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.

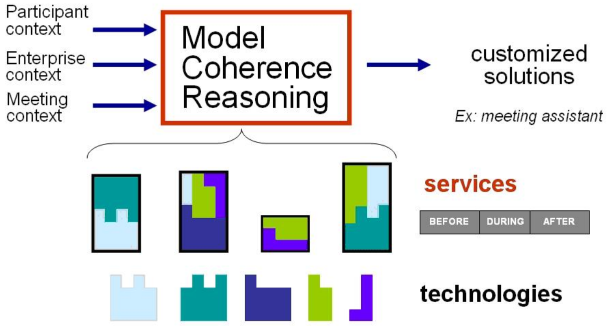


Figure 1: *This illustration conceptually depicts the architecture to create technology for synchronous collaboration. The reasoning piece of the architecture is built based on the participant's context and the task interaction so to provide the services needed to make the meeting more efficient. These 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 (N-GN).

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

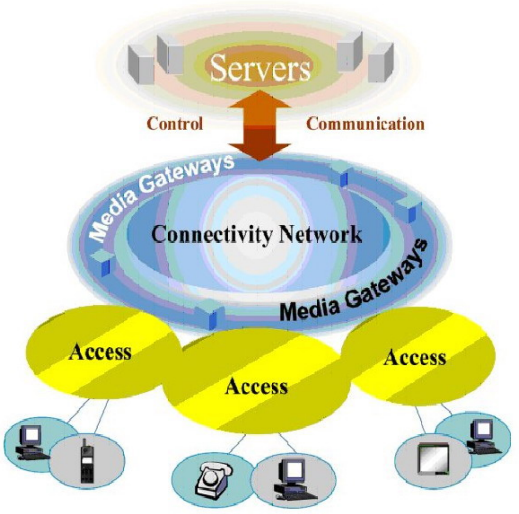


Figure 2: *Multi service networks (illustration courtesy of ITU-T).*

(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.

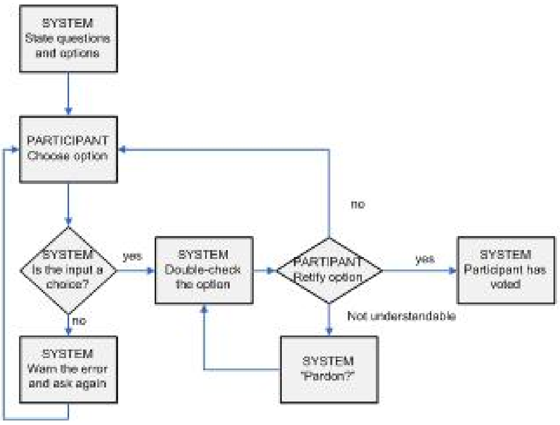


Figure 4: *This schema illustrates the interaction between the participant and the automatic text vote registration.*

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 auto-correlation 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 pre-defined 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.

