchers for choosing their own database structure to be used in their landmarking algorithms.*

In order to write a quick code in limited time we decided to use .net 2.0 framework and C# as the programming language. We couldn't try the cross platform compatibility of the code using Mono in Linux, but we've implemented the code in cross platform manner. The LandMark querying part of the software is finished for the use of participants during Manual - Automatic Landmarking comparisons. For landmarking users have no chance but add the destination directory names as input to their software. After using our software user will be able to query the system by gender, occlusion or any structure they want and save the results to a document which can be input to their implementations. After having such an input document they can easily test their algorithms on the data they exactly want. While querying and finding the data directory paths is working properly, exporting this structure to an input file will be implemented with the acknowledgements coming from the users who want to test their landmarking algorithms on the database. Some screen shots of the software is shown in Fig. 8, 9 and 10.

As a conclusion, the proposed database structure does not need any third-party component or querying structure. Since it is XML based, it is human readable, and can also be used in many acquisition database structures owing to its easy and generic implementation. We hope the structure we've proposed will evolve and become a database structure for similar research applications.

4. DATA ACQUISITION

Facial data are acquired using Inspeck Mega Capturor II 3D, which is a commercial structured-light based 3D digitizer device [6]. It is able to capture a face in less than a second. Our acquisition setup is displayed in Fig. 11. Subjects are made to sit at a distance of about 1.5 meters away from the 3D digitizer. To obtain a homogenous lighting directed on the face we use a 1000W halogen lamp without any other ambient illumination. However, due to the strong lighting of this lamp and the device's projector, usually specular reflections occur on the face. This does not only affect the texture image of the face but can also cause noise in the 3D data. To prevent it, we apply a special powder to the subject's face which does not change the skin color. Moreover, during acquisition, each subject wears a band to keep his/her hair above the forehead. This also simplifies the segmentation of face before 3D reconstruction.

The on-board software in the scanner is used for acquisition and 3D model reconstruction. Inspeck's 3D digitizer is based on an optical principle, which combines Active Optical Triangulation and Phase Shifted Moir (Fig. 12). The surface of the object must reflect light back to the 3D digitizer. Hence transparent surfaces (like glass) can not be scanned. Certain fringe patterns, that are shifted versions of each others, are projected on the 3D surface of an object, and as a result these patterns are deformed according to object surface. In the Inspeck's system, a fringe pattern is in the form of straight vertical lines. All the 3D information is extracted from these deformed patterns. During acquisition we employed usually four fringe patterns. However, sometimes we preferred to use three to diminish the effect of motion for some subjects.

Once the raw data are acquired, three stages of processing (called PreProcessing, Processing and PostProcessing in the de-

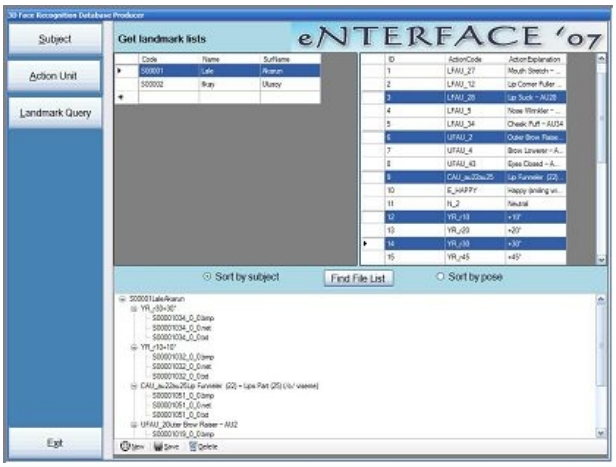


Figure 10: *Querying of the image data of manual landmarks and sorting due to subject or pose. User can query the data paths related to subject and their poses to be used in landmarking. Software user can select the subjects and the action units he/she wants to relate during landmarking. This query returns the user paths of the data they want to use.*

Figure 11: *Acquisition setup. At the left 3D Digitizer, which is mounted on a tripod, is seen. A mirror, an LCD display and a 1000W halogen lamp are placed below, above and behind of it respectively. A seat is positioned about 1.5 meters away from the 3D Digitizer. On the floor, just below the seat, yellow straps are placed according to rotation angles.*

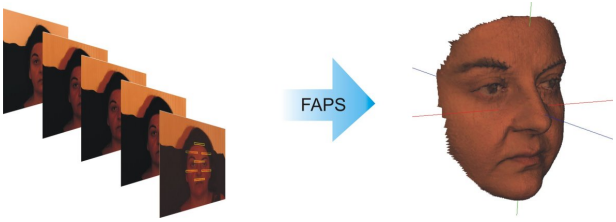


Figure 12: *3D model reconstruction. On the left are the raw data containing texture and 4 frames with shifted fringe pattern. The last picture shows parallax points. After processing by Inspeck Faps software we obtain the 3D model.*

vice's jargon) are carried out to obtain the 3D information (Fig. 13). In the PreProcessing stage, first the phase function is calculated (top right image in Fig. 13). Consequently the depth information is measured by using some reference points, called parallax points (top left image in Fig. 13). Most of the time we allowed the Inspeck software to detect these parallax points automatically; however, we reverted to manual marking in case of doubt. Next, an Interest Area is selected for the face image. We set this zone manually by creating one or more polygons that define it (top middle image in Fig. 13). A clearly defined Interest Area is very helpful to remove the background clutter. Thus we can cancel out extra streaks, which come from the digitizing of the background area.

After segmenting the face, phase unwrapping operation is performed to obtain 3D data in the Processing stage. However, during phase unwrapping some discontinuity errors, which are erroneous depth levels associated with discontinuous surfaces, may occur. The verified discontinuities are fixed by sequential push or pull operations.

Finally in the Post-Processing stage, the so-far processed data, currently in machine units, is converted to a data set in geometric parameters (millimeters). To convert the data into milimeters we determine a reference from a set of candidate reference (parallax) points. Generally, the point which has the highest confidence value is selected and the geometric parameters of the other points in the image are determined accordingly. Also, in this stage we sometimes perform simple filtering operations to correct slight optical distortions. Whenever serious distortions occurred we preferred to re-capture the data.

The result is exported in the form of ".net" file, which is a specific file format for the data-processing software of Inspeck Inc. [6]. The output file stores the 3D coordinates of the vertices together with color information as a related separate texture output file in ".bmp" format.

Since there might be problems for some of the scans, FAPS preprocessing has been done simultaneously with data acquisition. If a problem is observed on the data then that scan is repeated on the spot. Some commonly occurring problems are as follows (Fig. 14):

- *Movement during acquisition:* The Inspeck hardware is adjusted to capture four pictures that later constitute the raw data for 3D face images and one more picture of texture image. During acquisition, the subject must stay still (better hold the breath), so that the pictures are consistent with each other. Otherwise the wavelike defects occur globally around the face. In this case the acquisition is repeated.
- *Hair occlusion:* Fringes projected on hair becomes undetectable, hence the corresponding parts on the 3D image is noisy. Hair occluded parts are included in the face segmentation, since avoiding them is not possible. However, with smoothing operation we can obtain less noisy data.
- *Hand occlusion and open mouth:* In the case of hand occlusion and open mouth, depth level changes sharply over the edges, causing a discontinuous fringe pattern. In the 3D image, hand and tongue are constructed at wrong depth levels.
- *Facial hair and eyes:* Beards, eyebrows and eyelashes cause small fluctuations. Also eyes can not be kept constant because of high light intensity, resulting in a spiky surface on the eyes. These problems are ignored, not additional filtering processes are applied.
- *Eyeglasses:* Fringe pattern can not be reflected back from transparent surfaces. Eyeglasses seem to deform the area they cover on the face. The resulting noisy surface is included in the face segmentation.

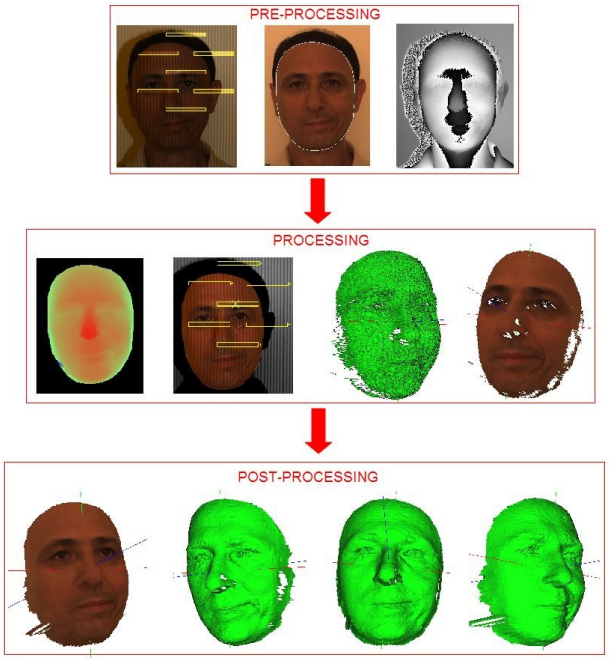


Figure 13: *Reconstruction by FAPS. Pre-processing (top row), processing (middle row) and post-processing (bottom row) stages. Top row: Parallax points, polygon to define the region of interest, phase image. Middle row: Depth image, parallax points, 3D shape, 3D texture. Bottom row: A scanned phase from different view points after post-processing such as hole filling.*

5. MANUAL DATA LANDMARKING

All of our face images are landmarked manually in order to be used in various training and testing methods as ground truth. On each face scan, 24 points are marked provided that they are visible in the given scan. The landmark points are listed in Table 4 and shown in Fig. 15.

1. Outer left eye brow	2. Middle of the left eye brow
3. Inner left eye brow	4. Inner right eye brow
5. Middle of the right eye brow	6. Outer right eye brow
7. Outer left eye corner	8. Inner left eye corner
9. Inner right eye corner	10. Outer right eye corner
11. Nose saddle left	12. Nose saddle right
13. Left nose peak	14. Nose tip
15. Right nose peak	16. Left mouth corner
17. Upper lip outer middle	18. Right mouth corner
19. Upper lip inner middle	20. Lower lip inner middle
21. Lower lip outer middle	22. Chin middle
23. Left ear lobe	24. Right ear lobe

Table 4: *Landmark points manually labeled on each scan.*

Landmarking software is implemented as a MATLAB GUI. This software is designed so that all landmark points can be marked guided by a reference template image for the landmarks in Fig. 15. This scheme helps us avoid errors which can occur when marking many landmark points over lots of images. The 3D coordinates corresponding to the 2D marked points on the texture image is extracted automatically using this software. Here we make the assumption that 2D image and 3D data are registered although they are not 100% registered.

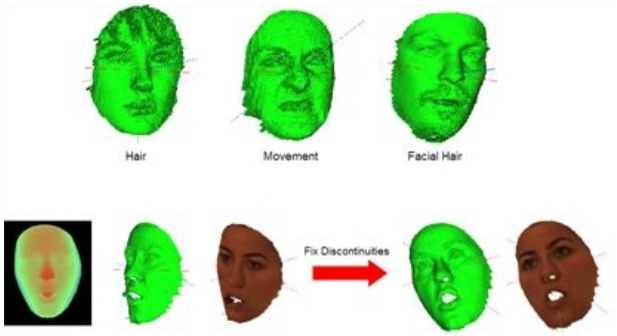


Figure 14: *Commonly occurring problems during image acquisition and face reconstruction. At top of the figure, noise due to hair, movement, and facial hair is seen on the face scans. At the bottom left, a mistake in the depth level of the tongue, and at the right, its correction is displayed.*

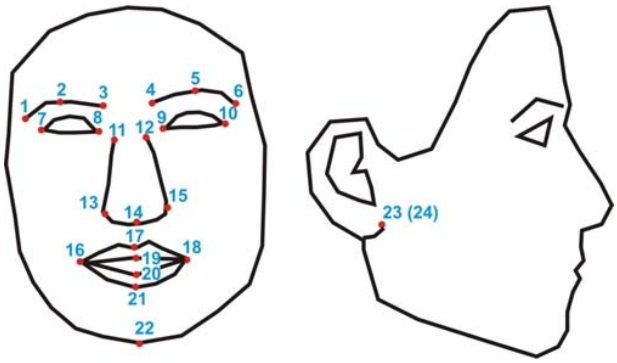


Figure 15: *Manually marked landmark points.*

In the end of manual landmarking, we obtain 2D and 3D locations of 24 facial fiducial points for every scan of every subject. The landmark data structure is stored in a MATLAB "mat" file, one for each subject. The same data is also saved in "text" file format in order not to force the users to MATLAB.

6. AUTOMATIC LANDMARKING

Face detection, normalization or registration should be performed before any face recognition algorithm. ICP (Iterative Closest Point) [10] and TPS (Thin Plate Spline) [11] are the mostly used 3D to 3D registration methods. However, they perform well only when the models being registered are very close to each other since these methods are sensitive to initialization conditions. Thus, a rough registration should be performed before the fine registration by ICP or TPS methods.

Colbry, Stockman and Jain [12] proposed a rough registration method based on 3D anchor points detected from the face data before application of ICP. They detected facial features based on their shape indexes, intensity values and spatial relations where they used both 3D shape and 2D image outputs of the scanner in order to locate anchor points. Similarly, Lu and Jain [13], Chang et. al. [14] and Boehnen and Russ [15] used 3D and 2D information together in order to extract facial anchor points. In most of these methods, detection usually starts from a single point and then followed by an order of other points. For example in [13], first nose was needed to be detected so that the eye and mouth corners could be detected next. Nose tip was detected as follows: First, pose was quantized into some number of angles. Then, the point with the maximum projection

value along the corresponding pose direction was found. Finally, based on the nose profile, the best candidate for a nose tip was selected. After locating the nose tip, other points are located by considering the spatial relationships among the points, shape indexes for these points obtained from 3D data and cornerness for these points determined from the intensity image. In [14], eye cavities were detected first by investigating the mean and Gaussian curvature values and by using some heuristics. Then, nose tip was detected as being a peak which has a spatial relationship with the eye pits. In [12], facial points are detected based on their mean and Gaussian curvature values independently and, again, to resolve between multiple detections, spatial relations between the anchor points were considered. This was done by a relaxation algorithm where some rules were used for anchor points and relations between them in order to eliminate incorrect sets of anchor points.

In this project we applied three different methods for automatic landmarking. There are 24 goal landmark points each of which is marked manually on the ENTERFACE07 database. Although seven of the landmark points are considered to be the most fiducial ones, each method deals with its own set of points. The three competitor methods are briefly described below and their test results are presented in Section 8 (Testing and Evaluation Section).

6.1. Statistical Automatic Landmarking (SAL)

This method exploits the face depth maps obtained by 3D scans. In the training stage, an incremental algorithm is used to model patches around facial features as mixture models. These mixture models are, in turn, used to derive likelihoods for landmarks during testing [16]. The likelihood tests yield the seven fiducial feature points (inner and outer eye corners, nose tip and mouth corners) and the remaining 15 landmarks are estimated based on the initial seven fiducial ones by back-projection. For enhanced reliability, the seven landmarks are first estimated on a coarse scale, and then refined using a local search [17]. Furthermore, incorrect localizations are detected and fixed with the GOLLUM Algorithm [18]. GOLLUM Algorithm uses smaller number of points to check all point localizations. The assumption is that these points contain statistically discriminatory local information, and can be independently localized. An example result is shown in Fig. 16.

In a second tier application of the GOLLUM algorithm, the locations of 15 additional feature points are estimated. These points are based on the coordinates of the seven fiducial ones. The initial seven points contain richer statistically discriminatory local information and can thus be independently localized, whereas the remaining 15 points have much weaker image evidence around them, hence they need the structural information of the former ones. In Fig. 17 all of these landmark points are shown.

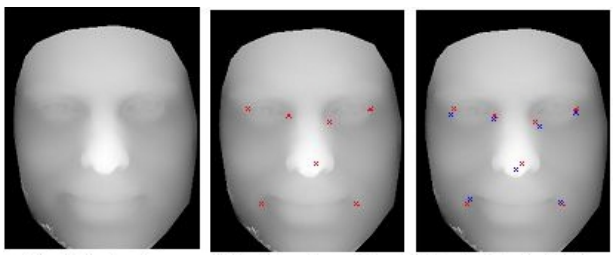


Figure 16: *Depth map of the face image (a), statistically located feature points (b) and corrected feature points on the face (c).*

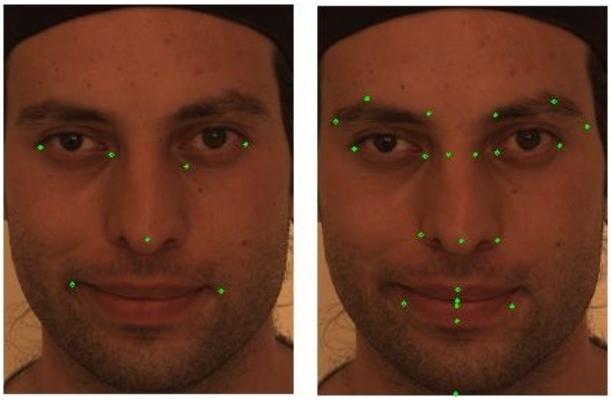


Figure 17: *A total of 22 feature points is landmarked (b) by using automatically landmarked 7 feature points (a).*

6.2. Robust 2D Landmarking (RL)

This method is quite similar to the method described in previous subsection, in that it is also a two-tier method which captures first the seven fiducial landmarks. It also proceeds from an initial coarse evaluation on 80x60 images to a refined version of 640x480 resolution images. The fiducial ones are the four eye corners, the tip of the nose, and the two mouth corners [2]. This method differs from the one in previous subsection as follows: we use templates for each landmark consisting of a subset of DCT coefficients and we obtain scores (likelihoods) employing SVM classifiers. In the coarse localization, 8x8 DCT blocks are extracted and trained separately for each facial landmark. We permit initially four candidates corresponding to the highest peaks in the matching score map of each landmark. This is in order not to miss any feature points. The 4x7 = 28 landmark candidates are reduced to a final seven based on an exhaustive graph search technique. Two methods are proposed for this task: Probabilistic Graph Model-I and Probabilistic Graph Model-II. In PGM-I, we first find a subset of the most reliable landmarks based on the anthropomorphic data learned during a training session and then estimate the position of the missing landmarks. On the other hand, PGM-II is designed for simpler scenarios, without variety in illumination and poses. It systematically searches for the seven feature landmark among the four peaks of the similarity score map. Facial feature landmarks are determined according to companion features, called the support set for that feature. Once the coarse-stage landmarks are located we proceed with the refinement stage on the original high-resolution image. The search proceeds with a 21x21 window from around the coarse localization points. Each configuration formed by reliable landmarks is set to origin, scaled to a fixed length and rotated. Since this procedure makes the configuration independent of the pose and scale of the face, this algorithm achieves high accuracy even under various poses and for this reason is called robust 2D landmarking. In our future work, we aim to apply this method to 3D scans and extend our algorithm to cover the remaining 15 features as the method in the previous subsection.

6.3. Transformation Invariant 3D Feature Detection (TIFD)

In transform and scale invariant feature detection algorithm, 3D fundamental elements such as peaks, pits and saddles are extracted with their scale information from the 3D scan data. With a predefined graph structure based on a group of fundamental elements and their special relationships, we can construct a topology of these elements which could then be used to define the

object. And this graph structure can be used for object detection, registration and recognition purposes.

The method finds the mean (H) and Gaussian (K) curvatures [10, 13] on the surface. Afterwards using Gaussian pyramiding, the H and K values are computed for the higher scales. In all of the scale levels, fundamental elements (peaks, pits and saddles) are found using H and K values as in [10, 13]. In Fig. 18, the higher levels of these fundamental elements are depicted with color (blue: peak, cyan: pit, red: saddle ridge, yellow: saddle valley) and are shown on the 3D surface. We name this 3D volume a "UVS space" where "u" and "v" are surface coordinates and "s" is for scale. By extracting the connected components inside this UVS volume, we obtain the fundamental elements with their position and scale information. The details of the algorithm are given in [19] and [20].

For the ENTERFACE'07 project, we have created a 3D topology for the human face as formed by one peak, two pits and one saddle which represent the nose, the eye pits and the nose saddle, respectively. A four node graphical model is constructed where each fundamental element is a node in the graph and spatial relationships between the fundamental elements are carried by the edges between the nodes. The spatial relations between a node couples are the position difference between the nodes normalized by scale, normal difference between the nodes and scale difference. These values are modeled statistically from the training scans. Then during testing, a graph search is accomplished on the topology obtained from the test scans. As the topology for a human face is formed as having two eye pits, one nose saddle and one nose peak, among the 1peak-2pits-1saddle combinations, the most probable combination is selected as the correct model.

Thus, using this restricted topology, we can detect two inner eye corners, middle of nose saddle and top of the nos. These differ from the labeled landmark points but the latter can be extrapolated easily. Since the location of face is found in the 3D scan data as well as landmark points, pose estimation and registration can be performed easily afterwards.

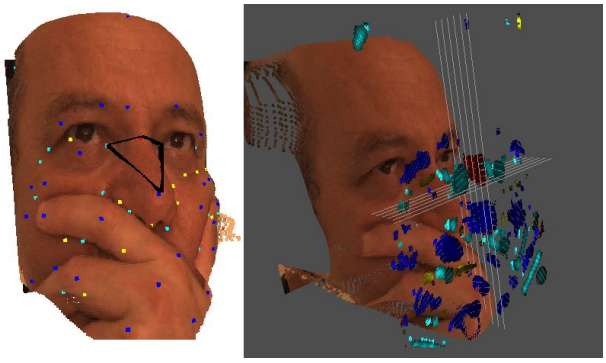


Figure 18: *Left: 3D surface structure centers colored by their types (blue: peak, cyan: pit, red: saddle ridge, yellow: saddle valley) and the graph representing the face, Right: UVS volume shown over the 3D data. Each level represents a layer in the scale space. Connected components are found inside this volume in order to extract the fundamental elements, namely peaks (blue), pits (cyan), saddle ridges(red) and saddle valleys (yellow).*

7. AFM BASED ICP REGISTRATION

In any 3D face processing, be it face recognition or expression understanding, the registration of faces is a crucial step. ICP (Iterative Closest Point) [10] and TPS (Thin Plate Spline) [11] are the most frequently used 3D-to-3D registration methods. However, these methods perform satisfactorily only when the models to be registered are very close to each other since these methods are initialization dependent. Thus, a coarse registration is a must before any further fine registration via ICP or TPS methods. And this coarse registration is usually obtained by automatic landmarking defined in the previous section.

Even though Iterative Closest Point algorithm has a high computational cost, its ease of implementation and accurate end-results make it a preferred method in the registration of 3D data. The ICP method finds a dense correspondence between a test surface and a gallery surface by transforming the test surface to fit the gallery while minimizing the mean square error [10]. The transformation is rigid, consisting of a translation and a rotation. Because this transform with six degrees of freedom is nonlinear, the error minimization problem has to be solved by means of iterative methods.

The most accurate results obtained so far in the literature, make use of a one-to-all registration in which a test shape is aligned separately to each of the faces in the gallery and the test face is compared to each gallery face separately. To overcome the computational cost of one-to-all registration, in [21] it was proposed to use an average face model (AFM). In the registration with AFM, a single registration of a test face is adequate to align it with all the pre-registered faces in the gallery. Motivated by this work on registration via AFM, in [22] it was proposed to use multiple AFMs for different facial categories to increase the accuracy of registration. Multiple AFMs can be achieved either by manually grouping the faces into different categories or by clustering them into classes in an unsupervised manner. The latter approach is a more accurate one, in which the discriminating facial properties existing in the dataset are acquired automatically. Although this method was not used during eNTERFACE'07, it is particularly suited to the dataset because of its many inherent variations. We will apply it in the future.

8. TESTING AND EVALUATION

The database of 3D face scans have been processed with conditioning operations. In addition, they contain ground-truth landmarks; hence they can potentially be used for testing in various tasks such as face detection, automatic facial feature detection, registration, face recognition and facial expression estimation. During the eNTEFACE'07 project we have limited ourselves to testing automatic landmarking and registration algorithms due to time limitations. Work on other tasks will be completed in the near future.

8.1. Landmarking Tests

For the comparative evaluation of landmark localizers, we have used a normalized distance. We divided the localization error, measured as Euclidean distance between the manually marked and automatically estimated landmark positions by the inter-ocular distance. A landmark is considered correctly detected if its deviation from the true landmark position is less than a given threshold, called the acceptance threshold. This threshold itself is typically a percentage of the inter-ocular distance. Curves obtained in this manner are used to present the success of the algorithms (Fig. 19, 21 and 22).

For automatic landmarking tests, we chose 66 subjects that possessed multiple frontal and neutral poses. The same subjects were used for both training and testing. More specifically, we trained our algorithms using one neutral frontal pose from each subject, and then tested it using alternate neutral frontal poses,

which were not used for training. In addition, we also tested the landmarking on three difficult categories: one from action units (open mouth, AU27), one from rotations (+20° yaw) and one from occlusion scans (mouth occlusion).

8.1.1. Statistical Automatic Landmarking (SAL)

In the training step of Statistical Automatic Landmarking, neutral poses of 40 subjects, in our 3D Face & Expression Database, have been used. To evaluate the success of the Statistical Automatic Landmarking algorithm, a test set has been prepared. The testing set consists of:

- 20 subjects for neutral pose
- 20 subjects for neutral pose with glasses,
- 20 subjects for neural pose with eye or mouth occlusion.

According to test results, it is seen that results with the neutral pose with occlusion is better (Fig. 19). Because of the statistical nature of the algorithm, initial landmarking step is processed similarly for all poses. In other words, occlusion in the pose does not affect statistical initial landmarking. After statistical initial landmarking, fine localization process corrects the misplaced feature points by using the depth map. In this step, a pose with occlusion can be corrected more efficiently than a neutral pose. Because there are several local minimas around the mouth in the neutral pose and this decreases the efficiency of the fine localization algorithm. Namely, mouth occlusion (by hand) provides to get rid off local minimas around mouth. As a result, we can say that if the pose and the pose angles are stable, this approach gives good results for the poses with occlusions.

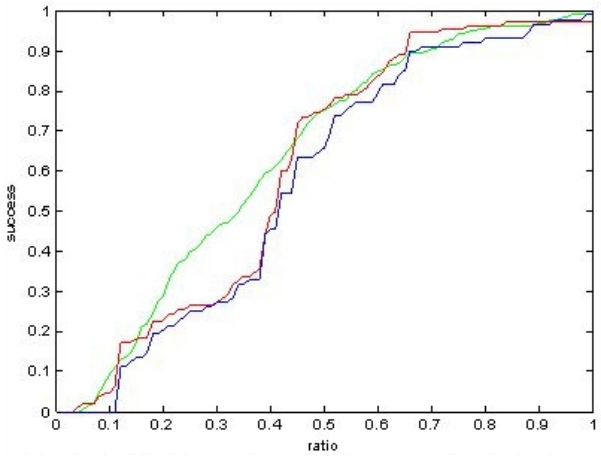


Figure 19: *Statistical Automatic Landmarking test results. Results for neutral pose, with mouth occlusion and with eye glasses occlusion is shown in as green, red and blue curves respectively.*

8.1.2. Robust 2D Landmarking (RL)

In this section, we presented results of our automatic landmarking algorithm tested on eNTERFACE Database. We have used 68 samples with neutral poses as training set. For each sample, DCT coefficients of manually landmarked points were computed and then, we trained an SVM classifier with these coefficients. In order to verify the robustness of proposed algorithm, we tested our classifier on samples with different poses. The testing set consists of:

- 12 subjects for rotation (+10° yaw rotation),
- 28 subjects for facial action unit (mouth stretch),
- 28 subjects for eye or mouth occlusion.

In Fig. 20, output of the algorithm is shown for some representing poses. As expected, the best results are obtained for neutral poses. In Fig. 21, performance of the feature localizer versus acceptance threshold is given for three testing set. According to results, the proposed method is more robust to facial expressions and occlusion rather than rotation.

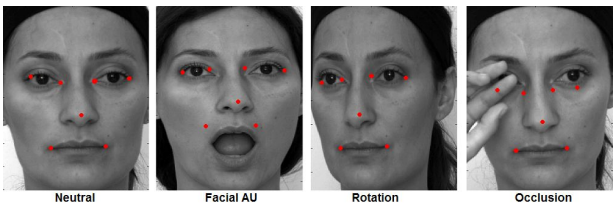


Figure 20: *Landmarking outcomes of the automatic feature localizer for neutral pose, yaw, rotation, mouth stretch and eye occlusion, respectively from left to right.*

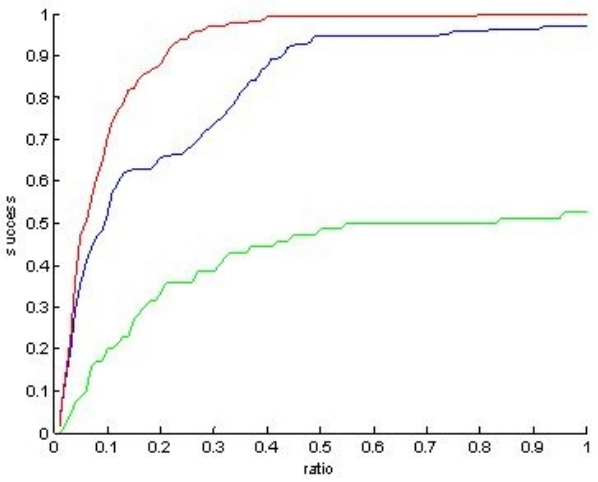


Figure 21: *Performance of 2D Landmarking. Red: pose with expression, Blue: mouth or eye occlusion, Green: rotation.*

8.1.3. Transformation Invariant 3D Feature Detection (TIFD)

In order to test the performance of the TIFD method on the eNTERFACE'07 scans we have made some of experiments. Beforehand, we have trained the system using 66 neutral scans taken from 66 different people. By training, we mean extraction of a statistical model for the relations between the nodes, namely the nose peak, the nose saddle and the eye pits by training a Gaussian Mixture Model [20]. During testing, the method searches for a quarted of a peak, a saddle and two pits which are statistically similar in terms of relative scale and relative orientation to the trained model of neutral scans.

For testing, we have used three groups of scans. The first group included the scans with the expression of mouth opening. Since opening the mouth does not change the locations of the four nodes or does not occlude them, the success rates were very high. We have used 25 models of different people having this expression. In Fig. 22, red curve depicts the success rate for this experiment. By success rate, we mean the average success among all feature points. If the difference between the four points of the detected quartet and their four originally marked points are below a threshold, the success is "one" for that model;

otherwise it is zero. The success rate is calculated among 25 models as an average and taking different error thresholds for success, the curve is depicted.

The second model group included the faces with poses where faces are rotated by 20° around the y-axis (yaw rotation). In this case, the right eye pit gets occluded. However TIFD algorithm still detects a pit near that region, but as center of the eye pit is occluded, the localization slightly fails. As expected, the success rates for the posed faces deteriorate due to this phenomenon. We have used 24 posed models of different people. In Fig. 22, green curve depicts the success rate for this experiment.

Finally we have tested the algorithm on mouth occlusion. Mouth occlusion profoundly distorts the geometry of the facial surface. On the other hand, since we have searched for the nose peak, nose saddle and the eye pits and since these regions are not occluded; the success rates for landmarking on these types of scans were very high. We have used 19 mouth occluded models of different people. In Fig. 22, blue curve depicts the success rate for this experiment.

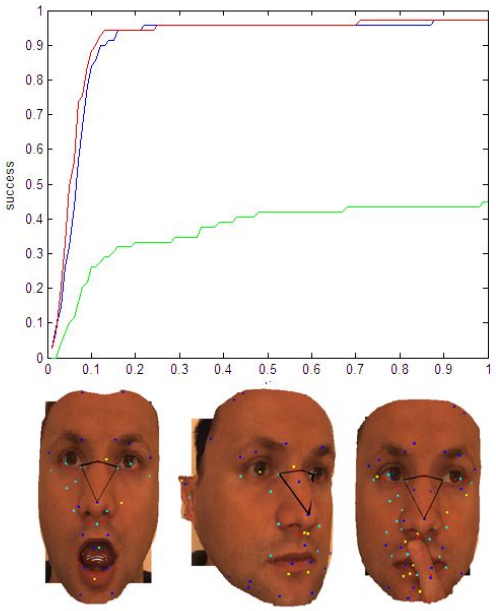


Figure 22: *Performance 3D feature detection results with (TIFD). Plot red: action unit (open mouth, AU27), green: one from rotations (+20° yaw) and one from blue: occlusion scans (mouth occlusion).*

8.2. Registration Tests

The ICP registration based on the use of an AFM can be summarized as in Fig. 23, where a test face is registered to an AFM before a comparison with the gallery faces that have already been aligned with the model.

The depth image of the AFM generated from the first neutral samples belonging to each subject is given in Fig. 24 (a). The construction algorithm proposed by [22] is based on initial alignment by Procrustes analysis and fine registration by TPS warping.

To explore the effect of pose and facial action variations on 3D face recognition, four different face subsets were used for registration and classification simulations. One subset consisted of faces with a pose variation of 10 degrees. Another one was

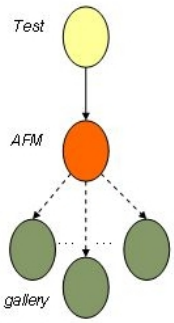


Figure 23: *The AFM-based registration approach.*

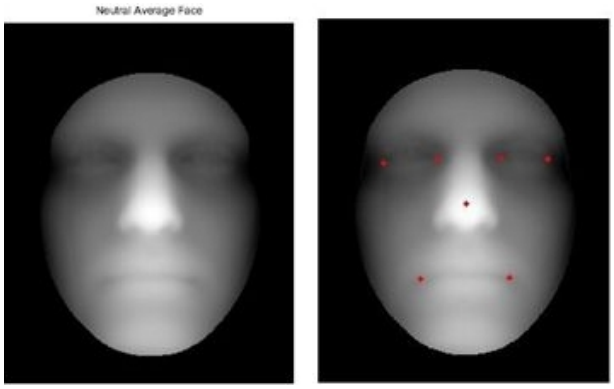


Figure 24: *The AFM constructed from neutral samples of FR database. (a) The depth image for the AFM, (b) the AFM with the calculated landmarks. These landmarks will be used for initialization in the registration phase.*

the subset of faces with "jaw drop" facial action unit. The other face subset had faces with eye(s) occluded. To compare the effects of these different facial variations, a subset of neutral faces was also used in simulations. The gallery set consisted also of neutral faces (but different samples from the neutral test subset). The set of landmarks used for the initial registration by Procrustes analysis can be seen in Fig. 24 (b). The initialization of faces with eye occlusion was handled with fewer landmark points that existed in each face (excluding the eye corners for the eye that was occluded by hand). The seven landmarks (inner and outer eye corners, the nose tip and the mouth corners) were also used for initialization before TPS warping that is used for AFM construction.

Faces in each test subset and the gallery were registered to the AFM by ICP. The registered faces were also cropped according to the average model. A classification simulation was performed based on point set distance (PSD), which is an approximation to the volume difference between faces after registration. Each test face was classified as the closest gallery face. The recognition rates for various ranks are plotted in Fig. 25. In Table 5, the rank-1 recognition rates for each face subset are given.

It is claimed that by using 3D data instead of 2D, pose variations can be eliminated. The registration is the key method to overcome the problems that will arise from pose changes. As seen from these results, a small pose variation can be eliminated by ICP registration. The slight drop in recognition performance is due to more noisy data when acquiring data with pose rotation: The nose occludes one side of the face; and the resulting holes are patched by post-processing, resulting in noisy data.

In the case of eye occlusion, the initial alignment was handled with fewer landmark points that were available in each face. Nevertheless the registration was successful. The low recognition rates were caused by the use of PSD for classification. It would be more appropriate to use a parts-based representation that eliminates the affect of the occluding hand. In the case of mouth opening, the registration was affected a little, while the movement of landmarks belonging to the mouth area also caused a transform for the whole face. Also the rigid structure of ICP is effective in this case. The test face will be aligned as much as possible to the AFM, but the open-mouth form will be kept. Therefore in PSD calculations, the mouth area will augment the distance values, causing the ill-effect on classification results. As in occlusion, a parts-based representation would have been more appropriate.

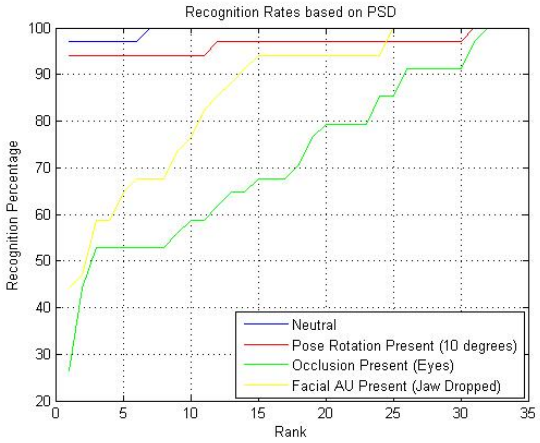


Figure 25: *The recognition rates for various ranks and for different subset simulations.*

	Neutral	Pose Rotation (10°)	Occlusion (Eye)	Facial Action (Jaw Drop)
Rank-1	97.06	94.12	26.47	44.12
EER	5.75	8.73	44.52	23.84

Table 5: *The Rank-1 recognition rates and EER values for each face subset*

9. CONCLUSION AND FUTURE WORK

As future work, we would like to investigate automatic methods for refining and validating the quality of hand-labeled landmarks. As with all human-produced annotations, the landmarks positioned by the labelers are subject to variation. This raises the issue of inter-rater agreement across different raters. We would like to perform appropriate inter-rater comparison and develop automated methods for refining rater input.

