Proceedings of the 4th International Summer Workshop on Multi-Modal Interfaces (eNTERFACE?08)

Christophe d'Alessandro

eNTERFACE'08

Proceedings of the 4th International Summer Workshop on Multi-Modal Interfaces

LIMSI-CNRS

Orsay, 4-29 august 2008

Printed in France

Distribution :
Available on order at
LIMSI-CNRS
BP 133
F91403 Orsay, France
Tel. 33 1 69 85 8X XX
Fax 33 1 69 85 80 88
bibliotheque@limsi.fr

Cover illustration: © Nathalie d'Alessandro

Edited by Christophe d'Alessandro
Audio & Acoustics Group, LIMSI-CNRS, BP 133, F91403 Orsay Cédex, France
cda@limsi.fr
www.limsi.fr

EDITORIAL

In this digital age, Multimodal Interfaces are extending and replacing mechanical and electrical instruments. Impressive results have been achieved and new challenging applications are foreseen in fields like medicine, assistance to disabled persons, communication, design, performing arts, games, and many other types of human activity involving computer interaction.

Research on Multimodal Interfaces is a challenging interdisciplinary field, blending perception and action, sensory modalities and cognition, physics and the humanities, computer science and signal processing, theory and experiments, and last but not least science and art.

The summer 2008 has been a bit rainy in Paris, like it was in Mons in 2005, the ideal weather to welcome at LIMSI researchers in Multimodal Interfaces worldwide for the eNTERFACE08 workshop, the 4th of a series of successful workshops initiated by SIMILAR, the European Network of Excellence (NoE) on Multimodal interfaces.

eNTERFACE'08 followed the fruitful path opened by eNTERFACE'05 in Mons, Belgium, and continued by eNTERFACE'06 in Dubrovnik, Croatia and eNTERFACE'07 Istanbul, Turkey.

The result was very successful, with 78 registered participants, working in 10 projects. It was also very international, as participants were coming from 20 different countries, and working in laboratories in 15 different countries.

The ultimate goal of this workshop was to make this four weeks event a unique opportunity for students and experts to meet and work together, and to foster the development of tomorrow's multimodal research community.

During the workshop we were delighted to welcome very distinguished speakers. We had the unique opportunity to listen to invited conferences and to discuss with the speakers in a friendly and studious atmosphere:

The 10 workshop projects explored many aspects of multimodal interfaces, including speech recognition and speech synthesis, embodied conversational agents, robots, haptic interface, various physiological signals, video cameras, body suit for motion capture. Smell and taste were the only missing sensory channels in these research projects. This was one of the main LIMSI staff's contribution, as we did our best to compensate this sensory deficiency by organizing barbecue parties

It is my pleasant duty to acknowledge the help and support of those people and institution that render this event possible.

The European Network of Excellence SIMILAR came to an end in 2007; this meant that we had no more financial support from SIMILAR for eNTERFACE'8. This year was the first post-SIMILAR workshop, with a different funding model. Launching out in this

project was challenging, because without the SIMILAR support it was difficult to guess what would be the response of the multimodal research community.

In this new situation, we received the assistance and financial support of

- LIMSI-CNRS (for organization, rooms and facilities, equipment and support, network, and all on-site details)
- The OpenInterface Foundation for invited speakers, grants and room reservation (thanks to the UCL people under the supervision o Pr. Benoît Macq: Olga Vybornova, Jean Deschuyter, and Colin Michel).
- the COST2101 and COST2102 actions (European Cooperation in the field of Scientific and Technical Research) for student grants,
- the eNTERFACE'08 steering committee and particularly its President, Thierry Dutoit.
- and last but not least, most of the support came from the participants' organizations for travel and leaving expenses.

We gratefully thank all our speakers for sharing their time and ideas with eNTERFACE'08 people. All their contribution is available on the eNTERFACE'08 web site: http://enterface08.limsi.fr/

Special thanks to those colleagues at LIMSI who committed themselves in eNTERFACE organization: Rami Ajaj, Albert Rilliard, Claudine Delcarte-Dang-Vu, Christian Jacquemin, Sylvain Le Beux , Nicolas Sturmel, Martine Charrue, Nadine Pain, Sophie Pageau-Maurice, Jean-Claude Martin, Nicolas Rajaratnam, Elisabeth Piotelat, Pierre Durand.

The invaluable and manifold help of Laurent Pointal is gratefully acknowledged.

And finally many thanks to all the participants, I hope you all had as much pleasure at working this summer at LIMSI as we had at organizing the event.

Chistophe d'Alessandro
Chair, eNTERFACE'08

Implementing a Multiparty Support in a Tour Guide system with an Embodied Conversational Agent

Aleksandra Čereković (1), Hsuan-Hung Huang (2), Takuya Furukawa (2), Yuji Yamaoka (3), Igor S. Pandžić (1), Toyoaki Nishida (2), Yukiko Nakano (4)
(1) Faculty of Electrical Engineering and Computing, University of Zagreb, Croatia, (2) Graduate School of Informatics, Kyoto University, Japan, (3) Department of Computer, Information and Communication Sciences, Tokyo University of Agriculture & Technology, Japan (4) Dept. of Computer and Information Science, Faculty of Science and Technology - Seikei University, Tokyo, Japan

Abstract

In this project we aim to implement multiparty support in a Tour guide system with an Embodied Conversational Agent (ECA). We studied multiparty dialogue issues and adapted some of them to our system. As the outcome of the project, the first prototype of a tour guide system has been developed. The system can automatically detect presence of humans and invite them to a tour session. During the session, according to the number of users and their positions, ECA uses a gaze model implemented from conversational theory to address a specific user. ECA also reacts to decreased level of attention and tries to keep the interest in the session. Although conversational model in our system is rather simple, we find this work a first step for further system development in the ECA-human multiparty domain. Experiments with such systems can provide useful research studies in the field of interaction between ECA and humans.

***Index Terms*— multiparty communication, multiparty issues, ECA gaze model**

I. PRESENTATION OF THE PROJECT

A. Introduction

Embodied Conversation Agents (ECAs) offer great promise to achieve natural interaction: they are anthropomorphic and have the potential to act like humans. Effects of ECAs are being studied through different roles they have in educational purposes, information delivery services, health care, sales and marketing. These studies indicate that ECAs can improve user engagement and significantly increase users' positive perceptions of learning experiences [10]. Despite all the positive effects, ECAs still mostly exist in research laboratories. As a recent state-of-the-art paper [14] claims, there is substantial research to be done on the way towards believable human-like ECAs.

So far, research community has mostly been oriented towards making ECAs capable of handling dyadic communication with a human or with another ECA. Proposed dialogue models have focused on monitoring just one conversational partner, which enables the ECA to participate in a "real" conversation. In multiparty ECA systems, most issues belong to the dialogue management domain. In [9] Traum identifies those issues and groups them into three parts:

- **Participants' roles:** identification of local roles of participants which shift during interaction (addressee, listener, speaker), responsibilities in the dialogue, social roles
- **Interaction management.** Managing a communication flow in a multiparty system is far more difficult than in a dyadic system. Some of the difficult issues are how to give and recognize a turn and how to handle different channels (and backchannels). Besides, conversations can be easily split and merged together and attention can be paid to several persons.
- **Grounding and obligations** are notions commonly used to model local state of dialogue. In multiparty communication usage of these models can become very complex; e.g. if there are more addresses, it can become unclear what a proper grounding should be.

The issues identified by Traum have been used in the MRE project [12] for creating a negotiation model for virtual agents. Although this model is being used for several agents and one human, the ideas and factors presented can also be used as general guidelines for implementing multiparty support in ECA systems. Since this research was published after the workshop, it was outside the scope when developing the eNTERFACE '08 project.

Apart from Traum [1, 9, 12] there are relatively few researchers who studied conversation of multiple users and ECA. In [11] Rehm and Wisnner point out that "the literature on the behavior of multiple users is sparse". Exception is the work from Vertegaal et al. [7] who studied gaze behavior shifts depending on conversational flow. They perform a real study on humans gazing in three-party conversation and propose a computational model of gazing which follows the findings from their study. They note that on average, subjects looked about 7 times more at the individual they listened to than at others and about 3 times more at an individual they spoke to than at others. They conclude that gaze, or looking at faces, is an excellent predictor of conversational attention in multiparty conversations. Rehm and Wissner [8] developed a

gambling system in which ECA plays a dice game with two humans. Game rules define the system scenario and make a simple dialogue model in which turn-taking mechanism and participants' roles are round-based. Their system lacks the active gaze model which follows human users. By using this system Rehm and Andre [11] investigated human gazing patterns in interaction with a real human and agent. Conclusions from their study are the same as Vertegaal's, however they also find that "People spent more time looking at an agent that is addressing them than at a human speaker." This phenomenon can be explained by the fact that prolonged eye contact in a social interaction can be found as impolite and rude; hence, the agent in this game might be considered an artifact and not a human.

Summarizing the state of the art, we conclude that the area of multiparty ECA systems is so far very little studied. A basis in communication theory has been set up by Traum, gaze behaviors have been studied in terms of conversational attention, but what lacks are systems which support it and which can serve as a test bed for identifying other issues or drawing interesting conclusions. In this project we aim to create a multiparty ECA system which will interact with two human users. Since we had a tour guide system at our disposal [6], we decided to use some of existing components (character animation player, platform) and investigate how the multiparty aspect can be incorporated into this system. Although it looked simple at first sight, multiparty communication support forced us to develop completely new components.

B. Domains and challenges

In the system design phase, we planned it in such a way that it can recognize and respond to typical situations in a multiparty conversation. We added certain additional features we found mandatory or interesting also for dyadic communication.

1) The multiparty aspect

Since our research group lacks knowledge in natural language processing and understanding we decided to predefine a system scenario and predict users' behaviours. We chose a dialogue model in which the tour guide ECA has a role of narrator and keeps the initiative during the tour session. Narration is related to Dubrovnik's cultural heritage, history, and monuments in each of the five specific scenes we use in the system. During interaction users can interrupt the ECA and ask questions or just make a comment about the current topic. We predict topic-related questions and define it by using specific keywords in the speech recognition engine (such as "where, "when", "how"). To keep attention during the session ECA can also ask users simple "yes/no" answer-based questions.