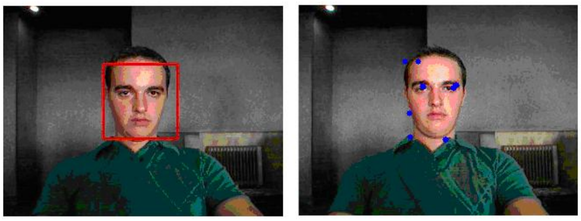


Figure 5: *Detected face & feature points that will be tracked.*

6.3. Experiments and Results

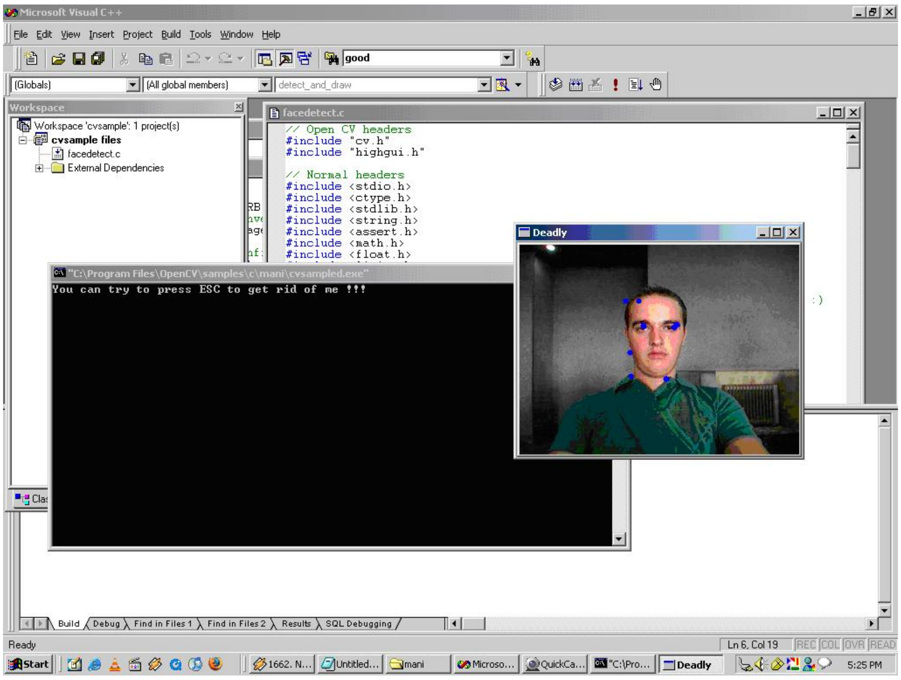
The development and testing was carried out on a basic 1.5 MHz laptop with 512 MB of RAM, without any specialized equipment. Video input of the frontal face was provided by a standard webcam with a resolution of 320X240 at 30 f/s.

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.



```
// Open CV headers
#include "cv.h"
#include "highgui.h"

// Normal headers
#include <stdio.h>
#include <ctype.h>
#include <stdlib.h>
#include <string.h>
#include <assert.h>
#include <math.h>
#include <float.h>
```

```
You can try to press ESC to get rid of me !!!
```

Figure 6: *Screen Shot of the System.*

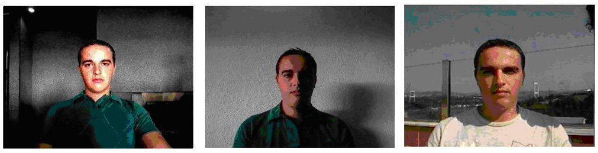


Figure 7: *Light Variation.*

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.

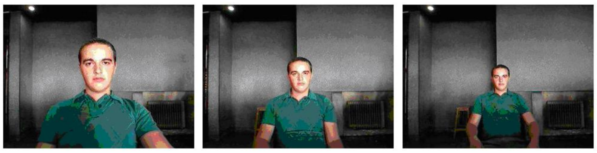


Figure 8: *Scale Variation.*

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.

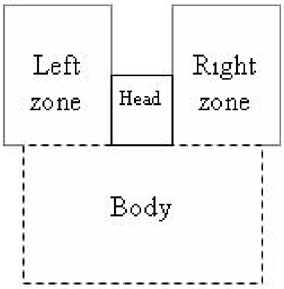


Figure 9: *Search zones.*

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.