The 3D database created during this month-long project will be a very valuable resource for our future work on expression analysis. Including many AUs from FACS and also the six emotions, this database encompasses a rich set of facial deformations. Existing research has focused on recognizing prototypical expressions (e.g., the so called six basic emotions). Poses containing various action unit combinations will help us extend the scope of automatic facial expression analysis. As future work, we intend to explore how wider range of expressions can be recognized using 3D surface data.

The present face database will be put to three different tests:

- 2D-3D Expression Understandings: 3D information can also be combined with the corresponding 2D texture to improve recognition. Moreover, this database can be very useful for developing algorithms that use 2D images for expression recognition. Most existing 2D systems perform well with frontal face images but fail under considerable out of plane head rotations. We would like to find ways of exploiting 3D information to develop pose independent systems. Consequently, we expect to advance the state of art in automatic expression recognition using our database.
- Automatic 2D-3D Landmarking: Facial feature point localization is an intermediate step for registration, face recognition and facial expression analysis. Since correct localization directly affects the performance of the face analysis algorithms, landmarking should be robust with respect to changes in illumination, pose and occlusions. Although there have been advances in automatic 2D/3D landmarking algorithms, there are still open problems. For example, existing algorithms generally work well with frontal and neutral poses; but fail in poses with rotation and facial expressions. We envision the 3D database to be a useful resource in the course of developing and testing 3D-aided 2D or 3D automatic landmarking algorithms that can deal with variations due to pose, illumination and occlusions.
- Registration: This database; including different poses, expressions and occlusions for each subject, is a good test case for registration of faces. The facial surfaces corresponding to different facial groups can be tested to examine the effect of registration of each facial surface difference. Also the multiple AFM-based registration approach from [22] can be adapted to the different groups in this database.

10. ACKNOWLEDGEMENTS

We would like to thank many participants of eNTERFACE'07 Workshop, actors and others who voluntarily let their faces to be scanned. We appreciate their help and patience, since they spent their valuable time during data acquisition process which takes at least half an hour. We would like to thank Niyazi Ölmez for his invaluable help. Without him, we could not find the opportunity of arranging many actors for our database collection. Also, his recommendations for our data capturing setup were very useful, and his personal assistance during acquisition was very valuable. Finally, we would like to thank organizers of eNTERFACE'07 for making collection of such a database possible.

The acquisition hardware in this project was made available through TUBITAK grant 104E080: 3D Face Recognition.

11. REFERENCES

[1] F. L. Bookstein, "The measurement of biological shape and shape change", *Lecture Notes Biomathematics*, vol. 24, 1978. 88

[2] H. Çınar Akakın and B. Sankur, "Robust 2D/3D Face Landmarking", tech. rep., 3DTV CON, 2007. 88, 94

[3] J. Wang, L. Yin, X. Wei, and Y. Sun, "3D Facial Expression Recognition Based on Primitive Surface Feature Distribution", in *IEEE International Conference on Computer*

Vision and Pattern Recognition (CVPR 2006), (New York, NY), IEEE Computer Society, June 17-22 2006. 88

[4] L. Yin, X. Wei, Y. Sun, J. Wang, and M. J. Rosato, "A 3D Facial Expression Database For Facial Behavior Research", in *7th International Conference on Automatic Face and Gesture Recognition (FGR06)*, pp. 211–216, April 10-12 2006. 88

[5] P. Ekman and W. V. Friesen, *Facial Action Coding System: A Technique for the Measurement of Facial Movement*. Palo Alto: Consulting Psychologists Press, 1978. 88

[6] "InSpeck". http://www.inspeck.com. 88, 91, 92

[7] P. Ekman and W. V. Friesen, "Constants Across Cultures in the Face and Emotion", *Journal of Personality and Social Psychology*, vol. 17, no. 2, pp. 124–129, 1971. 88

[8] C. Wallraven, D. Cunningham, M. Breidt, and H. Bülthoff, "View dependence of complex versus simple facial motions", in *Proceedings of the First Symposium on Applied Perception in Graphics and Visualization* (H. H. Bülthoff and H. Rushmeier, eds.), vol. 181, ACM SIGGRAPH, 2004. 88

[9] M. Kleiner, C. Wallraven, and H. Bülthoff, "The MPI VideoLab - A system for high quality synchronous recording of video and audio from multiple viewpoints", Tech. Rep. 123, MPI-Technical Reports, May 2004. 88

[10] P. Besl and N. McKay, "A Method for Registration of 3-D Shapes", *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 14, no. 2, pp. 239–256, 1992. 93, 95

[11] F. L. Bookstein, "Principal warps: thin-plate splines and the decomposition of deformations", *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 11, pp. 567–585, 1989. 93, 95

[12] D. Colbry, G. Stockman, and A. Jain, "Detection of Anchor Points for 3D Face verification", in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2005. 93, 94

[13] X. Lu and A. K. Jain, "Automatic feature extraction for multiview 3D face recognition", in *Proc. 7th International Conference on Automated Face and Gesture Recognition*, 2006. 93, 95

[14] K. Chang, K. W. Bowyer, and P. J. Flynn, "Multiple nose region matching for 3D face recognition under varying facial expression", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 10, 2006. 93, 94

[15] C. Boehnen and T. Russ, "A fast multi-modal approach to facial feature detection", in *Proc. 7. th. IEEE Workshop on Applications of Computer Vision*, pp. 135–142, 2005. 93

[16] A. Salah and E. Alpaydin, "Incremental Mixtures of Factor Analyzers", in *Int. Conf. on Pattern Recognition*, vol. 1, pp. 276–279, 2004. 94

[17] A. A. Salah, R. J. Tena, M. Hamouz, and L. Akarun, "Fully Automatic Dense Registration of 3D Faces: A Comparison of Paradigms", *submitted to IEEE Trans. PAMI*, 2007. 94

[18] A. A. Salah, H. Çınar, L. Akarun, and B. Sankur, "Robust Facial Landmarking for Registration", *Annals of Telecommunications*, vol. 62, pp. 1608–1633, January 2007. 94

[19] E. Akagündüz and İlkay Ulusoy, "Extraction of 3D Transform and Scale Invariant Patches from Range Scans", in *2nd Beyond Patches Workshop "Patches Everywhere" Workshop in conjunction with CVPR 2007*, 2007. 95

[20] E. Akagündüz and İlkay Ulusoy, "3D Object Representation Using Transform and Scale Invariant 3D Features", in *Workshop on 3D Representation for Recognition (3dRR-07), ICCV 2007 (Accepted)*, 2007. 95, 96

[21] M. O. Irfanoğlu, B. Gökberk, and L. Akarun, "3D Shape-Based Face Recognition Using Automatically Registered Facial Surfaces", in *Proc. Inf. Conf. on Pattern Recognition*, vol. 4, pp. 183–186, 2004. 95

[22] A. A. Salah, N. Alyüz, and L. Akarun, "Alternative face models for 3D face registration", in *SPIE Conf. on Electronic Imaging, Vision Geometry*, (San Jose), 2007. 95, 97, 98

12. APPENDICES

12.1. Database design

- Database:
  (Since the whole system may be composed of different databases we'll have a structure called database.)
  - DBID: Primary Key that will be given by system
  - DBName Folder - Mapping to the initial folder structure of the database
  - StartDate - If the different databases evolve this may be needed for date retrievals
  - EndDate
  - SessionNumber
- Subject:
  - SubjectID: Primary Key that'll be given by system
  - Name:
  - Middle Name:
  - Surname:
  - Birth date:
  - Gender:
  - Profession: Actor or a normal human
  - Approval for Publication:
- SessionList:
  - List ID: Primary key.
  - List of the sessions of the database. It will be situated under the database folder and will hold all the session list structure necessary to undertake session architectures.
  - Folder record path of related session.
- Session:
  - SessionID: Primary Key that'll be given by system
  - SubjectID: Refers to the subject table. (referential integrity a subject have many sessions / 1-n mapping)
  - Facial Hair: Yes / No If facial hair then Facial Hair situation Beard, Mustache, beard + mustache Facial Hair Degree Degree: 1 (Low) - 2 (Normal) - 3 (very Much)
  - Hair: 0 (No hair) - 1 (Normal) - 2 (Very Much)
  - Earing: Yes/No
  - Session Folder: Maps to the version folder where all the session data are located.

  - SessionDate
  - SessionCode
  - RecordFolder: Name of the folder holding the records of the session.
- Action Unit:
  - AUID - primary key
  - Code - Code of the action unit
  - Explanation
  - Image Path
  - Video Path
- Records:
  - RecordID: primary key
  - SessionID Referential integrity for finding the owner session.
  - AUID (The ID of the record that comes from Action Unit Table)
  - recordfile name of the corresponding record file
  - Applicable or Not? The stored 3D image is not proper.

12.2. XML Based Subject Table Example

```
<?xml version="1.0" standalone="yes"?>
<NewDataSet>
  <xs:schema id="NewDataSet" xmlns="" xmlns:xs="http
      ://www.w3.org/2001/XMLSchema" xmlns:msdata="urn
      :schemas-microsoft-com:xml-msdata">
  <xs:element name="NewDataSet" msdata:IsDataSet="true
      " msdata:UseCurrentLocale="true">
  <xs:complexType>
  <xs:choice minOccurs="0" maxOccurs="unbounded">
  <xs:element name="Subject">
  <xs:complexType>
  <xs:sequence>
  <xs:element name="ID" type="xs:short" minOccurs="0"
      />
  <xs:element name="Code" type="xs:string" minOccurs
      ="0" />
  <xs:element name="Name" type="xs:string" minOccurs
      ="0" />
  <xs:element name="SurName" type="xs:string"
      minOccurs="0" />
  <xs:element name="MiddleName" type="xs:string"
      minOccurs="0" />
  <xs:element name="BirthDate" type="xs:dateTime"
      minOccurs="0" />
  <xs:element name="Gender" type="xs:short" minOccurs
      ="0" />
  <xs:element name="Profession" type="xs:short"
      minOccurs="0" />
  <xs:element name="ApprovalCondition" type="xs:
      boolean" minOccurs="0" />
  </xs:sequence>
  </xs:complexType>
  </xs:element>
  </xs:choice>
  </xs:complexType>
  </xs:element>
  </xs:schema>
  <Subject>
    <ID>0</ID>
    <Code>S00001</Code>
    <Name>Lale</Name>
    <SurName>Akarun</SurName>
    <MiddleName />
    <BirthDate>2006-02-16T00:00:00+02:00</BirthDate>
    <Gender>0</Gender>
    <Profession>1</Profession>
    <ApprovalCondition>true</ApprovalCondition>
  </Subject>
  <Subject>
    <ID>1</ID>
    <Code>S00002</Code>
    <Name>Ilkay</Name>
    <SurName>Ulusoy</SurName>
    <MiddleName />
    <BirthDate>2006-02-16T00:00:00+02:00</BirthDate>
    <Gender>0</Gender>
    <Profession>0</Profession>
    <ApprovalCondition>true</ApprovalCondition>
  </Subject>
</NewDataSet>
```

13. BIOGRAPHIES

Arman Savran was born in Turkey, in 1981. He received the B.Sc. degree in Electronic and Communication Engineering from the İstanbul Technical University, İstanbul, Turkey, in 2002, and the M.Sc. degree in Electrical and Electronics Engineering from the Boğaziçi University, İstanbul, Turkey, in 2004. He worked at Sestek Inc., İstanbul, Turkey, between 2004 and 2006, where he did research on speech and text driven synthesis of facial animation. He is currently a Ph.D student at Electrical and Electronics Engineering Department of Boğaziçi University, BUSIM Laboratory, and. his research work is on analysis of facial expressions and 3D face modeling.
Email: arman.savran@boun.edu.tr

Oya Çeliktutan received the B.E degree in Electronic Engineering from Uludağ University, Bursa, Turkey, in 2005. She is now a teaching assistant and pursuing the M.S. degree in signal processing and communications at the department Electrical and Electronics Engineering at Boğaziçi University, İstanbul, Turkey. Her research interests include multimedia forensics, machine learning and pattern recognition, computer vision and image processing. Currently, she is working on automatic 2D/3D facial feature extraction.
Email: oyaxceliktutan@yahoo.com

Aydın Akyol was born in Afyon, Turkey in 1979. He received his B.Eng. degree at İstanbul Technical University, Computer Engineering Department, in 2001 and MSc. degree at Sabanci University, Computer Science Department, in 2003. After his graduation he worked as an engineer at R&D department of Verifone Inc. for 3 years. Currently he is pursuing the Ph.D. degree at İstanbul Technical University Vision Laboratory. His research interests include inverse problems in Computer Vision and Image Processing.
Email: akyol@su.sabanciuniv.edu

Jana Trojanová was born in Ústí nad Labem, Czech Republic in 1982. She received her M.Sc degree at the University of West Bohemia (UWB) in Pilsen in 2006. She is currently a PhD student at the UWB and a research engineer at project MUSSLAP. Her main research interest is in data mining, machine learning, artificial intelligence, pattern recognition and computer vision. The topic of the PhD thesis is Emotion detection in audio-visual expression of the person.
Email: jeskynka.jana@seznam.cz

Hamdi Dibeklioğlu was born in Denizli, Turkey in 1983. He received his B.Sc. degree at Yeditepe University Computer Engineering Department, in June 2006. He is currently a research assistant and a M.Sc. student at Boğaziçi University Computer Engineering Department Media Laboratory. His research interests include computer vision, pattern recognition and intelligent human-computer interactions. He works on his thesis with Professor Lale Akarun on Part Based 3D Face Recognition.
Email: hamdi.dibeklioglu@boun.edu.tr

Semih Esenlik was born in Tekirdag, Turkey in 1984. He is currently an undergraduate student in Bogazici University, Turkey. He is a senior student in Electrical & Electronics Engineering Department. His specialization option is Telecommunication Engineering.
Email: semihese@yahoo.com

Nesli Bozkurt was born in Izmir, Turkey in 1982. She received her B.Sc. degree at M.E.T.U. Electrical and Electronics Engineering Department, in June 2005. She worked as a research assistant during her M.Sc. education at M.E.T.U E.E.E. Computer Vision and Intelligent Systems Laboratory for more than a year. She is currently employed for a software company and in the mean time; she works on her thesis with Asst. Prof. İlkay Ulusoy on 3D Scan Data Analysis and Improvement technologies.
Email: e124410@metu.edu.tr

Cem Demirkır was born in Gölcük, Turkey in 1970. He received his B. Sc. at İ.T.Ü. Electronics and Telecommunications Engineering Department in 1991, M.Sc. degree at M.E.T.U Electrical and Electronics Engineering Department in 2000. He is currently a research assistant at Turkish Air Force Academy and a Ph.D student at Boğaziçi University, BUSIM laboratory. He works with Prof.Dr. Bülent SANKUR on image/video Processing and biometrics.
Email: cemd@boun.edu.tr

Erdem Akagündüz was born in Izmir, Turkey in 1979. He received his B.Sc. and M.Sc. degrees at M.E.T.U Electrical and Electronics Engineering Department, in June 2001 and January 2004. He is currently a research assistant and a Ph.D. student at M.E.T.U E.E.E. Computer Vision and Intelligent Systems Laboratory. He works with Asst. Prof. İlkay Ulusoy on 3D Computer Vision, 3D Pattern Recognition and 3D Facial Modeling.
Email: erdema@metu.edu.tr

Kerem Çalışkan was born in Izmit, Turkey in 1978. He received his B.Sc. and M.Sc. degrees at M.E.T.U Computer Engineering Department, in June 2000 and January 2005. He is currently a Ph.D student at M.E.T.U. Informatics Institute - Medical Informatics department. He works with Asst. Prof. Didem Akcay. He owns an R&D company InfoDif whose specialization is in RFID and Signal Processing applications. His main interest areas are Real Time Signal Processing and Medical Imaging Devices.
Email: kcaliskan@infodif.com

Neşe Alyüz was born in İstanbul, Turkey in 1982. She received her B. Sc. Degree at Computer Engineering Department, I.T.U, İstanbul, Turkey in January 2005. She is currently an M. Sc. student in Computer Engineering Department, Boğaziçi University, İstanbul, Turkey. She works with Prof. Lale Akarun on 3D Face Registration and Recognition techniques.
Email: nese.alyuz@boun.edu.tr

Bülent Sankur has received his B.S. degree in Electrical Engineering at Robert College, İstanbul, and completed his graduate studies at Rensselaer Polytechnic Institute, New York, USA. Since then he has been at Boğaziçi (Bosporus) University in the Department of Electric and Electronic Engineering. His research interests are in the areas of Digital Signal Processing, Image and Video Compression, Biometry, Cognition and Multimedia Systems. He has established a Signal and Image Processing laboratory and has been publishing 150 journal and conference articles in these areas. Dr. Sankur has held visiting positions at University of Ottawa, Technical University of Delft, and Ecole Nationale Supérieure des Télécommications, Paris. He also served as a consultant in several private and government institutions. He is presently in the editorial boards of three journals on signal processing and a member of EURASIP Adcom. He was the chairman of ICT'96: International Conference on Telecommunications and EUSIPCO'05: The European Conference on Signal Processing as well as technical chairman of ICASSP'00.
Email: bulent.sankur@boun.edu.tr

İlkay Ulusoy was born in Ankara, Turkey in 1972. She received her B.Sc. degree at Electrical and Electronics Engineering Department, M.E.T.U., Ankara, Turkey in 1994, M.Sc. degree at the Ohio state University in 1996 and Ph.D. degree at Electrical and Electronics Engineering Department, M.E.T.U. in 2003. She did research at the Computer Science Department of the University of York, UK and Microsoft Research Cambridge, UK. She has been a faculty member at the Electrical and Electronics Engineering Department, M.E.T.U. since 2003. Her main research interests are computer vision, pattern recognition and graphical models.
Email: ilkay@metu.edu.tr

Lale Akarun received the B.S. and M.S. degrees in electrical engineering from Boğaziçi University, Istanbul, Turkey, in 1984 and 1986, respectively, and the Ph.D. degree from Polytechnic University, Brooklyn, NY, in 1992. From 1993 to 1995, she was Assistant Professor of electrical engineering at Boğaziçi University, where she is now Professor of computer engineering. Her research areas are face recognition, modeling and animation of human activity and gesture analysis. She has worked on the organization committees of IEEE NSIP99, EUSIPCO 2005, and eNTERFACE 2007. She is a senior member of the IEEE.
Email: akarun@boun.edu.tr

Tevfik Metin Sezgin graduated summa cum laude with Honors from Syracuse University in 1999. He received his MS in 2001 and his PhD in 2006, both from Massachusetts Institute of Technology. He is currently a Postdoctoral Research Associate in the Rainbow group at the University of Cambridge Computer Laboratory. His research interests include affective interfaces, intelligent human-computer interfaces, multimodal sensor fusion, and HCI applications of machine learning.
Email: metin.sezgin@cl.cam.ac.uk

AUDIOVISUAL CONTENT GENERATION CONTROLLED BY PHYSIOLOGICAL SIGNALS FOR CLINICAL AND ARTISTIC APPLICATIONS

Mitchel Benovoy [1], *Andrew Brouse* [2], *Thomas Greg Corcoran* [3], *Hannah Drayson* [4], *Cumhur Erkut* [5], *Jean-Julien Filatriau* [2], *Christian Frisson* [2], *Umut Gundogdu* [6], *Ben Knapp* [7], *Rémy Lehembre* [2], *Christian Mühl* [8], *Miguel Angel Ortiz Pérez* [7], *Alaattin Sayin* [6], *Mohammad Soleymani* [9], *Koray Tahiroğlu* [10]

[1] Centre for Intelligent Machines, McGill University, Montreal, Canada
[2] TELE Laboratory, Université catholique de Louvain, Louvain-la-Neuve, Belgium
[3] Independent artist/researcher from Dublin, Ireland
[4] University of Plymouth, U.K.
[5] Helsinki University of Technology, Espoo, Finland
[6] İstanbul University Dept. of Electrical Eng., Turkey
[7] Queens University, Belfast, U.K.
[8] Institute of Cognitive Science, University of Osnabrück, Germany
[9] Computer vision and Multimedia Laboratory, University of Geneva
[10] University of Art and Design Helsinki, Finland

ABSTRACT

While an extensive palette of sound and visual generation techniques have been developed during the era of digital signal processing, the design of innovative virtual instruments has come to dramatic fruition over the last decade. The use of measured biological signals to drive these instruments proposes some new and powerful tools for clinical, scientific and artistic applications. Over the period of one month - during the eNTERFACE'07 summer workshop in İstanbul, Turkey - researchers from the fields of human-computer interfaces, sound synthesis and new media art worked together towards this common goal.

A framework for auditory display and bio-musical applications was established upon which were based different experimental prototypes. Diverse methods for the analysis of measured physiological signals and of mapping the extracted parameters to sound and visual synthesis processes were explored. Biologically-driven musical instruments and data displays for clinical and medical purposes were built. From this have emerged some worthwhile perspectives on future research. This report summarises the results of that project.

KEYWORDS

Multimodal interfaces – Biosignals – Brain-computer interfaces – Sonification – Auditory display – Interactive arts – Biologically-augmented performances

1. INTRODUCTION

There is already a rich history of people using measured biological signals to generate visual and auditory displays. Scientists, physicians and artists have been using these techniques to understand, analyse and gain insight into ongoing biological processes for almost 200 years. René-Théophile-Hyacinthe Laennec invented the stethoscope and the technique of mediate auscultation to listen to and diagnose maladies of the lungs and heart nearly 200 years ago [1].

Since that time, the techniques of visual and auditory display of biological signals have been important tools used in the interpretation and diagnoses of human biological processes as indicators of sickness and of health. The extent to which these tools have penetrated the day-to-day practices of scientists, clinicians and physicians is not really questioned. There is currently, however, a predilection for the use of visual techniques for such displays which is - given the overtly visual bias of our modern culture - not surprising. At the same time it is worth noting that one of the first techniques which aspiring young physicians learn to do is listening - mediate auscultation - to the heart, to the lungs, to the internal body processes. By the time these people become doctors they already have a finely tuned sense for the sounds of sickness and of health.

Many biological signals are indeed fascinating and challenging objects of study, but, somehow, those of the human brain prove to be the most promising yet problematic. The use of sonification to help understand brain activity goes back to the first confirmation of Hans Berger's tentative explorations into the human Electroencephalogram [2]. From the mid-1960s until the late 1970s a series of composers and artists including Alvin Lucier and David Rosenboom began experimenting with the use of brainwaves and other biological signals to drive sound and visual events [3, 4]. Then, in the early 1990s, Benjamin Knapp and Hugh Lusted began working on a human-computer interface called the *BioMuse* [5, 6]. This permitted a human subject to control certain computer functions via bioelectric signals. In 1992, Atau Tanaka [7] was commissioned by Knapp and Lusted to compose and perform music using the *BioMuse* as a controller. Tanaka continued to use the *BioMuse*, primarily as an EMG controller, during live performances throughout the 1990s. In the current project, we wish to continue the path initiated by these early pioneers and which we have explored during the two previous eNTERFACE workshops [8, 9] by investigating how ongoing measured biological signals can be used to provide insight and clarity to the -scientific and artistic - ways of understanding our own biological realities. An exhaustive review of the state of the art in both biosignal sonification and physiologically-driven musical interfaces is provided in [8] and [9].

In this current project, we intend to continue this established field of research and hopefully add some useful and innova-

tive perspectives to the ongoing practice [10]. Practically, the present work was conducted within the framework of the third eNTERFACE workshop, which took place in İstanbul, Turkey between July 16th and August 10th 2007. During four weeks, a multidisciplinary team, composed of experts from the fields of human-computer interaction, neurophysiology, and sound and music computing have worked together on the development of a robust and reusable framework for the capture and the processing of physiological signals geared towards sound and visual creation applications. With a total of 16 members from labs all over the world, our team had a very broad skill and interest range, from more technical aspects of signal acquisition, to musical performance. The team was divided into a number of smaller interest groups to tackle different aspects of the project. A large set of biosignals was considered and two main approaches were followed, implying different strategies for the design of the sonic interactions. In the first one, we attempted to develop applications aiming at exploiting sound to help the monitoring of the physiological state of a subject in a closed feedback loop, whereas in the second approach, relying on more aesthetic considerations, the objective was to create interactive art performances that fully depend on the physiological state of the performer.

This report presents the main outcomes of this project; it is composed of two main sections resulting of these two approaches: the first section presents the sonification-oriented part of our project aiming at developing a EEG-guided control system giving a visual and sonic feedback of a human brain activity. The second section of the report describes our artistic-oriented approach of the biologically-controlled interfaces, by detailing the *Bio-Music* platform we set up as the result of the workshop and the three artistic performances that were premiered at the end of the project.

2. AUDIO-VISUAL FEEDBACK OF BRAIN ACTIVITY

2.1. Introduction

Sonification is the use of non-speech audio to convey information [11]. *Interactive Sonification* (IxS) a more recent specialisation, takes advantage of the increasing availability of sensing and actuating technologies [12]. In IxS, the listener is actively involved in a perception/action loop, and the main objective is to generate sonic feedback which is coherent with interactions performed upon sonically-augmented artefacts. This allows active explorations of information via more engaging and meaningful modalities. A promising IxS approach is *Model-based Sonification* [13], in which the sound emerges as an organic product of interactions between a model and an external agent.

The sonification of EEG data for monitoring and offline analysis has been already investigated from a research point of view (see e.g. [14, 15] and the references therein). These implementations generally use either *audification* [14], or a kind of *parameter mapping*, i.e., controlling the synthesis parameters via arbitrary transformations of features derived from the data [11].

IxS has begun to exploit the fact that people can self-regulate their brain activity based upon auditory feedback [16]. This notion is essentially the same as "Biofeedback" - an idea which came to some prominence in the 1970s. The resulting device - parametric orchestral sonification of EEG in real-time (POSER) - allows auditory feedback of multiple EEG characteristics using the MIDI protocol. Six frequency bands of increasing centre frequency (slow cortical potentials (SCP), delta, theta, alpha, beta, and gamma) are assigned as instruments on a MIDI device. Two different classes of instruments are used: pitched percussive instruments (delta, theta, alpha) and smooth continuous instruments such as synthesiser, pads and ocarina (SCP, beta, gamma).

The POSER relies on parameter mapping sonification: the timing, pitch, and volume of the instruments are modulated by the features extracted from the frequency bands. The guidelines for sound design in auditory feedback are based upon a previous study [17], which compares the auditory, visual, and audiovisual feedback modes and concludes that audiovisual feedback mode shows no statistically-significant performance increase over simple audio or visual feedback.

In a pilot study with the POSER system, the subjects received full orchestral feedback of their brain activity from one channel of EEG. Initially, they were introduced to each instrument, they were then visually instructed to focus their attention upon either the rhythmic, percussive instruments or upon the continuous instruments for 8 seconds. After a 2-seconds resting interval, the next trial begin. 200 trials for 10 participants were collected. The results indicate that the participants exhibit impressive abilities for auto-regulation: five participants could regulate the amplitude of multiple frequency bands, four participants exhibited significant amplitude changes in the alpha band, and power spectra in beta and gamma wavelengths revealed significant differences in most of the subjects [16].

During this workshop, we set out to develop an EEG-guided control system in line with traditional brain-computer interface (BCI) technology ([18]). The latter uses the potentials emitted by certain pyramidal cell population of the grey matter of the brain to control computers (BCI) or machines (BMI). The development of such interface technology is hitherto mainly motivated as a rehabilitation and communication aid for paralysed patients and other motor-deficient patients. However, recent advances in the fields of EEG measurement and analysis tools along with increased computational processing power suggest possible uses of such techniques for non-medical applications. In this aspect our goal was twofold:

- Firstly, we wanted to build a system which enabled the user to control or manipulate auditory output using motor imagery and action. In order to have as much freedom as possible in the control of auditory output, we pursued the possibility of being able to discriminate between multiple, different tasks. This study was made offline and is described in section 2.2.
- Secondly, we aimed at developing a neuro-feedback system which reflects certain aspects of the users cerebral activity using visual and/or auditory means, and thus gives the user the opportunity to experience and manipulate these features of her brain activity more consciously. This study was made online and is described in section 2.3.

Multi-category classification of motor tasks gave results slightly better than pure chance, however, online control of a ball in 2-dimensions gave interesting results and proved that visual and auditory feedback can improve the user's discriminatory control.

2.2. Offline analysis

2.2.1. Materials

The system used for EEG acquisition was a Leonardo EEG/PSG system with up to 32 EEG channels and 6 PSG channels. A custom Matlab program acquired the data from an RS232 port. The cap is a low-cost swimming cap made of elastic nylon in which holes have been made according the 10/20 electrode placement system. Individual electrodes are then placed - via the given holes - on the scalp with conductive paste. For online real-time

purposes, the acquisition algorithm was written in C and compiled in the Matlab environment.

Pre-processing No muscular artefact removal algorithms were used, but the subject was asked to avoid eye blinking and extraneous movement as much as possible.

2.2.2. Experiment paradigm

The aim of this experiment was to discriminate between EEG patterns related to different motor tasks which would then be used to drive sound and visual synthesis processes. A visual stimulus was presented to the subject, displaying a word corresponding to a task which the subject is trying to achieve (i.e move left finger, right finger, left foot, right foot, tongue, relax). A reference cross was fixed in the centre of the display in order to give a point of reference to the subject and to limit EOG artefacts.

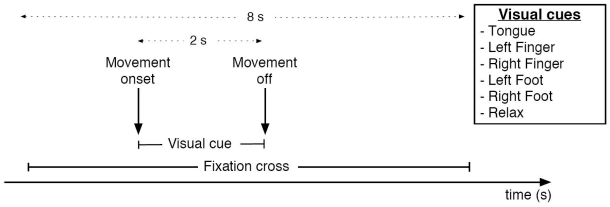


Figure 1: *EEG experiment paradigm. To avoid eye movements, a reference cross is displayed. Every six seconds, one of six motor tasks appear on the screen for two seconds.*

2.2.3. Event Related Potentials analysis

In order to investigate the potentials related to the different tasks which were actuated, we extracted either time or frequency features to train a classifier.

2.2.3.1. Event-related potential (ERP) analysis using time features

Pre-processing The trials were high-pass filtered with a cut-off frequency of 2 Hz. For each trial we applied a baseline correction with the mean value of the interval from 400 to 50 ms before stimulus onset. We applied a bootstrap significance test to investigate the differences between the conditions [19]. To remove excessive artefacts like eye-movements, we employed a threshold criteria of +/- 50 microV.

For the trial averages for channels F3, F4, C3, C4, P3, P4, O1, and O2 no significant differences between the control condition (relaxed) and each of the five other conditions (left finger, right finger, left foot, right foot, tongue) were found (see Fig.2 and 3 for examples).

2.2.3.2. Event-related potential (ERP) analysis using frequency features

Pre-processing Short Term Fourier Transform (STFT): STFT coefficients were computed over the 2 seconds following the stimulus onset. Each STFT coefficient is considered as a feature in our feature vector. A STFT with window length of 64 points and FFT length of 64 points was used for feature extraction. The STFT features from 8 channels were combined to shape the feature vector of each trial.

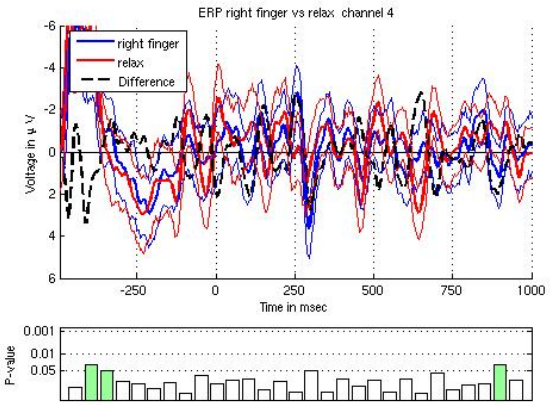


Figure 2: *The difference wave (black) for the right finger (blue) vs relaxed (red) condition at the left central electrode (C3). Standard deviation of mean is indicated by the thinner red and blue lines surrounding the respective signals.*

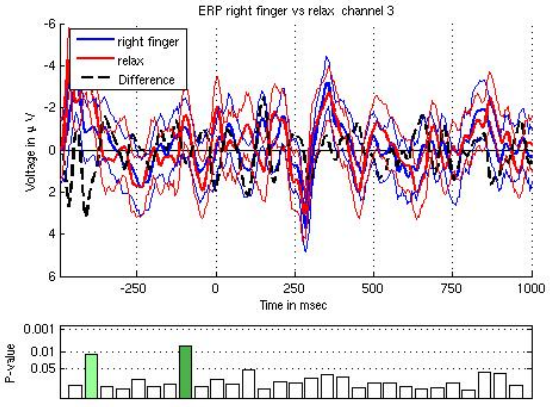


Figure 3: *The difference wave (black) for the right finger (blue) vs. relaxed (red) condition at the right central electrode (C4). Standard deviation of mean is indicated by the thinner red and blue lines surrounding the respective signals.*

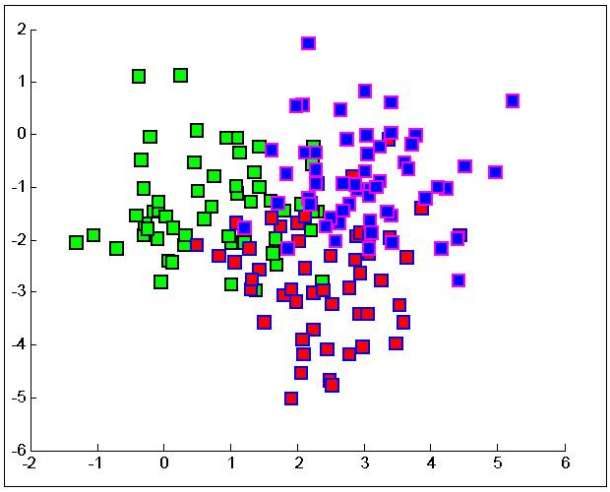


Figure 4: *Example of a 2-D feature vector for a training set. Each square shows a trial which is projected on two dimensional space by fisher projection. Three different classes are shown: state of rest (green), tongue movement (blue) and limb movement (red).*

2.2.3.3. Classification and results

Pre-processing In order to select the most relevant features, the fischer reduction technique was used. The high dimension feature vector for each trial was thus reduced to a two-dimensional vector (Fig.4). Three classifiers were trained on the reduced feature vector, Polynomial Support Vector Machines (P-SVM), Linear Support Vector Machines (L-SVM) and Linear Discriminant Analysis (LDA).

Differentiating all 6 tasks led to results only slightly better than chance, we therefore grouped all four limb movements into one category and decided to classify only three conditions : relaxed rest condition, tongue movement and limb movements. The results are shown in table 1. These outcomes are disappointing probably because of muscle artefacts stemming from the actual movements.

Table 1: *Classification results for three tasks: rest (R), tongue (T) and limb movements (M). The results are shown with and without Band Pass (BP) filtering. Three classification methods are used, Polynomial Support Vector Machines (P-SVM), Linear Support Vector Machines (L-SVM) and Linear Discriminant Analysis (LDA)*

	3 class (T,M,R) no BP	3 class (T,M,R) with BP
P-SVM	41.5%	43.1%
L-SVM	42.92%	45.45%
LDA	41.6%	43.7%

2.2.4. Discussion

To simplify the task for the experimental subject, we used real movement instead of imagined movement. This approach might be especially interesting for the implementation of BCI systems for real-world interactions and performance, which is usually accompanied by external artefacts (e.g. muscle activity, ocular artefacts). Therefore, a BCI system which can function in situations where the user is moving would be a big advantage over the typical systems which work well only in an artificial, almost noise-free context. However, we encountered many obstacles in our approach and there remain some key issues germane to the further development of an EEG-guided motion-resistant control system. The analysis of the data showed no significant differences between event-related potentials of the different conditions. The data suffered from variance introduced by motion artefacts, latency and jitter. To reduce the variance one should use electrodes (EMG) to detect the movements and then to average the signals aligned by the onset of movement to obtain the motor readiness potential [20]. Furthermore, the application of sophisticated artefact reduction techniques (e.g. ICA or regression-based methods for muscle and eye-movement artefact removal) might delay the recognition of mental tasks absolved but could increase the effective signal-to-noise ratio and thus the probability of correct recognition.

Further, a smaller variance between trials of one condition in motor imaginary approaches might be achieved by subject training. Here, sonification and visualisation techniques can give the subject a clearer understanding of her brain dynamics relative to the signal to produce. This is what we studied in the MiniBall experiment which is described in the following section.

2.3. Online Analysis

As stated in [18] *"BCI use is a skill"*. To master this skill, the user needs to train on a mental task. The use of feedback helps considerably and yields more accurate results.

Different types of feedback have been studied, for example, visual, haptic [21], vibrotactile [22] and auditory feedback [23]. Auditory feedback has been seldomly used but could prove very useful for blind disabled. A recent study [24] uses sounds coming from two different instruments as a feedback for the user. Although the learning time is longer for the auditory feedback, in the end results are similar with visual or auditory feedback.

The aim of this study was to compare visual and auditory feedback for a 2D cursor movement. Experiments were made on an untrained subject. The goal was to control a ball on the screen and drive it to a rectangular box in the top-right of the screen (Fig.5). Horizontal and Vertical movements were controlled by the spectral power in alpha and beta bands. The subject was not restrained by a particular protocol within the feedback, rather he was encouraged to find the states most effective for alpha and beta manipulation himself. This resulted in the vivid imagination of the subject flying through complex cloud formations. Indeed, this task not only induces alpha waves, because of the feeling of relaxation drawn from the sensation of flying, but also beta waves since the user achieves high cognitive visualisation tasks.

Results were measured in seconds, i.e the time needed for the user to reach the target. Either visual or auditory feedback gave similar results, less than five seconds. These results are significantly better than without feedback. However the combination of visual and auditory feedback led to poorer results, perhaps because of an inconvenient work overload.

The operating system and the audiovisual synthesis is presented in the next few paragraphs.