Regarding the multiparty/multimodal aspect, we reference Traum's paper [9], focus on a limited number of issues and propose their solution in our system. Since our system doesn't use advanced dialogue models, there is no grounding, goals, or natural turn-management:

- **Appearance of users. The** system can dynamically recognize how many users are present and where they are standing. Besides, it can invite users to join the session. Depending on the number and positions of users, we use the agent's gaze direction to maintain the conversational flow. If there is nobody present, the system will be restarted.
- **Channel management.** We combine users' verbal channels with nonverbal behaviours (face orientation, gazing) to resolve users' utterance. However, the system can understand only a limited number of users' behaviours: decreased level of attention, making a request and leaving the system.
    - **Speech collision.** During interaction it might occur that users ask questions at the same time. We handle this issue by using the agent to address one of the speakers and give him a turn, e.g. tour guide agent Dubravka says "You can speak one by one, I'll respond to you both. You can be the first one (addressing one user by pointing)"
- **Identification of conversational roles.** The system identifies local roles of participants during the session. Communication workflow is defined in the scenario; the agent keeps initiative and during the session he is the speaker most of the time, while the users are addressees. In the case when a user speaks, we identify him by using a localized microphone. It is a situation when the agent becomes an addressee, unless this user turns to another user and they start to communicate with each other.

Additionally in the multiparty domain we add two system features:

- **Handling users' mutual conversation.** This is a situation which we handle in "unnatural way". When users start to communicate with each other, the agent will try to get attention and continue his narration. Example is behaviour: "(waving) OK, let's continue with the tour". Since ECA can not understand the topic of their conversation we were forced to choose "a step back from empathy", but still retain ECA's politeness.
- **Inviting users to join the session. The** system can recognize human observers which are standing in the background and watching the interaction. They are treated as a group and, if present, the system will address them only if there are free places for users available. Example is the situation when user leaves the system and there is one person standing in the background. Then, Dubravka will look at the back and say "We have one free place, please come closer and join the tour".

2) Additional Features

We are also concerned with system features which can make interaction with the system fluent: the system can recover from failures, such as failure of speech recognition. In

case the user's speech is not recognized, we propose two-stage recovery. First, agent Dubravka will ask the user to repeat their question, and in case second failure occurs she will respond: "I'm sorry. I don't know a response to your question".

We also use the feature of Okao's vision [17] to detect gaze direction and recognize a situation in which the user is paying less attention to the session. If this occurs, agent Dubravka will speak until the end of her utterance, turn to the user and say "Seems you are not interested in this topic. Would you like to hear something else?"

3) Discussion

Issues addressed above are certainly not the only ones which must be solved to establish and maintain normal three-party interaction with ECA in a narration-based session. For this project these issues were general guidelines to implement a test bed for further research. After being integrated, it was planned to evaluate the system with random users (not during the workshop), and use the findings in further implementation.

C. Overview of the work done

The work has been organized along the system design: system input, central part and system output. For mediating and synchronizing messages between system components in real-time we use GECA framework [5]. System requirements are mapped to component functions and during the workshop they are developed distinctly. At the beginning we had to reject some prior "ambitious" ideas (e.g. usage of BML [13] realizer which was supposed to be completed before the workshop). By the end of the workshop we didn't complete all plans; but we managed to integrate the first prototype of the tour guide system. The system is capable of running the first scene in which we aimed to handle all planned issues.

Input part is bimodal: it captures users' verbal and nonverbal behaviours (face orientation and gazing direction). For verbal input we produced a component which detects specific keywords and defines the user's preferences. As for nonverbal input we produced two components which are combined to resolve the following situations: users' arrival and departure, numbers of users present, situations in which users are paying less attention. Unfortunately, one month of the workshop wasn't sufficient to complete a tracking algorithm which was mandatory to detect a situation when users are starting their own conversation (they rotate their faces to each other and activate verbal channels). Also, we didn't have time to deal with image processing and divide the space to distinguish between the users and audience. This issue turned out to be rather complex because the area where the system was installed wasn't large enough (we couldn't distinguish between users and audience members by using only size of the face since there was no space behind the users). That's the reason why we allow a maximum of two users in the system's scope.

To manage input multimodalities and choose an appropriate agent's response we initially intended to use MIDIKI's toolkit [4]. Idea was to integrate multiparty support in MIDIKI's state update mechanisms: e.g. nonverbal events and specification of gazing. During the workshop we realized that toolkit's documentation was inadequate for fast implementation and in the end we adapted an existing scenario component [6]. However, not all issues in the central part are implemented: it lacks speech collision handling and there is no smart scheduling and behaviour control (e.g. continuous control of gazing).

In the animation player we implemented patterned gaze behaviour towards users. Patterns follow findings from computational theory [7, 11]. We also modelled behaviours (gestures, facial expression) to increase ECA's believability. Animation player also lacks some features. We are not handling users' comments at the moment because there is no smart control for this level. We handle interruptions by deleting all queued behaviours and waiting for new behaviours from the central part of the system.

In the remainder of this paper, we describe the overall system. First, we depict and explain hardware configuration, while in the third chapter we explain system design and components' functionalities in detail. For the end, we discuss the system, evaluation and conclude the work completed on eNTERFACE '08 workshop.

II. Hardware Configuration

System was installed at LIMSI, which was venue of eNTERFACE '08 workshop. It is depicted on Fig. 1.

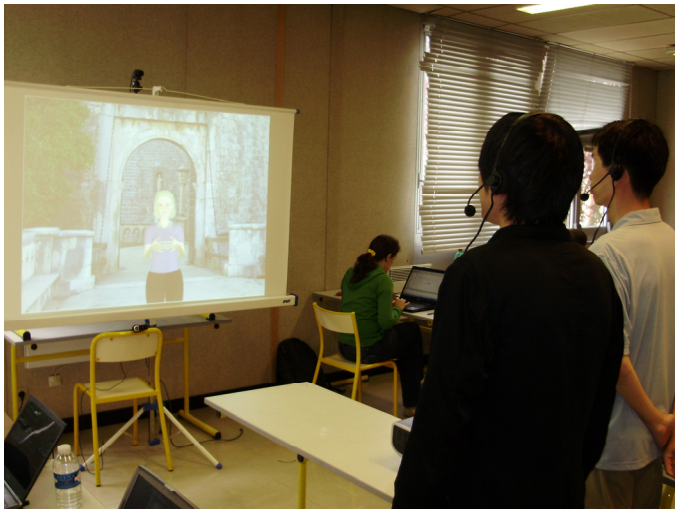


Figure 1. System installed at LIMSI

Two equal microphones are installed on two machines under the same settings. During installation we had to make sure that distance between microphones was great enough to avoid speech overlapping between microphones. Results of several experiments showed that Loquendo ASR [21] was stable enough not to recognize speech from approx. 0,5 m between users. We positioned a table between users to make sure they are not standing next to each other (Fig. 1).

System captures users' nonverbal behaviours by using two cameras; one sends data to Motion Detector which detects motions in the system's scope, while the other is connected to

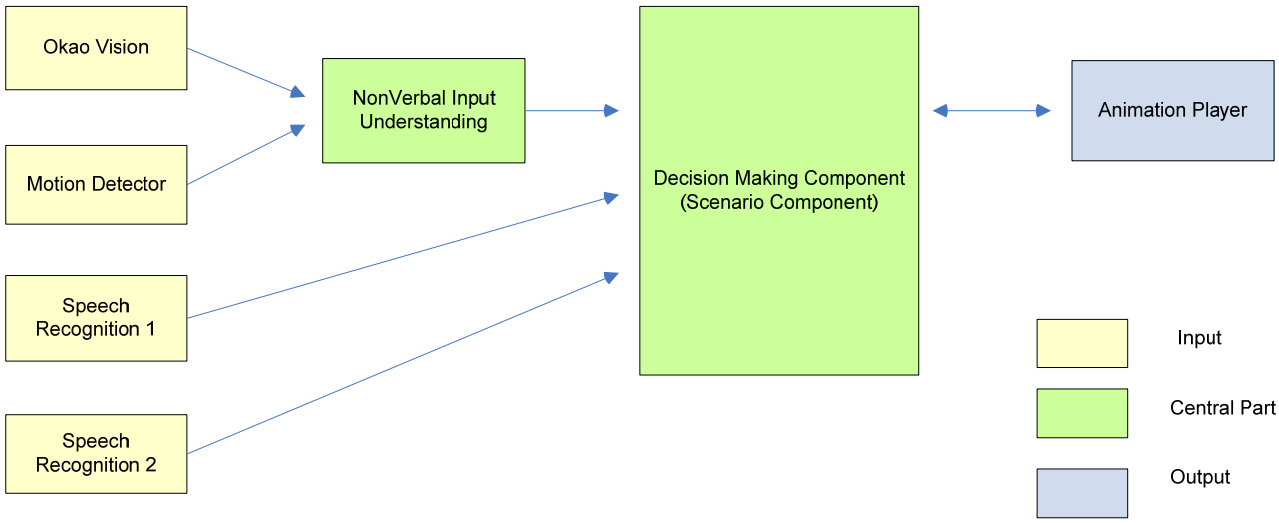


Figure 2. System design

Okao's vision component which can recognize face orientation and gaze direction if face doesn't exceed 60 degrees relative to the position of the camera (more details about how these components work will be presented in the third section.) To handle data from cameras we use another computer.

Scenario component runs on the third machine. Itdoes not inquire any additional hardware, and neither does GECA server.

For system output we use another machine, projector and a large panel to display the agent and its behaviours. We calibrate the window position to make the agent large enough to make the conversation circumstances more natural (Fig.1)

III. System Design and Components

System design is depicted in Fig. 2. System features explained in the first section are granulated and mapped into components which are divided into three categories: input, central part, and system output. All designed components are connected to GECA platform which can transport and mediate messages in real time.