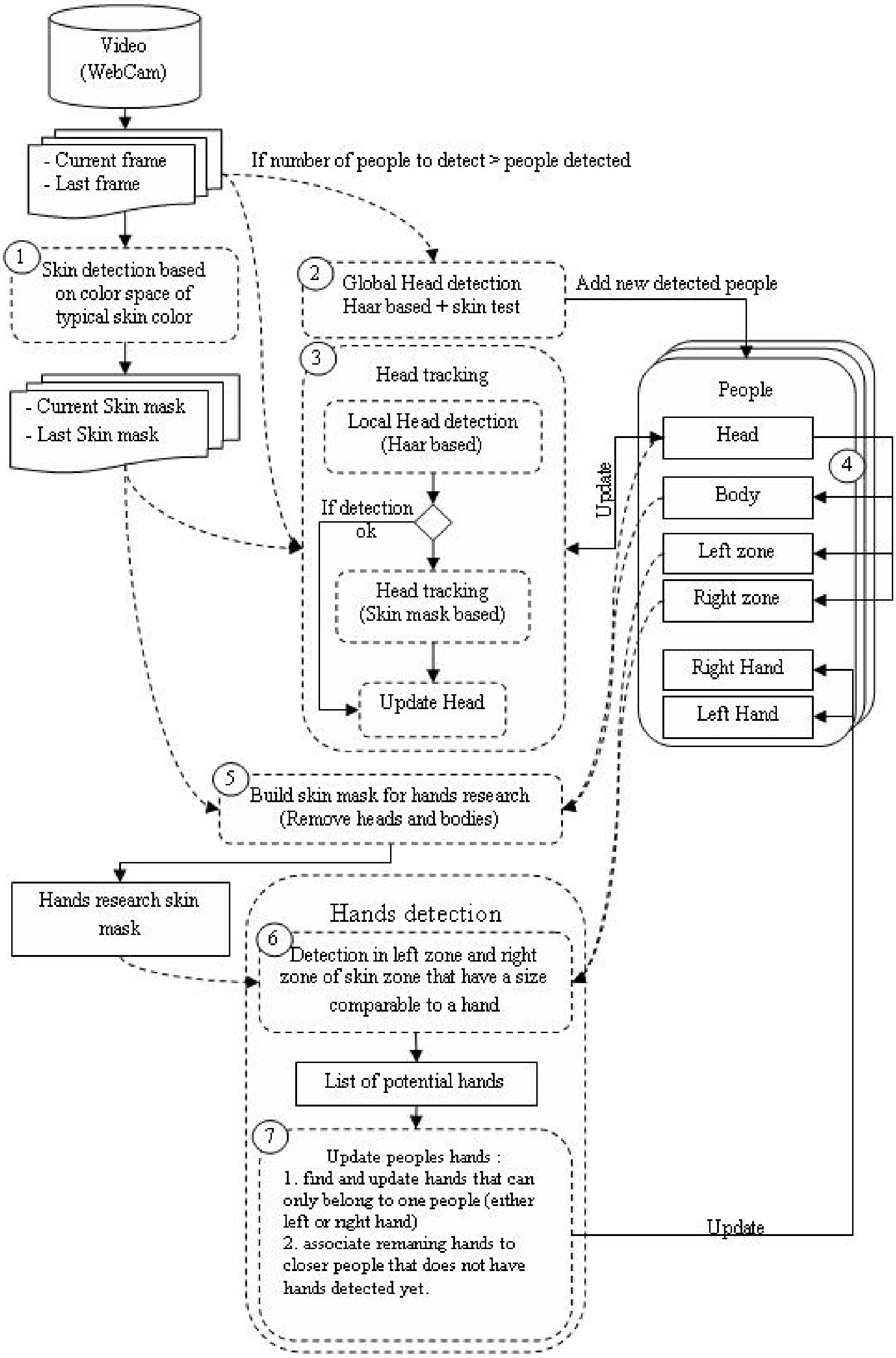

Video
(WebCam)
- Current frame
- Last frame
If number of people to detect > people detected
1
Skin detection based on color space of typical skin color
2
Global Head detection
Haar based + skin test
Add new detected people
- Current Skin mask
- Last Skin mask
3
Head tracking
Local Head detection
(Haar based)
If detection ok
Head tracking
(Skin mask based)
Update Head
People
Head
Body
Left zone
Right zone
Right Hand
Left Hand
4
Update
5
Build skin mask for hands research
(Remove heads and bodies)
Hands research skin mask
Hands detection
6
Detection in left zone and right zone of skin zone that have a size comparable to a hand
List of potential hands
7
Update peoples hands :
1. find and update hands that can only belong to one people (either left or right hand)
2. associate remaining hands to closer people that does not have hands detected yet.
Update

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.