2.3.1. Operating system

Data acquired by the system described in 2.2.1 is sent via UDP to Simulink (Fig. 6). Only one electrode (F3) was used for the following experiment. The sampling rate was 200Hz, and data from F3 was bandpassed filtered using a Simulink block provided by an open source software, rtsBCI available on the Biosig web pages[1]. Alpha and beta bands were chosen between

[1] http://biosig.sourceforge.net

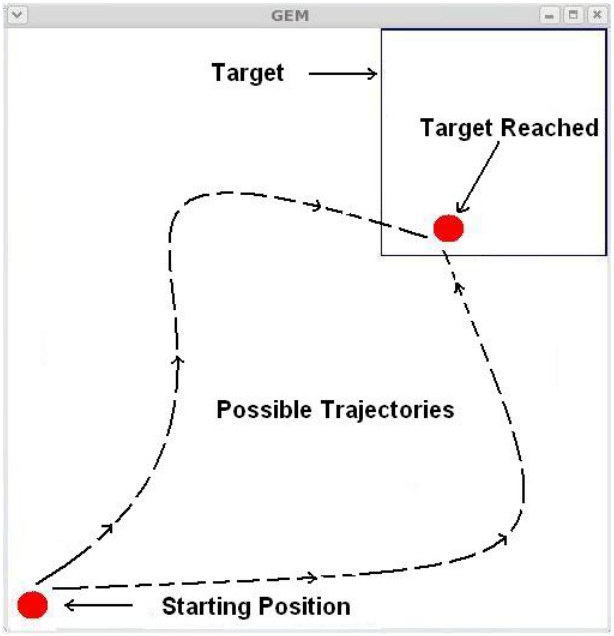


Figure 5: *Screen capture of the MiniBall interface. Horizontal and vertical movements of the ball are controlled via the spectral power in alpha and beta bands. The user is free to adopt any trajectory as long as the ball reaches the target.*

[8-12]Hz and [13-20]Hz respectively. A FIR filter of order 5 was used. Samples of spectral powers of alpha and beta bands are then sent with the OSC protocol with a hand-made level-2 m-file S-function.

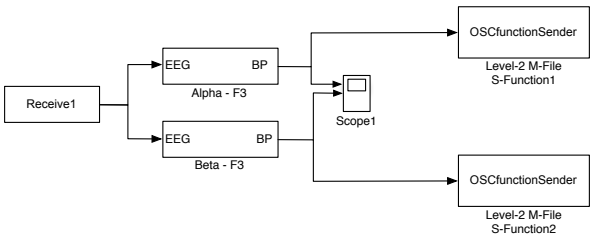


Figure 6: *Block diagram of the Simulink model. Data is received via UDP, band powers of alpha and beta bands are computed, then sent through OSC to the audiovisual synthesis.*

Simulink was chosen for its flexibility and rapid prototyping capabilities. In further work we will consider using more electrodes and more complex signal analysis.

2.3.2. *Description of the audiovisual synthesis*

We had an audiovisual operant conditioning training system in mind, and chose the pd-GEM environment as the implementation platform. The block-diagram in Fig. 7 illustrates the structure of our system `MiniBall_EEG`. `MiniBall_EEG` receives OSC streams or reads pre-recorded measurement data. The OSC streams follow our namespace convention. Currently, energy in the alpha and beta bands of a fixed channel is received. In the Source Selection block, other channels can be listened to by manual routing. Very primitive probing of the incoming data levels - akin to the parallel time-series visualisation - is provided. The probed signal levels inform the scaling values in the next block Preprocessing. The alpha-band energy is first hard-limited within the range $[\alpha_{min}\ ,\ \alpha_{max}]$ then linearly mapped to the normalised range $\bar{\alpha} \in [-1, 1]$. The beta-band is processed similarly. In both bands, good results are obtained with the values 0.01 and 512 as minimum and maximum values, respectively. These normalised instantaneous energy values are then smoothed with a moving average filter of length 20, and forwarded to the audiovisual rendering blocks.

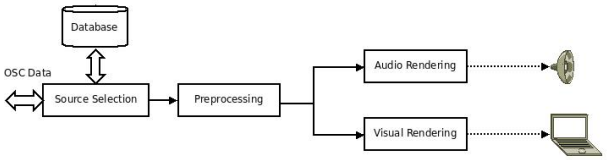


Figure 7: *Block diagram of* `MiniBall_EEG`.

The Visual Rendering block creates a region $\{(x, y) \in \Re^2 | -1 \leq x \leq 1; -1 \leq y \leq 1\}$ within a window of 500 x 500 pixels, which is updated by a rate of 20 fps. Two objects are created within this region; a red disc of radius 0.1 (ball), and a blue square of sides 0.75. The centre of the square is translated to (0.625, 0.625) so that it resides at the top left corner of the region. The normalised alpha and beta energy magnitudes determine the coordinates of the red ball along the x and y-axes, respectively.

The Audio Rendering block defines two sound generators, each generate the difference tones of two primitive waveforms. We have used two sine waves for alpha and two saw waves for beta band and tuned the difference tones to the centre frequencies of the corresponding band. These centre frequencies were 10 Hz (alpha) and 18 Hz (beta) in our experiment. The normalised alpha and beta values control the frequencies of the generators. For the alpha band, the mapping is defined by

$$f_\alpha = c_\alpha(2 - \bar{\alpha}) \tag{1}$$

where c_α is a scaling value. A similar expression gives f. We have used $c_\alpha = 32$ and $c_\beta = 4$. The normalised alpha and beta energy magnitudes determine the normalised gain according to $g = 0.25(|\bar{\alpha} - 1| + |\bar{\beta} - 1|)$ where $g \in [0, 1]$. This scaling ensures that the sound feedback becomes inaudible when the ball enters the target region.

We have experimented with two additional mapping strategies for rendering the alpha band only: the centre beating frequency and amplitude panning. The centre frequency is updated according to $f_{\alpha,c} = 10 - 2\bar{\alpha}$, the panning gains are set to

$$g_1 = (1 + \bar{\alpha})/2 \tag{2}$$

$$g_r = 1 - g_1 \tag{3}$$

for the left and right channel, respectively. In practice, these mappings are meant to sonify the coordinates of the red ball when there is no visual information available, e.g., when the subject closes his eyes.

The implementation of the `MiniBall_EEG` consists of a main control patch and visualisation window, as shown in Fig. 8. Once the data arrives and it is properly scaled, the feedback generation in the desired modalities (audio, visual, or audiovisual) can be switched on and off, and all the audio gains (alpha, beta, and the total mix) can be controlled in run-time. The goal is to move the red ball into the blue square. This translates to have brain activity in the alpha and beta bands simultaneously. We did not know how to construct a task description out of this

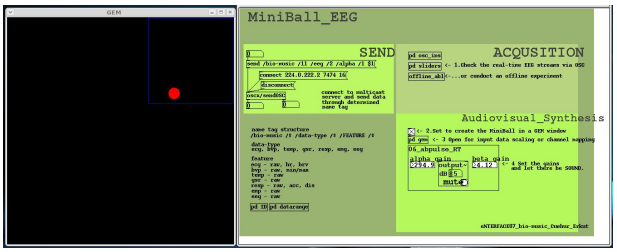


Figure 8: `MiniBall EEG` *visualisation window (left) and control patch (right).*

goal; fortunately our participant is a neuroscientist, and he constructed his strategies himself: he dreamed of flying. The experiments took place in the last days of the workshop. While the quantitative analysis of the recorded data is still in progress, the primary observations indicate that our subject exhibited a surprising skill in regulating the alpha and beta bands, and enjoyed this skill trough the `MiniBall_EEG`.

2.4. Conclusions and Future Work

In the future, we will implement automatic routing, online/offline selection, channel selection and mixing, and a data transmission port following the pre-processing or rendering operations in Source Selection block. We will determine the limiting values within the Pre-processing block by using the signal and the noise statistics. The current fixed-length of the moving-average filter can be made variable based on these statistics.

3. THREE PHYSIOLOGICALLY-DRIVEN ARTISTIC PERFORMANCES USING THE BIO-MUSIC PLATFORM

In this section we describe the second approach which was used during the workshop the objective of which was the design of musical interfaces controlled by the physiological states of a performer. For this purpose, we set up a platform which allowed us to handle a large range of biosignals as typically used during psychophysiological monitoring (i.e. electromyogram (EMG), electrocardiologram (ECG), galvanic skin response (GSR), blood volume pulse (BVP), temperature and respiration signals). In the following sections, we first describe the *Bio-Music* platform developed during the workshop and then we present three artistic projects which resulted from the work within the project and which are built upon the Bio-Music platform.

3.1. The Bio-Music platform

3.1.1. Overview of the biosignals

In order to assess the physiological/affective state of the performer, we developed a hardware and software framework, we named *Bio-Music*, providing a range of meaningful features extracted from the class of biosignals we considered (Fig. 9). Here are some precisions about the biosignals we worked with in this project:

3.1.1.1. Electromyograms (EMG)

Pre-processing Electromyography (EMG) is a technique for evaluating and recording physiologic properties of muscles: EMG measures the electrical potential generated by muscle cells at rest and while contracting. The resulting measured potentials range between 5 to 30 mV, and typical repetition rate of muscle is about 720 Hz, depending on the size of the muscle.

3.1.1.2. Electrocardiograms (ECG)

Pre-processing Electrocardiography (ECG), aims at acquiring and processing the electrical activity produced by the beating heart. By means of two electrodes coupled sequentially out of the total number of electrodes wired to the patient, in order to create current loops between the acquisition interface and the human body, potential differences can be detected at several locations. Many techniques have been developed throughout history [25], the current one standardised for clinical applications requires 12 leads and allows the extraction of many features, even the reconstruction of the tri-dimensional electrical and displacement fields of the heart. We chose to use a technique derived from the "Einthoven triangular method", bearing the name of the scientist that introduced it in the early twentieth century, where in our case the three electrodes are not positioned on limb endings as the original protocol recommends, but closer to each others, thus housed altogether more ergonomically fixed on the chest. Our motivations in the choice of this setup are: cost-effectiveness, non-invasive ergonomics and feature extraction economy. The latter illustrates the fact that we found relevant to extract a small number of features, all of which allowing artistic applications, such as heart rate and heart rate variability.

3.1.1.3. Galvanic skin response (GSR)

Pre-processing The galvanic response is a measure of electrical impedance of the wearer's skin surface, which reacts proportionally to emotional arousal i.e. stress level. The sympathetic system is responsible for activating the secretion of sweat on a person's hand or feet palms. The rise in surface sweat in turn increases the skin surface's conductance. GSR is typically measured with two skin surface electrodes placed on the palm side of two non-adjacent fingers. The baseline conductance levels, measured in μSiemens, are person specific; however noticeable changes in GSR are generally seen in emotional states such as anger, fear, excitation, startle and sexual arousal.

3.1.1.4. Blood volume pulse (BVP)

Pre-processing The blood volume pulse sensor uses photoplethysmography to detect the blood pressure in the extremities. Photoplethysmography is a process of applying a light source and measuring the light reflected by the skin. At each contraction of the heart, blood is forced through the peripheral vessels, producing engorgement of the vessels under the light source, thereby modifying the amount of light to the photosensor. BVP is measured with a sensor worn on the palmar side fingertip of the subject's non-dominant hand to minimise motion artefacting. Since vasomotor activity (activity which controls the size of the blood vessels) is controlled by the sympathetic nervous system, the BVP measurements can display changes in sympathetic arousal. An increase in the BVP amplitude indicates decreased sympathetic arousal and greater blood flow to the fingertips.

3.1.1.5. Phalange temperature

Pre-processing Phalange temperature is measured with a thermocouple fixed on the palmar side of one of the subject's fingers. The acral skin temperature experiences short changes related to the vasomotor activity of the arterioles. It is shown that the surface temperature of peripheral limbs vary as a consequence of blood flow [26].

3.1.1.6. Respiration

Pre-processing The respiration sensor is placed either over the sternum for thoracic monitoring or over the diaphragm for diaphragmatic monitoring. The sensor consists mainly of a large Velcro belt which extends around the chest cavity and a small

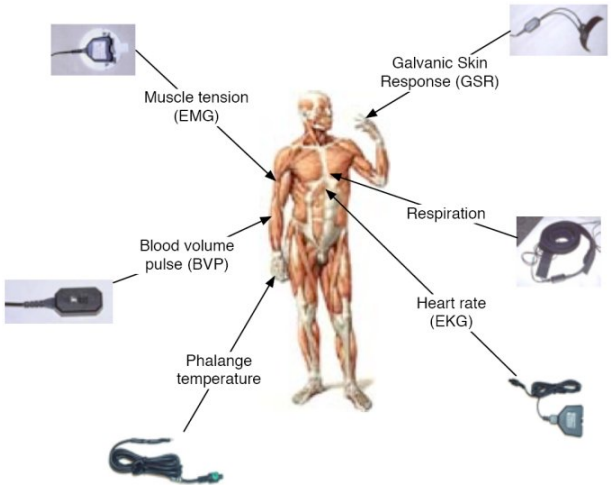


Figure 9: *Overview of the biosignals handled in the Bio-Music platform*

elastic which stretches as the subject's chest cavity expands. The amount of stretch in the elastic is measured as a voltage change and recorded. From the waveform, the depth the subject's breath and the subject's rate of respiration can be learned.

3.1.2. Description of the two setups

The *Bio-Music* platform which we have developed is actually composed of two different configurations:

3.1.2.1. First setup

Pre-processing The first uses Thought Technology's ProComp Infiniti biofeedback system[2] capturing EMG, ECG, BVP, GSR, respiration and temperature signals, all sampled at 256 Hz. For the processing of these signals, we used the primary analysis functions provided by the manufacturer's software API. Further, we also developed a function allowing us to send the raw signals to Matlab and Python allowing us to use our own more complex processing tools (see Tab 2). The sound and visual synthesis tools were implemented using the Max/MSP-Jitter[3] and Pure Data-GEM[4] software environments, which offer a large palette of sound and image processing tools for real-time artistic applications. The communication between all the software components of our platform rely upon the Open Sound Control protocol (OSC).

3.1.2.2. Second setup

Pre-processing The second hardware implementation we used consisted of two EMG sensor bands from Biocontrol Systems[5] which were connected to an Arduino[6] interface board with bluetooth networking. These hardware components interact with a computer running EyesWeb[7] software and a custom built patch for data acquisition. This setup provided a smaller number of biosignals but offered a more flexible and less intrusive way to measure muscle activity than the first setup. The analysis of the biosignals mainly consisted in this case of an envelope follower and was implemented within the same Max-MSP patch as the sound processing tools.

[2] http://www.thoughttechnology.com/proinf.htm
[3] http://www.cycling74.com
[4] http://www.puredata.org
[5] http://www.biocontrol.com
[6] http://www.arduino.cc
[7] http://www.eyesweb.org

3.1.2.3. OSC Multicast

Pre-processing The *Bio-Music* platform is envisaged as potentially accomodating a wide variety of data acquistion hardware devices for the capture of biological signals which can then be sent to a wide variety of software applications for further processing. In order to receive data, from diverse sources, we decided to set up a multicast OpenSoundControl protocol server, which gives us flexibilty in having fast and direct connections between a number of sender and receiver clients connected to a local area network (Fig.10). OpenSoundControl[8] (OSC) is a protocol for communication between host computers, sound synthesizers and other multimedia devices which is optimised for modern networking technology. For this project, we defined our own proper OSC namespace, allowing to us to formalize a common way of exchanging data between the various components of the *Bio-Music* platform. During this workshop, a rich dataset has been collected and has been used extensively to implement and tune the signal processing functions of our Bio-Music platform. These datasets - and some illustrative video documentation - are freely available on the public project wiki[9]. Please also refer to appendices A and B for detailed descriptions of our OSC namespace and our biosignal datasets.

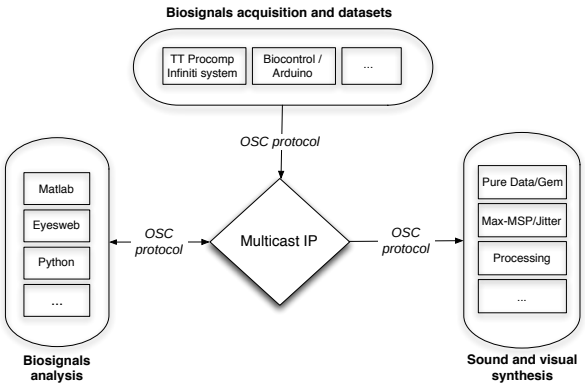


Figure 10: *The multicast OSC-based* Bio-Music platform

As an outcome of the artistic component of this project, we organised a Bio-Music performance night at a club in İstanbul where three musical and artistic performances utilising our platform were performed. The first one, performed by Koray Tahiroğlu and Selçuk Artut was an improvisation for digital instruments controlled by electromyogram (EMG), galvanic skin response (GSR) and respiration signals. The second performance *Carne*, the result of a collaboratiion between the composer Miguel Angel Ortiz Pérez and the visual artist Hannah Drayson, was an interactive piece controlled by two EMG sensors. The last performance *Time Series* by Hannah Drayson was an investigation into subjective visualisations and sonifications of human emotions based upon measured biosignals. Each of these performances are described in the following sections.

3.2. Improvisation for Live BioSignals / various interfaces

This was an improvisation for digital instruments controlled by biological signals of two performers, Koray Tahiroğlu and Selçuk Artut. Each performer was wired through different sensors in

[8] http://www.cnmat.berkeley.edu/OpenSoundControl/
[9] http://biomusic.wikidot.com

Table 2: *Features and characteristics extracted from biosignals*

Feature	Characteristics
Statistical	Running analysis of mean, median, variance, standard deviation, minimum and maximum.
Rate of change	First and second derivatives.
Envelope detection	Using a low pass filtering algorithm.
Power spectrum	Power amplitude in selectable frequency ranges. Implemented with the Discrete Fourier Transform. The buffer length is adjustable, allowing control of spectrum accuracy.
Heart rate	Heart rate (HR), HR acceleration/deceleration

order to control and master their instruments during a real-time musical activity. This improvisation was a pure biological activity where there was an infinite joy and aesthetic pleasure. *Improvisation for Live BioSignals* aimed at achieving ever-changing unexpected sound structures and giving a chance to audience to recognise new possibilities in sound. Instruments that were driven by biological signals were created regarding to the characteristics of the signals.

One type of the biosignals used during the performance was the galvanic skin response (GSR), which measures the resistance of the skin to the part of a very small electric current. The magnitude of this electrical resistance is affected by immediate emotional reactions. In order to map the galvanic skin response changes, Selçuk Artut, who played a traditional bass guitar instrument during the performance, placed GSR sensors on his left index and ring finger. Higher or reduced arousal states of the body causes a change in the skin's conductivity; therefore, galvanic skin response sensor was used together with ECG sensor to control an instrument that created constantly dynamic rhythmic patterns.

Electrocardiography (ECG) measures the heart's electrical activity over time. Within the received data flow, rhythmic period of heart was mapped by detecting the peak levels for the each beat. Koray Tahiroğlu positioned three ECG electrodes, two on the left and right side of the upper chest area and one on lower part of the rib cage closer to the heart. Changing patterns of the heart's electrical activity gave more chance to create a dynamic pattern with sound synthesis. Generating the common sonic representation of heart beat and ECG monitoring sound responses was avoid intentionally, instead noise sound sources were combined together with glitchy sound samples in order to create a rhythmic pattern. Heart beat structure could be traced in the sound structure. Glitchy sound samples, `[sampleosc center ~]` abstraction patch, were also controlled by galvanic skin response signals, so that rhythmic pattern that was constructed with `[sampleosccenter ~]` and `[noise4c ~]` abstraction was representing the overall heart beat activity of the two performers at the same time.

During the performance Koray Tahiroğlu strapped the respiration sensor around the chest area and the measurement of the chest expansion was creating respiration signals. Respiration signals were easily controlled by producing different breathing patterns and this led Koray Tahiroğlu to use respiration signals as a master controller for the overall sound structures in this improvisation process. Through this master control during the performance sonic structure of the improvisation process could be changed into three different modes.

3.2.1. Slow Respiration Mode

Slow Respiration mode activates the `[sin3---- ~]` abstraction in a structure where pitch values changes in a higher frequency rate and modulation varies 0-8.50 out of 0-20 amplitude range. `[sin3----- ~]` abstraction was created to control different sample position readings on a recorded sound material and also to have a modulation by sinusoid wave signs. Parameters of `[sin3----- ~]` abstraction is set through drunk walk probability method. Simply producing a slower breathing pattern and keeping the chest expansion in a lower state can keep the sonic structure of the improvisation in the slow respiration mode. `[instr6]` abstraction is the polyphonic sound synthesis of different sound samples driven by Electromyography (EMG) sensor. EMG measures muscle response or electrical activity in response to a nerve's stimulation of the muscle. First EMG sensor was taped on the inner side of the right arm of Koray Tahiroğlu and by moving his right wrist and his arm itself, it was possible to change the pitch information of the related sound sample. Master control through respiration signals also change the sound sample type in `[instr6]` abstraction into three different modes. Slow Respiration mode sets a sound sample, which responds with continues long and no sudden changing structures in polyphonic synthesis.

3.2.2. Normal Respiration Mode

In this mode, `[sin3----- ~]` abstraction generates a continuous low frequency pitch. Without changing the pitch value, only modulation parameters are generated between the range of 1.63-4.48 over 0-20 amplitude range through drunk walk probability method. Keeping the constant pitch information created a low frequency wall sounds with different modulation varieties. On the other hand normal respiration mode set a rhythmic sound sample in `[instr6]` abstraction, and by changing the right arm's muscle activity it was possible to control the pitch values of the second sound sample.

3.2.3. Fast Respiration Mode

Fast respiration mode modulation was set at 0.807 amplitude levels; however, pitch values were dynamically created through drunk walk probability method with `[sin3----- ~]` abstraction. `[instr6]` abstraction was creating heavy noise based structures with low frequency and glitchy sounds. This level was meant to be creating a chaotic sonic structure regarding to the noise based music.

EMG sensors had been used as a controller for `[instr6]` abstraction patch. The pitch values of sound samples with applied polyphonic sound synthesis within their related mode structure was controlled by the performer. Koray Tahiroğlu also taped a second EMG sensor on the inner side of his left arm.

Figure 11: *Miguel Angel Ortiz Pérez performing* Carne

He used the second EMG signals to control the instruments on and off situations. Related to the improvisation, with the second EMG signals some instruments could be turned off and some instruments could be turned on. This second main controller gave a chance to decide which one of the instruments should play together at the certain stage of the improvisation. Except the EMG sensor's signals, GSR, ECG, respiration had continuous streaming signals, because all were related with continues heart beat and respiration activities; therefore, there had been a need to control the instruments current on/off situations during the improvisation.

3.3. Carne

Carne is an interactive piece for two EMG sensors. It was composed as part of the activities on the eNTERFACE'07 summer workshop. It was premiered at the Boğaziçi University Music Club on August 8 2007. The piece is an audiovisual collaboration between Miguel Angel Ortiz Pérez (interface and sounds) and Hanna Drayson (visuals). Fig.11 shows the performer at the premiere.

3.3.1. Piece concept

Carne is loosely inspired by Terry Bisson's 1991 short story "*They're made out of meat*" [27]. The concept behind *Carne* is based on a very simplistic view of muscle activity as the friction between slices of meat. Taking this idea further, we could say that all types of arms movement from minimal arm gestures up to the highly complex synchronised movements of fingers during musical instrument performance, are simple variations of this meat grinding activity.

The sounds in *Carne*, evolve inside a continuum from imaginary muscle sounds to pre-recorded sounds of western bowed string instruments, while always keeping focus on friction as a unifying metaphor.

3.3.2. Piece architecture

This piece used the second setup of the Bio-Music platform, relying on the Biocontrol Systems unit and an EyesWeb data acquisition patch. EMG signals are then transferred in real-time through OSC protocol to a second computer running a slightly hacked version of the CataRT[10] application by Diemo Schwarz

[10] http://imtr.ircam.fr/index.php/CataRT

Figure 12: *Three dimensional sphere built from rotating disks, whose direction and speed of spin is dictated by the incoming physiological signals.*

[28]. Within this patch, a large database of samples are loaded, analysed and organised using psychoacoustic descriptors. The resulting sound units are laid on a two-dimensional descriptor space where the X axis represents noissines and the Y axis represents pitch. The EMG signals from each arm controls movement on one of these axes. The values from the EMG are dynamically scaled throughout the duration of the piece, allowing the performer to explore cluster areas of the sound corpus and giving a sense of structure and evolution to the piece.

3.4. Time Series; an uncontrolled experiment

This performance was presented by Hannah Drayson (experimenter/performer), Mitchel Benovoy (data acquisition) and Christian Mühl (subject) during our Bio-Music night session. The initial work on this piece evolved from an early investigation into capturing physiological data related to emotion and psychophysiology. The Thought Technology ProComp unit was used to capture six datasets consisting of five physiological signals, i.e. blood volume pulse, galvanic skin response, temperature, respiration and electrocardiogram. In order to stimulate various emotional states, the subject sat quietly and attempted to describe in writing a number of memories of experiences which correlated with the desired emotions. The texts were marked every 30 seconds to allow them to be aligned with the data recordings. The emotion datasets consisted of sadness, joy, anger, depression, calm and pain.

For this performance, we designed a Pure Data patch that transforms these emotion datasets into sounds and visuals. The final output of this patch was a three dimensional sphere built from rotating disks, whose direction and speed of spin was dictated by the incoming physiological signals. Each signal was also attached to a specific oscillator audio object, and as a group create a constantly evolving range of tones. The resulting visuals and sounds were projected and amplified, allowing the audience to experience the subjects data re-interpreted in real-time. Hannah Drayson presented a number of experiments on the subject, Christian Mühl. These were four light hearted "multimodal" sensory experiences, with the aim to provoke observable physiological change in the state of the subject, and to take advantage of the "non-scientific" experimental setup to try slightly absurd and fun combinations and stimuli.

These stimuli consisted of:

- Standing with your back to a room full of people while they all stare and point at the back of your head. The subject was asked to stand with his back to the room, and

the audience were invited to stare and point as hard as they could.

- Relaxing and eating cake. The subject was seated in a large comfortable armchair, and given a box of baklava to eat.
- Drinking beer and looking at kittens. The subject was given a cold can of beer to drink and asked to look at kittens, there was also a young cat sleeping on a chair in the performance space which was brought the the front to be looked at.
- Having all your arms and legs pulled by your friends. Members of the audience came forward and took hold of a limb each and pulled on the subject.

During the experiment changes in the auditory and visual signals were observed, one audience member remarked that at one point the sonification had sounded "a bit like a whale". Whilst the performance itself was lighthearted, it forms part of larger project and observation of the limits of human machine interaction, in that gestalt states such as emotion are easily observed by other humans, but extremely hard to define simply in terms of physiology. One interesting finding of this work is that highly abstract representations of a subject data could still reveal emotional states to human observers, and in fact give greater intimacy with physiological changes which are not overtly manifest in everyday contact.

4. CONCLUSION

In this paper we have surveyed some approaches which are possible in using biological signals as control sources to generate computer synthesis parameters for visual or auditory displays. We have also here presented some techniques, strategies and methods for using biological signals as motive forces for use in ongoing biologically generated visual, sound and music experiences. We feel we have introduced some innovative and useful techniques and expanded upon other established ones which will help give rise to nascent explorations as well as further ongoing research in this important yet less-travelled area of scholarship. We look forward to criticisms, additions and engagements from other interested parties who could help to further the goals of an open-source, free and functional system for the treatment of biological signals as meaningful drivers for sonification and visualisation processes.

Hopefully, we have also shown that there are diverse signal processing techniques which can be applied to raw biological signals, which, in due course, may help us to understand the semantics of their ontologies. The sonifications or visualisations of these signals may have the functions of scientific, medical, clinical or artistic expressions. The final manifestation is more a matter of end-use than method. We would hope that all potential methods for treating biological signals might be considered as useful to those who are engaged in this field, be their domain science, medicine or the arts.

5. ACKNOWLEDGEMENTS

This report, as well as the source code for the software developed during the project, is available online from the eNTERFACE'07 web site: http://www.enterface.net.

This research was partly funded by SIMILAR, the European Network of Excellence on Multimodal Interfaces, during the eNTERFACE07 Workshop in İstanbul, Turkey.

Rémy Lehembre is supported by a grant from the Belgian NSF(FRIA).

Cumhur Erkut's work is supported by the Academy of Finland (project 120583).

The authors would like to thank Prof. Bülent Sankur and all the eNTERFACE'07 organizing comittee for the excellent working conditions they have offered us. We are also grateful to Prof. Burak Güçlü for letting us use his EEG system. Many thanks to Mustafa Zahid Yildiz who spent some of his precious time helping us in our EEG experiments.

6. REFERENCES

[1] R. Laennec, "Traité de l'auscultation médiate et des maladies des poumons et du coeur", Paris J.S. Chaudé, 1826. 103

[2] H. Berger, "Über das Elektroenkephalogramm des Menschen", *Arch. f. Psychiat.*, vol. 87, pp. 527–570, 1929. 103

[3] A. Lucier and D. Simon, *Chambers : scores by Alvin Lucier ; interviews with the composer by Douglas Simon.* Wesleyan University Press, Middletown, Conneticut, USA, 1980. 103

[4] D. Rosenboom, *Biofeedback and the arts : results of early experiments*. Aesthetic Research Centre of Canada, Vancouver, Canada, 1976. 103

[5] B. Knapp and H. Lusted, "A Bioelectric Controller for Computer Music Applications", *Computer Music Journal*, vol. 14(1), pp. 42–47, 1990. 103

[6] B. Knapp and H. Lusted, "Controlling Computers with Neural Signals", *Scientific American*, pp. 82–87, 1996. 103

[7] A. Tanaka, *Trends in Gestural Control of Music*, ch. Musical Performance Practice on Sensor-based Instruments, pp. 389–406. IRCAM / Centre Pompidou, 2000. 103

[8] B. Arslan and al., "Biologically-driven musical instruments", in *Proceedings of the 1st summer workshop on Multimodal Interfaces (eNTERFACE'05),*, (Mons, Belgium), pp. 21–33, 2005. 103

[9] A. Brouse and al., "Instruments of sound and visual creation driven by biological signals", in *Proceedings of the 2nd summer workshop on Multimodal Interfaces (eNTERFACE'06),*, (Dubrovnik, Croatia), pp. 59–68, 2006. 103

[10] E. Miranda and M. Wanderley, *New Digital Musical Instruments: Control And Interaction Beyond the Keyboard (Computer Music and Digital Audio Series)*. AR Editions, Inc. Madison, WI, USA, 2006. 104

[11] G. Kramer, ed., *Auditory Display: Sonification, Audification, and Auditory Interfaces*. Reading, MA, USA: Addison Wesley, February 1994. 104

[12] T. Hermann and A. Hunt, "Guest Editors' Introduction: An Introduction to Interactive Sonification", *Multimedia, IEEE*, vol. 12, no. 2, pp. 20–24, 2005. 104

[13] T. Hermann and H. Ritter, "Model-based sonification revisited - Authors' comments on Hermann and Ritter, ICAD 2002", *ACM Trans. Appl. Percept.*, vol. 2, pp. 559–563, October 2005. 104

[14] A. de Campo, R. Hoeldrich, G. Eckel, and A. Wallisch, "New sonification tools for EEG data screening and monitoring", in *Proc. Intl. Conf. Auditory Display (ICAD)*, (Montreal, Canada), pp. 536–542, June 2007. 104

[15] G. Baier, T. Hermann, and U. Stephani, "Multi-channel sonification of human EEG", in *Proc. Intl. Conf. Auditory Display (ICAD)*, (Montreal, Canada), pp. 491–496, June 2007. 104

[16] T. Hinterberger and G. Baier, "Parametric orchestral sonification of EEG in real time", *Multimedia, IEEE*, vol. 12, no. 2, pp. 70–79, 2005. 104

[17] T. Hinterberger, N. Neumann, M. Pham, A. Kbler, A. Grether, N. Hofmayer, B. Wilhelm, H. Flor, and N. Birbaumer, "A multimodal brain-based feedback and communication system.", *Exp Brain Res*, vol. 154, pp. 521–526, February 2004. 104

[18] J. R. Wolpaw, N. Birbaumer, D. J. McFarland, G. Pfurtscheller, and T. M. Vaughan, "Brain-computer interfaces for communication and control", *Clin Neurophysiol*, vol. 113, pp. 767–791, 2002. 104, 106

[19] F. D. Nocera and F. Ferlazzo, "Resampling approach to statistical inference: bootstrapping from event-related potentials data.", *Behav Res Methods Instrum Comput*, vol. 32, pp. 111–119, 2000. 105

[20] C. Babiloni, F. Carducci, F. Cincotti, P. M. Rossini, C. Neuper, G. Pfurtscheller, and F. Babiloni, "Human movement-related potentials vs desynchronization of EEG alpha rhythm: a high-resolution EEG study", *Neuroimage*, vol. 10, pp. 658–665, 1999. 106

[21] L. Kauhanen, T. Palomäki, P. Jylänki, V. Aloise, M. Nuttin, and J. d. R. Millán, "Haptic Feedback Compared with Visual Feedback for BCI", *in Proceedings of 3rd internationnal conference on BCI*, vol. Graz, pp. 70–79, 2006. no.2. 106

[22] Cincotti and al., "Vibrotactile Feedback for Brain-Computer Interface Operation", *Computational Engineering and Neuroscience*, 2007. 106

[23] T. Hinterberger and G. Baier, "Parametric orchestral sonification of EEG in real time", *IEEE Multimedia*, vol. 12, pp. 70–79, 2005. no.2. 106

[24] F. Nijboer and al., "An auditory brain-computer interface (BCI)", *J. Neuroscience Methods, Vol. In Press, Corrected Proof*, 2007. 106

[25] C. Zywietz, "A Brief History of Electrocardiography-Progress through Technology", 2003. `http://www.openecg.net/Tutorials/A_Brief_History_of_Electrocardiography.pdf`. 108

[26] V. Shusterman and O. Barnea, "Sympathetic Nervous System Activity in Stress and Biofeedback Relaxation.", *IEEE Engineering in Medicine and Biology Magazine*, March/April 2005. 108

[27] T. Bisson, "They're Made Out Of Meat!", *Omni*, 1991. 111

[28] D. Schwarz, G. Beller, B. Verbrugghe, and S. Britton, "Real-Time Corpus-Based Concatenative Synthesis with CataRT", in *Proc. of the Int. Conf. on Digital Audio Effects (DAFx-06)*, (Montreal, Quebec, Canada), pp. 279–282, Sept. 18–20, 2006. 111

7. APPENDICES

7.1. Data Transport Protocol (OSC over UDP multicast)

The *Bio-Music* system is envisaged as potentially accommodating a wide variety of Data Acquisition (DAQ) hardware devices for the capture of biological signals which could then be sent to a wide variety of software (and potentially hardware) applications for further processing or display. Biosignal sensors require certain type of interfaces and software; in principle, the biosignals would be the only sources whilst other applications that consume this data would be sinks but could in fact also be sources of processed data analyses etc. Using OpenSoundControl over UDP multicast was a very specific design decision: practically it was not possible to run each sensor type with each computer during the workshop, not only because of operation system dependencies, but also because of computers data processing efficiency. In order to receive different data types from different sources, setting up multicast open sound control protocol server became an agreed alternative, which can maintain fast and direct connections between sender and receiver.

Open Sound Control (OSC) is a User Datagram Protocol (UDP) network based protocol for communication among computers with a low limitation on latency, which is optimized for modern networking technologies. It uses a URL style syntax to specify instructions or data for specific uses. OSC allows the user to define a namespace for their particular application purpose. OSC also supports Unix-style file globbing and pattern-matching mechanisms so that sophisticated and fine-grained selections can be made (similar to regular expressions). It is very network-friendly but is not tied to any one transport protocol, i.e. it could use TCP/IP as well as IEEE 1394/Firewire. Additionally, and very importantly, OSC is already supported by a wide range of hardware and software applications. Also, the bundle mechanism allows multiple messages to be sent with tone given time-tag so that all events can occur at the same time. Even through OSC protocol, users can interact clearly with the networked computers; however, regarding to the server-client communication type of the project, OSC based multicast server was chosen to be used as a communication protocol in local area network. The advantage of using multicast server is the possibility to transmit the same data to more than one host, and instead of broadcasting, only related host is receiving the message and the message is transmitted only once for many clients, which saves a lot of bandwidth in the network[11].

Building OSC multicast structure required defining the project name tags, which must to be specified by the sender and receiver for the requested data types. For the purpose of this project, we defined our own OSC namespace characterized thus: `/bio-music/#/source/#/type/#/`. As an example, a typical bio-music OSC message could look like this: `/bio-music/3/eeg/2/coherence/2 ...DATA... DATA...`, which would indicate that we have 2 channels of "`DATA`" from subject 3's EEG which is indicating coherence in their EEG. A Pure Data patch has been designed to automatically manage the formatting of OSC messages sent and received between each components of the *Bio-Music* platform (Fig.13).