A. System Input

Input data processed within the system are human speech, appearance, faces and direction of gazing. Combination of this data can resolve one the following situations:

- User arrival
- Number of users
- Interruptions
- Unrecognized speech
- Decreased level of attention
- Leaving the system

We also planned to detect and handle speech collision, conversation between users and presence of the audience (to invite them if there is a free place in the system). However, there was no time to complete the tracking algorithm and divide the space. Speech collision detection issue is simple from the system input point of view, but takes more effort on the central part.

1) Speech Recognition

Speech recognition component is developed with Loquendo ASR [21]. Experiments performed at the very beginning of the workshop showed that Loquendo is very satisfying for our needs; it is stable in noisy environments, speaker-independent, there is no significant speech overlapping at reasonable distance between users (0,5 meters approx.), and no keyword spotting issue, which was a problem with Microsoft SAPI. Besides, it has garbage-rules definition to reject words that are not defined in vocabulary, confidence score recognition and timing specification of recognized words, which are additional features used in our project.

Once the speech recognition component is started, it waits for users' speech input. When it receives a speech start signal, it starts recognizing the speech. The system also recognizes who is speaking. The results of recognizing the speaker and speech content are sent to OpenAirServer. The process flow is as follows.

1. Once the Speech Recognition component is started, this module starts listening to the audio stream.
2. When there is an utterance input from the user, this module gets the start time of utterance from Loquendo, as well as the utterance duration.
3. Finally, it transmits the results obtained in flow 2 as character string. It repeats flow2 and flow3.

SR sometimes reacts to voices which are not from the user. We think such an error can be prevented by using a Speaker Verification function in Loquendo.

2) Nonverbal Behavior Recognition

To detect nonverbal behaviours we developed two components: Okao Vision and Motion Detector. In cooperation these components have two functions: they can

detect the number of users and calculate the user's attention level towards the agent.

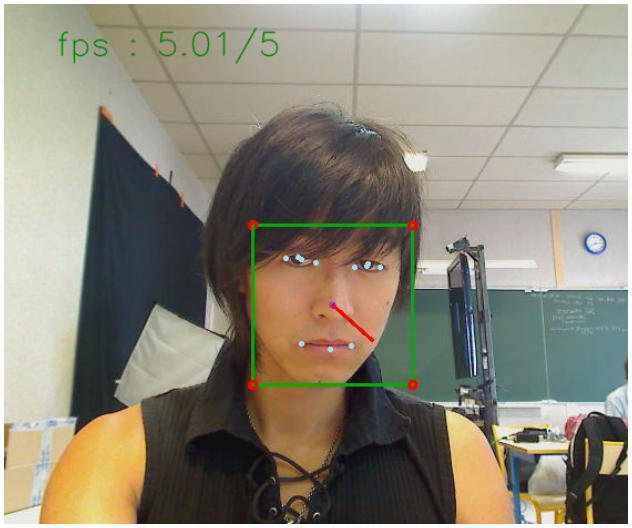


Figure 3. Detection of user's face and estimation of gaze directions with Omron's Okao Vision

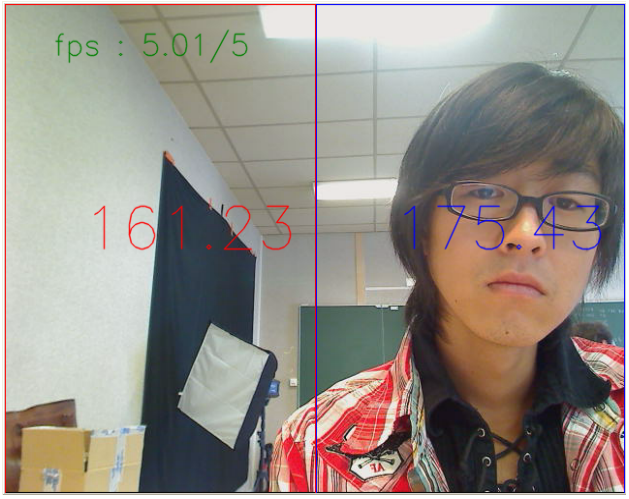


Figure 4. Input image processing with Motion Detector

Okao Vision is a commercial product made by Omron [17]. It is a library that provides accurate face detection and extra functions like face orientation, gaze direction, the positions and openness of eyes and mouth, gender detection, age identification and face identification from one single image. Fig.3 is an example output of a program written with Okao Vision. Requirements and limitations vary with functions. Face detection requires a 10x10 pixel square image of the face and the head rotation must be within 60 degrees in left-right direction (yaw) and 30 degrees in up-down direction (pitch). Face orientation and eyes-mouth openness detection require a 40x40 pixel image, and the rotation of the head must still be within 60°(yaw) and 30°(pitch). Gaze direction requires an 80x80 pixel image and is limited to 30°(yaw) and 15°(pitch). In our test program with a 960x720@15fps web cam, accuracy of face detection is sufficiently high and undoubtedly usable, but most of the other functions are not reliable and can only be treated as an extra bonus. Since Okao Vision does not require stereo cameras, it is an acceptable result. We thus decided to use the face detection, face orientation and eye-mouth openness. Face orientation is also used to approximately determine the users' gaze direction. This is not very accurate but should be sufficient if we only need to distinguish rough directions like the screen or another user.

Motion Detector is a component that can detect a moving object in distinct areas, and is developed by using OpenCV library [20]. The component divides the viewed region into two distinct areas and detects motions inside each area. Fig. 4 shows video image data divided into two areas surrounded with a blue and a red square. Moving objects are recognized when the sum of differences in pixel values between the current and previous frame is above a certain threshold.

To determine the number of users we use both Okao Vision and Motion Detector. Okao Vision has a function to determine the number of user faces, but it cannot detect a user's departure because it fails in face detection when rotation of user head exceeds 60 degrees. Hence, we use Motion Detector to track movements in the area.

The number of users should be equal or less than two and they stand at positions decided beforehand. Therefore we can know the user existence in the area and the number of users by detecting the movement in regionby using Motion Detector. The result of this function is the greater of the values for the number of people provided by Motion Detector and OkaoVision, respectively. This function has been implemented and tested.

Although the topics that the users are talking about serve as a very important cue for knowing whether the users are interested in the system, it is not practical to use speech recognition in our noisy real-world application. Therefore, we hypothesized that users' interest can be approximated as the users' attention toward the system during a relatively long interval of time, such as dozens of seconds. Attentions of the users can be approximated as the patterns at which users are looking at a high ratio of time. The meaningful patterns we consider are either the agent or other users. The attention of the users toward the system or the agent can also be discovered by tracking whether there is a change in the users' activities immediately after a remarkable action done by the agent.

B. Central Part

1) Nonverbal Input Understanding

This component combines input data coming from Motion Detector and Okao's Vision and uses simple heuristic methods to resolve the number of users and level of interest in the system. As Okao's Vision fails in detection when users rotate their head beyond 60 degrees, it is important to save data in short time periods and combine it with information from Motion Detector. For example, two users talk/listen to the agent and the left user turns his head to see who entered the room behind him. In this situation Okao sends value of `UserNumber` variable set to one, as if there is just the right

user in the system. At the same time, Motion Detector detects motions in the left area, and Nonverbal Input Understanding component sets `UserNumber` value to two and sends it to the Decision Making Component.

2) Decision Making Component

At the beginning of the workshop we planned to implement a central part following the Information State Theory [2] [3], which is a general theory of human-agent dialogues. A dialogue based upon this theory is defined as a set of variables (or information) that describe the current state of the dialogue. According to the modeled dialogue, the information state is updated by dialogue moves that compose the agent's and the user's utterances. The agent's dialogue strategies which trigger particular dialogue moves in response to current conditions that are caused by both of the user's and the agent's dialogue moves are defined as plans. The agent then behaves as a regular rational autonomous agent that chooses and executes plans to achieve its goals. This is the general idea of the dialogue theory; dialogue moves, dialogue plans and lexicons are user-defined and require intensive work. The prototype was being implemented based on one of the implementations of information state dialogue move engine [4] to be capable to deal with multi-modal, multi-party conversations, dynamically changing behaviors accompanying the emotion dynamics simulating component, etc.

However, due to lack of time to complete the dialogue manager by the end of the workshop, instead of a fully functional dialogue manager, a power-up version of existing GSML (Generic Embodied Conversational Agent Scenario Language) [5] script interpreter was developed.

Compared to the previous system which merely matches recognized speech inputs and non-verbal inputs with predefined patterns, a variable system is introduced. Following information state theory, interaction between the agent and one of two users is described with a set of variables like a snapshot. For example, `SpeechInput` represents the most recent result from one of the speech recognition components, `Speaker` represents the id of the speech recognition component, `UserNumber` represents the number of users who are standing in the user area, `UserStatus` represents the arability of the users, `UserAttention` represents how much the users are paying attention to the system, `Addressee` specifies who should be the addressee of the agent's next utterance, etc.

The values of these variables are updated with the agent system's internal status and perception events sent from the speech recognition components and non-verbal input interpretation component. How the value of the variables should be updated can also be specified by the script designer in the script as the effects of particular input patterns. `Effect` element is introduced into `Template` element for this purpose. An input event can cause the values of particular variables to be bound to, added with, or subtracted from certain values.

The syntax of the patterns defined in GSML scripts is also extended. `Predicate` element is introduced to represent a test on the values of a variable. The value of the variables can be tested to be equal to, lesser or greater than certain values.

The chatbot-like ECA system is then extended to a more powerful rule-based autonomous system. The agent or the script execution engine updates its internal status variables via the perceptions from the outside world or the users, and picks the first valid template which made all of the conditions (predicates) true to perform. Therefore, the rules like that the tour guide agent should walk to the front to greet when there are users present in the user area, say goodbye to the user and go back to the initial position when the user has left the user area and so on can be specified in the script.