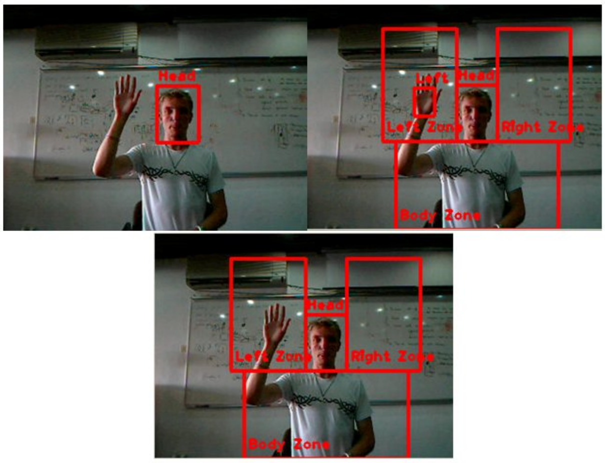


Figure 11: *Single-user: step 2(head), step 4(Research zones) and step 7(Hands).*

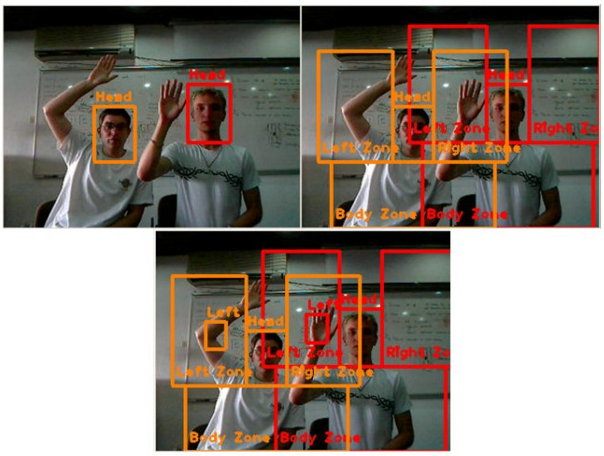


Figure 12: *Multi-user: step 2(head), step 4(Research zones) and step 7 (Hands).*

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

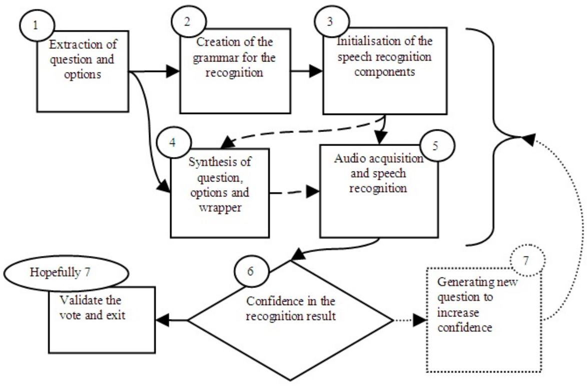

1
Extraction of question and options
2
Creation of the grammar for the recognition
3
Initialisation of the speech recognition components
4
Synthesis of question, options and wrapper
5
Audio acquisition and speech recognition
6
Confidence in the recognition result
Hopefully 7
Validate the vote and exit
7
Generating new question to increase confidence

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

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.

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.

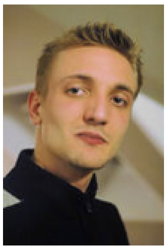
Therefore, protocols for real-time applications must be worked out before the real multimedia time comes. As a solution to this matter the Integrated Services working group in the Internet Engineering Task Force developed an enhanced Internet service model that includes best-effort service and real-time 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

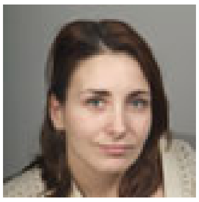
Contents of the Source Code folders:

- **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.

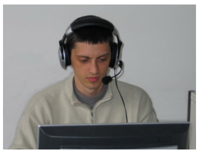
14. BIOGRAPHIES



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

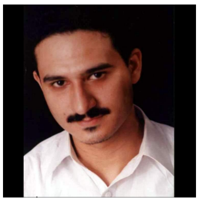


Ana C. Andrés del Valle (M'04) was born in Barcelona, Spain, in 1976. She received the M.Sc. degree in telecommunications from the Technical University of Catalonia, Barcelona, in 1999 and the Ph.D. degree in image and signal processing from Tlcom Paris, Paris, France, in 2003. She is currently a Senior 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