All devices (host computers, servers, capture stations etc.) are connected to a local network, preferably using 100baseT or gigabit ethernet, which has UDP multicast enabled, as most LANs do. In principal, all nodes can act as sources and/or sinks although in most cases they will be one or the other. Some method of auto-discovery for the bio-music nodes is preferred (could likely be the ZeroConf protocol) such that namespace collisions can be avoided whilst also allowing bio-music nodes to discover each other and auto-configure themselves. Some possible bio-music nodes could include:

- biosignal source
- DSP analysis
- server stockage of data (recording)
- sonification or music output
- visualisation of visual artistic output

[11] http://www.tack.ch/multicast/

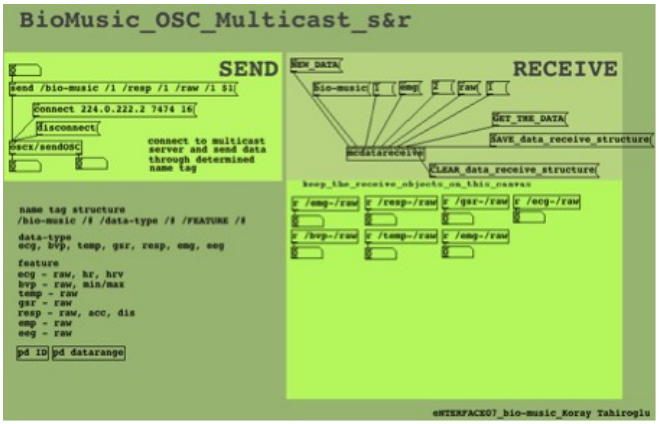


Figure 13: *Snapshot of the OSC Multicast Pure Data patch*

So, in a sense, all nodes on the bio-music network can provide and have access to any information simultaneously if they wish. Sampling rate of biosignals are fairly low (64 - 256 Hz typically) and thus it seems that issues of data collisions, timing and jitter on the network should not be significant if a fast and reliable ethernet LAN is used.

7.2. Biosignals datasets collection

A rich dataset was collected using GSR, BVP, ECG, phalange temperature and respiration signals and was extensively used to implement and tune the signals processing functions. Dataset generation was conducted in a quiet, comfortable lab environment. The subject either remained seated or standing and was instructed to limit his body movement to minimize motion artefacts in the collected signals. The biosignals were recorded using Thought Technologys ProComp Infiniti biofeedback system all sampled at 256 Hz. Two types of scenarios where recorded: emotional states and electrical muscle activity. The different emotional state data permits training and classifier validation of online affective state recognition. Emotions were elicited using self-generation such as event recollection and personally chosen music listening. As for the muscle activity, four sites were selected while a musician performed on an accordion: bicep, pectoral, forearm and calf muscle activity was measured in synchronous with video recording. Fig. 14 shows a report of a biosignals-based emotional state recording session achieved during the project:

Biosignals Dataset v.0.1
Date of collection: 23/07/07
Acquired by: Mitchel Benovoy
Environment: Air conditioned lab, noisy

- Subject
  - Gender: female
  - Age: 26
  - Background: English, artistic training, Ph.D student
- Acquisition interface
  - Hardware: Thought Technology ProCom Infinity
  - Software: Biograph Infinity
  - Signals acquired: GSR, BVP, ECG, Temp, Resp
  - Signal characteristics: raw
  - Sampling rate: subsampled at 256 Hz
- Number of trials: 6
  - Elicitation paradigm: Past event recollection (emotionnally charged), subject used hand writting as expressive aid.
  - Emotions/states captured: sadness, joy, anger, depression, relaxation, physical pain
- Trial specifics
  - Trial 1.xls: sadness
  - Trial 2.xls: joy
  - Trial 3.xls: anger
  - Trial 4.xls: depression
  - Trial 5.xls: relaxation (using breathing techniques)
  - Trial 6.xls: pain (cold water drinking with wisdom tooth pain)

File format: Excel worksheets. column separated.

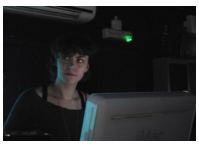
Figure 14: *Report of a biosignals-based emotional state recording session.*

8. BIOGRAPHIES

Mitchel Benovoy is a Ph.D student at the Centre for Intelligent Machines at McGill University in Montreal, Canada. He received his B.Sc.A in Electrical Engineering with Systems specialization at the University of Ottawa in 2006. His current work focuses on machine learning applied to human physiological signals. The primary application of this research is automatic and real-time emotion recognition using pattern recognition. He is also studying the link between the brain and the peripheral nervous system when emotions are experienced.
Email: benovoym@cim.mcgill.ca

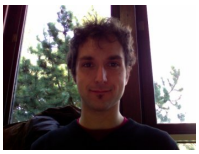
Andrew Brouse is a multidisciplinary musician, composer, artist, and technologist. He has worked in the contemporary intermedia arts and music for over fifteen years. He currently divides his time between the UK, Belgium and Montreal, Canada.
Email: andrew.brouse@uclouvain.be

Thomas Greg Corcoran was born in 1982 in Cork, Ireland. He received his BA Mod in Theoretical Physics (Hons) from Trinity College Dublin in 2006. He now works as an artist in Dublin. His current art research involves drawing, painting, and notions of graphical linguistics in the computational analysis and processing of drawings as datasets.
Email: thomascorcra@gmail.com



Hannah Drayson is an artist and doctoral candidate currently working within Trans-technology Research at Plymouth University in the UK. After receiving a BA hons in Critical Fine Art Practice at Brighton University in 2003, she continued her creative practice using a combination of theoretical research and media ranging from web and graphic design to visual performance, video and digital audio production. In 2006 she gained an MSc in Digital Art and Technology from Plymouth University. Her current research interests are the relationship between medical instrumentation and subjective perception of the body, and the history and philosophy of technology and medicine. http://www.trans-techresearch.net/ http://x2.i-dat.org/~hd/
Email: hannah.drayson@plymouth.ac.uk

Cumhur Erkut was born in İstanbul, Turkey, in 1969. He received B.Sc. and the M.Sc. degrees in electronics and communication engineering from the Yildiz Technical University, İstanbul, Turkey, in 1994 and 1997, respectively, and the Dr.Sc.(Tech.) degree in electrical engineering from the Helsinki University of Technology (TKK), Espoo, Finland, in 2002. Between 1998 and 2002, he worked as a researcher, and between 2002 and 2007 as a postdoctoral researcher at the Laboratory of Acoustics and Audio Signal Processing of the TKK, and contributed to various national and international research projects. 2007 onwards, he is appointed as an Academy Research Fellow, he conducts his research project Schema-SID [Academy of Finland, 120583] and contributes to the COST IC0601 Action "Sonic Interaction Design" (SID). His primary research interests are sonic interaction design, and physics-based sound synthesis and control.
Email: cumhur.erkut@tkk.fi



Jean-Julien Filatriau was born in Lille (France) in 1980. He received the Engineering degree from Telecom Lille I in June 2002, and graduated a MSc in Acoustics at University of Marseille in June 2004. Since October 2004, he is research assistant at the Communications and Remote Sensing Lab of Université Catholique de Louvain, Belgium (UCL-TELE, prof. Benoît Macq). His research interests include sound processing, computer music and gestural control of audio systems; he started in 2005 a Ph.D. Thesis at UCL, dealing with the synthesis and control of sonic textures. Sonic textures is a class of sounds, including most of the environmental sounds (waterfall, fire, wind etc...), characterized in terms of both microscopic and macroscopic features, and requiring a novel approach for digital synthesis. Furthermore, he was involved in the eNTERFACE'05, '06 and '07 summer workshops as a leader of projects related to the "physiologically-controlled musical interfaces". This research aimed at investigating the use of physiological signals analysis (EEG, EMG, EKG, EOG etc...) to perform and control sonic and visual process for both sonification and art performances oriented applications.
Email: jehan-julien.filatriau@uclouvain.be

Christian Frisson was born in Thionville, France, in 1982. He received his M. Eng. degree from Ecole Nationale Supérieure d'Ingénieurs du Mans (ENSIM) at Université du Maine, France, in 2005, where he studied acoustics and metrology; and graduated a M. Sc. entitled "Art, Science, Technology" from Institut National Polytechnique de Grenoble, France, in 2006. Since October 2006, he has been a PhD student with Prof. Benoît Macq at the Communication and Remote Sensing Lab (TELE) of Université catholique de Louvain (UCL), Belgium. His current research interests include the use of multimodal interfaces for the semi-automatic annotation of 1D signals (such as biological and audio signals).
Email: christian.frisson@uclouvain.be

Ben Knapp leads the Music, Sensors, and Emotion (MuSE) research group at SARC. His research at SARC focuses on the understanding and measurement of the physical gestures and emotional states of musical performers and their audience. For over 20 years, Ben has been researching and developing user-interfaces and software that enable the composer and performer to augment the physical control of a musical instrument with more direct neural interaction. From his first article on the BioMuse system, "BioController for Computer Music Applications" in the 1990 Computer Music Journal to his 1996 Scientific American article on "Controlling Computers using Neural Signals" to his 2006 presentation at the International Computer Music Conference introducing the concept of an Integral Music Controller (a generic class of controllers that use the direct measurement of motion and emotion to augment traditional methods of musical instrument control), Ben has focused on creating a user-aware interface based on the acquisition and real-time analysis of biometric signals. Ben received his Master of Science and Ph.D. in Electrical Engineering from Stanford University in the area of speech and sound coding for a multi-channel cochlear prosthesis. At the same time he collaborated with researchers at the Center for Computer Research in Music and Acoustics (CCRMA) and began developing novel musical interfaces based on physiological signal acquisition. Ben then became a faculty member and eventually Professor and Chairman of the Computer, Information, and Systems Engineering Department at San Jose State University. In 1999, he left academia to become a director at the consulting firm MOTO Development Group where he helped develop new human-computer interaction systems for companies such as Sony, Microsoft, Logitech, and Teledyne. Ben is also co-founder of BioControl Systems, a company that has commercialized some of his interface designs. In the Spring of 2006, he served as a Fulbright Senior Specialist at University College, Dublin where he is now a Senior Research Fellow working as part of a multi-disciplinary team of researchers helping elderly people live independent lives. Ben's most recent work at SARC includes collaboration with the HUMAINE (Human-Machine Interaction Network on Emotion) EU Network of Excellence where he is working with the group in Genoa to learn more about the emotional state of world-class violinists during performance.
Email: b.knapp@qub.ac.uk

Umut Gundogdu was born in Denizli, Turkey in 1980. He received his BS in Electronical Engineering from Istanbul University, in 2004. He is currently working as a research assistant at Istanbul University, Turkey and continues his education as a MS student at the same university. His interests include time-frequency analysis, statistical signal processing, and biomedical signals.
Email: gunumut@istanbul.edu.tr

Rémy Lehembre received the diploma in electrical engineering from the Université catholique de Louvain in June 2005. Since October 2005, he is "boursier FRIA", i.e. research fellow funded by the Fonds pour la formation à la Recherche dans l'Industrie et dans l'Agriculture, at the TELE lab in UCL. His main research area is the study of electroencephalograms (EEG) for different applications including Brain Computer Interfaces (BCI), brain pathologies, and also musical compositions based on EEG (Brain Computer Musical Interfaces : BCMI).
Email: remy.lehembre@uclouvain.be

Christian Mühl was born in Cottbus, Germany in 1979. He received his M.Sc. degree in cognitive science from the University of Osnabrueck, Germany in 2007. He is currently a Ph.D. student in the Human Media Interaction lab at the University of Twente (NL) under the supervision of Prof.Dr. Anton Nijholt. His research interest is the application of brain-computer interface technology for computer games. He is part of the BrainGain project which is dedicated to the development of BCI technology and applications in general.
Email: christian.muehl@uos.de

Miguel Angel Ortiz Pérez is a mexican composer and sound artist based in Belfast. His works range from pieces for solo instruments, chamber ensembles and orchestra to tape, electroacustic and installation works. He has composed for theatre and film, and has received numerous awards and schoolarships. Miguel graduated from the Conservatorio de las Rosas in Morelia, México under his mentor Eduardo Solís Marín before pursuing a Masters degree at the Sonic Arts Research Centre at Queen's University Belfast under the guidance of Ricardo Climent and Pedro Rebelo. He is currently a PhD candidate at SARC, focusing on spatialisation/sound diffusion as aesthetic means for composition and the use of biosignal interfaces for musical applications under the supervision of Professor Michael Alcorn and Dr. R. Benjamin Knapp.
Email: mortizperez01@qub.ac.uk

Alaattin Sayin was born in Eregli, Turkey in 1980. He received his BS degree from the Department of Electronic Engineering of İstanbul University, in 2004. He is currently a MS student at the Department of Biomedical Engineering, and working as a research assistant at Biomedical Equipment Technology Program, at İstanbul University. His current interests are time-frequency analysis, biomedical signal processing and brain-computer interface.
Email: sayina@istanbul.edu.tr

Mohammad Soleymani was born in Tehran-Iran in 1980. He received a B.Sc. and M.Sc. degrees in electrical engineering (Biomedical signal and image analysis) in 2006 from the University of Tehran. From 2000 to 2002, He was a signal processing programmer at the Paya software company. At those years he worked on optical mark reader software and non-cursive handwritten Persian OCR (optical character recognition) systems. He worked on functional magnetic resonance imaging group analysis during his masters. His current research interests include physiological signal analysis for brain computer interaction, emotion assessment, and multimedia affective analysis. He is currently working toward his Ph.D. in Computer Science at computer vision and multimedia Laboratory in the University of Geneva, Switzerland. http://vision.unige.ch/MMI/
Email: mohammad.soleymani@cui.unige.ch

Koray Tahiroğlu is a sound artist, performer, researcher and lecturer who grew up in İstanbul. He is a doctoral student and conducting research on modeling human observable musical activity at Media Lab TAIK. He is developing interactive performance systems with experimental musical instruments. He has been performing with these interactive systems at various sound art events. Since 2004, he has been also teaching courses at University of Art and Design Helsinki, Academy of Fine Arts in Helsinki, Finland, Music department and Visual Communication Department at İstanbul Bilgi University. These courses are aimed to give opportunity to anyone to learn and to experience how to make electronic music, organize sounds, mapping physical interaction, video processing and networking possibilities with open source environments. He is performing noise and electronic music collaborating with different sound artists and performers as well as with solo performances. Koray Tahirolu performed at various festivals in Finland, Turkey, Norway, Italy, Austria, Sweden and Canada. http://mlab.taik.fi/~korayt/
Email: koray.tahiroglu@taik.fi

USIMAGTOOL: AN OPEN SOURCE FREEWARE SOFTWARE FOR ULTRASOUND IMAGING AND ELASTOGRAPHY

Rubén Cárdenes-Almeida [1], *Antonio Tristán-Vega* [1], *Gonzalo Vegas-Sánchez-Ferrero* [1], *Santiago Aja-Fernández* [1], *Verónica García-Pérez* [1], *Emma Muñoz-Moreno* [1], *Rodrigo de Luis-García* [1], *Javier González-Fernández* [2], *Darío Sosa-Cabrera* [2], *Karl Krissian* [2], *Suzanne Kieffer* [3]

[1] LPI, University of Valladolid, Spain
[2] CTM, University of Las Palmas de Gran Canaria
[3] TELE Laboratory, Université catholique de Louvain, Louvain-la-Neuve, Belgium

ABSTRACT

UsimagTool will prepare specific software for the physician to change parameters for filtering and visualization in Ultrasound Medical Imaging in general and in Elastography in particular, being the first software tool for researchers and physicians to compute elastography with integrated algorithms and modular coding capabilities. It will be ready to implement in different ecographic systems. *UsimagTool* is based on C++, and VTK/ITK functions through a hidden layer, which means that participants may import their own functions and/or use the VTK/ITK functions.

KEYWORDS

Software – Medical Imaging – Ultrasound – Elastography

1. INTRODUCTION

The extraordinary growth experimented by the medical image processing field in the last years, has motivated the development of many algorithms and software packages for image processing. One of the most important set of algorithms for image processing can be found in the Insight Toolkit (ITK) [1] open source libraries for segmentation and registration of multidimensional data. At the same time, several software packages for medical image visualization and analysis have been developed using ITK and the visualization toolkit (VTK) [2] libraries. There are also some efforts to develop software that could be easy to modify by other researchers, for example the Medical Imaging Interaction Toolkit (MITK) [3], is intended to fill the gap between algorithms and the final user, providing interaction capabilities to construct clinical applications.

Therefore, many software packages for visualization and analysis of medical data are available for the research community. Among them we can find commercial and non commercial packages. Usually, the former ones are focused on specific applications or just on visualization being more robust and stable, whereas the latter ones often offer more features to the end user, and they are usually available as open source projects. Some of the non commercial packages equipped with graphic user interface (GUI), are 3D Slicer [4, 5], Julius [6], Osirix [7], ITK-SNAP [8], Medical Studio [9], MedInria [10], Amide [11] and FusionViewer [12], to cite some of them. Other important commercial software packages are Analyze [13], Amira [14] and 3D Doctor [15].

The requirements that we look for in a medical imaging software package for researchers can be summarized in the following list:

- Open source code: to be able for everyone to modify and reuse the source code.
- Efficiency, robust and fast: using a standard object oriented language such as C++.
- Modularity and flexibility for developers: in order to change or add functionalities as fast as possible.
- Multi-platform: able to run in many Operating systems to be useful for more people.
- Usability: provided with an easy to use GUI to interact as easy as possible with the end user.
- Documentation: it is important to have a well documented web site as well as a complete user and developer manuals or tutorials.

Among the non commercial packages mentioned before, Amide and FusionViewer do not accomplish all these features because they are not specially designed for developers, and they are not well documented for that, Osirix is only available for Mac OS, MedInria is not open source, Medical Studio is in an early stage of development, and only ITK-SNAP, Slicer and Julius accomplish almost all of the features mentioned, but their structure is relatively complex compared with the architecture of *UsimagTool*, thus they do not offer a flexible architecture for researchers. We propose here a package that fulfills all these requirements and at the same time is equipped with a set of algorithms for ultrasound (US) image processing, which makes this package one of the few software packages developed for US image processing.

2. INTERFACE

This section describes the new interface developed for *UsimagTool*, as well as the new ways for the interaction between the user and the tool.

2.1. General Structure

The general view of the *UsimagTool* interface appears in Figure 1, where it can be seen that the interface is divided into two main parts: the processing and configuration part and the visualization part. The first one is on the left-hand side of the application window, and it is divided into three subparts:

1. **Type of Processing**. It is a menu for selecting the type of processing (see Figure 2). When a type of processing is selected, multiple filters and algorithms related to the selected type appear in the options panel. These options are thought to be reconfigurable.

2. **Options Panel**. This panel presents the most usual algorithms for each processing type.
3. **Configuration Panel**. Here all the parameters of each filter can be selected, as well as the input/output origin or visualization destinations.

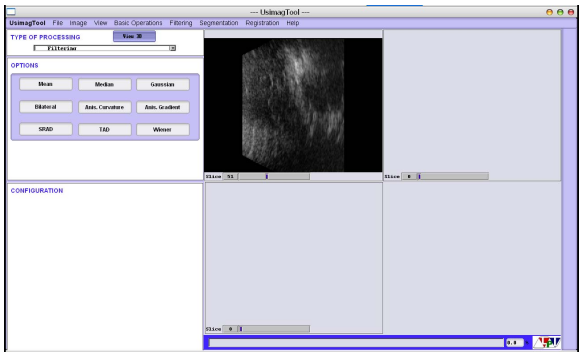


Figure 1: *General view of the UsimagTool graphic user interface*

The visualization part is on the right-hand side of the application window. There are multiple visualization options available: 3D Surface Rendering, Orthogonal visualization and Split visualization. Each visualization mode can be selected pressing the corresponding button.

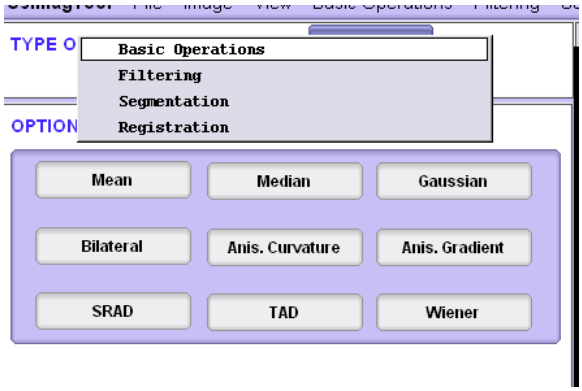


Figure 2: *Type Processing menu*

2.2. Menus

In order to provide an easy and fast access to all the basic operations as well as all the image processing algorithms available, *UsimagTool* is organized using different menus for different group of operations. Figure 3 shows the area of the interface where the menus are placed. Next, we briefly describe the structure and options of the different menus:

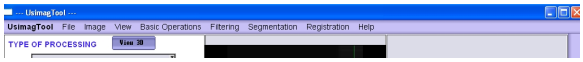


Figure 3: *Detail of the UsimagTool interface showing the different menus.*

1. *UsimagTool*: This menu allows to access to three general options of the program:
   - **About UsimagTool:** This operation is intended to show information about the program release and versions. It has not been implemented yet.
   - **Preferences:** Using this option, the user can access the general setting of the program, which will appear in the *Configuration Area*.
   - **Quit UsimagTool:** This option exits the program.
2. *File*: This menu gives access to file input and output operations:
   - **Open:** This option opens a selection window for choosing an image or volume file in several file formats: metaheader, raw data (in another menu item), voluson, DICOM, JPEG, PNG, TIFF.
   - **Open Raw:** This operation opens a selection window for raw data, and also another window for the introduction of the necessary parameters (see Figure 4).

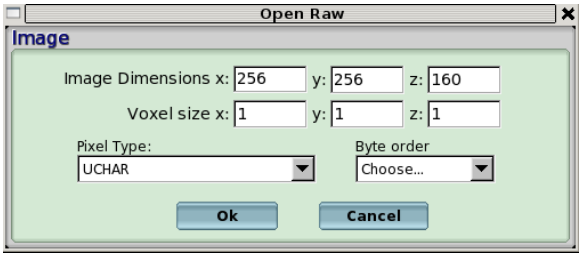


Figure 4: *Parameter window for the "Open Raw" operation.*

   - **Save:** Using this option, the current image in the selected viewer is saved. A window appears where the file name needs to be chosen. Metaheader, voluson, DICOM, JPEG, PNG and TIFF are supported.
   - **Save as:** This option is not yet implemented.
3. *Image*: This menu has been designed for easy access to some useful operations in the visualization of the images and volumes. Currently, these options are not accessible using this menu and can only be employed using keyboard buttons.
4. *View:* In this menu, three different visualization options can be selected:
   - **Split:** Four viewers allow the independent visualization of four 2D images. This is the default option.
   - **Orthogonal:** For a 3D volume, three orthogonal views are shown: sagittal, axial and coronal.
   - **Surface Rendering:** For a 3D surface polydata file in VTK, this option shows the surface in 3D. Zoom in, zoom out and 3D rotation can be then performed.
5. *Basic Operations:* This menu provides access to some basic image processing operations, such as addition, difference, multiplication (these three operations need two input images), erosion, dilation, gradient, rescaling, relabeling, absolute difference and inverse image. Some of these operations need the configuration of some parameters. If this is the case, a configuration window will appear to this end.

6. *Filtering:* Within this menu, different filtering algorithms are placed, from the most basic mean, median or Gaussian filtering to the most advanced bilateral or Wiener filtering. Besides, two more groups of filtering algorithms are implemented: Edge detection algorithms (zero-crossing edge detector and Canny detector), and anisotropic filtering (with five different variants).
7. *Segmentation:* This menu offers three different segmentation algorithms. First, a basic thresholding algorithm is proposed. Second, a level set segmentation algorithm has been implemented, and finally a supervised kNN algorithm can be employed. All three methods need different parameters to be tuned and therefore open a configuration panel in the configuration area.
8. *Registration:* Two registration algorithms are to be implemented in this menu. First, a demons registration algorithm (which opens a configuration panel). Second, a TPS (Thin Plate Splines), which is not implemented yet.
9. *Help:* This menu provides access to different help options, which are not implemented yet.

2.3. Configuration Area

As roughly indicated in Section 2.1, the left side of the application window is dedicated to configuration purposes. When a general type of processing is selected in the top part, the most common processing operations belonging to that type appear in the Options Panel, as was explained before and showed in Figure 2. Then, when a particular operation is chosen by pressing the corresponding button or by selecting it in the different operation menus explained in Section 2.2, the different configuration options of this particular processing operation will appear in the Configuration Panel. Besides these options, which depend on the particular choice of the operation, the user can also select which viewer will be selected as an input for the processing algorithm[1], and which viewer will be selected for the visualization of the algorithm output.

2.4. Visualization Area

The right part of the application window is the Visualization Area. *UsimagTool* provides three main visualization modes: Split Mode, Orthogonal Mode (not implemented yet) an 3D Rendering Mode. Depending on the visualization mode, which can be chosen in the right end of the application window, the Visualization Area changes its appearance.

2.4.1. Split Mode

Once the data is loaded, the image appears in one of four main viewers of the interface. These windows always show 2D images, but they are capable of representing the different orthogonal cuts of a volume (pressing the buttons 0,1,2 when the cursor is on them). Furthermore, there is a *slider* at the bottom of every viewer that allows to change the cut showed when the data is a volume. The visualization of the processed information can be done in any of the four main viewers.

The destination viewer for the output of every filter is controlled in the Configuration Panel.

There are different visualization operations that are available for each visualization window, and are accessible pressing the corresponding button as explained in Table 1.

[1]For certain types of algorithms, several images or volumes must be selected as inputs.

0	See cuts in the x axis
1	See cuts in the y axis
2	See cuts in the z axis (by default)
¡ ,	See the following cut
¡ ,	See the previous cut
r	Reset all the options
h	Help (this summary)
x	Flip x axis
y	Flip y axis
z	Flip z axis
q w	Diminishes, increases the top limit of the window in intensity
e	alternates the values of the top limit windows in intensity between *clipping* and *setting-to-black*
a s	Diminishes, increases the lower limit of the window in intensity
d	Alternates the values of the lower limit windows in intensity between *clipping* and *setting-to-black*
+ =	Zoom-in a factor 2
- _	Zoom-out a factor 2
i m j k	Moves the image int he directions up, down, left, and right.
t	Transposes the axis of the cut showed
A	See axis labels: P=posterior, L=left, S=superior
C	See crosses where the cursor was last clicked
I	See image values where the cursor was last clicked
T	Show or hide the clicked points
P	Change the coordinates between index and physical units
D	See image details superposed
O	See a superposed color
p	Save to a file the clicked points
l	Change the the screen data showing way Ways change cyclically in the following views: Intensity Inverse Log Derivative respect to x Derivative respect to y Derivative respect to z Mix with the previous and following cut MIP

Table 1: *Available options in the visualization windows*

2.4.2. Orthogonal Mode

This visualization mode is not implemented yet.

2.4.3. 3D Rendering Mode

This mode allows the visualization of a 3D surface. When it is selected, the corresponding data file can be loaded, and then the 3D surface is rendered. Zoom and rotation can be applied in this visualization mode (see Figure 5).

2.5. Points management

We have included a module for points management, useful in many situations such as landmark based registration, initialization of region growing of level set algorithms and other supervised algorithms like the k nearest neighbors algorithm. The functionalities included in this module are:

- Add point.

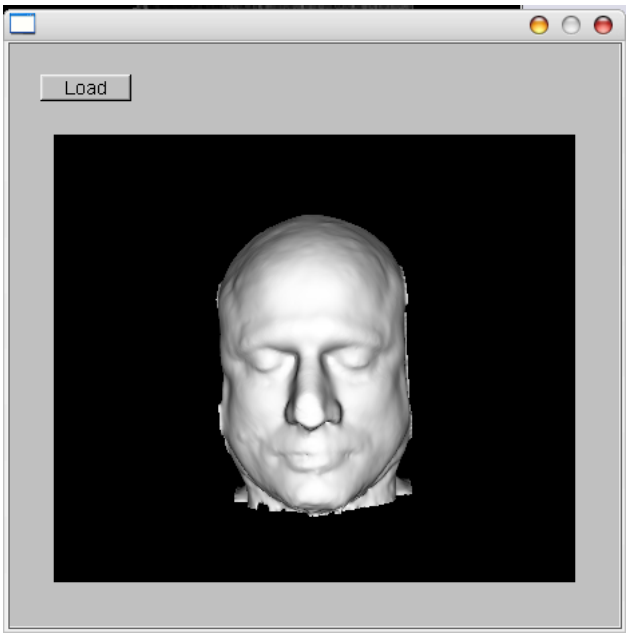


Figure 5: *Rendering volume window*

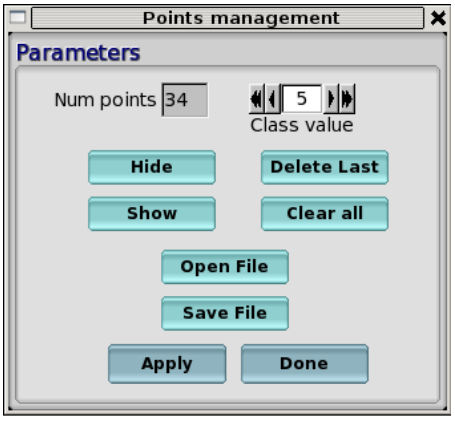


Figure 6: *Points management module*

- Open an existing points file.
- Save the points selected to a file.
- Clear all points.
- Delete last point.
- Change the id of the points to add.
- Show and hide the points in the viewer.
- List number of points selected.

2.6. Preliminary integration of an intuitive device for navigation

Successful image manipulation requires an intuitive user interaction. We propose an original navigation device which allows the user to interact with and manipulate items on screen through the use accelerometers and optical sensors. A wireless one-handed remote control is presented instead of the traditional cursor-based navigation of current medical consoles.

Figure 7: *Wii remote control device*

Nintendo's Wii Remote has been used as a low-cost prototype, see fig 7. It has the ability to sense acceleration along three axes through the use of an Analog Devices ADXL330 accelerometer. The Wii Remote also features an infrared (IR) optical sensor, allowing it to determine where the Wii Remote is pointing at if an external infrared bar is used. It includes eleven buttons for user-interaction. It communicates wirelessly with the PC via short-range (10 meters) Bluetooth radio.

A customized 'wiiuse' library [16] has been used. It is completely written in C and meant to be portable between Linux and Windows platforms. It has been preliminary tested on Linux hosts, making use of BlueZ bluetooth drivers that permit communication with Bluetooth HID-capable devices. It is capable of capturing button press events, motion sensing (tilt), and will include support for IR pointing in the future.

User interaction has been prototyped as follows: pitch and roll angles control Y and Z navigation, respectively. Zoom level is controlled by means of the Wiimote's cursor pad. Blue Light Emitting Diodes (LED) show different user modes or the active window.

3. WORKPACKAGES

3.1. Basic Operations

In order to support pre and/or post processing of the data images, some basic operations are included in this tool. They include gradient magnitude computation, addition and multiplication of images, relabeling of segmented images, connected components labeling, inversion of images, Canny edge detection filtering, and binary morphological operations (erosion, dilation, opening and closing). All of them are implemented using standard filters of ITK.

3.2. Segmentation

We include in *UsimagTool* some classic algorithms such as basic thresholding and level sets previously implemented in the ITK. The main algorithms implemented in our tool, that are not part of the official ITK release are the k nearest neighbors segmentation and a segmentation method based on a MRF model.

3.2.1. K nearest neighbors (kNN)

We have included the k nearest neighbors segmentation (kNN) algorithm in *UsimagTool*. The kNN algorithm is a supervised non parametric technique used for years mostly to segment Magnetic Resonance Images (MRI). Its main drawback was its computational complexity to find the neighbors in the feature space, but this issue was drastically reduced by using lookup tables based on high order Voronoi Diagrams, as proposed in [17] and [18]. The implementation used here is based on those publications, and the algorithm is coded in ITK, and it is also publicly available in [19]. As this is a supervised method, the user needs to initialize or to train the classifier, by clicking several pixels in the image, and assigning to each pixel the class it should belong to. For this reason we can use the point management module explained above for this training step. We have designed a wizard (see Figure 8) to guide a non expert user to use this algorithm. It consists of three windows: first window to choose parameters and input/output, the second window to choose between load existing points or train the classifier and the third one for landmark manipulation.

In Figure 9 we show the result of this algorithm applied to a MRI of a human brain.

3.2.2. Kidney segmentation

This is a semiautomatic method of segmentation of the kidney external surface in US images. The main objectives of this task are to improve the common measurements of the renal volume and the visualization and analysis of the organ shape. In the manual part the practitioner inserts six points that correspond to the main axes of the best-fit ellipsoid for the kidney surface [20]. This simple insertion is extended to those cases in which the whole kidney volume is not acquired, where the user must introduce twelve points that correspond, to the main axes of the visually estimated fit of an ellipse to the extremal and central kidney slice contours.

The automatic method is a model based segmentation one. It is a 3D extension of the method in [21]. It takes the points and builds an ellipsoid. Then it discretizes this surface by means of a cylindrical coordinate system induced parametrization. So the initial surface is described by a set of $S = P \times J$ points that correspond respectively to the set of cuts and rays of the model. This template surface is refined by allowing its deformation into a discrete space defined by the set $\Omega = \Lambda \times S$, where Λ refers to a set of possible radial deformations in the surroundings of the template surface.

Each of the elements in Ω define two regions in the volume image, namely a region of closeness and an inner region. The algorithm is going to construct the two external potential terms with the pixels included in these two regions. In our framework these terms contribute to the *a posteriori* likelihood terms. They are based respectively on an exponential distribution modeling of the image gradient and a beta distribution modeling of the image intensity. On the other side, from the locations of the possible deformation points and their neighbors, we can construct a predefined *a priori* model or the internal energy of the model that tries to maintain the smoothness of the surface by favoring those solutions in which the first and second finite radial deformation derivatives are small. Both trends are joined together into a Markov Random Field of deformations. The estimation of the relative importance of each of them as well as the estimation of the deformed surface are integrated in a Maximum a Posteriori framework [22, 23].

A result of a segmented contour in one slice of a healthy kidney is shown in figure 10. The code of the module is based on VTK and it is integrated with the rest of the functionalities in

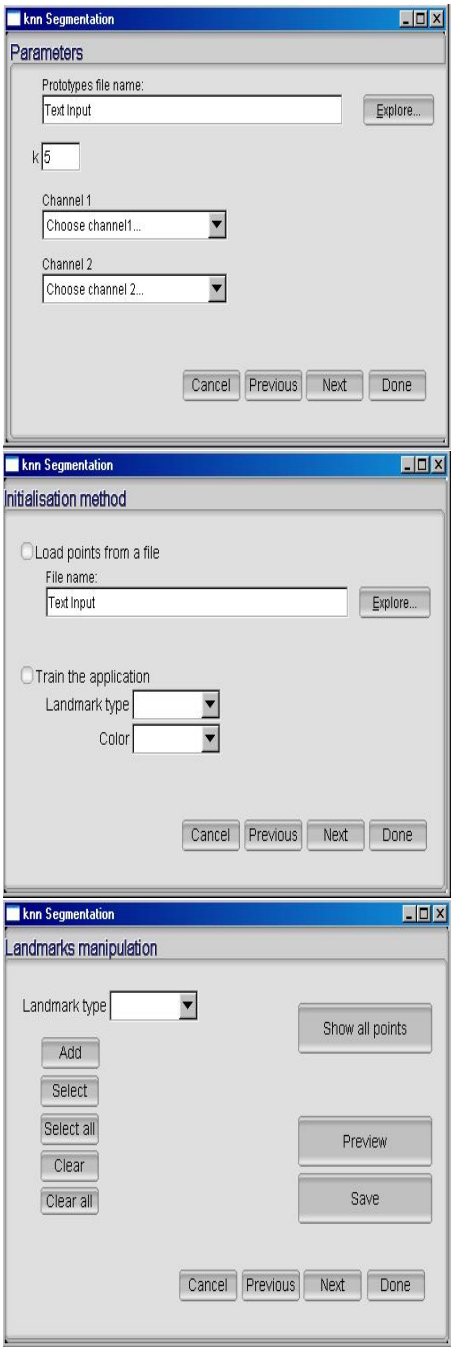


Figure 8: *Wizard windows*

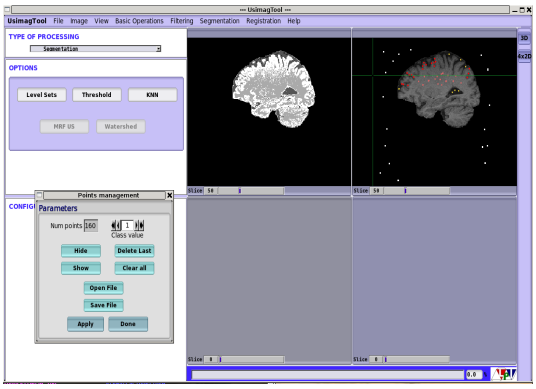


Figure 9: *kNN Segmentation in UsimagTool*

Figure 10: *2D kidney segmentation*

UsimagTool, although it is at this time a beta version. It seems clear for us that the method could be adapted for the segmentation of another abdominal organs as the liver or the prostate as pointed out in [21]. This could be interesting for taking a better advantage in the visual interpretation of 3D images acquired with modern echographic techniques.

3.3. Registration for Elastography

The Demons ITK algorithm has been added to the tool with general purpose. This workpackage is very important for the elastography intended part of *UsimagTool*, therefore own software has been implemented and modified to be adapted for elastography purposes as this modality has its specific needs. In this section, a brief description of elastography and the registration method used is described. This filter is in process to be added to the *UsimagTool*. This process will be finished after the eNTERFACE due to its difficult integration as it is not an ITK filter but a C++ module that needs to be compiled together with the main code.

Changes in tissue stiffness correlate with pathological phenomena that can aid the diagnosis of several diseases such as breast and prostate cancer. US elastography measures the elastic properties of soft tissues using US signals.

The standard way to estimate the displacement field from which researchers obtain the strain in elastography is the time-domain cross-correlation estimator (TDE). Optical flow (OF) methods have been also characterized and their use keeps increasing.

We have introduced the use of a modified demons algorithm (MDA) to estimate the displacement field. A least-squares strain estimator (LSE) is applied to estimate the strain from the displacement. The input for the algorithm comes from the US scanner standard video output; therefore, its clinical implementation is immediate. The resulting elastogram and the isoechoic input precompression image for the simulated phantom are presented in Figure 11. It is observed how an inclusion 3 times stiffer than the background can not be appreciated in the B-mode image of the experiment presented. However, at the elastogram, at the right of the same figure, the inclusion is clearly visible.