States limit possible patterns that will be used in matching in the current conversation situation and thus isolates the interference from other states which may happen to have the same triggering patterns. Because of the absence of context management mechanism in the agent's behavior control, there is no way to determine whether a user answer is related to the last question asked by the agent. However, for example, when the agent is going to ask a yes/no question like "Do you need a tour guide? ", a transition to a specific state representing the question can isolate the question from other yes/no questions.

`GlobalState` is introduced for error and interruption handling. When a failed or unknown recognition occurs, appropriate response will be searched among the categories defined in the global state. When interruptions from the user like "excuse me" or "pardon" occur, they are also matched with the patterns defined in this state.

Unlike a full dialogue-managing central component, the disadvantage of this approach is: the agent does not conduct a plan that contains multiple steps to achieve a certain goal. The agent's behaviors are driven by the events that occurred in the outside world. The management mechanism for information like grounding or topics is not included in the script execution kernel. These features are still implementable but are left as script programmer's responsibility.

During the period of eNTERFACE'08 workshop, only a few issues specific to multi-party conversation could be addressed. Because it is possible that there can be two users standing in front of the tour guide agent, the gaze direction of the agent becomes essential in three-party dialogue. Because we do not do natural language processing and have no way to understand the meaning of the conversation, our simplified model is: the agent should distribute different ratio of its gaze depending on who the addressee is. It will look at both users at the same frequency or look at the user standing at the left hand side more frequently if he is the addressee. Detailed gaze control is not done by the central component but is by the player. An `Addressee` attribute is introduced in the `Utterance` element to specify the addressee of this utterance is to the left, right or both. Addressee specification is done by the rules in the scenario script by the script programmer; for example, this utterance of the agent should be directed towards the last speaker and so on.

C. *Animation Player*

The component which displays the city of Dubrovnik and tour guide agent Dubravka was developed by using Visage|SDK tool from Visage Technologies company [15]. As this product does not have support for virtual environments, we used a normal image as background. Position of the agent was calculated by using ARToolkit software from eNTERFACE '06 workshop [6]. Size of the agent was adjusted to large in order to efficiently address the user by gazing (Fig. 5).

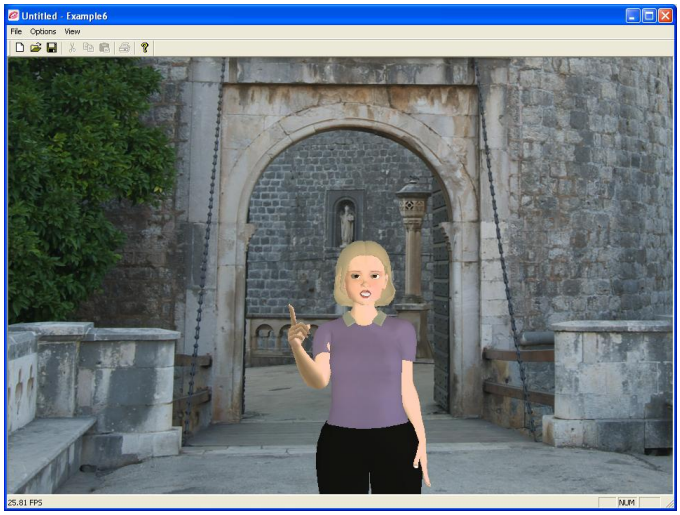


Figure 5. Example of agent's behaviours in Animation player

Agent's behaviors defined in system scenario are produced and tested by using GSML syntax [5] which synchronizes nonverbal behaviors (gestures, locomotion, facial expressions) with speech. Animations for the character were either modeled manually in 3ds Max or procedurally in Visage.

Gazing model runs on the utterance level and is controlled by the central part of the system according to the `Addressee` attribute. E.g. if the addressee is the left user, the agent will gaze first at the left side, then it will glance to the right for a while (if there are two users) and gaze back to the left user again. Since we cannot predict the duration of the utterance, we repeat patterns and stop when the utterance is finished.

Animation Player receives messages from Decision Making component and runs them in an appropriate way. We distinguish between four types of messages:

- **Utterances.** Contain GSML behavior description to be performed by an agent. After they have been received, they are put into queue.
- **Interruptions.** Messages which stop the agent from performing "planned" behaviors. When received, the behavior queue is cleaned.
- **Stop messages.** Force player to shut down and save current content.
- **Reset messages.** Reset the player to the initial state. This message is received when there are no users within the system scope.

It was also planned to add message priority within Utterance category to avoid unnecessary re-planning in case of interruptions like comments. Example is a situation when the agent is speaking and the user says "That's very nice". From the system input it is recognized as interruption, however in system output we treat it in this way: the agent stops with speech, responds to the user ("Yes, it is"), and then continues with the behavior which was running before the comment occurred. A level of "smart scheduling" according to the utterance priority should be achieved for this feature.

IV. Discussion and Conclusion

A. *Summary of software produced.*

A first prototype of ECA tour guide system which can communicate with two users has been developed. The system supports some of the multiparty aspects that were obtained from the literature and planned from the beginning of the workshop. The system can:

1. Recognize presence of people and invite them to the session
2. Resolve the number of users and their level of attention
3. Handle interruptions, such as the users' requests, and respond to these requests by using the agent's gaze direction
4. Recover from failure (e.g. if a user's speech isn't recognized)
5. Recognize users' departure and automatically finish the session.

It was also planned to have a system with five different scenes, but as writing a scenario and manual modeling of the agent's behavior was rather time-consuming, we focused only on first scene. Extension of the system on five scenes doesn't require modification of existing components- it is enough to write a new script and, if necessary, add new animations to the player.

B. *Towards system evaluation*

Since some of the planned multiparty aspects were not yet implemented and tested, up to now no system evaluation has been performed.

After we have completed our work, the first thing to do is to see how our system acts in normal interaction with one and two humans (what lacks at the moment is that system doesn't handle users' mutual conversation). As it cannot "understand" natural language, we are in particular interested in learning what type (and level) of interaction can be achieved with this scenario. Of course, we want to improve it, if necessary and possible. Compared to the previous dyadic tour guide system which was only served for demo purposes, this system is several steps ahead.

On the system-level, evaluation involves three things: evaluation of user-ECA performance, e.g., how fluent and efficient it is, evaluation of the user experience from a

subjective standpoint and evaluation of the effectiveness of the ECA-enabled application in achieving its goals [19]. At this point we are interested in the first and second type of evaluation. Usually, results of experiments are collected by using questionnaires or monitoring users' bio-signals in interaction, such as gazing or speech. Overviews of relevant factors which influence human-character interaction are given in [18], and will serve us as general guidelines to set up an experimental system and choose an appropriate system study.

Compared to dyadic ECA systems our system has one strength from evaluation point of view: it provides a possibility to measure and compare human-human interaction and ECA-human interaction at the same time. Although the full system does not "allow" communication between humans, measurements can be simplified by observing attempts of humans to establish communication and neglect ECA. It would also be interesting to compare evaluation results to the recent work by Rehm [16], who performs evaluation study on an emotionally tense lying game. The agent in this system is even less "smart" than ours (it can understand some digits and "yes/no"-variants as well) and is treated in three different ways by human users: as a normal conversational partner, as an interlocutor and as an artifact. This is, however, expected, but the question is; what are actually variations among these categories and what are their values? What is the level of empathy and effectiveness which an ECA has to have to keep attention, and what is the duration of this attention? How to model an ECA which can be "interesting" for prolonged communication? Our system can serve as a useful test bed for further studies.

C. Conclusion

Implementing multiparty support in ECA systems partly relies on communication theory which has set up several bases (e.g. [9]). However, systems which can establish and maintain multiparty communication with two humans are scarce.

During this workshop we attempted to implement a first prototype of our system which handles this type of communication. Although some of the tasks were not completed, first experiments in interaction with the system are encouraging. For example, the system can detect the number of users, distinguish local roles of users in a conversation and use ECA's gazing direction to specify an addressee.

The weakest point in the system is the conversational model which is rather simple, but overall outcome of this workshop is just a step towards further experiments and studies in the field of multiparty communication in ECA systems.

ACKNOWLEDGEMENTS

This research is partially supported by the Ministry of Education, Science, Sports and Culture of Japan, Grantin-Aid for Scientific Research (S), 19100001, 2007,"Studies on Construction and Utilization of a Common Platform for Embodied Conversational Agent Research" and the Ministry of Science Education and Sports of the Republic of Croatia, grant nr. 036-0362027-2028" Embodied Conversational Agents for Services in Networked and Mobile Environments." We would also like to thank Dr. Lao Shihong from OMRON Corporation Research & Development Headquarters, regarding licensing of OKAO Vision.

REFERENCES

[1] D. Traum and J. Rickel, "Embodied Agents for Multi-party Dialogue in Immersive Virtual Worlds," in AAMAS 2002, vol. 2.

[2] Traum, D., Bos, J., Cooper, R., Larsson, S., Lewin, I., Matheson, C., and Poesio, M. (1999). A model of dialogue moves and information state revision.

[3] Larsson, S., Berman, A., Gronqvist, L., and Kronlid, F. (2002). TRINDIKIT 3.0 Manual. Trindi Deliverable D6.4

[4] The MITRE Corporation (2005). Midiki User's Manual, version 0.1.3 beta edition.

[5] Huang, H., Cerekovic, A., Pandzic, I., Nakano, Y, and Nishida, T.: The Design of a Generic Framework for Integrating ECA Components, Proceedings of 7th International Conference of Autonomous Agents and Multiagent Systems (AAMAS08), Estoril, Portugal, pp128-135, May, 2008.

[6] Huang, H., Cerekovic, A., Tarasenko, K., Levacic, V., Zoric, G., Pandzic, I., Nakano, Y., and Nishida, T.: An Agent Based Multicultural Tour Guide System with Nonverbal User Interface, the International Journal on Multimodal Interfaces, Vol. 1, No. 1, pp. 41-48. 2007.

[7] Vertegaal, R., Slagter, R., van der Veer, G., and Nijholt, A. 2001. Eye gaze patterns in conversations: there is more to conversational agents than meets the eyes. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Seattle, Washington, United States). CHI '01. ACM, New York, NY, 301-308

[8] Rehm, M., André, E., and Wissner, M. 2005. Gamble v2.0: social interactions with multiple users. In *Proceedings of the Fourth international Joint Conference on Autonomous Agents and Multiagent Systems* (The Netherlands, July 25 - 29, 2005). AAMAS '05.

[9] David Traum. 2003. Issues in multi-party dialogues. In *Advances in Agent Communication (F. Dignum, ed.)*. Springer-Verlag LNCS.

[10] Lester, J.C. and Converse, S. A.: The Persona Effect: Affective Impact of Animated Peda-gogical Agents. In S. Pemberton, ed. Human Factors in Computing Systems: CHI+97 Con-ference Proceedings, pp. 359-366. New York: ACM Press (1997)

[11] Matthias Rehm and Elisabeth Andre. Where do they look? Gaze Behaviors of Multiple Users Interacting with an Embodied Conversational Agent. In Proceedings of Intelligent Virtual Agents (IVA), 2005.

[12] David Traum, Stacy Marsella, Jonathan Gratch, Jina Lee, and Arno HartHolt. Multi-party, Multi-issue, Multi-strategy Negotiation for Multi-Modal Virtual Agents. In Proceedings of Intelligent Virtual Agents (IVA), 2008.

[13] Kopp, S. , Krenn, B. , Marsella ,S. , Marshall, A., Pelachaud, C. , Pirker, H., Thórisson, K. and Vilhjálmsson, H.:"Towards a Common Framework for Multimodal Generation: The Behavior Markup Language", in Proceedings of the IVA '06 conference, Marina del Rey, USA, 2006

[14] Vinayagamoorthy, V. and Gillies, M. and Steed, A. and Tanguy, E. and Pan, X. and Loscos, C. and Slater, M. (2006) *Building Expression into Virtual Characters.* In: Eurographics 2006, 04-08 Sept 2006, Vienna, Austria.

[15] http://www.visagetechnologies.com

[16] Rehm, M. 2008. "She is just stupid"-Analyzing user-agent interactions in emotional game situations. Interact. Computing. 20, 3 (May. 2008), 311-325

[17] http://www.omron.com/r_d/coretech/vision/okao.html

[18] G. Ruttkay, Z., Pelachaud, C., 2004. From Brows till Trust: Evaluating Embodied Conversational Agents. Kluwer Academic Publishers, Dordrecht / Boston / London.
[19] Evaluating Embodied Conversational Agents" seminar, 15-19 March, 2004 Organized by Z. Ruttkay, E. André , K. Höök, W. Lewis Johnson, C. Pelachaud (http://wwwhome.cs.utwente.nl/~zsofi//eeca/")
[20] http://sourceforge.net/projects/opencvlibrary/
[21] http://www.loquendo.com/en/

Aleksandra Čereković is a Ph.D. student at the Department of Telecommunications, Faculty of Electrical Engineering and Computing in Zagreb, Croatia. She obtained her diploma degree on audio-visual speech processing (lip-sync) in year 2006 from the same University. Currently she is working on topics of body animation and control for Embodied Conversational Agents (ECAs). Her research interests are: body communication, gesture theory, animation modeling, and computational learning.

Hung-Hsuan Huang graduated from the Computer Science Department of National Chen-Chi University, Taiwan in 1998 and obtained his master degree of computer science and information engineering from National Taiwan University, Taiwan in 2000. After a two-year military service where he was the political warfare director and the co-commander of an army company, he went abroad to Japan. After the learning in a Japanese language school for one year, he entered the Ph.D. course of the Graduate School of Informatics of Kyoto University, Japan in 2003. His research interests include intelligent software agent, information visualization, photo management, gesture interface.

Takuya Furukawa (Japan, October 1984) graduated from the Human and AI Systems Department of Fukui University, Japan in 2007 and entered the master course of the Graduate School of Informatics of Kyoto University, Japan in 2007. His research interests include multimodal conversational interfaces, and human behaviors in communication with embodied conversational agents. He has an experience of fellowship job in Works Applications is Japanese ERP software company. Mr. Furukawa is a member of JSAI (Japanese Society for Artificial Intelligence).

Yuji Yamaoka is a student of a Master course program of Ubiquitous and Universal Information Environment, at the Graduate School of Informatics, Tokyo University of Agriculture and Technology, Japan. His research interests include culture in nonverbal behaviors of ECAs.

Toyoaki Nishida is a professor of Department of Intelligence Science and Technology, Graduate School of Informatics, Kyoto University. He received the Doctor of Engineering degree from Kyoto University in 1984. His research centers on artificial intelligence and human computer interaction. In 2001, he founded a series of international workshops on social intelligence design (see http://www.ii.ist.i.kyoto-u.ac.jp/sid for more details). Then, he broadened the scope of research to include understanding and augmenting conversational communication, and opened up a new field of research called Conversational Informatics. Currently, he leads several projects on social intelligence design and conversational informatics. He is a member of the board of directors of IPS (Information Processing Society) of Japan and JSAI (Japanese Society for Artificial Intelligence). He serves as an editorial board member of several academic journals, including Web Intelligence and Agent Systems, AI & Society, and Journal of JSAI (editor-in-chief).

Associate Prof. **Igor S. Pandžić** received his BSc degree in Electrical Engineering from the University of Zagreb in 1993, and MSc degrees from the Swiss Federal Institute of Technology (EPFL) and the University of Geneva in 1994 and 1995, respectively. He obtained his PhD from MIRALab, University of Geneva, Switzerland in 1998. In the same year he worked as a visiting scientist at AT&T Labs, USA. In 2001-2002 Igor was a visiting scientist in the Image Coding Group at the University of Linköping, Sweden, and in 2005 visiting researcher at Kyoto University, Graduate School of Informatics, Nishida & Sumi Lab.
He is now 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. 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.

Asst. Prof. **Yukiko Nakano** received her bachelor degree in psychology from Tokyo Women's Christian University, Japan in 1988 and one master degree in educational psychology from the University of Tokyo and one in media arts and sciences from Massachusetts Institute of Technology, USA in 1990 and 2002, respectively. She obtained her Ph.D. in information science and technology from the University of Tokyo, Japan in 2005. Yukiko was a researcher of NTT Research Laboratories from 1990 to 2000 and was a sub-leader researcher of Research Institute of Science and Technology for Society from 2002 to 2005. She was working at the Department of Computer, Information and Communication Sciences of Tokyo University of Agriculture and Technology as an associate professor from 2005 to 2008. From summer 2008 she has been working at Dept. of Computer and Information Science Faculty of Science and Technology, Seikei University. With her special interest in Embodied Conversational Agents (ECA), she has been studying human face-to-face communication in psychology and communication science, and creating multimodal conversational interfaces based on a model of human communication behaviors.

Multimodal High Level Data Integration

Olga Vybornova, Hildeberto Mendonça (1), Daniel Neiberg (2), David Antonio Gomez Jauregui (3), Ao Shen (4)
(1) UCl-TELE, Louvain-la-Neuve, Belgium, (2) TMH/CTT, KTH Royal Institute of Technology, Sweden, (3) TELECOM and Management SudParis, France, (4) University of Birmingham, UK

Abstract

A method of multimodal multi-level fusion integrating contextual information obtained from spoken input and visual scene analysis is defined and developed in the context of this research. The resulting experimental application poses a few major challenges: the system has to process unrestricted natural language, free human behaviour, and to manage data from two persons in their home/office environment. The application employs various components integrated into a single system: speech recognition, person identification, syntactic parsing, natural language semantic analysis, video analysis, human behavior analysis , a cognitive architecture that serves as the context-aware controller of all the processes, manages the domain ontology and provides the decision-making mechanism.

***Index Terms*—multimodal data integration, semantic representations, cognitive architecture, spoken input, video input, decision-making.**

I. Presentation of the Project

A. Motivation

ince Bolt's groundbreaking work on speech- and gesture-based interaction [2], multimodal systems and fusion engines have progressively evolved towards more robust semantic interpretation and they all include some fusion mechanism to combine the separate data streams by reducing uncertainty. The interest to jointly process audio and visual information related to human activities, and to extend the technological developments in individual modalities for human-computer interaction has been increasing. To improve naturalness and robustness of user interfaces they should be able to automatically recognize human identity, intentions, speech, actions in computing environments, define the person localization, perform source separation, media synthesis and media content mining. A critical factor that allows to make multimodal processing in speech-based interactions effective is the robust integration or fusion of information from multiple modalities.

Multimodal fusion is a central question to solve and provide effective and advanced human-computer interaction. To understand and formalize the coordination between modalities involved in the same multimodal interface it is required to extract relevant features from signal

representations and thereafter to proceed to high-level (semantic-level) fusion. Signal- and semantic-level processing is deemed tightly interrelated since signals are seen as bearers of the meaning. To provide natural and robust interaction with the user(s) the system must be able to interpret as a whole the various modalities used by a human to interact, i.e., to extract information arriving simultaneously from different sources and to combine them into one or several unified and coherent representations of the user's behavior. A challenge is to fuse these different input modalities effectively to augment their natural strengths and use redundancy to increase robustness. Moreover, it is critical to decide what information should be fused and what should not.

We see multimodality as a set of variables describing states of the world (user's input, an object, an event, behavior, etc.) represented in different media and through different information channels. In this respect the goal of data fusion (merging semantic content from multiple streams) should give an efficient **joint interpretation** of the multimodal behavior of the user(s) – to provide effective and advanced interaction.

We have been working with multimodal fusion considering two main modalities, which are speech recognition and human behavior analysis through image processing tools. We have conducted experiments in a hypothetical context, with two or more people interacting with each other and with objects in the scene. We observed that when intending to perform an action, the user(s) might:

- speak about it describing their actions or intentions (the ideal case, then the linguistic and action recognition data are complementary);
- be doing things, but speaking about something absolutely irrelevant to those actions (in this case the data from the two modalities should be analysed separately, without merging);
- the users' words might contradict the actions performed (then the actions have priority, since "actions speak louder than words");
- be just silent when performing actions (in this case the unimodal operation mode is required and no data fusion is needed).