Although other researchers have used registration methods for elastography, as far as we know, it have not been used as stand alone but together with elastic modulus reconstruction or FE which iteratively varies material properties to improve registration. Our method does not requires any assumption about the isotropy or initialization for the algorithm, reducing the computational cost with respect to the other registration methods used for elastography. Compared to OF methods, we obtain a more regularized displacement field reducing the artifacts in the strain image. The advantage of these method based on the B-mode ultrasound image compared to those based on the RF A-lines (TDE) is the flexibility to include into commercial US scanner because the formerinput comes from the standard video output of the equipments. Decorrelation effects, mainly when dealing with pronounced strains, compromises the accuracy of the TDE methods. The obtained results have been already sent for its presentation in an international congress.

3.4. Filtering

Several classic filtering methods have been implemented in *UsimagTool* such as Gaussian, median, Wiener, or anisotropic filters, as well as a set of edge-detection techniques. Additionally we have developed new specific filters for US images: a Speckle Reducing Anisotropic filter (SRAD) [24], Detail Preserving Anisotropic Diffusion (DPAD) filter for speckle [25], a Tensor guided Anisotropic Diffusion filter (TAD) based on [26] and an anisotropic Wiener filter for speckle removal based on [22].

3.4.1. Classic filtering methods

The following well-known methods has been included in the current *UsimagTool* version:

- Gaussian filtering: smoothing through a convolution with a Gaussian kernel of variance σ^2.
- Median filtering: block-wise median filtering [27].
- Mean filtering: sample average in a neighborhood.
- Wiener filter: Classic implementation of the Wiener filter; i.e. the Linear Minimum Mean Squared Error (LMMSE) estimator assuming Gaussian additive noise [27].
- Anisotropic Diffusion filtering schemes [28, 29, 30, 26].
- Edge-detection techniques: Zero crossing edge detector and Canny filter [27].

3.4.2. Speckle Reducing Anisotropic Diffusion (SRAD) and Detail Preserving Anisotropic Diffusion (DPAD)

Anisotropic diffusion is a well known and widely used technique in image processing, since it offers a good compromise between noise removal and edge preservation; the original formulation by Perona and Malik [28] was based on the heat diffusion equation, where the gray level of the image, I, acts as the temperature that drives the heat conduction over the image.

This approach has shown a very good performance when dealing with additive noise. However, it seems to fail when working with US imaging, where the additive model does not hold. We use here the ideas proposed in [24], substituting this gradient-driven diffusion coefficient with a coefficient based on adaptive filtering theory; the coefficient of variation used in the LMMSE filtering for speckle removal [31, 32] is casted into the

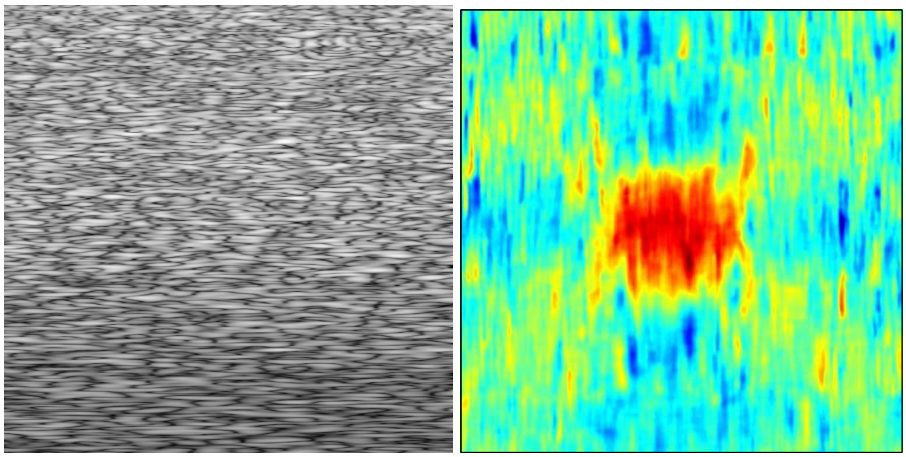


Figure 11: ***Left:*** *Precompression B-mode image.* ***Right:*** *Strain image resulting from the MDA registration.*

PDE environment, achieving a similar formulation to the original anisotropic one, but with the ability to cope with multiplicative noise.

Although the formulation of SRAD is deduced in [24] only for the 2D case, we have extended it to a 3D case. Moreover, given the templated nature of the ITK software, the same source code, with no modifications, is valid for both the 2D and 3D scenarios.

In [25] authors propose a more robust way to estimate the coefficient of variation in a SRAD scheme. This, together with a more complete solution produces an alternative scheme called Detail Preserving Anisotropic Diffusion (DPAD) filter. A new ITK class has been developed and included into *UsimagTool*.

3.4.3. Tensor Guided Anisotropic Diffusion Filtering (TAD)

One of the main issues with anisotropic diffusion filtering is the numeric scheme used to solve the PDE. The most usual way is the explicit implementation by means of finite differences: at a given iteration, the right hand side of the equation is computed given the current estimates of $I(x, y, t)$, and then the partial derivative in time is approximated by $\partial I/\partial t \simeq (I(x, y, t+\tau) - I(x, y, t))/\tau$ so that $I(x, y, t+\tau)$ can be easily found by forward differences as:

$$I(x, y, t + \tau) = I(x, y, t) + \tau \nabla \cdot (c\,(I(x, y, t)) \cdot \nabla I(x, y, t)) \quad (1)$$

The problem with this scheme is that it needs a very low value of τ (typically $\tau < 1/4$ in the 2D case) for the system to be stable, so many iterations are needed to converge to the steady state. The solution given in [26] is to use a semi-implicit scheme, which has been proved to unconditionally converge for any value of τ.

The method is said to be semi-implicit because one uses the gradient in the next time sample but the diffusion coefficient is assumed to be the one computed in the current iteration. Each step of the algorithm requires to solve a linear system of equations which is highly sparse, so a recursive algorithm may do this task with low computational cost [26].

On the other hand, traditional anisotropic diffusion is based on the fact that diffusion is simply stopped near relevant structures, so edges are preserved, but not enhanced. The approach of [33] generalizes the heat diffusion equation by the introduction of a diffusion tensor instead of the diffusion parameter, so $c(x)$ is a 2×2 matrix (in the 2D case); the eigenvectors of this matrix are set to be the same as those of the structure matrix of the image at this point. Regarding the eigenvalues, the one corresponding to the direction parallel to the contour is kept constant, meanwhile the eigenvalue along the orthogonal direction is also computed.

This way, near edge pixels there still exists a diffusion process, but it happens in a direction parallel to the contour. Far from the edges, the diffusion is isotropic to favor noise removal. This kind of filters not only preserve the contours, but they are able to enhance them as well. A semi-implicit implementation is derived once again to obtain an adequate computational behavior.

Although the work in [33] is intended for speckle removal, the problem is the same as in the case of scalar anisotropic diffusion: speckle is a multiplicative noise, so the additive model cannot be assumed. For this reason, we have adapted the algorithm to overcome the problem in a similar way to that introduced in [24] for the scalar case. The diffusion tensor is here computed based on local statistics specific for the US imaging case. The direction of the eigenvector parallel to the contour is found using the sticks model proposed by Czerwinski [34].

For the eigenvalues, a simple speckle classifier [35, 22] based on the moments of the homodyned k-distribution is used to discriminate fully formed speckle, where the diffusion is purely isotropic, and resolved speckle, where the diffusion is driven only following the sticks direction, which are ideally parallel to the contours of the US image.

3.4.4. Anisotropic Wiener Filter with Bias Correction

The Wiener filter, that is, the optimum linear filter in the minimum mean square error (LMMSE) allows to estimate a signal $\mathbf{X}$ from its noisy observation $\mathbf{Y} = \mathbf{X} + \mathbf{N}$, where N is the additive noise [27]. Wiener filter is known to be optimal under the assumption of Gaussian distributed data. However, this assumption is not valid for US images, that are better modeled by a Rice distribution [36]. For this reason, the direct application of Wiener filtering would introduce a bias in the results. Therefore, our Wiener filter implementation includes a previous bias correction step based on the Rice distribution. It is based on the fact that the SNR (Signal to Noise Ratio) for a Rice distribution is a monotonic function that depends only on one parameter $\gamma = \frac{s}{\sigma}$ [22], where S is the signal envelope and σ its variance. The SNR can be estimated from the observation by the method of the mo-

ments. Since the SNR function is monotonic, we can estimate the parameter γ from the estimated SNR value by means of the inverse function. On the other hand, the second order moment of the Rice distribution is given by $EY^2 = 2\sigma^2 + s^2 = s^2\frac{2+\gamma^2}{\gamma^2}$. Therefore, the signal envelope can be estimated by the eq. (2).

$$\bar{s} = \sqrt{\frac{\bar{\mathbf{Y}}^2\bar{\gamma}^2}{2+\bar{\gamma}^2}} \tag{2}$$

In such a way, we compute the bias corrected signal $Y'(n) = Y(n) - \bar{Y}(n) + \bar{s}(n)$ and the anisotropic Wiener filtering is applied over it. The resulting approach eliminates the bias, maintains the underlying signal structure unaltered and at the same time reduces the local variability palliating the speckle effect.

4. FUTURE OF *USIMAGTOOL*

Although *UsimagTool* is now in an advance stage, there is still too many work to do. We need to complete some of the tasks in the immediate future such as: integration of the elastography algorithms and other filtering algorithms developed during eNTERFACE 07, addition of vector image processing, complete the on-line help and the user and developers manual, and redesign some of the features of the GUI in order to make them more intuitive and easy to use. Another immediate task is the integration and testing of the wiimote control, in order to improve the interaction with the user.

Additionally we plan to restructure part of the main classes of the code to make it even more modular than now, and therefore more easy for a new user to include new functionalities. In the long term, we will include tensor data support: input and output using the tensor standard developed in similar, and also processing and visualization capabilities.

5. CONCLUSIONS

We can say that the work done in eNTERFACE 07, has been quite useful for the development of a new graphical user interface designed to help researchers in the development and use of algorithms for US image processing, that can be also used for other image modalities. The features of this tool makes it quite useful for researchers, due to the flexibility offered by its simple architecture, that allows to include new algorithms very fast. It is also important to highlight that this is an open source project which is available on line in *http://www.lpi.tel.uva.es/usimagtool*.

The work done in the redesign of the interface has been intensive from the whole eNTERFACE period, and the final result although not finished is quite satisfactory. We have improve the 3D viewer for surface rendering, and we improve the 2D viewers in order to include support for points management.

We have also developed some algorithms for US elastography although they are not included in *UsimagTool*. We have also implemented several algorithms not included in the official ITK release: one supervised segmentation method (kNN), one segmentation algorithm for US kidney based a MRF model, and the filtering algorithms: TAD, DPAD and SRAD, they all based on well known published works. We have also included in this tool many other well known algorithms from the ITK official release that are very useful for developers.

In order to improve interaction with the user we have studied the Wiimote control which provides many capabilities to the user and that we consider quite interesting also because it is a low cost device. We have developed a simple interaction mode with this device in a prototype with successful results, but more work is still needed.

We have also taken especial care in the multi-platform compatibility (windows and linux), so the same source code can be compiled and executed in both operating systems.

We can state that with all these accomplished tasks, the work done for *UsimagTool* has been quite satisfactory. We began with a prototype in a very early stage of development and now we can say that this tool has the features of an advanced software package for US image processing.

6. ACKNOWLEDGEMENTS

This work have been funded by CEE SIMILAR NoE (FP6-5076-09) and Spanish Government Project USIMAG (TEC2004-06647-C03-02).

7. REFERENCES

[1] L. Ibañez, W. Schroeder, L. Ng, and J. Cates, *The ITK Software Guide*. Kitware, Inc., second ed., 2005. ISBN 1-930934-15-7. http://www.itk.org/ItkSoftwareGuide.pdf. 117

[2] W. Schroeder, K. Martin, and W. Lorensen, *The Visualization Toolkit: An Object Oriented Approach to Computer Graphics*. Kitware, Inc, third ed., 2004. http://www.vtk.org. 117

[3] I. Wolf, M. Vetter, I. Wegner, T. Böttger, M. Nolden, M. Schöbinger, M. Hastenteufel, T. Kunert, and H. Meinzer, "The Medical Imaging Interaction Toolkit", *Medical Image Analysis*, vol. 9, pp. 594–604, 2005. 117

[4] MIT Artificial Intelligence Lab and the Surgical Planning Lab at Brigham & Women's Hospital, Harvard Medical School, *3D Slicer. 3D Slicer software*. http://www.slicer.org. 117

[5] D. Gering, A. Nabavi, R. Kikinis, W. E. L. Grimson, N. Hata, P. Everett, F. Jolesz, and W. W. III, "An Integrated Visualization System for Surgical Planning and Guidance using Fusion and Interventional Imaging", in *MICCAI*, (Cambridge, England), pp. 808–819, Sept 1999. 117

[6] Research center caesar, Bonn, Germany, *The Julius software platform*. http://www.julius.caesar.de. 117

[7] A. Rosset, L. Spadola, and O. Ratib, "OsiriX: An Open-Source Software for Navigating in Multidimensional DICOM Images", *J Digit Imaging*, vol. 17, pp. 205–216, Sep 2004. 117

[8] P. Yushkevich, J. Piven, H. Hazlett, R. Smith, S. Ho, J. Gee, and G. Gerig, "User-Guided 3D Active Contour Segmentation of Anatomical Structures: Significantly Improved Efficiency and Reliability", *Neuroimage*, vol. 31, no. 3, pp. 1116–1128, 2006. 117

[9] V. Nicolas, UCL, *MedicalStudio*. http://www.medicalstudio.org. 117

[10] P. Fillard, N. Toussaint, and X. Pennec, "MedINRIA: DT-MRI Processing and Visualization Software", November 2006. Guest paper at the Similar Tensor Workshop, Las Palmas, Spain. 117

[11] A. Loening and S. Gambhir, "AMIDE: a free software tool for multimodality medical image analysis", *Molecular Imaging*, vol. 2, no. 3, pp. 131–137, 2003. 117

[12] Y. Lu, S. Pathak, L. Gong, J. Goldschneider, and P. Kinahan, "Open Source Medical Image Fusion Display". http://fusionviewer.sourceforge.net. 117

[13] Mayo Clinic, *ANALYZE software*. http://www.mayo.edu/bir/Software/Analyze/Analyze.html. 117

[14] Mercury Computer Systems Inc., *Amira*. http://www.amiravis.com. 117

[15] Able Software Corp., *3D Doctor*. http://www.ablesw.com/3d-doctor. 117

[16] "C Wiimote Library". http://www.wiiuse.net. 120

[17] S. Warfield, "Fast k-NN Classification for Multichannel Image Data", *Pattern Recognition Letters*, vol. 17(7), pp. 713–721, 1996. 121

[18] O. Cuisenaire and B. Macq, "Fast k-NN Classification with an Optimal k-Distance Transformation Algorithm", *Proc. 10th European Signal Processing Conf.*, pp. 1365–1368, 2000. 121

[19] R. Cárdenes, M. R. Sánchez, and J. Ruiz-Alzola, "Computational Geometry Computation and KNN Segmentation in ITK", *Insight Journal, ISC/NA-MIC Workshop on Open Science at MICCAI 2006*, Oct 2006. http://hdl.handle.net/1926/204. 121

[20] J. Bakker, M. Olree, R. Kaatee, E. de Lange, K. Moons, J. Beutler, and F. J. Beek, "Renal volume measurements: accuracy and repeatability of US compared with that of MR imaging", *Radiology*, vol. 211, pp. 623–628, 1999. 121

[21] M. Martin and C. Alberola, "An approach for contour detection of human kidneys from ultrasound images using Markov random fields and active contours", *Medical Image Analysis*, vol. 9, no. 1, pp. 1–23, 2005. 121, 122

[22] M. Martin-Fernandez, E. Muñoz-Moreno, and C. Alberola-Lopez, "A Speckle Removal Filter based on an Anisotropic Wiener Filter and the Rice Distribution", in *IEEE International Ultrasonics Symposium*, (Vancouver, Canad), pp. 1694–1697, Oct 2006. 121, 122, 123

[23] S. Lakshmanan and H. Derin, "Simultaneous parameter estimation and segmentation of Gibbs Random Fields using Simulated Annealing", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 11, pp. 799–813, Aug. 1989. 121

[24] Y. Yongjian and S. Acton, "Speckle reducing anisotropic diffusion", *IEEE Transactions on Image Processing*, vol. 11, pp. 1260–1270, Nov 2002. 122, 123

[25] S. Aja-Fernandez and C. Alberola-Lopez, "On the Estimation of the Coefficient of Variation for Anisotropic Diffusion Speckle Filtering", *IEEE Transactions on Image Processing*, vol. 15, no. 9, pp. 2694–2701, 2006. 122, 123

[26] J. Weickert, B. ter Haar Romeny, and M. Viergever, "Efficient and reliable schemes for nonlinear diffusion filtering", *IEEE Transactions on Image Processing*, vol. 7, pp. 398–410, Mar. 1998. 122, 123

[27] J. Lim, *Two Dimensional Signal and Image Processing*. Prentice Hall, 1990. 122, 123

[28] J. M. P. Perona, "Scale-space and edge detection using anisotropic diffusion", *PAMI*, vol. 12, pp. 629–639, July 1990. 122

[29] R. K. G. Gerig, O. Käbler and F. Jolesz, "Nonlinear anisotropic filtering of MRI data", *IEEE Transactions on Medical Imaging*, vol. 11, pp. 221–232, June 1992. 122

[30] F. Catté, P. Lions, J. Morel, and T. Coll, "Image selective smoothing and edge detection by nonlinear diffusion", *SIAM J. Numer. Anal.*, vol. 29, pp. 182–193, 1992. 122

[31] J. S. Lee, "Digital image enhancement and noise filtering by using local statistics", *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 2, no. 1, 1980. 122

[32] V. S. Frost, J. A. Stiles, K. S. Shanmugan, and J. C. Holtzman, "A model for radar images and its application to adaptive filtering of multiplicative noise", *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 4, no. 1, pp. 157–165, 1982. 122

[33] K. Z. Abd-Elmoniem, A. M. Youssef, and Y. M. Kadah, "Real-Time Speckle Reduction and Coherence Enhancement in Ultrasound Imaging via Nonlinear Anisotropic Diffusion", *IEEE Trans. on Biomedical Engineering*, vol. 49, no. 9, pp. 997–1014, 2002. 123

[34] R. N. Czerwinski, D. L. Jones, and W. J. O'Brien, "Line and Boundary Detection in Speckle Images", *IEEE Trans. on Image Processing*, vol. 7, no. 12, pp. 1700–1714, 1998. 123

[35] R. W. Prager, A. H. Gee, G. M. Treece, and L. H. Berman, "Decompression and speckle detection for ultrasound images using the homodyned k-distribution", *Pattern Recognition Letters*, vol. 24, pp. 705–713, 2003. 123

[36] T. A. Tuthill, R. H. Sperry, and K. J. Parker, "Deviation from Rayleigh Statistics in Ultrasonic Speckle", *Ultrasonic Imaging*, vol. 10, pp. 81–89, 1988. 123

8. BIOGRAPHIES

Rubén Cárdenes-Almeida received his M.S. degree in physics from the University of Zaragoza in 1998, and he received his PhD degree in 2004 from the University of Las Palmas de Gran Canaria (ULPGC), at the Telecomunication Engineering School. He has been teaching and working in several research projects in the Center for Technology in Medicine (CTM), at the ULPGC, and he is currently working as a postdoc in the University of Valladolid, in the Laboratory of Image Processing (LPI), where he participates in various national and international research projects since March 2006. From June 2000 he also collaborates with the Surgical Planning Lab (SPL), at Brigham & Women's Hospital, (Harvard Medical School), where he has been invited researcher several times. His PhD thesis and current research work are related to medical image processing and discrete computational geometry.
Email: ruben@lpi.tel.uva.es

Antonio Tristán-Vega received his M.E. degree on Telecommunications engineering in 2004, at the University of Valladolid. He is currently a PHD student at this same University, and participates in a number of national and regional research projects as a member of the Laboratory of Image Processing (LPI). His research interests comprise multimodal and interpatient image fusion, image registration, and computer aided diagnosis based on medical imaging.
Email: atrive@lpi.tel.uva.es

Gonzalo Vegas-Sánchez-Ferrero is a PhD student at University of Valladolid, Spain. He received his Engineering of Telecommunication degree from that university in October 2006. He is a research affiliate at the Laboratory of Image Processing (LPI), Valladolid, Spain. Currently is working in tensor field processing in resonance imaging of the left-ventricular endocardium and epicardium of the heart.
Email: gvegsan@lpi.tel.uva.es

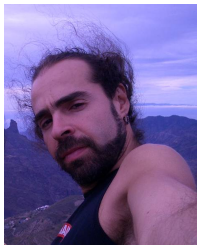
Santiago Aja-Fernández is currently an Assistant Professor with the ETSI Telecomunicación, the University of Valladolid, and is also with the Laboratory of Image Processing (LPI) at the same university. He received the Engineering of Telecomunication and Ph.D. degrees, both from the University of Valladolid, Valladolid, Spain, in 1999 and 2003 respectively. His research interests are fuzzy techniques for image and signal processing. In 2006 he was been awarded with a Fulbright Scholarship for a one year stay as a Research Fellow in the Laboratory of Mathematics in Imaging at Brigham and Women's Hospital (Boston).
Email: sanaja@tel.uva.es

Verónica García-Pérez received her Master Engineering in Computing Science degree from the Computer Science School, at the University of Valladolid, Spain, in 2004. Currently, she is PhD student in the Telecommunications School, in the signal processing area, where she also participates in a number of national and international research projects. She is a research fellow in the Laboratory of Image Processing (LPI) at this same University. Her main research interests comprise fuzzy systems, virtual reality and surgical simulation.
Email: veronica@lpi.tel.uva.es

Emma Muñoz-Moreno received her M.S. degree in Electronical Engineering from the University of Valladolid in 2003, and currently she is a PhD student at the same received University. She has participated in several national and international research projects as member of the Image Processing Laboratory at the University of Valladolid. Her research work is related to medical image processing, specially registration algorithms of DT-MRI and surgical simulation.
Email: emunmor@lpi.tel.uva.es

Rodrigo de Luis-García graduated in Telecommunications Engineering from the University of Valladolid in 2002, and received his PhD degree in Signal Processing at the Laboratory of Image Processing (LPI) from the same university in 2007. Most of his research activity focuses on the development of novel segmentation techniques for Medical Imaging applications. His work has been published in several international journals and conferences. Currently, he is an assistant professor at the Department of Signal Processing and Communications of the University in Valladolid, and has been recently awarded with a posdoctoral Fulbright scholarship.
Email: rluigar@lpi.tel.uva.es

Javier González-Fernández received the Degree in Industrial Engineering in 2000 from the University of Las Palmas de Gran Canaria (ULPGC), Spain. He is now a PhD candidate at the Center for Technology in Medicine, ULPGC. His research interests include Ultrasound Elastography, hardware design for biomedical embedded systems, and DSP.
Email: jgonzalez@ctm.ulpgc.es



Darío Sosa-Cabrera received the Mechanical Engineer degree from the University of Las Palmas de Gran Canaria (2001), is currently a PhD student in the Telecommunications Engineering School, focusing on US Elastography. He has studied in France at IFMA. After working in Latvia, he has coursed a Management of Technology Master at UTSA and joined the Department of Radiology, University of Texas, Houston, for one semester.
Email: dario@ctm.ulpgc.es

Karl Krissian received his M.S. in computer science and artificial intelligence from the university of Paris XI and the Ecole Normale Supérieure de Cachan in 1996. He received his Ph.D. degree in Computer Vision from the university of Nice-Sophia Antipolis and the INRIA in 2000. His main research topics are filtering and segmentation of three-dimensional vascular images using Partial Derivative Equations. He is currently working at the university of Las Palmas of Gran Canaria in Spain.
Email: krissian@dis.ulpgc.es

Suzanne Kieffer was born in Thionville, France, in 1977. She graduated in Computer Science at the University of Metz, France, in 1998 and she obtained MPhil in Computer Science in 2001. She defended her PhD thesis entitled "Multimodal assistance in the visual exploration of 2D interactive visualisations" in 2005 at the Université Henri Poincaré of Nancy, France. At the present time she is post-doctoral researcher at the Université Catholique de Louvain at the Communication and Remote Sensing Laboratory (TELE) and she is also member of the Belgium Computer-Human Interaction Laboratory (BCHI). Her research topics are focused on Human-Computer Interaction (HCI). In particular, they are related to Ergonomics, Graphical User Interfaces (GUI), Interaction Techniques and Usability Evaluation of Multimodal Interfaces.
Email: suzanne.kieffer@uclouvain.be

RAMCESS FRAMEWORK 2.0 REALTIME AND ACCURATE MUSICAL CONTROL OF EXPRESSION IN SINGING SYNTHESIS

Nicolas d'Alessandro [1], *Onur Babacan* [2], *Barış Bozkurt* [2], *Thomas Dubuisson* [1], *Andre Holzapfel* [3], *Loïc Kessous* [4], *Alexis Moinet* [1], *Maxime Vlieghe* [1]

[1] Signal Processing Laboratory, Polytechnic Faculty of Mons (Belgium)
[2] Electrical and Electronics Engineering Dpt, Izmir Institute of Technology (Turkey)
[3] Computer Science Dpt, University of Crete, Heraklion (Greece)
[4] Speech, Language and Hearing, Tel Aviv University (Israel)

ABSTRACT

In this paper we present the investigations realized in the context of the eNTERFACE 3rd summer worshop on multimodal interfaces. It concerns the development of a new release of the RAMCESS framework (2.x), preserving the accurate and expressive control of voice quality dimensions available in the existing system (from 2006), but adding coarticulated speech possibilities. This work will be achieved by the analysis, separation and modeling of the glottal source component from a limited-size and adapted singing voice database.

1. INTRODUCTION

Expressivity is nowadays one of the most challenging topics studied by researchers in speech processing. Indeed, recent synthesizers provide acceptable speech in term of intelligibility and naturalness but the need to improve human/computer interactions has brought researchers to develop systems that present more "human", more expressive skills.

Speech synthesis research seems to converge towards applications where multiple databases are recorded, corresponding to a certain number of labelled expressions (e.g. happy, sad, angry, etc.). At synthesis time, the expression of the virtual speaker is set by choosing the units in the corresponding database, then used in well-known unit selection algorithms.

Recently remarkable achievements have also been reached in singing voice synthesis. We can highlight naturalness and flexibility of Bonada *et al.* [1] algorithms where singing frames are structured at a high performance level. These kinds of technologies seem mature enough to allow for the replacement of human vocals with synthetic, at least for backing vocals.

However existing singing systems suffer from two restrictions: they are aiming at mimicking singers rather than offering real creative voice timbre possibilities, and they are generally limited to note-based MIDI controllers.

In this context, we propose to investigate an original option. We postulate that, even if the use of databases is strategic in order to preserve naturalness, voice modeling has to reach a higher level. These improvements have to meat particular needs, such as more realistic glottal source/vocal tract estimation, manipulation of voice quality features at a perceptual and performance level, and strong real-time abilities. The other issue concerns mapping strategies that have to be implemented in order to optimize the performer/synthesizer relation.

In this paper we present the work that have been achieved during the 3rd eNTERFACE workshop. This work is in the continuity of both 2005 and 2006 work about voice quality manipulation. Thus, after a short introduction of the mixed-phase model of voice production (cf. section 2), we present the existing framework, called RAMCESS 1.x, which was used as a starting point for our investigations (cf. section 3). Then we describe the analysis routines, implemented in order to separate glottal source and vocal tract contributions on a given database (cf. section 4). We also give a comment about the use of SDIF encoding (cf. section 5). Finally we give an overview of the RAMCESS 2.x framework (cf. section 6).

2. MIXED-PHASE MODEL OF VOICE PRODUCTION

The mixed-phase speech model [2] [3] is based on the assumption that speech is obtained by convolving an anti-causal and stable source signal with a causal and stable vocal tract filter. The speech signal is a mixed-phase signal obtained by exciting a minimum-phase system (vocal tract system) by a maximum-phase signal (glottal source signal). (It should be noted that the return phase component of the glottal source signal is included in the vocal tract component since it also has minimum-phase characteristics.) The mixed-phase model assumes that speech signals have two types of resonances; anti-causal resonances of the glottal source signal and causal resonances of the vocal tract filter.

The mixed-phase modeling plays an important role in both analysis and synthesis in our study. In analysis, estimation of glottal flow is achieved by ZZT (Zeros of Z-Transform) decomposition (cf. section 4) which decomposes the signal into anti-causal and causal components. In synthesis, mixed-phase signals are synthesized by exciting minimum phase filters with maximum phase excitation (cf. sections 3 and 6). These result in achieving a natural timber for the output sound which is very important for a digital instrument.

3. RAMCESS 1.X SYNTHESIS ENGINE

In this section we describe the existing framework, developped during preceding eNTERFACE workshops. This project, called RAMCESS (i.e. *Realtime and Accurate Musical Control of Expression in Singing Synthesis*) can be used as an original tool for studies on voice quality, analysis by synthesis or by gesture, and the performance of high quality singing voice. This section exposes new elements of the synthesis software library (which becomes progressively a collaborative project, called VQCLIB [4]), improvements achieved in the modeling of glottal source and vocal tract in order to preserve expressivity and achieve realtime production, and a comment about dimensionnal control of voice quality.

3.1. Voice Quality Control Library

The *Voice Quality Control Library* (or VQCLIB) [4] is a collaborative project that aims at developing a large set of modules for realtime environments (at this time MAX/MSP, and PURE DATA in the pipeline), focusing on the accurate and expressive realtime control of voice timbre features. Indeed voice production patches can be implemented step-by-step, opening large possibilities on glottal signal configuration, vocal tract filtering, or multiple representations of voice signal features. This initiative aims at attracting scientific research or computer music communauties in interactive forums, in order to share knowledges and know-how about realtime voice synthesis. The first version has been released, and a major update is in the pipeline. VQCLIB serves now as the basis of all further investigations on realtime voice synthesis. An example patch is illustrated in the Figure 1.

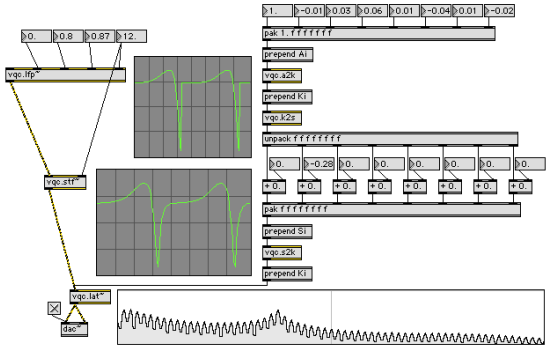


Figure 1: *Example of* MAX/MSP *patch using several objects of* VQCLIB.

3.2. Computation of the Glottal Flow Signal

Glottal pulse signal can be synthesized with many different models. Typical temporal and spectral representations of one period of the derivative of the glottal flow (with usual descriptive parameters: T_0, GCI, O_q, α_m and T_L) are illustrated in Figure 2. In term of flexibility and quality, we can particularly highlight LF [5] and CALM [6] models. However none of them is really suitable for realtime processing. On the one hand, LF parameters are the solution of a system of 2 implicit equations[1] which is known to be unstable. On the other hand, CALM is linear filter processing but one of the filters has to be computed anticausally. This is possible in realtime but with a limited flexibility [7].

The improvement that we propose can be seen as a compromise between both LF and CALM models, or a kind of *spectrally enhanced LF model*. In order to avoid the resolution of implicit equations, only the left part of the LF model is used. It is computed using the left part (cf. equation (1)) of the normalized glottal flow model (GFM) described in [8].

$$n_g(t) = \frac{1 + e^{at}(a\frac{\alpha_m}{\pi} sin(\pi t/\alpha_m) - cos(\pi t/\alpha_m))}{1 + e^{a\alpha_m}} \quad (1)$$

where t evolves between 0 and 1, and is sampled in order to generate the $O_q \times \frac{F_s}{F_0}$ samples of the opened phase (O_q : open quotient, F_0 : fundamental frequency, F_s : sampling rate); α_m

[1]The LF model gives equations of temporal shapes of both curves on the left and the right of the GCI. The conditions are then 1) the integration of the whole period has to be 0, and 2) left and right curves have to be connected at the position of the GCI.

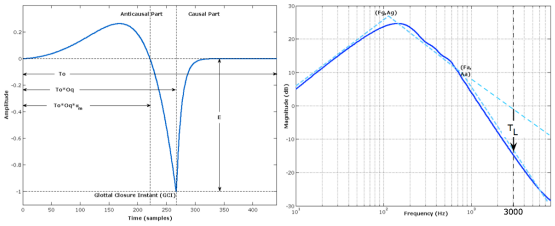


Figure 2: *Temporal (left) and spectral (right) representation of the derivative of one period of glottal flow, with usual descriptive parameters: T_0 (fundamental period), GCI (glottal closure instant), O_q (open quotient), α_m (asymetry coefficient) and T_L (spectral tilt).*

is the asymetry coefficient and $a = f(\alpha_m)$ is the pre-processed buffer of solutions of the equation (2).

$$1 + e^a(a\frac{\alpha_m}{\pi} sin(\frac{\pi}{\alpha_m} - cos(\frac{\pi}{\alpha_m}))) \quad (2)$$

Then the right part (the return phase) is generated in spectral domain, which means that the left LF pulse is filtererd by the spectral tilt low-pass first order filter presented in [6]. This option is also prefered because a long filter-generated return phase smoothly overlaps with following pulses, thus avoiding discontinuities. The complete process, also integrating the derivation of the pulse and the normalization (in order to control separately spectral tilt and energy of the pulse), is illustrated in Figure 3.

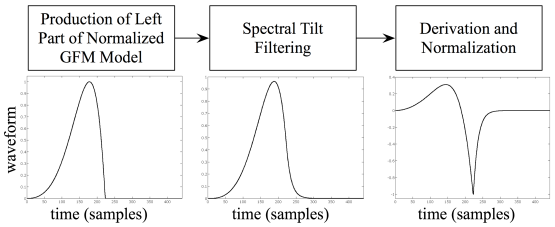


Figure 3: *Synthesis of the glottal pulse by combination of LF left part time-domain generation, spectral tilt filtering, derivation and normalization.*

3.3. Computation of the Vocal Tract Filter

The vocal tract is computed with a simple tube model. LPC coefficents a_i are converted into reflection coefficients k_i, and then into area (or section) coefficients S_i, defining geometrical properties of vocal tract. A complete coefficient conversion framework have been developed in order to jointly manipulate multiple representations (spectral and physical) of the vocal tract. This approach is powerful in order to create typical voice quality effects: vowel interpolation, obstructions, singer formant, etc [7]. A representation of the vocal tract by its sections (S_i) is illustrated in Figure 4.

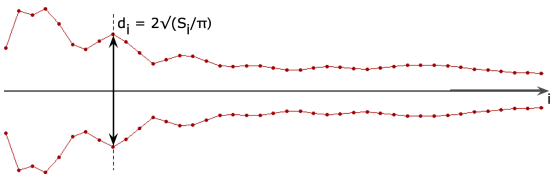


Figure 4: *Geometrical representation of the vocal tract, thanks to its S_i.*

3.4. Dimensional Study of Voice Quality

On the top of typical synthesis parameters that VQCLib can manipulate (F_0, O_q, α_m, T_l, spectral and geometrical features of vocal tract), it is interesting to build a layer which is able to provide more perception-based control to the performer. This dimensional study of voice quality have been achieved, resulting in a set of dimensions and their corresponding mapping equations with synthesis parameters [7]. As an example, we describe here the way we define the behavior of the open quotient (O_q) as a function of *tenseness* (T), *vocal effort* (V) and *registers* (M_i) in equations (3), (4) and (5).

$$O_q = O_{q_0} + \Delta O_q \tag{3}$$

$$O_{q_0} = \begin{cases} 0,8 - 0,4\,V & \text{if } M_i = M_1 \text{ (modal)} \\ 1 - 0,5\,V & \text{if } M_i = M_2 \text{ (chest)} \end{cases} \tag{4}$$

$$\Delta O_q = \begin{cases} (1 - 2\,T)\,O_{q_0} + 0,8\,T - 0,4 & \text{if } T \leq 0,5 \\ (2\,T - 1)\,O_{q_0} + 2\,T + 1 & \text{if } T > 0,5 \end{cases} \tag{5}$$

4. EXPRESSIVE VOICE ANALYSIS

This section describes investigations that have been realized in order to work on database contents instead of pure rule-based modeling. Different steps are described. First we discuss choices that have been made in the design and the segmentation of the first database. Then we propose the use of a sequence of different algorithms (ZZT, LF fitting, ARX-LF and spectral compensation) in order to achieve a high-quality decomposition of speech signals in its glottal source and vocal tract contributions, and the reliable modeling of these contributions. Finally we propose a suitable format, called SDIF for the encoding of the parameters of these models.

4.1. Database Design

In order to achieve coarticulated speech synthesis, we need to record a database containing at least one instance of each diphone (i.e. a two-phoneme combination) of a language, whatever it is. However, as we are investigating first steps in this research, our target is to be able to synthesize a few sentences with high quality and realtime control over pitch and voice quality parameters (as a proof of concept). Therefore we think that we do not need to record now a whole database but only sentences we have decided to produce. Consequently database design is made as follows.

- Each sentence is recorded with constant pitch and with voice quality as constant as possible, so as to constrain the range of subsequent analysis steps.
- Each sentence is recorded two times. In the first recording, the speech is slightly slower than in natural speech while the second recording is made at a normal rate. The first recording is made to ensure that each vowel has a non-coarticulated central region which will allow us to easily connect diphones together during synthetic voice production without loosing or mismatching the natural coarticulation.
- Each sentence is a succession of voiced and unvoiced phonemes and, due to their complexity, we don't use voiced consonants (e.g. /v/, /z/, /b/, /g/, ...) nor liquids (e.g. /R/, /l/, ...).
- We choose two phonemic sequences, one with different plosive and fricative consonants and one with the same consonant interleaved with various vowels. Their SAMPA transcriptions are respectively:

$$/t\,a\,k\,a\,p\,a\,f\,a\,S\,e/$$
$$/t\,a\,t\,2\,t\,i\,t\,o\,t\,u/$$

The recordings are made on one channel at 44100 Hz and each sample is stored in 16 bits PCM format. A MAX/MSP application which is illustrated in the Fig 5 helps the speaker maintaining pitch, voice quality and syllabic rate as constant as expected.

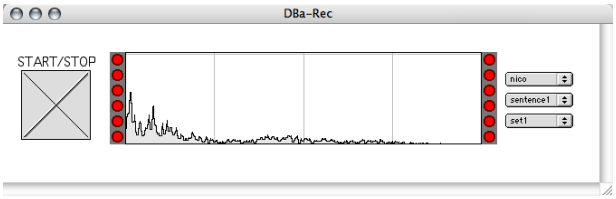


Figure 5: *Graphical User Interface for the recording of the database, with an auditory feedback helping the speaker maintaining the pitch, voice quality and syllabic rate as constant as possible.*

4.2. Database Segmentation

The four phonemic sequences are manually segmented into two sets: the vowels and the consonants. Each vowel is in turn divided in three parts. The first and third parts are the regions of the vowel that are coarticulated respectively with the preceding and following consonant (or silence). These are the left and right transient parts of the vowel. These coarticulated parts of speech usually contain a few (less than 10) periods of the vowel. The central part, also called the steady part is the region of speech that we can consider as coarticulation-free and thus actually quasi-stationnary (cf. Figure 6).

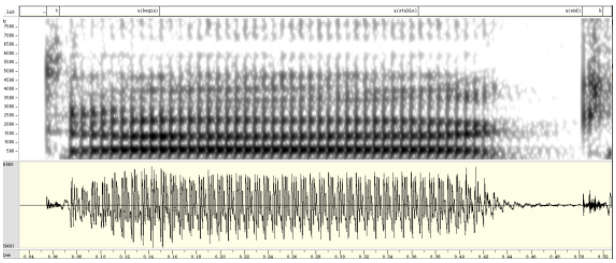


Figure 6: *Typical example of segmentation on a syllable of the database.*

Next, for each voiced region, the first GCI (GCI_1 : corresponding to the GCI of the first glottal pulse of a voiced island, meaning not overlapped with preceding vocal tract response, and thus clearly observable) is automatically approximated as the position of the minimum of the speech waveform in a region of $1.2 \times T_0$ after the unvoiced-to-voiced (e.g. /t/ to /a/) segmentation point. Finally positions of these GCIs are checked and corrected if necessary.

4.3. ZZT Representation

In order to find precise position of following GCIs and a first estimation of the glottal flow parameters, the ZZT representation

[3] of speech is used. For a series of N samples $(x(0), x(1), \cdots, x(N-1))$ taken from a discrete signal $x(n)$, this representation is defined as the set of roots (zeros of the polynomial) $(Z_1, Z_2, ..., Z_m)$ of its corresponding Z-Transform $X(z)$ (cf. equation (6)).

$$X(z) = \sum_{n=1}^{N-1} x(n) z^{-n} = x(0) z^{-N+1} \prod_{m=1}^{N-1} (z - Z_m) \quad (6)$$

This representation implies to compute roots of polynomials [9] of which the degree increases with the sampling frequency, introducing errors on the estimation of zeros in high frequencies. That is why we perform the analysis at 16000 Hz, thus first downsampling the waveforms of the database.

According to the mixed-phase model of speech, the ZZT representation of a speech frame contains zeros due to the anticausal component (mainly dominated by the glottal source) and to the causal component (mainly dominated by the vocal tract response) [10]. Consequently zeros due to the anticausal component lie outside the unit circle, and zeros due to the causal component inside the unit circle. Under some conditions about the location, the size and the shape of the window analysis, zeros corresponding to both anticausal and causal contributions can be properly separated by sorting them according to their radius in the z-plane. The waveform and the spectrum of these contributions are then computed by the Discrete Fourier Transform (equation (7)).

$$X(e^{j\phi}) = G e^{(j\varphi)(-N+1)} \prod_{m=1}^{N-1} (e^{(j\varphi)} - Z_m) \quad (7)$$

A typical windowed speech frame is displayed in Figure 7, and results of decomposition based on zeros separation are displayed in Figure 8, illustrating relevant shapes of derivative of glottal flow and vocal tract impulse response.

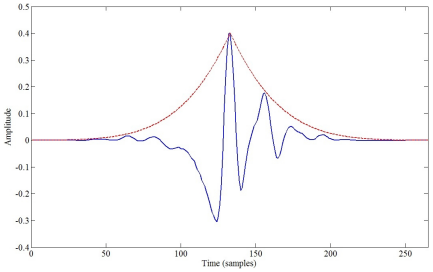


Figure 7: *Example of a typical speech frame (blue), weighted by a Hanning-Poisson window (red). With that shape, the signal is prepared to be processed by the ZZT-based decomposition algorithm.*

4.4. GCI Adjustment

The location, the size and the shape of the window chosen for the analysis have a huge effect on the ZZT representation, thus on the estimation of anticausal and causal components. A review of these different conditions and their effects can be found in [3, 11]. We can highlight that the analysis window has to be centered on a GCI in a really precise way, as the ZZT-based decomposition is really sensitive to wrong positioning.

In the framework of this project, we first use this sensitivity in order to get precise location of GCIs as instants where the ZZT-based decomposition is well achieved. As illustrated in Figure 9, a first approximation of GCIs can be extrapolated from

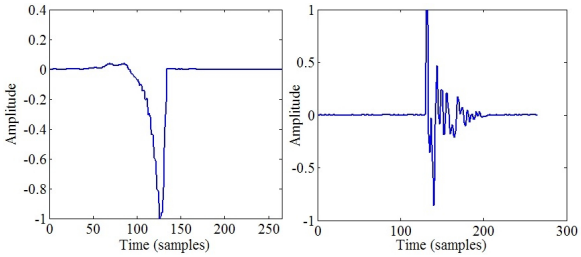


Figure 8: *Results of ZZT-based decomposition for a speech frame: the anticausal component, close to the derivative of glottal flow (left) and the causal component, close to the vocal tract impulse response (right).*

the first GCI of a voiced island marked in the segmentation task (GCI_1) and by marking next ones thanks to an estimation of the pitch contour (computed e.g. by autocorrelation).

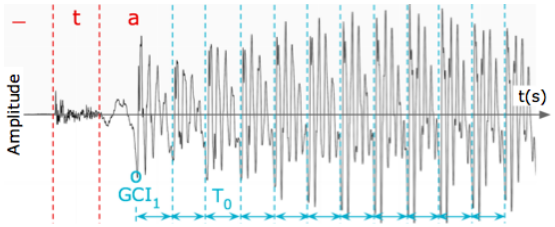


Figure 9: *First approximation of GCI positions taken from the first onset of the voice island (with a appearing GCI: GCI_1) and extrapolation thanks to a pitch contour estimation (e.g. by autocorrelation).*

Recent experiments have shown that it is possible to obtain reliable glottal source and vocal tract estimates by shifting few samples around each estimated GCI [12]. If the maximum shift is set e.g. to 4 samples, it gives us, for each estimated GCI, 9 candidates for the glottal source. The challenge is to find which shift gives the best decomposition, in terms of proximity to the LF model. By comparing a correct glottal pulse and a wrong one in the spectral domain (cf. Figure 10), we can observe that their behaviour is quite the same below 2 kHz and significantly different in higher frequencies.

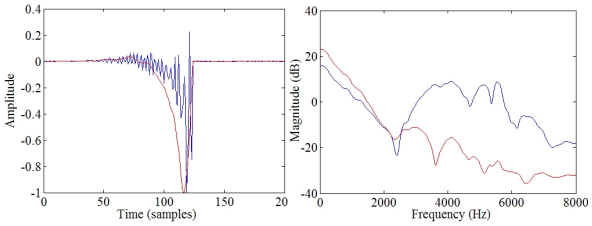


Figure 10: *Good glottal pulse (red) vs wrong glottal pulse (blue): illustration of the spectral criterion, which is based on the clear increasing of high frequencies when decomposition fails.*

In order to choose the best one among all candidates, we define a spectral criterion as the ratio between the energy in 0-2 kHz frequency band and the energy in the whole spectrum (0-F_s/2).

$$Criterion = \frac{Energy_{[0-2000Hz]}}{Energy_{[0-8000Hz]}} \quad (8)$$

Thus for each GCI, the best glottal pulse is chosen as the one which maximises this criterion among all the possible candidates. The location of every GCI can thus be refined (and the pitch estimation as well). Figure 11 shows the result of decomposition after shifting and taking the best candidate.

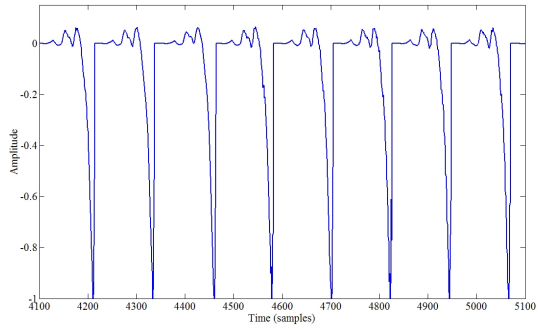


Figure 11: *Result of decomposition (anticausal contribution) after the procedure of GCI shifting (4), the computation of the spectral criterion, and the choice of the best candidate.*

4.5. Estimation of O_q and α_m

Once the best glottal sources are obtained for all the GCIs, the open quotient (O_q), the asymmetry coefficient (α_m) and the onset time of the glottal source can be estimated by the anticausal component to the LF model. For each GCI, this fitting is performed as:

- Computation of all the LF waveforms with the same pitch period and energy as the one obtained from the analysed speech frame at the considered GCI, and a defined range for O_q (0.3-0.9) and α_m (0.6-0.9), thus defining a codebook of LF waveforms.
- Windowing of the LF codebook in order to take into account the effect of the window analysis on the estimated glottal source;
- Synchronization between the estimated glottal sourcet obtained by the ZZT-based decomposition and the LF codebook, performed by an alignement between the GCI of the glottal source and the GCI of the LF waveforms;
- Computation of the square error between each windowed LF waveform of the codebook and the estimated glottal source;
- O_q and α_m are defined as the couple of parameters corresponding to the LF waveform closest to the glottal source;
- Once O_q is computed, the onset time (defined as the beginning of the opening phase of the vocal folds) can be estimated.

The result of the fitting and the estimation of the onset time are displayed in Figure 12 for the glottal sources shown in Figure 11. We can observe that estimated LF waveforms are quite close to the estimated glottal sources. Then Figure 13 shows the evolution of O_q and α_m for the whole part */ta/* of the database. These evolutions also confirm that the constant voice quality we had targeted at recording time has been somehow achieved by our speaker.

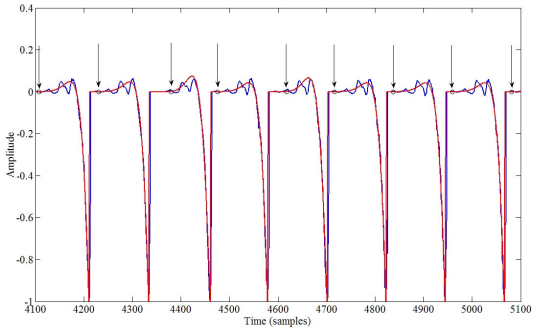


Figure 12: *Fitting in time domain between the anticausal component coming from ZZT-based decomposition (blue) and the LF model of glottal pulse (red). Black circles and arrows show onset instants.*

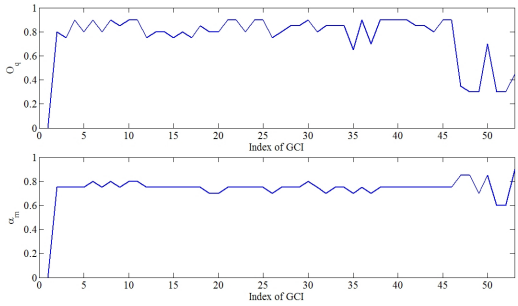


Figure 13: *Evolution of estimated open quotient (O_q) (upper panel) and asymetry coefficient (α_m) (lower panel), coming from the time-domain fitting, for the /ta/ sound in the database.*

4.6. ARX-LF Optimization on Sub-Codebook

In the source/filter model [13], a sample $y(n)$ of speech is modeled by the AR equation :

$$y(n) = -\sum_{i=1}^{p} (a_n(i)y(n-i)) + b_n x(n) + e(n), \qquad (9)$$

where $x(n)$ is the source, $e(n)$ is the residual and $a_n(i)$ and b_n are the AR filter coefficients (these coefficients represent the vocal tract which is varying over time instants n and thus their values evolve accordingly).

In the usual LP modelization, the source $x(n)$ is supposed to be impulse train or white noise and, given this assumption, some mathematical developments lead to the Yule-Walker equations which can be solved to obtain the AR filter coefficient. This is the classical LP analysis.

In [14] and [15] Vincent *et al.* developped a more complex model which assumes that the source $x(n)$ is a glottal flow in the voiced segments of speech. Therefore the hypothesis on the nature of the source has been changed and Yule-Walker equations cannot be used anymore. Instead, one can write equation (9) for successive values of n and solve the resulting set of linear equations the obtain a set of prediction coefficients minimizing the residual energy.

As explained in subsection 4.4 the results of analysis on voiced speech are strongly dependent upon the choice made to build the analysis window. Accordingly we work with particular position and length configurations for that window. It is GCI-centered and has a length equal to two periods ($2\,T_0$). Moreover, the analysis window is Hanning-weighted.

Consequently, for the k^{th} GCI, we can write equation (9) for 2 T_0 values of n ($[GCI_k - T_0 + 1 \dots GCI_k + T_0]$) and we solve the system of equations in $p+1$ unknowns : $a_k(i)$ and b_k, considered as constant around GCI_k, instead of varying from sample to sample:

$$Y = MA + E, \tag{10}$$

where Y is a vector of $y(n)$, M is the concatenation of a matrix of $-y(n-i)$ values and a vector of $x(n)$, all of them for $n = [GCI_k - T_0 + 1 \dots GCI_k + T_0]$. A is the vector of unknown values $a_k(i)$ and b_k. E is a vector of residuals $e(n)$, that is to say a vector of modelization errors that we want to minimize when computing A.

We then solve these equations for different glottal flows x_w (particular values of x, called the *codebook* hereafter). The glottal flow minimizing the modelization error (see below) is considered as the most correct estimate of the actual glottal flow produced by the speaker. In the next paragraphs we will explain how we create the codebook and compute the corresponding errors.

The glottal flows are built using a spectrally enhanced LF model [16] and are driven by three parameters: O_q (open quotient), α_m (asymetry coefficient) and T_l (spectral tilt). Nevertheless a codebook of glottal flows based on the possible variations of these three parameters would be rather bulky and solving (10) for all the waveforms stored in that codebook would be CPU expensive.

Fortunately O_q and α_m estimates are already known (thanks to ZZT analysis and LF fitting techniques) which allows us to select a part of the codebook that we call a sub-codebook. T_l is the only varying parameter of that sub-space of LF pulses. Moreover, although we are confident in the estimate of O_q and α_m, we can refine these results by selecting a somehow larger sub-codebook, allowing some slight variations of O_q and α_m around the initial estimates.

Let us say there are W glottal flows x_w in the sub-codebook. As said above, for each one of them we can compute the $a_k(i)$ and b_k coefficients and therefore re-synthesize an approximation y_w of the speech frame y by applying equation (9).

The error for word x_w is then measured as the Euclidean distance between the re-synthesized speech frame y_w and the original analysis window y. Note that both y_w and y are Hanning-windowed.

$$E_w = \sqrt{\sum_{n=1}^{2T_0} (y(n) - y_w(n))^2} \tag{11}$$

However before actually computing errors, two important points remain: the GCI position and the filter stabilization.

Indeed, the estimate of each GCI position is provided by the ZZT analysis. Although that position fits very well for ZZT decomposition, it's not necessarily the best one for ARX decomposition. For that reason one more step is added to the algorithm explained above: we do not consider only the analysis window y centered on the GCI approximation coming from ZZT but also windows centered a few points on the left and on the right of that location.

In our implementation we look three points before and after the position of the current GCI. Henceforth we will have $7W$ error measurements and not only the minimum error will give us the best guess for the glottal flow parameters but also for the GCI optimal position.

Finally, although the Levinson-Durbin method that solves the Yule-Walker equations guarantees that the AR filter has all of its poles inside the unit circle and therefore is stable, this is no longer the case when solving equation (10). Consequently, the last step before synthesizing any of the y_w is to reflect the outside poles inside the unit circle and adapting the value of parameter b accordingly [13].

All these steps are performed at a sample rate of 8kHz which allows us to get reliable estimates of T_l and the positions of GCIs (as well as an estimate of the filter parameters for that rate). However high quality singing synthesis is produced at higher sampling rate such as 44.1 or 48 kHz.

The glottal flow parameters O_q, α_m and T_l are independent of the sampling frequency and therefore they can be used as is. On the contrary the filter coefficients rely upon the sampling rate and need to be recomputed. The task is fast and easy since we just have to solve equation (10) once with a different sampling frequency for $y(n)$ (for which we have the original recording at 44.1kHz) and $x(n)$ (which is the LF/Klatt model and thus can be produced at any given rate for the fixed O_q, α_m and T_l).

To make things short, equation (10) is first solved at 8kHz for 24 $a(i)$ parameters ($p = 24$) and considering a sub-codebook with O_q and α_m constant and T_l varying between 3dB and 20dB (with a 1dB step). Then it is solved at 44.1kHz for $p = 46$ and O_q, α_m and T_l constant. Results are illustrated in the Figure 14.

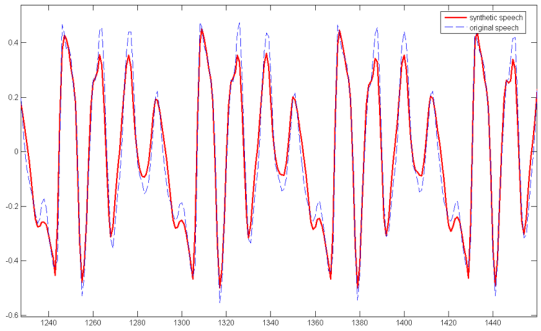


Figure 14: *Superposition of original (blue) and resynthesized (red) signals, after the computation of ARX-LF on a sub-codebook defined by ZZT-based parameters.*

4.7. Vocal Tract Response Compensation

It's observed that synthesis obtained by exciting the ARX filter with the glottal flow results in a certain loss of high frequency components. To compensate for this effect, we devised a simple compensation method via AR filtering. For this, the AR compensation filter is obtained by linear predictive analysis of an impulse response obtained in the following way. The frequency response of the original signal is divided by the frequency response of the synthetic signal, and the inverse Fourier transform of the result is taken. A sample result of the compensation is presented in Figure 15. The obtained AR compensation filter is combined (by cascading) with the ARX filter to obtain a single filter that will perform the synthesis in one stage.

5. SOUND DESCRIPTION INTERCHANGE FORMAT

The analysis tool being MATLAB and the realtime synthesis tool being MAX/MSP, a compatible format must be used. It was found that SDIF tools exist for both of these softwares. SDIF means *Sound Description Interchange Format*. This kind of file does not contain the sound itself, but sets of descriptive parameters e.g. coefficients of an estimated AR filter for speech or values of harmonics for additive synthesis.

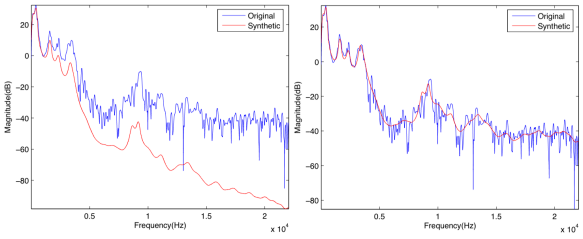


Figure 15: *Spectral envelope of original (blue) and synthetic (red) signals before (left) and after (right) the HF compensation.*

A SDIF file is divided in many simple parts. The biggest entity in such a file is called a *stream*. Streams are series of time-tagged frames and are identified by a stream ID (integer). Frames are identified inside a stream by their corresponding instants on the timeline. They have to be stored in increasing time sequence. Two frames of the same stream may not have the same time tag. In each frame, it is possible to store multiple matrices which differ from each other by their type.

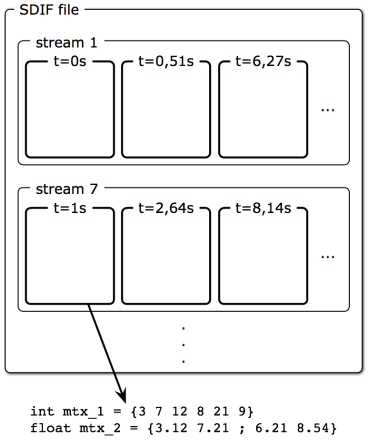


Figure 16: *Structure of a* SDIF *file: streams, time-tagged frames and matrix.*

5.1. Storage of MATLAB Analysis Results

Different functions are available in order to deal with SDIF files in MATLAB [17]. There are two possible ways to read and verify SDIF contents: *loadsdif* and *loadsdiffile*. However the most interesting function in this project is the *sdifwrite*. It allows to encode easily SDIF files frame by frame. Parameters to be specified in each frame are the stream ID, the frame type, the frame time, and type and content for each of the matrices. MATLAB SDIF encoder presented some unexpected results as there was always an unintended stream with a negative stream ID in addition to encoded streams. Knowing this problem and adapting reading/writing algorithms to it, encoded streams work properly.

In this project, we decide to create, for each GCI-centered analysis window, a SDIF frame containing a single vector with parameters of source and tract structured as follow (for $t = t_k$): reflection coefficients (K_i), fundamental frequency (F_0), open quotient (O_q), asymetry coefficient (α_m), spectral tilt (T_L) and and arbitrary index informing about the tag that has been used in the segmentaion file (S_k).

$$\{K_1 ... K_n \; F_0 \; O_q \; \alpha_m \; T_L \; S_k\}|_{t=t_k}$$

Regions of the database corresponding to unvoiced speech are represented in the SDIF file as a island of ten frames, with time tag distributed on the whole unvoiced period, and identified by their S_k values.

5.2. SDIF Interface in MAX/MSP

The voice synthesis is performed on MAX/MSP. Here we present a list of the existing MAX/MSP objects existing to deal with SDIF files [18].

The first SDIF object to use is the *sdif-buffer* object. It allows storing temporarily the content of a SDIF file. There are three main messages understood by *sdif-buffer*. `[streamlist myfile.sdif]` lists the streams existing in the file. `[frame list myfile.sdif]` lists and prints all the existing frames in myfile.sdif. Finally `[read-stream-number myfile.sdif StreamID]` allows to load into the buffer a whole stream, including all the frames and all the matrixes it contains. Once a stream has been read, it is possible to know his properties through the message `[print]`.

Other objects allow manipulating the stream contained in the *sdif-buffer* object. First *sdif-tuples* is able to output the data content of a specified frame. The message which allows outputing such data is `[tuple]`. Arguments that may be specified are the frame time, the matrix type, indexes of columns of the matrix to output, and the format of the output. Possible formats are data outputed by row or concatenated data. The *sdif-range* external is able to scan the stream contained in a buffer to know the maximum number of columns in the matrices of a specified type or the maxima and minima of each column of a specified type of matrix. The *sdif-menu* external is a graphical interface to manage different streams and obtain information about each of them. Finally the *sdif-listpoke* allows the user to write his own matrices in frames and put these together to form a stream in a buffer. At this time it is not possible by a simple way to write data into a sdif file but it should soon be possible through the FTM library [19].

The only objects needed to use a SDIF file inside MAX/MSP are *sdif-buffer* and *sdif-tuples* objects. The first step is to load a stream into a buffer. Then, at each new frame instant, all the parameters are transmitted in a list by the *sdif-tuples* object. Some objects then slice this list in a list of reflection coefficients redirected to the lattice filter (*vqc.lat~*), anticausal glottal parameters redirected to the *vqc.lfp~* object and the spectral tilt redirected to the *vqc.stf~* object. Finally the equalization gain is isolated and used as a multiplier after the lattice filter.

6. OVERVIEW OF RAMCESS 2.X FRAMEWORK

Finally these investigations in expressive voice analysis allow us to draw the first contours of the next step in the development of the RAMCESS framework (version 2). This structure combines the realtime synthesis layer of RAMCESS 1.x system with now the encapulsation of lots of analysis and decomposition processes in a SDIF database, imported in the performance application, with modules communicating by OSC [20].

A gestural sollicitation is captured by sensors, driving the behavior of a typical digital instrument. In our experimentations, the HandSketch [21] has been used. Typical mappings related to phonetic contents and dimensions of voice quality are applied. At this level, information extracted from the database is used in order to generate more relevant syntheis parameters, especially in the realtime production of the coarticulation. Finally sound is produced by the VQCLIB layer. This process is summarized in Figure 17.

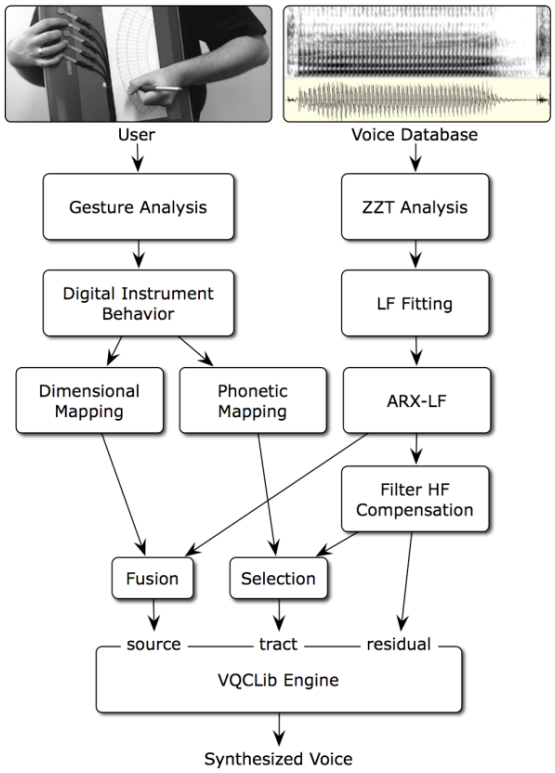


Figure 17: *An overview of RAMCESS 2.x framework.*

7. CONCLUSION

In this paper we described main improvements that have been achieved in order to transform a vowel-only system into a framework which is able to produce coarticulated speech expressively. This work illustrated that the deep analysis of glottal source features and the separation with vocal tract components were achievable on a prepared database. We were also able to validate the use of VQCLIB, and more generally RAMCESS elements in a complex voice production context.

On the top of this current work, two significant axis will be developed further: the achievement of a first (limited but functional) diphone synthesizer, and the extension of these concepts to musical signal synthesis, as recent work have shown that it was clearly possible [22].

8. ACKNOWLEDGMENTS

Authors would first like to thank the organization committee of eNTERFACE'07 summer workshop in Istanbul (Boğaziçi University) for their great involving and support. We also would like to highlight a role of the European Union (through FP6 and SIMILAR) in the achievement of all these tools. Finally we would like to thank our respective laboratories and supervisors for their advices, trust and support.

9. REFERENCES

[1] J. Bonada and X. Serra, "Synthesis of the Singing Voice by Performance Sampling and Spectral Models", *IEEE Signal Processing*, vol. 24, no. 2, pp. 67–79, 2007. 129

[2] B. Bozkurt and T. Dutoit, "Mixed-Phase Speech Modeling and Formant Estimation, using Differential Phase Spectrums", in *Proc. of ISCA ITRW VOQUAL*, pp. 21–24, 2003. 129

[3] B. Bozkurt, *New Spectral Methods for the Analysis of Source/Filter Characteristics of Speech Signals*. PhD thesis, Faculté Polytechnique de Mons, 2005. 129, 132

[4] "VQClib". http://vqclib.blogsite.org. 129, 130

[5] G. Fant, J. Liljencrants, and Q. Lin, "A Four-Parameter Model of Glottal Flow", *STL-QPSR*, vol. 4, pp. 1–13, 1985. 130

[6] B. Doval and C. d'Alessandro, "The Voice Source as a Causal/Anticausal Linear Filter", in *Proc. of Voqual'03, Voice Quality: Functions, Analysis and Synthesis, ISCA Workshop*, 2003. 130

[7] N. d'Alessandro, B. Doval, S. L. Beux, P. Woodruff, Y. Fabre, C. d'Alessandro, and T. Dutoit, "Realtime and Accurate Musical Control of Expression in Singing Synthesis", *Journal on Multimodal User Interfaces*, vol. 1, no. 1, pp. 31–39, 2007. 130, 131

[8] B. Doval, C. d'Alessandro, and N. Henrich, "The Spectrum of Glottal Flow Models", *Acta Acustica*, vol. 92, pp. 1026–1046, 2006. 130

[9] A. Edelman and H. Murakami, "Polynomial Roots from Companion Matrix Eigenvalues", *Mathematics of Computation*, vol. 64, no. 210, pp. 763–776, 1995. 132

[10] B. Bozkurt, B. Doval, C. d'Alessandro, and T. Dutoit, "Zeros of the Z-Transform Representation with Application to Source-Filter Separation in Speech", *IEEE Signal Processing Letters*, vol. 12, no. 4, pp. 344–347, 2005. 132

[11] B. Bozkurt, L. Couvreur, and T. Dutoit, "Chirp Group Delay Analysis of Speech Signals", *Speech Communication*, vol. 49, no. 3, pp. 159–176, 2007. 132

[12] T. Dubuisson and T. Dutoit, "Improvement of Source-Tract Decomposition of Speech using Analogy with LF Model for Glottal Source and Tube Model for Vocal Tract", in *(to appear in) Proc. of Models and Analysis of Vocal Emissions for Biomedical Application Workshop*, 2007. 132

[13] G. Fant, *Acoustic Theory of Speech Production*. Mouton and Co. Netherlands, 1960. 133, 134

[14] D. Vincent, O. Rosec, and T. Chonavel, "Estimation of LF Glottal Source Parameters Based on ARX Model", in *Proc. of Interspeech*, (Lisbonne), pp. 333–336, 2005. 133

[15] D. Vincent, O. Rosec, and T. Chonavel, "A New Method for Speech Synthesis and Transformation Based on an ARX-LF Source-Filter Decomposition and HNM Modeling", in *Proc. of ICASSP*, (Honolulu), pp. 525–528, 2007. 133

[16] N. d'Alessandro and T. Dutoit, "RAMCESS/HandSketch: A Multi-Representation Framework for Realtime and Expressive Singing Synthesis", in *(to appear in) Proc. of Interspeech*, (Anvers), 2007. 134

[17] D. Schwarz and M. Wright, "Extensions and Applications of the SDIF Sound Description Interchange Format", in *Intl. Computer Music Conf.*, 2000. 135

[18] "SDIF for Max/MSP". http://www.cnmat.berkeley.edu/MAX/downloads/. 135

[19] N. Schnell, R. Borghesi, D. Schwarz, F. Bevilacqua, and R. Muller, "FTM - Complex Data Structures for Max", in *Proc. of Intl. Computer Music Conf.*, 2007. 135

[20] "OpenSoundControl". http://opensoundcontrol.org. 135

[21] N. d'Alessandro and T. Dutoit, "HandSketch Bi-Manual Controller", in *Proc. of NIME*, pp. 78–81, 2007. 135

[22] N. d'Alessandro, T. Dubuisson, A. Moinet, and T. Dutoit, "Causal/Anticausal Decomposition for Mixed-Phase Description of Brass and Bowed String Sounds", in *(to appear in) Proc. of Intl. Computer Music Conf.*, 2007. 136

10. BIOGRAPHIES

Nicolas d'Alessandro holds an Electrical Engineering degree from the Faculté Polytechnique de Mons (FPMs) since 2004. He did the master's thesis in the Faculty of Music of the University de Montréal (UdeM) (supervisor: Prof. Caroline Traube). That work gathered the development of applications based on perceptual analogies between guitar sounds and voice sounds, and a study of mapping strategies between (hand-based) gestures and speech production models (source/filters and concatenative approaches). He started a PhD thesis in September 2004 in the Signal Processing Lab of the Faculté Polytechnique de Mons (supervisor: Prof. Thierry Dutoit) related to the real-time control of unit-based synthesizers.
Email: nicolas.dalessandro@fpms.ac.be

Onur Babacan was born in Izmir, Turkey in 1986. He is currently a senior undergraduate student at Izmir Institute of Technology (IYTE, Izmir, Turkey), studying Electronics and Telecommunications Engineering.
Email: onurbabacan@gmail.com

Barış Bozkurt is currently employed as an Assistant Professor in Izmir Institute of Technology (IYTE, Izmir, Turkey) where he teaches electronics and digital audio/speech/signal processing and continues his research in the same fields. He obtained his Electrical Engineering degree in 1997 and Master of Science degree in Biomedical Engineering in 2000 both from Boğaziçi University, Istanbul, Turkey. After obtaining his PhD degree (Electrical Engineering (speech processing)) in 2005 from Faculté Polytechnique De Mons, Mons, Belgium, he worked as a research engineer in Svox AG, Zurich. He is an associate editor of Journal of Interdisciplinary Music Studies.
Email: barisbozkurt@iyte.edu.tr

Andre Holzapfel received the graduate engineer degree in media technology from the University of Applied Sciences in Duesseldorf, Germany, and the M.Sc. degree in computer science from University of Crete, Greece, where is currently pursuing the Ph.D. degree. His research interests are in the field of speech processing, music information retrieval and ethnomusicology.
Email: hannover@csd.uoc.gr

Thomas Dubuisson was born in Tournai, Belgium, in 1983. He received the Electrical Engineering Degree from the Faculty of Engineering, Mons (FPMs) in 2006. He is currently PhD Student in the Circuit Theory and Signal Processing (TCTS) Lab, FPMs, where he is working on assessment of speech pathologies and source-tract separation of speech, in the framework of the WALEO II Research Project ECLIPSE (Walloon Region, Belgium).
Email: thomas.dubuisson@fpms.ac.be

Loïc Kessous was born in Marseille, France. He received the Master of Science degree in Physics from the Aix-Marseille University and obtained is PhD in 2004 from Paris IX University, and was affiliated at this time to the Mechanics and Acoustics Laboratory at CNRS in Marseille (LMA-CNRS). He realized during his PhD several digital instrument including an gesture controlled singing voice synthesizer. During the same period he was also a research and development associate at GMEM, (National Center for Music Creation) in Marseille. From 2004 to 2007 he was working in the framework of the european HUMAINE (Human-Machine Interaction Network on Emotion) project at Tel Aviv University, Israel. He is currently working at The Computer Sciences Laboratory for Mechanics and Engineering Sciences (LIMSI), CNRS laboratory, in Paris, France.
Email: loic.kessous@limsi.fr

Alexis Moinet holds an Electrical Engineering degree from the FPMs (2005). He did his master thesis at the T.J. Watson research Center of IBM (2005). He is currently working on the IRMA project in the Signal Processing Lab of the FPMs. He is particularly interested in source/filter decomposition of speech and music, phase vocoder, voice activity detection, voice conversion and HMM synthesis.
Email: alexis.moinet@fpms.ac.be

Maxime Vlieghe was born in 1984 in Belgium. When he was 7, he learned academic music for 5 years and played piano for 3 years at the "conservatoire Royal de Tournai". After finishing his college classes in sciences, he wanted to go on in this direction and decided to start studies at the "Faculté Polytechnique de Mons". He studied Electric Engineering for 5 years. In his last year, joining the fields of music and sciences, he wrote his master thesis at the Université de Montréal with Pr. Caroline Traube about the synthesis of singing voice and particularly rules synthesis of plosive consonants. As he was writing this work, he was invited by Pr. Thierry Dutoit to be member of a team of eNTERFACE'07 on a singing voice synthesis project.
Email: maxime.vlieghe@gmail.com

MOBILE-PHONE BASED GESTURE RECOGNITION

Barış Bahar [1], *Işıl Burcu Barla* [1], *Ögem Boymul* [1], *Çağlayan Dicle* [1], *Berna Erol* [2], *Murat Saraçlar* [1], *Tevfik Metin Sezgin* [3], *Miloš Železný* [4]

[1] Boğaziçi University, İstanbul, Turkey
[2] Ricoh California Research Center, CA, USA
[3] University of Cambridge, England
[4] University of West Bohemia in Pilsen, Czech Republic

ABSTRACT

Mobile phones are increasingly being used for applications beyond placing simple phone calls. However, the limited input capabilities of mobile phones make it difficult to interact with many applications. In this work, we present a mobile gesture recognizer where camera input of a mobile device is analyzed to determine the user's actions. Simple user gestures such as moving the camera from left to right and up and down are employed in two mobile applications: map navigation and a drawing program. In addition, more complex user gestures are converted into commands with an HMM based algorithm. These commands allow higher level interaction with the applications, such as moving to a specific location on the map or changing the drawing color. Index Terms mobile devices, computer vision, mobile phones, camera phones, motion estimation, camera-based interaction, HBMA, command recognition, gesture recognition

KEYWORDS

Mobile devices – Computer vision – Mobile phones – Camera phones – Motion estimation – Camera-based interaction – HBMA – Command recognition – Gesture recognition

1. INTRODUCTION

The simple keypad present in many mobile phones is adequate for placing voice calls but falls short when interacting with mobile applications such as web browsers, image and video browsers, and location services. Because mobile phones are handheld devices, user's gestures can be easily utilized as additional inputs (e.g. moving the phone to the right resulting in panning to the right on a map on the screen). User gestures can be recognized by sensors such as accelerometers and gyros on the device. However, currently mobile devices with such sensors is not common. Most mobile phones are equipped with video cameras which can be used to perform gesture recognition by analyzing the optical flow. Simple gestures such as moving the phone up and down can then be converted into commands that are interpreted by the application running on the phone.