Everything what is said or done is meaningful only in the particular context. To accomplish the task of semantic fusion we should take into account the information obtained at least in the following three types of context [7]:

- domain context: meaning prior knowledge of the domain, semantic frames with predefined action patterns, user profiles, situation modelling, a priori developed and dynamically updated ontology defining subjects, objects, activities and relations between them for a particular person;

- conversation context: derived from natural language semantic analysis ;

- visual context: capturing the user's gesture/action in the observation scene and allowing eye gaze tracking to enable salience models.

Basically, in our approach the high level fusion of the input streams can be performed in three stages: (i) early fusion that merges information already at the signal or recognition stage to provide reliable reference resolution, (ii) late fusion that integrates all the information at the final stage to give interpretation of the human behavior [26], and (iii) reinforcement fusion that execute one or more cycles of verification, reevaluating all identified meanings to improve the overall integration. We argue that in order to make the multimodal semantic integration more efficient and practically real, special attention should be paid to the stages preceding the final fusion stage.

During the early fusion, some data received from each modality, which meaning is redundant in two or more modalities, should be synchronized in order to strengthen the link between data and optimize the late fusion. An example of early fusion done in our experiment is the person identification from the speaker and from the image features. This singular meaning helps us to complement a recognized intention from the speech with the behaviour of the speaker, predicting what decision is more appropriated for the situation.

When the early fusion finish, the search for semantics continues with the application trying to identify plurality of meanings. This is the input of the late fusion, which deals with higher symbolic elements abstracted by the modality recognizers. These elements are first mapped between the different modalities and then integrated during the final decision stage to resolve uncertainty and produce the user's intention interpretation. One example of plurality of meanings on speech recognition is to detect the human intention linking the person (subject) with a concrete or abstract things (object) through a action or relation (predicate). In the sentence "*I want to call Nick*" there is a clear intention represented by links between "*I*", "*want to call*" and "*Nick*", which are subject, predicate and object respectively. These meanings will be combined with other meanings during the late fusion.

In addition to the early and late fusion, we are contributing with the *reinforcement learning*, which is a cyclical process that uses all found meanings to reinforce or redefine them selves, producing better inputs for the late fusion. The need for reinforcement came from a reflection about how a existent modality can help the analysis of other modalities preserving the modality cohesion, where a modality must not be aware of the presence of other modalities in order to avoid some level of coupling. We could observe this need in our experiment, when we implemented n-best reranking functions for the speech recognition analysis to find the best hypothesis about what was said. These functions need direct access to the list of concepts from the ontology and what was analysed from other modalities, in order to probabilistically rank the best hypothesis possible. The reinforcement fusion is more robust than early fusion, less conclusive than late fusion, and it is not mandatory, but recommended when there are clear possibilities of improvement.

Developing optimal strategies for fusion is still a crucial issue as none of the known approaches has so far emerged as the outstanding solution. Various different formalisms have been proposed and mainly consist of deterministic fusion schemes:

- rule-based implementations: including variants such as unification grammars used in the multidimensional parsing of typed feature structures, finite state transducers supporting tighter coupling of parsing and input recognition, finite state machine for biologically inspired rhythmical synchronization;
- statistical methods: such as integration based on maximum entropy;
- mixed approaches: probabilistic agent-based resolution and integration performed with simple voting mechanism based on confidence values of resolved concepts, probabilistic salience weighting of action-based concepts that are bound to speech concepts according to a predefined set of rules.

II. Application, scenarios and modalities of interest

In order to develop and validate the whole experiment, we defined five different scenarios where two people interact mutually in a room, talking in a natural way and behaving without restrictions. Each scenario tries to explore specific combinations of speech and behavior to increase the robustness of the system. This system is able to track people, analyze their behavior, movements, speech, and takes decisions about how to prompt them necessary information when required or provide any other assistance.

The main challenges that we face in this application are unrestricted human language and free natural behavior within home/office environment. In order to analyze behavior efficiently, the system has to correctly process, interpret and create joint meaning of the data coming from speech analysis and video scene analysis. We consider that human behavior is goal-oriented, so our main aim is to recognize the users' plan, to see what they want to achieve.

The system manages the data streams arriving from two sources – video scene and speech. In particular, we show a technique distinguishing between the data from different modalities that should be fused and the data that should not be fused but analyzed separately.

III. Project architecture and implementation

The application employs various components integrated into a single system: speech recognition, speaker identification, syntactic parsing, natural language semantic analysis, video analysis, human behavior analysis component and a cognitive architecture that serves as the context-aware controller of all the processes, manages the domain ontology and provides the decision-making mechanism.

Figure 1 below describes how all components collaborating mutually when connected to compose the overall solution. The process starts when an audio or a video signal is detected by the first time. There is no restriction about what signal should start first because all modalities can be processed in parallel and independently. When an audio stream is received, the speech recognition component processes it, generating a string of what was said. The same signal is sent to the speaker identification component, which will associate what was said with who said that. The string is sent to the syntactic parsing component to identify the syntax of each word, which is important for the natural language semantic analysis component, responsible for the identification of the subject, the agent, the predicate, the object of interest and other elements. From the semantic analysis, it is possible to extract semantic structures very similar to the structure of the knowledge base, represented by ontology. If we find the identified semantic in the ontology then it means that sentence is valid inside of the context and can be useful to fuse with other meanings coming from other modalities.

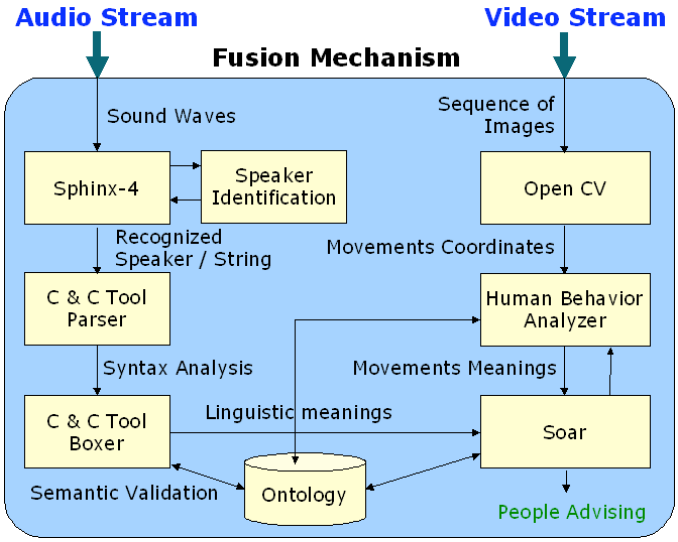


Figure 1. Project architecture

On the other side, when a video stream is detected, the image is processed by image processing component called Open CV. This component analyses some image features to calculate the position of each person on the horizontal plan of the scene, their movements direction and identifies who is each person according to a predefined profile. It is also important to identify people through this modality because we have to know the position of who is talking in order to associate the user intention with his actions. The synchronization of detected users in each modality is done during the early fusion.

The next step is to analyse the human behaviour, comparing the movements of the user with a set of rules. The behaviour is relative to fixed objects in the scene, which are defined in the context domain and are directly associated with the aid to be given by the system. The rules define the boundaries of what is near of far from a certain object. Then the result of the rule processing at this stage is: [person] is near to [telephone], [person] is far from [computer] or [person] is moving to [library]. This result is produced for each person in each frame of the video. Individually, these results are not significant enough for fusion. We have to analysis the movements in many frames in order to have final conclusions. For instance: if in the last 80 frames the rule engine produced "[person] is moving to [library]" then we can conclude that there is a real intention to achieve the library, considering some variables of the environment, such as area of the room. The late fusion occurs when we identify a person moving to the library and we also detect the intention to find a book in a sentence like this: "I can find a book about it in the library.". Therefore, we can conclude the person is moving to the library because he wants to find a certain book there. Previous sentences, already analysed, indicates that the article "it" in the current sentence actually means "French wines", an object identified through the semantic analysis.

Many existent components were just reused, such as Sphinx, C & C Parser, C & C Boxer, Open CV, Protegé and Soar. Other components were developed from the beginning, which was the case of the Human Behavior Analyzer, the Fusion Mechanism and the interoperability between all components, which is implemented using sockets, a simple network communication strategy.

Speech Recognition

1) Speech Data

The speech was recorded by two non-native subjects, one 23 year Chinese male and one 32 year Swedish male. The data was recorded using 16kHz, 16 bit audio. The 5 scenarios consisted of a total of 72 sentences and 148 seconds. For development of speaker identification, we used ten phonetically rich sentences for training, and for parameter tuning we used another ten phonetically rich sentences and ten words of different length.

2) Speech Recognition

For speech recognition we used Sphinx 4, which is an open source, Java-based speech recognizer [29]. For acoustic modeling, we used the 8 Gaussian triphone models, trained on the Wall Street Journal Corpus, which are supplied along with Sphinx. Since we wanted to allow the system to monitor a discussion between two or more people, we want to have a large vocabulary language model. For this purpose, 3-grams with a maximum of ~5000 words were trained using the orthographic transcriptions from the Wall Street Journal Corpus. The 5000 words were selected as the most common ones plus the ones that are present in the scenarios.

In order to enhance understanding of concepts, we wanted to evaluate different methodologies to pick hypotheses from the N-best list which Sphinx can produce. For this purpose, Sphinx was configured to produce as long N-best lists as possible given a memory constraint of 1 Gb. This resulted in list lengths of ~100 hypotheses for short utterances with good accuracy and ~10000 list lengths for long utterances with poor accuracy. Two different approaches may be used. The first is to estimate a re-scoring function on data, the second is to apply re-ranking heuristics. The first approach is likely to be most efficient, but it requires enough data to be done

optimally. The second approach is easier to apply and is likely be more robust to less knowledge about the users may say.

By making an assumption about the domain, a word list consisting of all nameable things in the ontology was queried. Three kinds of re-ranking heuristics were tried out. In each of these, a score were assigned to each hypothesis and then N-best list was re-sorted using this score in such way that the initial ranking was preserved if two or more hypotheses had equal score.

Method A: The score is set to one if one or more words from the ontology are found.

Method B: One scoring point is added for each matching word in the ontology word list.

Method C: A score is added, which is equal to the word length in characters, for each matching word in the ontology word list.

In Tables 1 and 2, the performance gain by this approach is compared to selecting the 1sf top hypothesis and by letting an oracle select the hypothesis which maximizes Accuracy. These results where obtained using a preliminary version of the ontology. While Accuracy takes substitutions, insertions and deletions into account, Correctness does only take substitutions and deletions into account. In table Y, Accuracy and Correctness is computed by only using the words from the ontology as a reference, forming Concept Accuracy and Concept Correctness. The idea is to examine to what extent information, which the cognitive modules can process, is present.

Table 1: Word accuracy and correctness for different hypothesis selection methods

Hypothesis selection method	**Accuracy**	**Correctness**
1st	56,30	64,52
Method A	50,13	62,98
Method B	46,53	63,50
Method C	47,81	64,01
Oracle	74,81	78,15

Table 2. Concept accuracy and correctness for different hypothesis selection methods

Hypothesis selection method	**Accuracy**	**Correctness**
1st	54,12	67,06
Method A	27,06	68,24
Method B	10,59	78,82
Method C	12,94	81,18
Oracle	78,82	74,12

From the results presented in Table X, it is clear that the methods don't improve over the baseline (1st hypothesis). It should also be noted, that these results are obtained using a preliminary version of the ontology. From the results in Table Y, it should be noted that while there is a large drop in Accuracy, there is a clear increase in Correctness, which means that insertions are present. This means that the information content in terms of words known to the ontology is increased on the expense of word insertions.

3) Speaker Identification

Speaker Identification is the task to determine who is speaking. For the application is described in this report, a standard speaker identification system was considered. It is based on Mel-Frequency Cepstral Coefficients (MFCCs) and Gaussian Mixture Models (GMMs), as in Renynold et. al. We used 28 log-Mel filters between 300 and 8000 Hz, cosine projected to 24 dimensions, where the first 12 with their delta where used. A simple data driven procedure for speech detection was tried: The MFCCs where clustered using the k-means algorithm with euclidean distance for two clusters. Then the cluster which had the highest energy was marked as "speech" and the other one as "non-speech". Then these two clusters where used for frame based segmentation. Visual inspection showed that the approach seemed reasonable. No channel compensation was used. This system was implemented in Matlab.

The two speakers in the scenarios where enrolled, using ten phonetically rich sentences. For parameter tuning, another ten phonetically rich sentences and ten words were used. The performance measured in classification accuracy for various numbers of Gaussians are shown in Table 3. Further analysis showed that the errors occurred for test utterances containing very short words, such as "no" and "hi". Using speech detection didn't improve accuracy. Based on these results, 16 Gaussians where chosen for evaluation using the recorded data for the scenarios. The accuracy for the evaluation data was 94%.

Table 3. Speaker identification accuracy, with and without speech detection (SD)

Gaussians	No SD	SD
2	85,00%	92,50%
4	90,00%	92,50%
8	97,50%	92,50%
12	97,50%	92,50%
16	97,50%	90,00%
20	97,50%	95,00%
24	92,50%	95,00%
32	97,50%	92,50%

4) System integration of the speech components

A speech module, consisting of Sphinx 4 and the Speaker Identification system described above was fully implemented

in a TPC/IP client. Also on the server side, the re-ranking method A was implemented, in way which makes it easy to extend to Method B or C.

B. Syntactic and Semantic analysis of spoken input

The next stage after speech recognition is syntactic and semantic analysis of the discourse. For our purposes we use the CCG (Combinatory Categorial Grammar) parser, Release 0.96, developed by S. Clark and J. Curran [9]. The grammar used by the parser is taken from CCGbank developed by J. Hockenmaier and M. Steedman [3].

CCGbank is a treebank containing phrase-structure trees in the Penn Treebank (WSJ texts) converted into CCG derivations. It allows easy recovery of long-range dependencies, provides a transparent interface between surface syntax and underlying semantic representation, including predicate-argument structure. The grammar is based on 'real' texts, and that is why it has wide-coverage, thus making parsing efficient and robust.

The CCG parser has Boxer [8], as add-on to generate semantic representations - Discourse Representation Structures (DRSs), the box representations of Discourse Representation Theory (DRT) [13]. DRSs consist of a set of discourse referents (representatives of objects introduced in the discourse) and a set of conditions for these referents (properties of the objects). Our initial experiments with the CCG parser together with Boxer showed that it suits well for our application in parsing speech of elderly people which is on one hand somewhat restricted to their environment, background and social relationships (the domain model and user profile help us cope with the challenge), but on the other hand is naturally broad enough to need wide-coverage tools for processing. DRSs can be generated in different output formats in Prolog or XML. To link the DRSs output with the ontology we use the XML format.

C. Visual scene analysis

1) Introduction

The video information provides a description of what happens in the environment at a given time. In this project a video sequence is made for each possible scenario. The video sequences were recorded using a distributed 8-camera voxelised visual hull [22]. The description of the environment is obtained by processing each image of the videos using computer vision algorithms.

2) Problem

In these video sequences, there are 3 types of fixed objects (a telephone, some books and a computer) located in different positions inside the scene, there are also two persons that are moving and interacting with these fixed objects. In order to have a good description of the environment for each one of these scenarios is necessary to extract the information of the position of each fixed object, the position of each person at any moment and also the motion direction of each person. Using this information it's possible to know if a person is near or far from each object or if a person is moving to an object. By this way, the system will have good information about the human behavior in order to make better decisions.

3) Procedure

Extracting all this information of the video sequences are common problems of computer vision field. These problems are mainly:

- Object detection (to find the position of each fixed object)
- People detection and tracking (to find the position of each person)
- Motion analysis (to find the motion direction of each person)

The computer vision system to extract this information was implemented in C++ using OpenCV library [12] developed by Intel. In the next sub-sections, will be described the procedure and algorithms to resolve the problems already mentioned.

a) Finding the position of each object

The objects (the telephone, the books and the computer) in these video sequences are always fixed (in the same position), also they don't change size or rotate because the camera's viewpoint is always the same. Hence, the detection of these objects can be easily done by using a template matching algorithm [7]. These algorithms compare a template with a region of an image in order to determinate a similarity measure, wherein the similarity measure is determined using a statistical measure.

To obtain these templates, a sample picture of each object is captured from any image of the video sequence. The OpenCV operator *cvMatchTemplate* was used, this operator returns the probable positions in the image where the template can be located. By this way, the most probable position corresponds to the area where the object is detected. The similarity measure that provides better results for this problem was the correlation coefficient normalized. Because each object never changes position, this template matching step is only done once in the first captured image of the video sequence. Then, these positions are used for all the images in the video sequence.

b) Finding the position of each person

There has been much work in the area of people detection and tracking using computer vision techniques. This problem has been a big challenge since many real-time vision systems require robustness in different environments. In the video sequence of this project, two persons are talking with each other and also moving randomly inside the scenario in order to interact with these objects. The problems here are mainly:

1. The shape and size of each person can change over time because the persons can move far or close to the viewpoint in different parts of the scenario.

2. The persons can approach too much from each other in the video, making difficult the identification of each one.
3. One person can be partially occluded by the other person.
4. Some body parts of each person can be outside the scenario because of the viewpoint of the camera.
5. Each person moves in a random way.

In order to resolve this problem, a color-based tracking was used. Principally the color of the clothes of each person was used, assuming that both people in the video will have different color clothes. However, in order to make this people detection robust to the cluttered environment, a background subtraction technique and also blob detection was implemented in order to discriminate noise. The procedure can be described in this way:

1. Firstly, the first frame in the image was used in order to estimate the background image. This, of course, assumes that nobody is going to be present in the first frames of the video sequence. In order to detect the foreground, it was calculated the absolute difference of the pixel values between and image with the people present and the background image (using operator *cvAbsDiff* from OpenCV) [9]. Then, a threshold in the absolute difference image was applied. By this way, the foreground detected will be the silhouettes of each person. This technique, although is easy to implement and it has a very fast processing time, have the disadvantage of not discriminating the shadows of people.
2. Secondly, a Single-Gaussian Model [5] of each color was applied to the silhouettes in order to obtain the probability that the pixel values corresponds to a color. This Gaussian Model (mean and covariance matrix) was learned in an initialization step using samples manually taken of the clothes color of each person. The result of this process is a binary image where the pixels that are above a threshold are marked as belonging to each color.
3. After, a morphological opening operation was applied to the result image of the previous step. By applying this, we can obtain a more regular shape in the binary image. This was implemented using OpenCV operator *cvMorphologyEx* with a 3 x 3 square structuring element.
4. In this step, a blob detection and analysis was implemented in order to find the region of interest (ROI) that corresponds to each person. The analysis is made by using the size information of each blob in order to discriminate noise or false detections. After this analysis is done, we can estimate the bounding box of each person using the body human proportions.
5. In order to follow each person, the size of the bounding box of each person is increased for the next frame and the person is searched inside this area. This is a very simple tracking, however, it gives good results since the persons are not far from the camera and also they don't change position too fast.

This procedure, although, it works very well for all the video sequences of this project maybe it can not give the same result for other types of video sequence. Consequently, some experiments were done using features tracking from each person, by this way, the idea is to identify and track using optical flow (pyramid-based KLT feature tracking) [15] a group of features conserving the distance properties between each feature of each group [14]. In these experiments some features were lost because of the occlusions, making it difficult to identify each person. However, this can be a very good approach for future research in order to improve the robustness of this system.

c) Finding the motion direction of each person

In order to find the motion direction of each person, Motion Templates algorithms were used based on papers by Davis and Bobick [10] and Bradski and Davis [4]. These algorithms are very fast and robust. The implementation was done using OpenCV Motion Template functions. This functions can determine where a motion occurred, how it occurred, and in which direction it occurred. To calculate the motion direction of each person, the silhouettes (obtained by background subtraction as described in section 3.2) are updated in time using *cvUpdateMotionHistory* operator, after the motion gradient is calculated using the information of the temporal silhouettes (applying *cvCalcMotionGradient*), finally connected regions of Motion History pixels are found using OpenCv operator *cvSegmentMotion*.