In this paper, we present an implementation of mobile camera phone based gesture recognition and its applications. We recognize simple gestures such as leftright and up-down by analyzing optical flow and utilize these gestures for panning on map images, Figure 1, and in a drawing application. In addition, a sequence of gestures, such as up-down followed by left-right, are converted into high level commands that can be used for map navigation or for customizing the parameters of a paint applications (e.g., specify pen thickness).

Although other camera phone based gesture recognizers exists in the literature [1][2], our system is the first vision based gesture recognizer designed to run on a windows mobile phone and first to incorporate high-level command recognition from gestures. In the following sections, we first present an overview of our system, then give details of the camera phone based gesture recognizer in Section 2. Command recognizer is described in Section 4, and applications are presented in 5. Finally, conclusions and future work is presented in Section 6.

Figure 1: *Using gesture recognition users can easily navigate maps and images on mobile phones.*

2. SYSTEM OVERVIEW

Figure 2 given an overview of the camera phone based gesture recognizer, which is implemented in C++ using Windows Compact Framework and Pocket PC SDKs. A Windows mobile device, Palm Treo 700w [3] is used for deploying and running the gesture recognizer.

Direct Show filters are employed for communication and frame capture from the device camera. Frames are captured in YUV12 format. Only luminance (Y) component is used for analyzing the motion for real-time processing. Motion estimator uses hierarchical block matching algorithm to determine the global motion direction. We currently assume motion is limited to panning (left-right-up-down). Once the global motion vector $[\Delta x, \Delta y]$ is computed, it is directly used as input for applications, such as displacement on a map and moving the cursor in a paint application. In addition, it is sent to a **command recognizer** to determine the high level meaning of gestures.

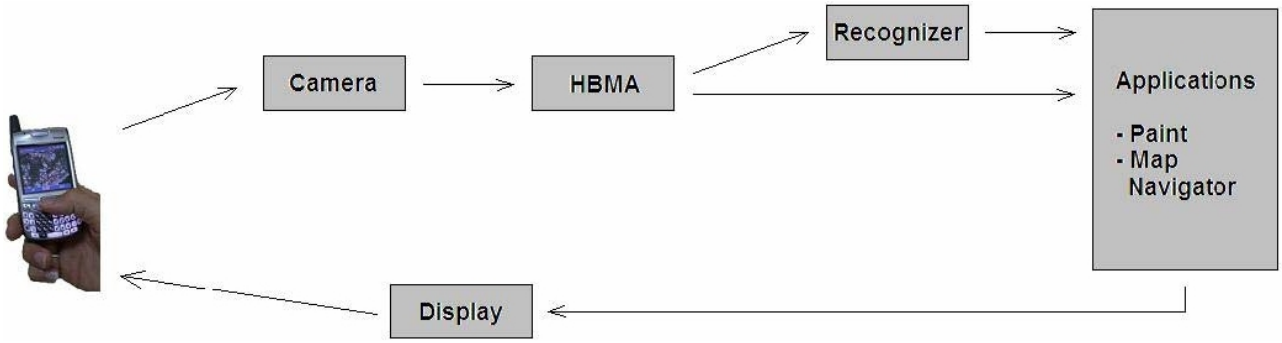


Figure 2: *An overview of the gesture recognition system and its interaction with the applications.*

3. MOTION RECOGNITION

In order to estimate the motion, the mobile phone's onboard camera is employed as the source of input. The direction and the magnitude of movement in consecutive video frames are used to infer the physical movement of the device by the user. Illustrative screen-shots captured while performing various gestures are presented in Figure 3. As seen in the figure, our algorithm is aimed to work with a variety of indoor and outdoor backgrounds.

Proper estimation of the observed motion from camera's onboard camera requires fast and accurate tracking. In order to fulfil the speed and accuracy criteria, hierarchical block matching algorithm (HBMA) is used to determine the motion direction and magnitude. The difference between two adjacent frames is used to estimate motion direction and magnitude, where direction values represent the x-y plane and magnitude is scaled according to dimension of the video frame.

Since colour image processing is computationally more expensive, we work with gray level images. For each consecutive image, we estimate motion matrices for the x and y directions. The means of these matrices give the direction and magnitude of the motion in x-y plane. In general, HBMA tries to align an anchor image to a target image while minimizing their difference, as illustrated in Figure 4. In order to achieve this, frames are compared at different scales. First, the frames are down sampled and smaller images are compared with little computation. Then, at each level, images are enlarged in order to make accurate estimates based on the ones made in the previous level. This low computational complexity algorithm results in accurate motion estimation.

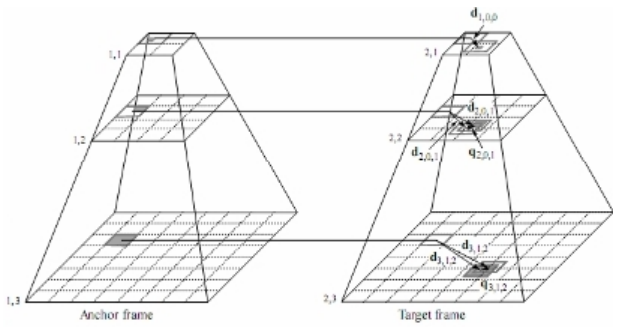


Figure 4: *3-D illustration of HBMA process.*

At every stage of the above mentioned computations, the anchor frame is divided into blocks according to the level, and then each block is compared with the corresponding ones in the target frame as shown in Figure 5. The correspondence is computed within a fixed search radius with the assumption that a target block exists in this specified area. Although this limits the maximum interframe device movement to fall below this specified radius, it does not cause any deficiency in practice. When matching two adjacent blocks, sum of Minimum Absolute Difference method is employed and block mean values are subtracted from this sum to reduce the lightning disturbance on matching. The target block is chosen to be the one that has the minimum difference. The distance and the position of the computed target block with respect to the anchor block give the magnitude and direction of the motion. By applying this technique for each block in the anchor frame, motion matrices of the images in x and y directions are computed. At the subsequent level, larger blocks are compared in a wider search area using a larger radius where the previous motion matrices are used as a starting point in order to reach a better estimation. Searching at multiple levels and reusing estimates from coarser levels results in an efficient and accurate matching. Motion vectors estimated using HBMA for consecutive levels is presented in Figure 6.

4. COMMAND RECOGNITION

Command gesture recognition system processes the motion vectors coming from the motion tracking part and decides on the most probable command. This command is then sent to and realized by the GUI. All possible commands, which are seven letters namely "b, r, g, t, e, w and d" in this project, are defined to the system. These commands and their corresponding meeting in our two applications, paint and map navigation, are presented in There is a database consisting of short movies, which are recordings with the mobile device, where each subject has imitated the letter by moving the device on the air. For building the database all movies are processed by the motion tracker, then vector quantization is applied for feature extraction and the system is trained using this data.

By using the command recognizer in the demo applications, the system will work real-time. In these applications, while one certain button is pressed, the mobile device captures the video. Then the motion tracker is activated. For each processed frame, the motion vectors are sent to the command recognizer. By releasing the button, the system finishes getting the observation. After that, the command recognizer decides on the command and sends this information to the running application GUI in order to be realized.

In the following parts, first some general information about the recognition theory will be given. Then, the implementation of this theory will be mentioned, and lastly some test results will be shown.

Figure 3: *Actual captured image sequences for various gestures.*

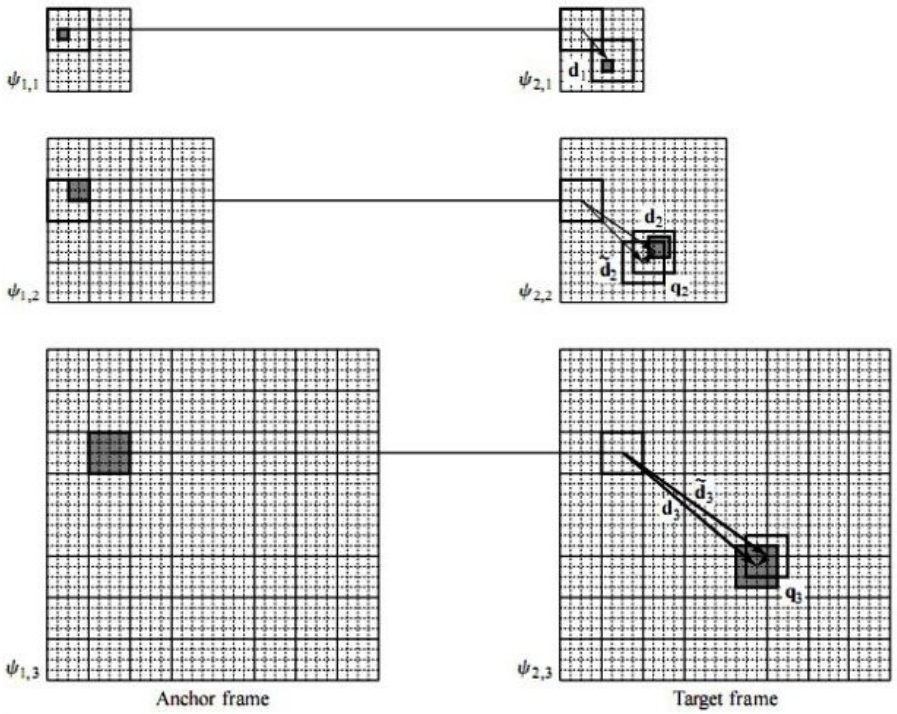


Figure 5: *In the first level of the pyramid, corresponding target frame block is estimated. In the subsequent levels, frame is up sampled, previous motion vectors are corrected.*

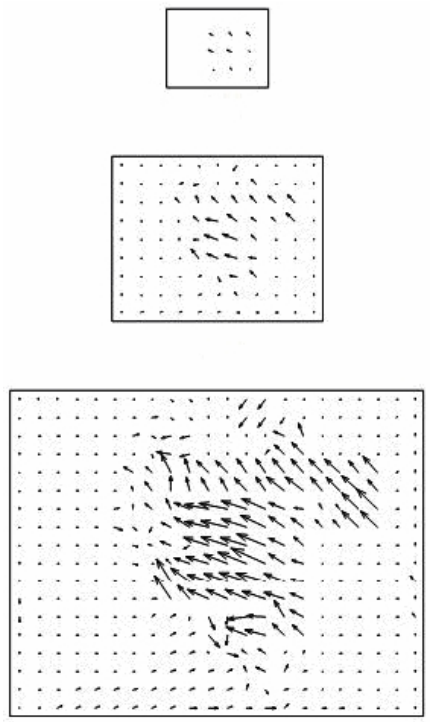


Figure 6: *Motion vectors corresponding to the estimates of HBMA for consecutive levels. The final output of HBMA is the image at the bottom from which the final motion is computed.*

Command s	Meaning	
	In Paint Application	In Map Navigation Application
b	Set the pen color to blue	Go to BUMED
r	Set the pen color to red	Go to Revir
g	Set the pen color to green	Go to Green Court
t	Make the pen thinner	Go to Teras Canteen
e	Enable eraser	Go to ETA Building
w	Make the pen thicker (wider)	Go to Washburn Hall
d	Set the pen color to black (default)	Go to Dorm

Table 1: *High level commands and their interpretations in two different applications.*

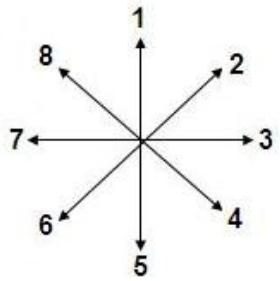


Figure 7: *The enumeration of the directions.*

4.1. Theory

All recognition systems are based on similar theory. This type of command recognition, which is described above briefly, can be used in many applications, where the user wants to control a mobile device without being limited with the keyboard.

In this project by the command recognition, Hidden Markov Models (HMMs) are used as the underlying technology. Generally, modeling and recognition process requires four steps [4], namely, feature extraction, determining the model parameters, computing the probabilities and determining the most probable state sequence.

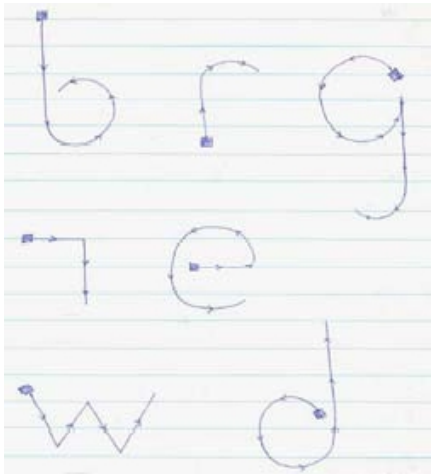
Let's examine the above mentioned steps in more detail. In our project, by feature analysis vector quantization is used. A vector quantizer maps k-dimensional vectors in the vector space R_k into a finite set of vectors. The motion vector pairs are grouped into three and these small groups are mapped in the observation space to some centroids and according to the database of all these observations a codebook is determined and defined to the system. The incoming data is then quantized according to this codebook, which contains the values of the 32 centroids, and each data group is mapped to the centroids.

The other three steps are the main problems in recognition. First of all the model parameters should be determined according to the nature of the commands, which will be recognized. The model parameters should be chosen such that the probability of the observation sequence given the model is locally maximized and then the iterative Baum-Welch algorithm or the Viterbi algorithm will be used by computing the reestimated model parameters. In this project the first approach was using the eight directions as the base of the model parameters. The advantage of this choice is that these model parameters are used by all of the letters, so the need on huge amount of database is decreased because by this method more samples for each model parameter could be observed. The definitions of the letters and the model parameters are shown in Figure 7. The model parameters were these eight directions, long 1, long 5 (these were needed according to the used letters) and the stabile situation. Long 1 was labeled as 1+, long 5 as 5+ and the stabile standing as 9. So, the letters coded with these labels are shown in Figure 8.

The test results with these model parameters were not satisfactory. One reason for that was that the samples were very different than each other. So, sharing the model parameters between the letters caused a large error rate. Hence, it was decided to define 7 different parameters for each command. By running the training algorithm, the system will learn the probability values corresponding to these parameters automatically.

There are two estimation algorithm candidates: Baum-Welch and Viterbi algorithms. The difference of these two algorithms is that the first one computes probabilities of all paths for all times, where in Viterbi only the most probable path survives to the next time instant. Hence, using Viterbi in this step decreases the needed memory and improves the speed of the system, which is very important in this project, since the applications will be run on a mobile device, which has a low CPU and limited memory capacity. Therefore Viterbi algorithm is chosen as the used algorithm in estimation part. Moreover, the number of the iterations should be chosen such that an optimum value for the model is determined but one should take care not to cause the model to completely depend on the training data. That will cause the model to give false results by a data outside the training set.

By computing the probabilities and building the file contain-



(a)

b	: 9 5+ 4 3 2 1 8 7 6 9
r	: 9 1+ 2 3 4 9
g	: 9 8 7 6 5 4 3 2 1 5+ 6 7 8 9
t	: 9 3 5+ 9
e	: 9 3 2 1 8 7 6 5 4 3 9
w	: 9 4 2 4 2 9
d	: 9 8 7 6 5 4 3 2 1+ 9

(b)

Figure 8: *Letter and model parameter definitions.*

ing the probability mass functions of the model parameters the training algorithm is used. As stated above the Viterbi algorithm is used by determining the most probable state sequence in each iteration. According to the state sequence and corresponding observation in the data, the probability values are updated and used in the next iteration.

When the system is trained according to the database, the final step is determining the most probable state sequence where any observation sequence and the determined model are given. By this decoding task the Viterbi algorithm is used. Given the state transitions, the first thing to do is building the trellis. Each cell of the Viterbi trellis, $v_t(j)$ represents the probability that the HMM is in state j after seeing the first t observations and passing through the most likely state sequence $q_1 \cdots q_{t-1}$, given the model λ. The value of each cell $v_t(j)$ is computed by recursively taking the most probable path that could lead to this cell. Formally, each cell expresses the following probability [5]:

$$v_t(j) = P(q_0, q_1 \cdots q_{t-1}, o_1, o_2 \cdots o_t, q_t) = j|\lambda) \quad (1)$$

Like other dynamic programming algorithms, Viterbi fills each cell recursively. Given that the probability of being in every state at time $t-1$ is already computed, the Viterbi probability is computed by taking the most probable of the extensions of the paths that lead to the current cell. For a given state q_j at time t, the value $v_t(j)$ is computed as [5]:

$$v_t(j) = \max_{1 \le i \le N-1} v_{t-1}(i) a_{ij} b_j(o_t) \quad (2)$$

In the above equation, a_{ij} gives the transition probability from state i to state j and $b_j(o_t)$ is the probability that at time t in state j observation ot occurs.

Explaining shortly the algorithm, it can be stated that for time $t = 1$, the metrics values in the trellis corresponding to the states to which a transition from the beginning state is allowed, are computed. After that according to the transitions the metrics are computed at each time slot. By these calculations only the path with the lowest cost can survive, others are eliminated, which is the key point in the Viterbi algorithm. Moreover, at each time slot the most probable path for each state from the previous time slot is memorized in a back pointer. At the end, the back pointer is read from end to beginning starting by the state for $t = T$ with highest probability. The recognition system used in this project is designed in the light of this theory using the programming languages C/C++.

4.2. Implementation

4.2.1. Search and Decoding

This part is done by using the Viterbi algorithm as explained above. The main function is called `Recognizer()` and is designed as an API, which should be called by the application. While a certain button is pressed, the movie is captured, processed and given as input to the recognizer after the features are extracted. Feature extraction is done by quantizing the incoming vectors according to the predefined codebook. This program needs four input files, namely, `pmf.txt`, `durations.txt`, `labels.txt` and `transitions.txt`. The first file contains the probabilities of the model parameters. The second one gives the probability of staying in the same state for each model parameter. In this project seven commands are used and 49 model parameters are defined. The third file contains the labels of these parameters and the last one determines the model, where each possible transition and its probability are defined.

4.2.2. Model parameter estimation

It is used for estimating the pmf values and the duration probabilities. As output it gives the minus likelihoods and the updated versions of `pmf.txt` and `durations.txt` at each iteration. The `Count()` function counts the occurrence rate of each model parameter for the complete command space. Using this information and the result of the Viterbi algorithm at each step, new probabilities are calculated. It takes as input the initial probabilities and the transition files corresponding to each separate command.

5. APPLICATIONS

Possible applications of the gesture recognition include: browsing maps, images, videos, web; playing games; handwriting... For demonstration of gesture recognition we implemented two applications: a map navigation application and a drawing application.

5.1. Map Navigation

A map navigation application is implemented utilizing gesture recognition as illustrated in Figure 9. In the map navigation application three control modes are assumed: panning the map, zooming the map and navigating to a particular place. To be able to process these three control modes, we designed main variables for the control: X and Y coordinates of the center of the map and zoom factor. Map control can be seen as watching bigger map (generally bitmap image) through a small window. X and Y coordinates control the position of the center of this window and zoom factor controls the magnification of the viewable part.

To control the map the algorithm is as follows:

- to navigate the map to certain position means set X and Y to that position (in image coordinates)

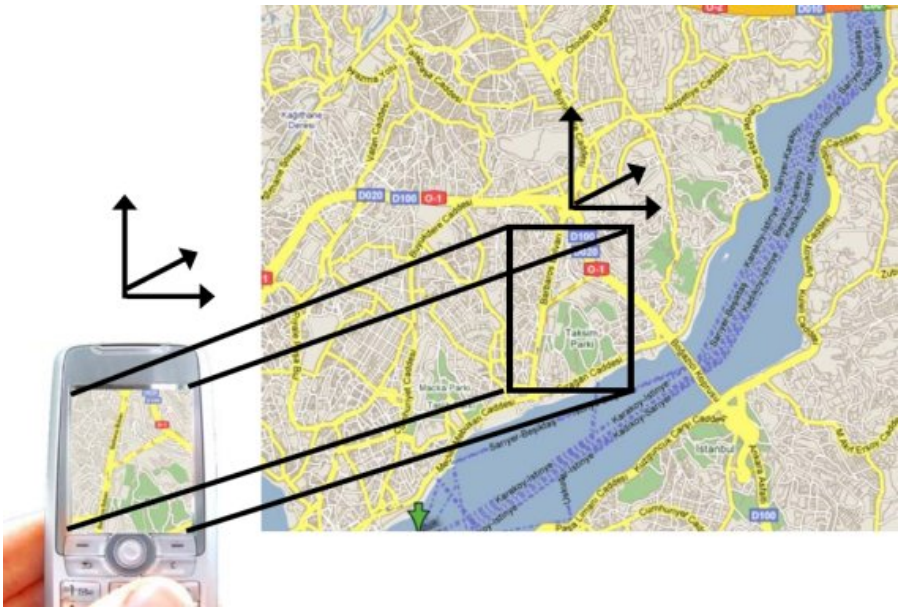


Figure 9: *Map application that interprets hand gestures for moving the viewing window on a larger map image in the memory.*

- to pan the map means modify X and/or Y according to desired pan motion
- to zoom the map means to increase or decrease the zoom ratio by amount specified by desired zoom change

It is supposed to control these three functionalities by device gestures. Navigating the map by letter gesture command (or speech), panning the map by the movement of the device in front of a stable view, and zooming the map by the movement in the direction towards the user or farther from the user.

To simulate the pan and zoom functionality before gesture input could be connected, mouse (stylus) input was implemented. The Palm application had two main modes, the pan and zoom mode. In the pan mode, the touch and movement of the stylus caused panning the map in the corresponding direction. In the zoom mode, the touch and movement of the stylus caused the zooming (movement upwards - zoom in, movement downwards - zoom out).

To demonstrate the navigate functionality there also exist several modes which zoom into the predefined places of the map. They are activated by drawing the first letter of the place in a predefined manner with the device in the air, or by choosing the name from the menu with the stylus. In the latter case the functionality could be debugged before connecting the real gesture input.

5.2. Paint Application

Paint application assumes simple drawing functions: Starting to draw a stroke, drawing a stroke by device movements, finishing to draw a stroke, changing the colour and thickness of the pen and erase. Then user may want to change position and start to draw another stroke. To accomplish this functionality, we must retain the list of stroke point coordinates and draw a line to the screen between points given by these coordinates. Dynamic array of coordinates is a good structure for this purpose, since the length of a stroke is not known in advance and is dependent on how long the user draws the stroke. To indicate any change of colour or thickness, again dynamic arrays are used.

Functionality that has to be implemented is thus

- start to draw
- draw
- finish drawing
- add color and width options (as eraser is imitated with a considerably thick white pen.)

Again, this will be controlled by gestures of the device, starting and finishing by pushing the button on the device, drawing by movements of the device. Another button is reserved for identifying the color and thickness options which are specified by drawing the first letter of the command in the air with the device in a predefined manner. For simulation, stylus events were processed and used as an alternative control.

5.2.1. Mouse/stylus input and windows messaging

To process mouse/stylus input in MS Windows environment, windows messaging system is used. In the main processing function of an application, we must process the windows messages. For mouse input, the main three messages that were used for simulation functionality are `WM_LBUTTONDOWN`, `WM_LBUTTONUP`, and `WM_MOUSEMOVE`. The first message is sent whenever user presses the left mouse button (or touches the touch-screen with the stylus), the second one is sent once the button is released (the stylus leaves the touch-screen), the third when the mouse coordinates have changed (in the case of stylus, we can of course follow the coordinates only when the stylus touches the screen).

All three messages are sent together with screen coordinates of the mouse/stylus. Thus, we have to process these coordinates to control the application. In the case of map panning functionality, on `WM_LBUTTONDOWN` we set `panning=TRUE`, and remember the coordinate, then on each `WM_MOUSEMOVE` we check if `panning==TRUE`, then get coordinates, count difference from last stored coordinates and control panning by this difference. Then we store new last coordinate. On `WM_LBUTTONUP` we finish panning by setting `panning=FALSE`. Similar approach is used for other functionalities. In zooming mode we also retain the last coordinates, but by the coordinate difference we control the zoom factor. In paint functionality, on `WM_LBUTTONDOWN` we start to draw (`drawing=TRUE`) and retain coordinates of the first point of a stroke, on `WM_MOUSEMO`

VE we check whether drawing==TRUE, then get coordinates and store them in memory. On WM_LBUTTONUP we finish the stroke and set drawing=FALSE.

For drawing to the device screen, another windows message is used: WM_PAINT. We have to process this message and paint the screen according to given control variables. It means that all the painting (showing the desired map portion or drawing a stroke) is done in this part of code whenever the Windows ask for it. In processing of the mouse messages (or the gesture input) we only modify control variables and tell the Windows that there is a change that needs repainting the window (calling InvalidateRect(...);)

Appendix 9.1 contains part of the code that illustrates the windows messaging.

6. CONCLUSIONS AND OUTLOOK

In this paper we presented our work on mobile phone based gesture recognition and its applications. The video input of a camera phone is analyzed with HBMA in order to determine simple user actions. These simple actions then analyzed further with command recognizer to generate high level commands.

Improvements to the motion recognizer is possible by incorporating tracking of image features, such as SIFT features, and prediction of motion with Kalman filtering. We presented two sample applications, map navigator and a drawing program. Many other applications, such as video games, web browsing, and handwriting recognition are possible using mobile gesture recognizer.

7. ACKNOWLEDGEMENTS

Authors thank Ricoh Innovations California Research Center for donating the mobile device for development.

This report, as well as the source code for the software developed during the project, is available online from the eNTERFACE'07 web site: http://www.enterface.net

8. REFERENCES

[1] J. Wang, S. Zhai, and J. Canny, "Camera Phone Based Motion Sensing: Interaction Techniques, Applications and Performance Study", in *ACM UIST*, 2006. 139

[2] A. Haro, K. Mori, V. Setlur, and T. Capin, "Mobile Camera-based Adaptive Viewing", in *4th International Conference on Mobile Ubiquitous Multimedia (MUM)*, 2005. 139

[3] *Palm Treo 700 w mobile device*. http://www.palm.com/us/products/smartphones/treo700w/. 139

[4] L. R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition", in *Proceedings of the IEEE*, vol. 77, feb 1989. 142

[5] D. Jurafsky and J. H. Martin, *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition*. Prentice-Hall, 2000. 143

9. APPENDIX: SOFTWARE NOTES

9.1. Applications

Below is a part of the code that illustrates the windows messaging:

```
LRESULT CALLBACK WndProc (
  HWND hWnd, UINT message,
  WPARAM wParam, LPARAM lParam)
{
 int wmId, wmEvent;
 PAINTSTRUCT ps;
 HDC hdc;
// RECT rc_inv;
 int xPos;
 int yPos;
 double half_x;
 double half_y;
(...)
 switch (message)
 {
(...)
  case WM_PAINT:
   hdc = BeginPaint(hWnd, &ps);
   DrawImage(hdc);
   EndPaint(hWnd, &ps);
  break;
(...)
  case WM_LBUTTONDOWN:
//WM_LBUTTONDOWN fwKeys = wParam;
   last_mouse_x = LOWORD(lParam);
   last_mouse_y = HIWORD(lParam);
   drawing = TRUE;
  break;
  case WM_LBUTTONUP:
   drawing = FALSE;
  break;
  case WM_MOUSEMOVE:
   if (drawing==TRUE)
   {
    xPos = LOWORD(lParam);
    yPos = HIWORD(lParam);
    shift_x = xPos - last_mouse_x;
    shift_y = yPos - last_mouse_y;
(...)
    InvalidateRect(hWnd, &rc_wnd, FALSE);
    UpdateWindow(hWnd);
   }
  break;
(...)
 }
 return 0;
}
```

9.2. Pseudo code of HBMA and EBMA

```
Input:
A = Anchor Image
T = Target Image
B = Block Size
R = Radius
N = Number of Pyramid Levels

Output:
Mv = Motion Vectors

Initialize motion vectors Mv to zero.
For n = N to 1
 Downsample anchor image A and target image T by 2^(n
      -1)
 Pad zero pixels to fit the block size
 Set the search radius r = R/2^(n-1)
 Set the block size b = B
 Apply EBMA to update the motion vectors Mv
 Upsample Mv by 2
End
```

```
Input:
A = Anchor Image
T = Target Image
B = Block Size
R = Radius
Mv = Initial Motion Vectors

Ouput:
Mv = Updated Motion Vectors

For each block in the anchor frame
 Take the next bxb anchor block Ba
 Shift to the previous estimate Mv
 For each block Bt of target frame in the radius r
  Find the block with minimum MAD error
```

```
    Update motion vectors Mv
  End
End
```

10. BIOGRAPHIES

Barış Bahar finished high school in Bursa in 2003. He is now an undergraduate student of the Faculty of Computer Engineering at Boğaziçi University, İstanbul.
Email: barisbahar86@yahoo.com

Işıl Burcu Barla graduated from the Faculty of Electrical and Electronic Engineering of Boğaziçi University, Turkey in 2007. She is now continuing her study at Technical University Munich, Germany. She has worked on speech and gesture recognition systems.
Email: burcubarla@gmail.com

Ögem Boymul finished high school in Adana in 2003. She is currently an undergraduate student of the Faculty of Electrical and Electronic Engineering at Boğaziçi University, İstanbul.
Email: ogem.boymul@boun.edu.tr

Çağlayan Dicle received his B.Sc. degree from Computer Science Department at Yeditepe University in 2004. He is now an M.S. student at System and Control Engineering at Boğaziçi University. His main research interests are computer vision and machine learning applications specifically on deformable object tracking and non-parametric classification.
Email: caglayan.dicle@boun.edu.tr

Miloš Železný was born in Plzen, Czech Republic, in 1971. He received his Ing. (=M.S.) and Ph.D. degrees in cybernetics from the University of West Bohemia, Plzen, Czech Republic in 1994 and in 2002 respectively. He is currently a lecturer at the University of West Bohemia, Plzen, Czech Republic. He has been delivering Digital Image Processing, Structural Pattern Recognition and Remote Sensing lectures since 1996 at the University of West Bohemia. He is working in projects on multi-modal speech interfaces (audio-visual speech, gestures, emotions, sign language). He publishes regularly and he is a reviewer of the INTERSPEECH conference series.
Email: zelezny@kky.zcu.cz

Berna Erol received her B.Sc. degree in Control and Computer Engineering at İstanbul Technical University and M.Sc. and PhD. Degrees in Electrical and Computer Engineering at the University of British Columbia, in 1998 and 2002, respectively. Since September 2001 she has been a senior research scientist at Ricoh California Research Center, USA. Dr. Erol has authored or co-authored more than 35 journal and conference papers, two book chapters, and more than 40 patent applications in the area of multimedia signal processing. Her main contributions to research and development consist of content-based video and image retrieval, image and video coding and representation, E-learning and E-meeting systems, text retrieval, new era multimedia systems, and applications, and multimedia processing for mobile devices. She has served in the program and the technical committees of leading ACM and IEEE conferences such as ACM Multimedia and IEEE ICME. She is an associated editor of the IEEE Signal Processing Magazine and a co-chair in the SPIE Electronic Imaging organizing committee. She had been an active participant in the video coding standardization activities such as ITU-T H.263 and MPEG-7.
Email: berna_erol@rii.ricoh.com

Murat Saraçlar received his B.S. degree from Bilkent University, Ankara, Turkey in 1994. He earned both his M.S.E. and Ph.D. degrees from Johns Hopkins University, Baltimore, MD, USA in 1997 and 2001 respectively. He worked on automatic speech recognition for multimedia analysis systems from 2000 to 2005 at the AT&T Labs Research. In 2005, he joined the Department of Electrical and Electronic Engineering at Boğaziçi University as an assistant professor. His main research interests include all aspects of speech recognition, its applications, as well as related fields such as speech and language processing, human-computer interaction and machine learning. He currently leads a TUBITAK funded project on Turkish Broadcast News Transcription and Retrieval. He authored and co-authored more than two dozen papers in refereed journals and conference proceedings. He has filed four patents, both internationally and in the US. He has served as a reviewer and program committee member for various speech and language processing conferences and all the major speech processing journals. He was the Area Chair for the Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP) in 2005. He is currently a member of IEEE and ISCA. He was recently elected to serve as member of the IEEE Signal Processing Society Speech and Language Technical Committee (2007-2009).
Email: murat.saraclar@boun.edu.tr

Tevfik Metin Sezgin graduated summa cum laude with Honors from Syracuse University in 1999. He received his MS in 2001 and his PhD in 2006, both from Massachusetts Institute of Technology. He is currently a Postdoctoral Research Associate in the Rainbow group at the University of Cambridge Computer Laboratory. His research interests include intelligent human-computer interfaces, multimodal sensor fusion, and HCI applications of machine learning.
Email: metin.sezgin@cl.cam.ac.uk

BENCHMARK FOR MULTIMODAL AUTHENTICATION

Morgan Tirel [1], *Ekin Olcan Şahin* [2], *Guénolé C. M. Silvestre* [3], *Clíona Roche* [3], *Kıvanç Mıhçak* [2], *Sinan Kesici* [2], *Neil J. Hurley* [3], *Neslihan Gerek* [2], *Félix Balado* [3]

[1] University of Rennes, France
[2] Boğaziçi University, Turkey
[3] University College Dublin, Ireland

ABSTRACT

We report in this document on the development of a multimodal authentication benchmark during the eNTERFACE' 07 workshop. The objective of creating such a benchmark is to evaluate the performance of multimodal authentication methods built by combining monomodal authentication methods (i.e., multimodal fusion). The benchmark is based on a graphical user interface (GUI) that allows the testing conditions to be modified or extended. It accepts modular monomodal authentication algorithms (feature extraction, robust hashing, etc) and it allows them to be combined into multimodal methods. Attacks and benchmarking scripts are similarly configurable. An additional output of the project is a multimodal database of individuals, which has been collected in order to test the benchmark.

KEYWORDS

Benchmarking – Multimodal authentication – Feature extraction – Robust hashing

1. INTRODUCTION

Traditional authentication of individuals has usually been focused on methods relying on just one modality. Typically these modalities can be images of faces, hands (palms), irises or fingerprints, or speech samples. For instance, one may take a photo of the face of a person and obtain from it a nearly unique low-dimensional descriptor that identifies that person. Depending on the particular application targeted, this identifier can be obtained by means of different types of methods. Typical examples are feature extraction methods or, under some conditions, robust hashing methods, e.g. [1], [2]. The identifiers thus obtained can be compared to preexisting ones in a database for a match. Authentication systems based on multimodal strategies – that is, joint strategies– combine two or more monomodal methods into a multimodal one. For instance, it is possible to combine one method to hash an image using face images and another method to obtain a feature vector from a palm image. This is sometimes referred to as multimodal fusion. The aim is to increase the reliability of the identification procedure when combining different sources of information about the same individual (see [3], for example). As we will see, some other considerations are necessary in order to optimally undertake the merging of different multimodal methods.

Over the last number of years, many algorithms applicable to authentication have been proposed. Although some of these methods have been partially analyzed in a rigorous way, in many cases it is not feasible to undertake exhaustive analytical performance analyses for a large number of scenarios. This in part due to the sheer complexity of the task. Nevertheless, it is necessary to systematically evaluate the performance of new methods, especially when they are complex combinations of existing methods and used in a variety of scenarios. With such an evaluation it becomes possible to determine the best authentication strategies.

One way to tackle this problem is by means of benchmarking. Benchmarks have been proposed in the past for performance evaluation of many technologies, ranging from CPU units to watermarking technologies [4]. An advantage of benchmarks is that they see methods for testing as black boxes, which allows a high degree of generality. Despite this great advantage, one must be aware that benchmarks also entail issues such as how to choose fair (unbiased) conditions for benchmarking without an exponential increase in the associated computational burden.

The main goal of the eNTERFACE Workshop Project number 12 has been to create a GUI-driven benchmark in order to test multimodal identification strategies. This technical report contains information on the planning and development of this project. The remainder of this document is organized as follows. In Section 2 we describe the basic structure of the benchmark. In Section 3 we give the benchmark specifications which have been used as guidelines for implementing the benchmark, while Section 4 describes the methods and functions implemented to be tested within the benchmark. Finally, Sections 5 and 6 describe the database collection effort and the tests undertaken, while Section 7 draws the conclusions and future lines of this project.

2. DESCRIPTION OF THE BENCHMARK

Early in the project preparations, it was decided to implement the benchmark prototype in Matlab. This decision was taken in order to speed up the development time, as Matlab provides a rather straightforward procedure to build GUI applications, and it is faster to write Matlab code for the development of methods to be included in the benchmark. The downside is inevitably the execution speed, which can be critical for completing benchmark scripts within a reasonable timeframe. Nevertheless C code can also be easily interfaced to Matlab, using so called Mex files. The prototype is meant to be both usable and extendable, in order to facilitate the inclusion of new items and features. The interface has been designed so that extension or modification of the benchmark is almost completely automated. An exception is the addition of new benchmarking scripts (see Section 2.4), in order to keep the benchmark implementation simple. This means that it is possible to do most operations through the GUI, and manual adjustments of the source code are only necessary for the less frequent action of adding new types of benchmarking scripts. A scheme showing the relationships between the different parts of the benchmarking system is shown in Figure 1.

The benchmark relies on a database storing all relevant data. This is implemented in MySQL and interfaced to Matlab. The purpose of this database architecture is two-fold. Firstly, it is

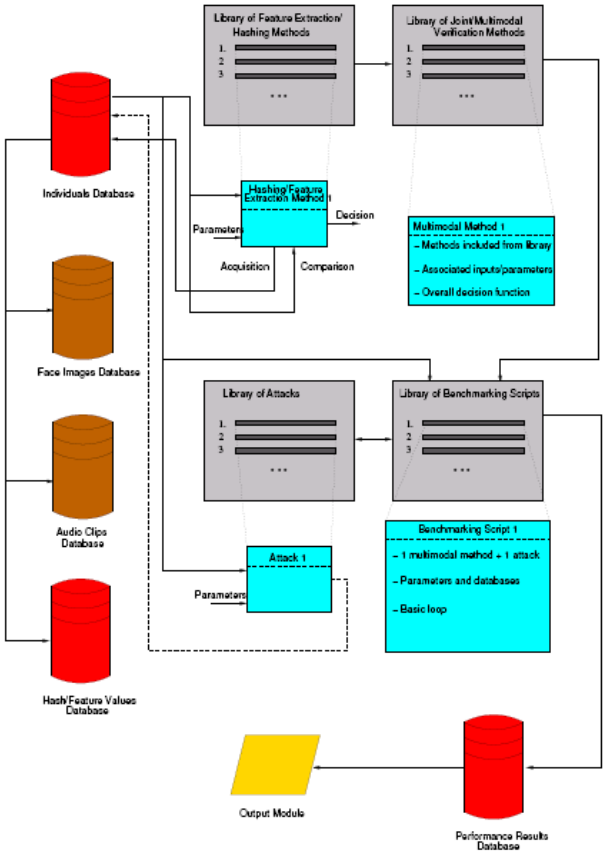


Figure 1: *Relationships between the main parts of the benchmark.*

an efficient way to store and access the information; secondly, it allows easy sharing of the data over a network in order to parallelize the benchmark in the future, thus distributing the unavoidable computational burden of the benchmark.

The project requires a database of individuals featuring signals such as face images, hand images and speech. The details on the database collection task are given in Section 5. All this information is stored in the MySQL database together with the identifiers (i.e., extracted features, hash values) obtained from the individuals, and all libraries of methods and functions. In order to minimize the effects of intra-individual variability, which especially affects some robust hashing algorithms (see for instance [5]), the database of individuals includes several instances of each identifier corresponding to a given individual.

The benchmark admits new modules through four libraries (see Figure 1) whose function we describe next.

2.1. Library of monomodal methods

This library contains standard monomodal methods which can be added, removed or edited through the GUI (see Section 3.6). For each method two functions are defined:

- An acquisition function, that takes as input a file containing a signal of the given modality (e.g., an audio clip or image) associated with a particular individual, as well as function-dependent parameters, such as thresholds and other. It outputs an identifier vector, binary or realvalued, depending on the method. The output identifier is stored in the database associated with the individual whose signal has been used.
- A comparison function, which takes as input two identifier vectors plus any necessary parameters, and outputs both a Boolean (hard) decision of similarity between them, and a (soft) reliability measure. This reliability shows the degree of confidence we put in the decision which is put forward by the function. As we will discuss in the next section, it is a key element in order to optimally combine two different modalities.

2.2. Library of multimodal methods

This library contains methods which, relying on the library in Section 2.1, specify ways to combine two (or more) monomodal methods in order to create multimodal identifiers. We may view this operation as an instance of multimodal fusion. For instance, the system allows the combination of a method to robustly hash face images with a method to extract features from a fingerprint; the newly created method is stored in the library as a multimodal method.

As already discussed, it is fundamental that each multimodal method implements an overall comparison function, able to break ties between possibly contradictory monomodal decisions when looking for matches in the database. Let us denote by e_1 the difference between the two input identifiers to the comparison function for modality type 1, and let us call d_1 the outcome of the monomodal binary decision, mapped without loss of generality to $+1$ and -1. If D_1 represents the random variable associated with that decision, with possible values $D_1 = +1$ (the two input identifiers correspond to the same individual) and $D_1 = -1$ (otherwise), the optimal monomodal decision is given by:

$$d_1 = \text{sign}\left(\log\frac{Pr\{D_1 = +1|e_1\}}{Pr\{D_1 = -1|e_1\}}\right). \quad (1)$$

We may see the log-likelihood ratio as the reliability of the decision. We propose to obtain the overall decision d_F for the fusion of M modalities as

$$d_F = \text{sign}\left(\sum_{k=1}^{M} w_k \cdot \log\frac{Pr\{D_k = +1|e_k\}}{Pr\{D_k = -1|e_k\}}\right), \quad (2)$$

where the subindex k refers to the modality k used in the fusion, and w_k is a set of positive weights such that $||\mathbf{w}||^2 = 1$. These weights reflect the importance that we wish to grant to each modality in the multimodal fusion. Note that in order to implement Eq. 1 accurate statistical modelling is required in order to obtain the conditioned probabilities, which may not always be feasible. In fact, many feature extraction and robust hashing methods implement this comparison function in a mostly heuristic way. If the reliability measures above are not available, it is always possible to implement a weaker version of Eq. 2 using the hard decisions:

$$\tilde{d}_F = \text{sign}\left(\sum_{k=1}^{M} w_k \cdot d_k\right). \quad (3)$$

2.3. Library of attacks

It accepts attack functions on the signals stored in the individuals database. Attacked signals are used to assess how robust multimodal methods perform in two different situations:

1. The inputs are distorted versions of the authentic signals.
2. The inputs are non-authentic (malicious) signals, aiming at being wrongly verified as authentic.

2.4. Library of benchmarking scripts

It lists scripts which may be run in batch mode (i.e., autonomously), using signals from the database, a multimodal method, and attacks suitable to the modalities involved. Performance measures such as the rates of detection and false alarm (obtained by comparison with the authentic identifiers) will be computed during the execution of the script. In the scripts there may be loops where some attack parameters are generated pseudo-randomly.

3. BENCHMARK SPECIFICATIONS

We describe next the specifications that were used as technical guidelines to implement the benchmark. The most important structures and functions are described with some level of detail.

3.1. Individuals database

The basic structure of an entry in the individuals database is given by the following structure:

```
struct(
  'name',{},
  'authenticated',{},
  'file_list',struct(
    'name',{},
    'path',{},
    'type',{}
  ),
  'hash_list',struct(
    'method_name',{},
    'h_value',{}
  )
)
```

`h_value` may contain `double` or `char` values depending on the particular output of the method: some authentication methods methods output binary vectors, whereas others output real vectors.

Example: the 3rd individual `dbi(3)` in the database `dbi` with the structure above could be

```
dbi(3).name='joe'
dbi(3).authenticated=1
dbi(3).file_list(1).name='joe1.jpg'
dbi(3).file_list(1).path = '/tmp/'
dbi(3).file_list(1).type = 'face'
dbi(3).file_list(2).name='joe2.jpg'
dbi(3).file_list(2).path = '/tmp/'
dbi(3).file_list(2).type = 'face'
dbi(3).file_list(3).name='hand1.jpg'
dbi(3).file_list(3).path = '/tmp/'
dbi(3).file_list(3).type = 'hand'
dbi(3).file_list(4).name='joe1.jpg'
dbi(3).file_list(4).path = '/tmp/'
dbi(3).file_list(4).type = 'wav'
dbi(3).hash_list(1).method_name='philips_method'
dbi(3).hash_list(1).h_value='adsfdasbasdfsdsafsa'
dbi(3).hash_list(2).method_name='mihcak_method'
dbi(3).hash_list(2).h_value='qqvx&3242rew'
```

Notice that two hash string values are associated to this individual, corresponding to the output of the corresponding functions in the library of hashing/feature extraction methods (see next section). The `dbi` variable is duly stored in the MySQL database.

3.2. Library of monomodal methods

The basic structure of entries in this library is:

```
struct(
  'method_name',{},
  'media_type',{},
  'hash_function',struct(
    'name',{},
    'parameters_list',{}
  )
  'comp_function',struct(
    'name',{},
    'parameters_list',{}
  )
)
```

As discussed in Section 2.1, every monomodal method will have a hash function and a comparison function associated. The benchmark accepts functions whose prototype for the acquisition is

```
string h_value = function hash_f(string file,
    parameters)
```

and for the comparison

```
[boolean decision,double reliability] = function
    comp_f(string h_value1,string h_value2,parameters
    ).
```

If `decision=1` then the hash strings `h_value1` and `h_value2` match according to the comparison function, whereas `decision=0` means they do not. The `reliability` parameter ranges indicates how good the decision is.

Example: the 2nd method in a monomodal library `mml` with the structure above could be:

```
mml(2).method_name='philips_method'
mml(2).media_type='audio'
mml(2).hash_function.name='philips_hash'
mml(2).hash_function.parameters_list={0.37,0.95}
mml(2).comp_function.name='philips_comp'
mml(2).comp_function.parameters_list=.9
```

The files `philips_hash.m` and `philips_comp.m`, which must be in the path, implement the corresponding acquisition function

```
h_value = function philips_hash(file,frame_size,
    overlap),
```

and comparison function

```
[decision,reliability] = function philips_comp(
    h_value1,h_value2,threshold).
```

The `mml` array variable is stored in the MySQL database.

3.3. Library of multimodal methods

The basic structure of entries in this library will be:

```
struct(
  'method_name',{},
  'monomodal_methods_list',{},
  'comp_weights',{},
  'attack_list',{}
)
```

The generation of a multimodal hash entails the execution of all the monomodal methods whose names are listed in `monomodal_methods_list` on all corresponding file types of a given individual (image, audio). This generates a series of monomodal identifiers which are incorporated into the structure in Section 3.1.

As discussed in Section 2.2, the comparison of multimodal identifiers requires an overall function in order to break ties between two (or more) monomodal comparison functions (e.g. two monomodal methods that are fused into a multimodal one can give contradictory decisions when using the monomodal comparison functions). According to that discussion we implement this function using the `reliability` parameter furnished by monomodal comparison function, and using a set of weights `comp_weights`. This set is a list of values between 0 and 1 that adds up to 1; each value corresponds to a function in `monomodal_methods_list`, in order to weight the importance of the monomodal methods in the overall comparison. The multimodal decision will be 1 if the weighted sum of monomodal reliabilities is greater than 0.5, and 0 otherwise (note that we have mapped for convenience $\{+1,-1\}$ to $\{1,0\}$ with respect to Section 2.2).

Example: the 1st entry in the multimodal library `MM1`, with the structure described above, could include two methods from the monomodal library. The first method was described above. Let us assume that the second method is of `media_type='image'`.

```
MM1(1).method_name='MM_first'
MM1(1).monomodal_methods_list={'philips_method','
    mihcak_method'}
MM1(1).comp_weights={.45,.55}
MM1(1).attack_list={'gaussian','random'}
```

The `MM1` array variable is stored in the database. The overall comparison for the multimodal function `MM_first` will be 1 if (cf. Eq. 2)

```
r_1*comp_weights(1)+r_2*comp_weights(2)>0.5
```

where `r_1`,`r_2` are the reliabilities given by the comparison functions of the two monomodal methods.

3.4. Library of attacks

The basic structure in this case is

```
struct(
  'media_type',{},
  'attack_function',struct(
    'name',{},
    'parameters_list',{}
  )
)
```

Each element `parameters_list(i)` is a triplet indicating a range {`starting_value,step,end_value`}. The prototype of an attack function is

```
string attacked_file = function attack_function_name(
    string file,parameters)
```

where `file` is the full patch of a file of type `media_type`.

Example: a simple unintentional attack can be Gaussian noise addition on audio (or image) files. For instance, assume that the first element `atl(1)` in the array of attacks `atl` with the structure above implements Gaussian noise addition for audio files:

```
atl(1).media_type='audio'
atl(1).attack_function.name='g_noise'
atl(1).attack_function.parameters_list={{.5,.1,2}}
```

The function `g_noise.m` which must be in the execution path will have a header

```
attacked_file = function g_noise(file,power)
```

More complex attack functions can be defined after this type of simple attacks is properly implemented.

The `atl` array variable will be stored in a MySQL database and interfaced to the Matlab code.

3.5. Library of scripts

Benchmark scripts undertake simulations of the effect of attacks on the performance of multimodal methods, relying on the database of individuals and on the multimodal and attacks libraries. Scripts are implemented as loops sweeping the parameter range of a given attack, while computing the rates (i.e., empirical probabilities) of miss/false alarm when using a given multimodal method and attack:

- The rate of miss is computed as the percentage of authenticated individuals not correctly matched.
- The rate of false alarm is computed as the percentage of non-authenticated individuals (incorrectly) matched to authenticated individuals.

In order to simplify the GUI implementation, the structure of benchmark scripts is defined by templates. For the creation of a new script, a list of predefined templates is offered to the user. Upon choosing a multimodal method and suitable attacks from the corresponding lists, a script is created based on the template chosen. The newly created script is stored in the library of scripts. The basic structure to add a script to the library is

```
struct(
  'script_name',{},
  'template_name',{},
  'script_path',{},
  'run_status',{},
  'multimodal',{}
)
```

A resettable Boolean variable indicates whether the script has been run by the benchmark already.

`script_path` gives the full name of the `.m` benchmark script file and `run_status` indicates whether the script hasn't been run yet, it is currently running, or it has been run. The output of the script will be found by default in a file with extension `.output.mat`, with the same name without extension as `script_path`. The output file containing the results from running the benchmarking script is timestamped and included in the database.

Example: the pseudocode of a script template may be:

```
- acquire 'multimodal hash'
for all individuals for all authenticated individuals
     in database
  for all 'ranges' of 'attack'
    - 'attack' individual
    - compute 'multimodal hash' of attacked individual
      for all hashes in the library
      - 'compare hash' with attacked hash
      - compute rate of miss
      end
  end
end
```

Using this particular template, the creation of a benchmark script would require to fill in the terms in inverted commas, that is, basically the multimodal method and the attack from the corresponding libraries. Templates will be Matlab files with dummy strings placed where the functions or parameters must be filled in.

For instance, the first method in the variable `scl`, containing the scripts library with the structure defined above, could be

```
scl(1).script_name='gaussian'
scl(1).template_name='template_1'
scl(1).script_path='/home/scripts/gaussian_script.m'
scl(1).run_status=2
scl(1).multimodal='newhand_face'
```

The output of this script will be found by default in the file `gaussian_script.output.mat`. The `scl` array variable is stored in the MySQL database.

3.5.1. Output module

Completed tasks will allow the user to plot the output resulting from running the benchmark script. The output file will store a fixed structure that will allow the output module to produce plots. It is the responsibility of the template to produce the right output file. This output file will contain a structure variable called `output` with the following form:

```
struct(
  'plot_list',struct(
    'xlabel',{},
    'ylabel',{},
    'title',{},
    'x',{}
    'y',{}
  )
)
```

Note that the vectors `plot_list.x` and `plot_list.y` must have the same size. An output plot will typically show ROC plots (probability of false alarm versus probability of detection), or these probabilities for different thresholds or noise levels.

Text reports about the benchmarking results are also produced. A text report may include details such as functions and parameters used, number of iterations, database signals used, and quality measures obtained.

Example: In `gaussian_script.output.mat` we may find the structure

```
output.plot_list(1).xlabel='Probability of Miss'
output.plot_list(1).ylabel='Noise Variance'
output.plot_list(1).title='Gaussian Additive Noise'
output.plot_list(1).x=[0.1 0.2 0.3 0.4 0.5]
output.plot_list(1).y=[0 0.01 0.05 0.075 .1]
```

More than one plots may be found in `plot_list`, and the user should be able to browse all of them.

3.6. Workflow

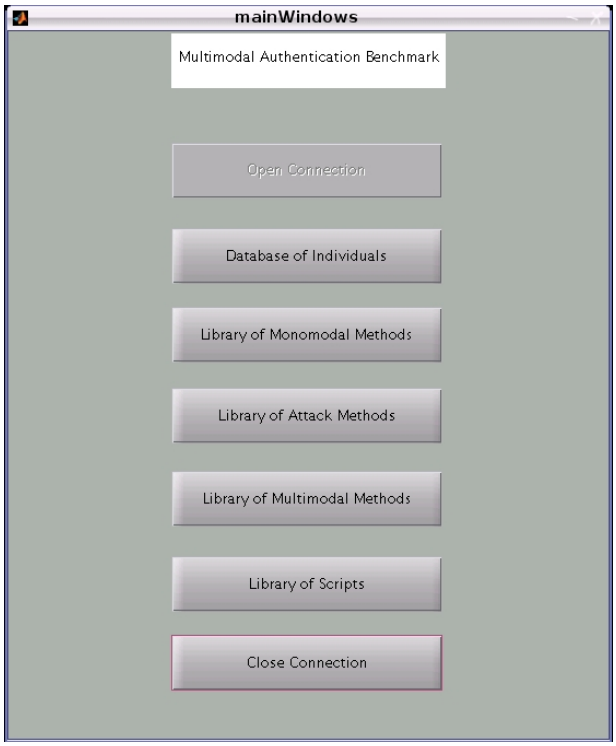


Figure 2: *Main window of the benchmark GUI.*

The main benchmark window in Figure 2 features several buttons which allow access to the subwindows which are described next as well as providing the interface for connecting and disconnecting the GUI from the database. The windows were designed keeping simplicity in mind and using the `guide` tool of Matlab. This tool generates a standard `.m` file associated to each window (`.fig` file). This file can be edited in order to implement the callback functions required by the buttons and other window objects.

3.6.1. *Database of individuals window*

The interface allows the user to:

- Browse, add and remove audio clips associated with each face image (these face images and audio clips must come in pairs).
- Generate hashes for images and audio clips as they are added to the database.

3.6.2. *Library windows*

These windows allow the user to browse the corresponding libraries and to add and remove functions. The libraries of monomodal methods and attacks accept names of external Matlab functions, whose headers we have defined above. It is also possible to enter the desired parameters for these functions. The library of multimodal methods accepts combinations of functions in the library of monomodal methods and an associated weight and attack function for each of these monomodal methods.

The libraries follow the structures defined in Section 3.

The library of scripts also allows the user to

- Generate a new script using an existing template and multimodal function.
- Run one script or all of the scripts, preferably as background processes. The window displays the `run_status` of each script - 0 if not run, 1 if currently running and 2 if run.
- Plot the outputs of scripts with `run_status=2`. Plots are generated from the files `*.output.mat` as described in Section 3.5.1.
- Generate a report detailing the inputs and outputs of the script e.g. the multimodal, monomodal and attack functions used.

4. METHODS AND FUNCTIONS IMPLEMENTED

In this section we briefly review the features of the methods and attacks that were implemented in order to test the benchmark capabilities.

4.1. Monomodal methods

4.1.1. *Image Hashing*

- Iterative Geometric Hashing [6]. Two algorithms are proposed. The first one (algorithm A) initially shrinks the input while keeping its essential characteristics (low frequency components). It is recommended in [6] to use to this end the discrete wavelet transform (DWT). However, a three-level DWT takes quite a long time in Matlab. Instead, we shrink the image linearly. Next, geometrically significant regions are chosen by means of simple iterative filtering. The reason for keeping geometrically strong components while minimizing geometrically weak ones is that a region which has massive clusters of significant components is more resilient to modifications. The second algorithm proposed in [6] (algorithm B) simply applies algorithm A on pseudorandomly chosen regions of the input.
- NMF-NMF-SQ. This algorithm is based on a dimensionality reduction technique called *nonnegative matrix factorization* (NMF) [7]. The NMF method uses nonnegative constraints, which leads to a parts-based representation of the input. The algorithm implements a two-stage cascade NMF, because it is experimentally shown in [7] that this serves to significantly enhance robustness. After

obtaining the NMF-NMF hash vector, a statistics quantization (SQ) step is undertaken in order to reduce the length of the hash vector.

- PRSQ (Pseudo-Random Statistics Quantization). This algorithm is based on the assumption that "the statistics of an image region in a suitable transform domain are approximately invariant under perceptually insignificant modifications on the image" [7]. After shrinking the input (i.e., obtaining its low-frequency representation), a statistic is calculated for each pseudo-randomly selected and preferably overlapping subregions of the gist of the input. Scalar uniform quantization on the statistics vector yields the final hash vector.

4.1.2. Audio Hashing

If we assume that the conditions are such that a speaker is able to approximately repeat the same utterance (as when a fixed text is read aloud), then audio hashing algorithms can be used for identifying voice clips.

- Microsoft Method [8] (also known as Perceptual Audio Hashing Algorithm). It computes the hash value from robust and informative features of an audio file, relying on a secret key K (seed to pseudorandom generators). An algorithmic description is given below:

  1. The input signal X is put in canonical form using the MCLT (Modulated Complex Lapped Transform) [9]. The result is a time-frequency representation of X, denoted by T_X.
  2. A randomized interval transformation is applied to T_X in order to estimate statistics, μ_X, of the signal.
  3. Randomized adaptive quantization is applied to μ_X yielding $\hat{\mu}_X$.
  4. The decoding stage of an error correcting code is used on $\hat{\mu}_X$ to map similar values to the same point. The result is the intermediate hash, h_X.

  The estimation of the signal statistics is carried out using Method III (see [8]), which relies on correlations of randomized rectangles in the timefrequency plane. For perceptually similar audio clips, estimated statistics are likely to have close values, whereas for different audio clips they are expected be different. The method applies frequency cropping to reduce the computational load, exploiting the fact that the Human Auditory System cannot perceive frequencies beyond a threshold.

- Boğaziçi Method [5]. This algorithm exploits the time-frequency landscape given by the frame-by-frame MFCCs (mel-frequency cepstral coefficients) [10]. The sequence of matrices thus obtained are further summarized by choosing the first few values of their singular value decomposition (SVD) [5]. The actual cepstral method implemented is an improvement on [11].

- Philips Fingerprinting [12]. This method is an audio fingerprinting scheme which has found application in the indexing of digital audio databases. It has proved to be robust to many signal processing operations. The method is based on quantizing differences of energy measures from overlapped short-term power spectra. This staggered and overlapped arrangement allows for excellent robustness and synchronization properties, apart from allowing identification from subfingerprints computed from short segments of the original signal.

4.1.3. Hand Recognition

The benchmark includes one algorithm for recognition of hands, based on [13]. The algorithm takes as input images of hands captured by a flatbed scanner, which can be in any pose. In a pre-processing stage, the images are registered to a fixed pose. To compare two hand images, two feature extraction methods are provided. The first is based on measuring the distance between the contours representing the hands being compared, using a modified Hausdorff distance. The second applies independent Component Analysis (ICA) to the binary image of hand and background.

4.2. Attack functions

4.2.1. Image Attack Functions

- Random Bending Attack. This attack distorts the image by modifying the coordinates of each pixel. A smooth random vector field is created and the pixels are moved in this field. The vector field must be smooth enough so that the attacked image is not distorted too much. An iterative algorithm is applied to create the horizontal and vertical components of the vector field separately. In each iteration, a Discrete Cosine Transform (DCT) is applied and high frequency components removed. The attack function is designed for grayscale images; color images are tackled using the luminance. The parameters of the attack are the strength of the vector field, the cutoff frequency for the DCT filtering, the maximum number of iterations, and a smoothness threshold.

- Print Scan Attack. Floyd and Steinberg's [14] error diffusion algorithm is applied to transform each of the components of a color image to bilevel values (0 or 1). The algorithm processes the pixels in raster order. For each pixel, the error between the bilevel pixel value and the image pixel value is diffused to the surrounding unprocessed pixel neighbours, using the diffusion algorithm. After processing all pixels, the image is filtered by an averaging filter.

- Contrast Enhancement. This function increases the contrast of the input image using the `histeq` histogram equalization function of Matlab. An input parameter specifies a number of discrete levels N, and the pixel values are mapped to these levels to produce a roughly flat histogram. Histogram equalization is applied separately to the three components of a color image.

- Rotation and Crop Attack. This function rotates the input image by a specified angle, relying on a specified interpolation method. Because we include crop in `imrotate` function we just have the central portion of the rotated image in the output. The input parameters are the rotation angle and the interpolation type (bilinear, nearest neighbor or bicubic interpolation).

- Noise Attack. This function adds noise of a specified variance to the input image using the `imnoise` function of Matlab. Four different types of noise are supported, namely Gaussian noise, Poisson noise, salt & pepper noise, and speckle noise.

- Simple Chimeric Attack. An image is pseudo-randomly selected from the database and a weighted average of the image with the input image is created, using weights given as input to the attack function. The two images are not registered before the averaging, and hence the resulting image does not correspond to a true morphing of the

images. Nevertheless, if the weight of the randomly selected image is sufficiently strong in comparison to that of the input image, it can be expected that this attack may be useful to benchmark recognition algorithms.

4.2.2. Audio Attack Functions

Three audio attacks have been included in the benchmark:

- Noise Attack. This attack adds noise to the audio signal. The strength of the added noise is determined by an input parameter that represents the signal to noise ratio. Another parameter specifies the distribution of the noise, which can be Gaussian, uniform, Gamma or Rayleigh.
- Delay Attack. The audio signal $x(t)$ is summed with a delayed version of itself, $x'(t) = x(t)+\alpha x(t-t_0)$ which produces an echo effect. The delay t_0 and the weight α of the delayed signal are input parameters.
- Pitch Bending Attack. This attacks modifies the pitch of the audio signal without changing its duration. Firstly, the length of the audio signal is changed while retaining the original pitch, using the Matlab toolbox "phase vocoder" [15]. Then the signal is decimated/interpolated in order to recover the original length, what changes the pitch. An input parameter determines whether the pitch is compressed or stretched.

4.3. Templates

Two basic templates are defined in the prototype of the benchmark:

- The one described in the example in Section 3.5
- A template able to generate and test attacks that generate modified/chimeric characters. It is essentially the same as the previous one, but, as it generates non-authenticated individuals, it focuses on computing the probability of false alarm instead of the probability of miss.

```
- acquire 'multimodal hash' for all individuals
for all authenticated individuals in database
  for all 'ranges' of 'attack'
    - 'attack' individual (i.e. non-
        authenticated individual)
    - compute 'multimodal hash' of attacked
        individual
    for all hashes in the library
      - 'compare hash' with attacked hash
      - compute rate of false alarm
    end
  end
end
```

5. MULTIMODAL DATABASE

Data was collected from 92 subjects. For each individual, six face images were collected, two scans of the palm of the left hand and three videos. Subjects stood in front of a blue background canvas and the room was illuminated by artificial light. The face images were taken with the subject's head in the following positions:

1. Directly facing the camera;
2. Directly facing the camera, with glasses removed, if subject was wearing glasses in the first picture;
3. Head turned $-90°$;
4. Head turned $-45°$;
5. Head turned $+45°$;
6. Head turned $+90°$.

The three videos recorded the subject carrying out the following activities:

1. Reading a fixed text;
2. Performing four gestures;
3. Moving head while speaking freely.

The only part of the video data used in the multimodal benchmark is the audio portion of the first video. In this audio signal, the subject is recorded reading a fixed English text of around 100 words. The participants consisted of an international group of mainly non-native English speakers. Subjects were asked to ignore any mistakes made in pronunciation and to continue reading to the end of the text. The final part of the text is the numbers one to ten, which were read twice. The rest of the video data was collected for use in another eNTERFACE project incorporating gesture recognition.

6. TESTS

6.1. Database of Libraries

The database consists of both the data collected and libraries of the various monomodal, multimodal and attack functions. It has been written as a simple relation database. The database has been integrated into the Matlab code by means of a Mex file in which a number of standard SQL commands have been defined e.g. `insert`, `select`. The benchmark is capable of running, and has been tested, under Linux, Mac OS and Windows.

6.2. Testing & Debugging

To illustrate the use of the benchmark, a multimodal identifier was created through the benchmark consisting of a face identifier and a hand identifier, both constructed using the iterative geometric image hashing algorithm A. Each monomodal method was set to report a match when the reliability of the comparison function was above a threshold of 0.8. The face identifier was weighted 0.6 in the multimodal combination. In a benchmark script, a print scan attack was applied to both the face and hand images, varying the window size of the averaging filter and tested on a small database of 5 individuals. The script produced a report containing the probability of a miss i.e. the probability that the multimodal identifier failed to correctly identify the individual, for different window sizes. Note that since there are 6 face images associated with each individual and two palm scans, 12 combinations of face and hand images could be tested per person. The script output the probability of mis-identification which is plotted in Figure 3.

The output plot in Figure 4 shows the results of simulating the effect of rotation and crop on a multimodal fusion using algorithms A and B in Section 4.1.1, applied to face images and hand images respectively. The database of 92 individuals was used. We see here that the benefit of the multimodal method over either of the modalities on its own.

7. CONCLUSIONS AND FUTURE WORK

Although the basic structure of the benchmark is fully functional, some issues have inevitably arisen during the short period of time allowed for the development of this project. The main issue faced at the end of the workshop were the computational problems posed by the Matlab implementation of the methods. The computational burden associated to the amount of iterations within a benchmarking script may pose difficulties to complete

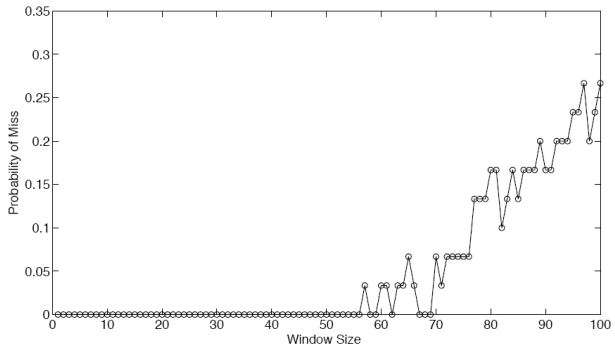


Figure 3: *Probability of a miss for a multimodal face and hand identifier under the print & scan image attack.*

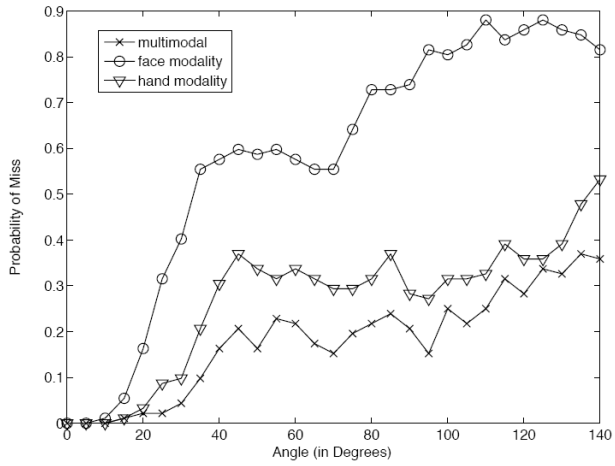


Figure 4: *Probability of a miss for a multimodal face and hand identifier under the rotation and crop attack.*

the simulations within a reasonable amount of time unless the methods used are optimized.

In a similar fashion as other benchmarks [16], it would be interesting to endow this benchmark with a certification procedure. This procedure would entail defining a fixed set of attacks – obviously dependent on the modalities – and benchmarking scripts. The report obtained from this certification procedure would be used to rank methods in a more systematic way. However establishing what type of attacks and scripts will be include can be controversial, as the rankings obtained may be biased if they are not carefully chosen.

8. ACKNOWLEDGEMENTS

The members of project #12 thank Prof. B. Sankur for contributing the source code of the hand feature extraction method developed by himself and his group.

This report, as well as the source code for the software developed during the project, is available online from the eNTERFACE'07 web site: http://www.enterface.net

9. REFERENCES

[1] F. Ahmed and M. Siyal, "A secure biometric authentication scheme based on robust hashing", in *Proceedings of the 5th International Conference on Information, Communications and Signal Processing*, (New York, USA), pp. 705–709, Dec 2005. 147

[2] Y. Sutcu, H. T. Sencar, and N. Memon, "A secure biometric authentication scheme based on robust hashing", in *Procs. of the 7th workshop on Multimedia and security*, (New York, USA), pp. 111–116, October 2005. 147

[3] A. Ross and A. Jain, "Information fusion in biometrics", *Pattern Recognition Letters, Special issue: audio- and video-based biometric person authenticaion (AVBPA 2001)*, vol. 34, pp. 2115–2125, September 2003. 147

[4] F. A. Petitcolas, "Watermarking schemes evaluation", *IEEE Signal Processing*, vol. 17, pp. 58–64, September 2000. 147

[5] H. Özer, B. Sankur, N. Memon, and E. Anarım, "Perceptual audio hashing functions", *EURASIP Journal on Applied Signal Processing*, no. 12, pp. 1780–1793, 2005. 148, 152

[6] M. K. Mıhçak and R. Venkatesan, "New iterative geometric methods for robust perceptual image hashing", in *Procs. of ACM Workshop on Security and Privacy in Digital Rights Management*, (Philadelphia, USA), 2001. 151

[7] V. Monga and K. Mıhçak, "Robust image hashing via non-negative matrix factorizations", in *IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 2, pp. 14–19, May 2006. 151, 152

[8] M. K. Mıhçak and R. Venkatesan, "A perceptual audio hashing algorithm: A tool for robust audio identification and information hiding", in *Procs. of the 4th Information Hiding Workshop*, vol. 2137 of *Lecture Notes in Computer Science*, (Pittsburgh, USA), pp. 51–65, Springer, April 2001. 152

[9] H. S. Malvar, "A modulated complex lapped transform and applications to audio processing", in *IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 3, pp. 1421–1424, March 1999. 152

[10] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*. Prentice Hall, 1978. 152

[11] B. Logan, "Mel frequency cepstral coefficients for music modelling", in *Procs. of International Symposium on Music Information Retrieval*, (Indiana, USA), October 2000. 152

[12] J. Haitsma, T. Kalker, and J. Oostveen, "Robust audio hashing for content identification", in *Procs. of the International Workshop on Content-Based Multimedia Indexing*, (Brescia, Italy), pp. 117–125, September 2001. 152

[13] E. Yoruk, E. Konukoglu, B. Sankur, and J. Darbon, "Shape-based hand recognition", *IEEE Trans. Image Processing*, vol. 15, pp. 1803–1815, July 2006. 152

[14] R. Ulichney, *Digital Halftoning*. The MIT Press, 1987. 152

[15] D. P. W. Ellis, "A phase vocoder in Matlab", 2002. http://www.ee.columbia.edu/~dpwe/resources/matlab/pvoc. 153

[16] J. Vorbruggen and F. Cayre, "The Certimark benchmark: architecture and future perspectives", in *IEEE Intnl. Conf. on Multimedia and Expo*, vol. 2, pp. 485–488, 2002. 154

10. BIOGRAPHIES

Morgan Tirel is a M.Sc. student at the University of Rennes, France.
Email: morgan.tirel@etudiant.univ-rennes1.fr

Ekin Olcan Şahin received in 2006 his B.Sc. in Electrical Electronics Engineering at Bilkent University, Ankara, Turkey. Currently he is a M.Sc. student at the Electrical & Electronics Engineering Department of Boğaziçi University, Turkey.
Email: ekin.sahin@boun.edu.tr

Guénolé C. M. Silvestre received the M.Sc. degree in electronic and electrical engineering in 1993. In 1996, he received the Ph.D. degree from the University of Dublin, Trinity College, Ireland, for his work in silicon-on-insulator materials. As a post-doctoral fellow with Trinity College, Dublin, he pursued research on digital image processing and watermarking. In 1997, he was appointed Research Scientist at the Philips Research Laboratories, Eindhoven, The Netherlands, and his research focus switched toward the study of polymer light-emitting materials. In 1999, he joined the National University of Ireland (University College Dublin), where his present research activities lie in the area of digital communications, data-hiding, and signal processing. Dr. Silvestre was the 1995 recipient of the Materials Research Society Graduate Student Award.
Email: guenole@ihl.ucd.ie

Clíona Roche received her Bachelor of Arts (Hons) in Computer Science (major) and History of Art, at University College Dublin (UCD), Dublin, Ireland. Currently she is a Ph.D. student in Computer Science at the same university on the topic of High throughput comparative modelling of protein structure by machine learning.
Email: cliona.roche@ucd.ie

Kıvanç Mıhçak was born in Turkey in 1974. He received his B.S. degree from Electrical and Electronics Engineering Department, Bilkent University, Ankara, Turkey in 1996 and the M.S. and Ph.D. degrees from Electrical and Computer Engineering Department, University of Illinois, Urbana-Champaign (UIUC), in 1999 and 2002 respectively. At UIUC, he was in the Image Formation and Processing Group; his thesis advisors were Pierre Moulin and Kannan Ramchandran. Between 2002 and 2005, he was a researcher, with the Cryptography & Anti-Piracy Group at Microsoft Research, Redmond, WA. Currently, he is assistant professor with the Electrical and Electronic Engineering Department of Boğaziçi University.
Email: kivanc.mihcak@boun.edu.tr

Sinan Kesici was born in 1984, in Erzincan, Turkey. He is currently an undergraduate student in Boğaziçi University, Turkey. He is a senior student in Electrical & Electronics Engineering Department. His specialization option is Telecommunication Engineering.
Email: sinan940@yahoo.com

Neil J. Hurley received the M.Sc. degree in mathematical science from University College Dublin (UCD), Dublin, Ireland, in 1988. In 1989, he joined Hitachi Dublin Laboratory, a computer science research laboratory at the University of Dublin, Trinity College, from which he received the Ph.D. degree in 1995, for his work in knowledge-based engineering and high-performance computing. He joined the National University of Ireland, University College Dublin, in 1999 where his present research activities lie in the areas of data-hiding, signal processing, secure and robust information retrieval and distributed computing.
Email: neil.hurley@ucd.ie

Neslihan Gerek is a M.Sc. student at Boğaziçi University, İstanbul, Turkey.
Email: neslihan.gerek@gmail.com

Félix Balado graduated with an M.Sc. in Telecommunications Engineering from the University of Vigo (Spain) in 1996, and received a Ph.D. from the same institution in 2003, for his work in data hiding. He then joined the National University of Ireland (University College Dublin) as a post-doctoral fellow at the Information Hiding Laboratory. Previously he worked as a research and project engineer at the University of Vigo, in different research projects funded by the Galician and Spanish Governments, and by the European Union. His research interests lie in the areas of multimedia signal processing, data hiding, and digital communications.
Email: fiz@ihl.ucd.ie

Index of Authors

ISBN : 978-2-87463-105-4

UCL PRESSES
UNIVERSITAIRES
DE LOUVAIN