With this result, we have region of motions with their gradient directions in the foreground image. In order to find what region of motion corresponds to each person, it was used the result of the procedure described in section 3.2. By this way, the gradient direction of the biggest region of motion inside the bounding box of each person corresponds to the motion direction of the person.

4) Human behavior analysis

Using the information described in the section 3, it's easy to know the human behavior at any time in the video sequence. In this step, we need to know principally:

1. If every object is near or far from each person.
2. The object toward which each person moves.

In order to find this information we only apply some rules and define thresholds using the result of the computer vision system (position of each fixed object and each person and motion direction of the persons). This human behavior information is sent by a socket TCP/IP connection to the multimodal high level data integration system in order to be fused with the speech recognition information (speech modality).

5) Results and conclusions

The computer vision system was executed in a laptop with an AMD Turion 64 2000 MHz processor, 1024 MB of RAM and a graphic card NVIDIA GForce Go 6150 (256MB). The computer vision system is capable of providing the required

information of all the 5 scenarios in real-time processing (25 frames by second). The system is robust to partial occlusions and it can detect and track both persons in any part of the scenario providing their position in all frames. In Figure 2, we can see the result of the computer vision system. The red circles with the red line inside is the motion direction (0 – 360 degrees) of each person, the blue pixels represent the Motion History of the human silhouettes, the white blob is the estimated area where each person is identified inside the scenario and the red small squares is the position of each fixed object (the computer, the books and the telephone).

In order to improve the robustness of the system in different illumination conditions, an approach of features tracking with optical flow can be used, by this way, the identification of each person is not going to rely only on the color and the size, also we're assuming that each person has different color clothes, this is not always true in a real scenario. Maybe for future research it can be used also particles filter approach in order to estimate the position of each person in the next frame improving the robustness of the tracking.

Figure 2: Video sequence of the first scenario.

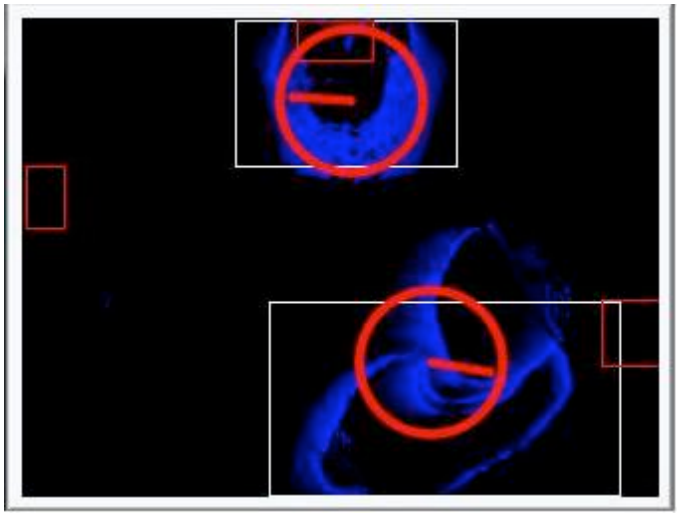


Figure 3: Results of the computer vision system

D. Ontology design and reasoning

In order to have a database for semantic reasoning, an ontology with comprehensive modeling pattern of multimodal actions and prior knowledge about the objects and user is necessary. We used open source tool Protégé [20] to create the ontology in classes, properties and individuals. Figure 4 is the looking of a developed ontology in Protégé 4.0 alpha environment.

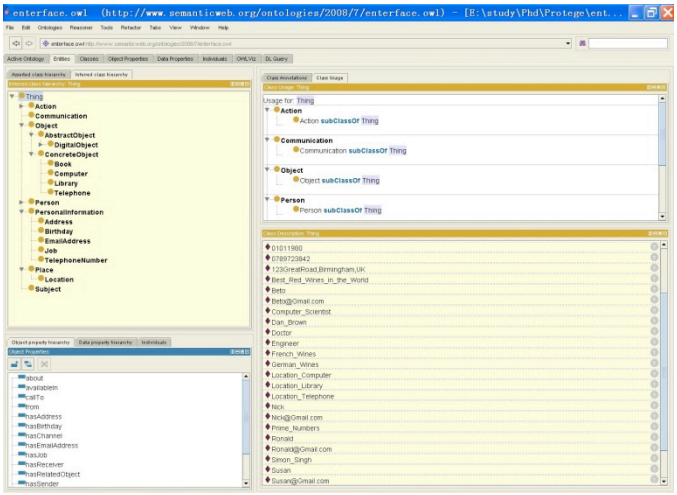


Figure 4. Ontology in Protégé 4.0 alpha environment

We use Soar [24] cognitive architecture in order to create and apply rules to query semantically the modality triples in the ontology. The ontology is re-implemented for adapting the specific working memory structure in Soar environment. To a certain range, class, property and individual in Protégé 4.0 alpha correspond to identifier, attribute and constant. Figure 5 shows an example of graph structure of the knowledge base. In order to have this knowledge base for reasoning, the ontology is created as states when the agent is initialized. After receiving the input from external application, an operator will be proposed and applied to map the input triple to the knowledge base then give out the result to the output link.

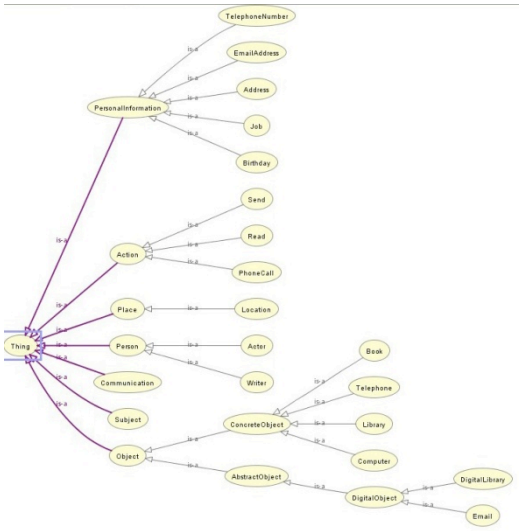


Figure 5. An example of graph structure

For the interacting between Soar and external world, we use SML (Soar Markup Language) to send and receive commands packages as XML packets. The agent created by soar is loaded by the SML codes and run till there is any outcome or run for a certain number of circles which is set by codes too.

The outcomes are checked by SML code as well and kept for further actions. See Figure 6 for the execution cycle.

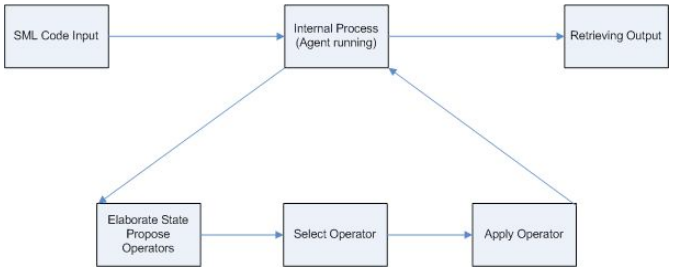


Figure 6. Execution cycle

In our project, when we detect there is an input triple in the input-link, we will try to map it to ontology. Different operators will be proposed corresponding to the conditions. Then in the application stage of the operator, output will be generated and retrieved by the SML code. Further actions are taken according to the outputs.

IV. Experiments and Results

Let us consider one of the annotated example scenarios:
(1) [Beto] Hi Ronald! How is the life going?
(2) [Ronald] I am fine.
(3) [Ronald] I want to call Nick.
(4) [Beto] What for?
(5) [Ronald] He mentioned that he attended a wine tasting course.
(6) [Beto] It sounds interesting, I like wine.
(7) [Ronald] Actually I plan to join the next class. He also mentioned a book about French wines, but I cannot recall the name of the author.
(8) [Beto] Why don't you send a mail to Nick?
(9) [Ronald] Maybe I can find a book about it in the library.
(10) [Beto] Yes, you are right.
(11) [Beto] Did you find it?
(12) [Ronald] Yes, I did.

In our experimental work we used scenarios only with natural human language – we did not work with isolated words, commands, restricted language or something like that. I goal was to experiment with normal complete utterances expressed in a natural way. It should be noted here that we did not do spontaneous language like with pauses, interjections, hesitations, overlaps in speech, etc., because the wide-coverage CCG parser [8] that we used, is trained on newspaper texts, cannot analyze fragments of sentences, it accepts only well-formed complete utterances.
For us, to make multimodal data fusion means to interpret human behavior, to identify the users' plan, infer their intentions. Having understood what the people in the scene want to do, the system takes the decision about how to assist them in the given case. Looking at our challenging example scenario, we can see that there are 4 points in this dialog where the speakers express their plan to do something. In (3) Beto wants to call Nick, in (7) Ronald plans to attend the next class, in (8) there is a possible plan to send a mail and then this possible path of decision is changed for another route - in (9) there is an intention to find a book in the library. How shall the system realize what plans to take into account and what not? When to react and when not? This is why we employ multimodal information – when the plan is identified from the user's words, we look at the other modality data to see, if the person is going to "confirm" his words with the corresponding actions or not. In (3), (7) and (8) the persons expressing the said intentions were standing still, just continuing the talk. And only in the (9) utterance Ronald moved to the book-shelves. That is why only in this last case of plan expression the system reacted and prompted where to find the desired book. By the way, in the phrase "Maybe I can find a book about it in the library" we have to resolve ambiguity between the library in the room and a library on the web. We again do that using information from the other modality – we look if the person is moving to the books in the room or if he is moving to the computer.
Schematic examples of decision making process:

```
(3) If (Ronald) [wants to send] {email to Nick} &
        (Ronald [is moving to] {the computer} | He
[is close to] {the computer}) then
  >> open the mail client with the "to" field filled
with nick@uclouvain.be
(9) If (Ronald) [can] find {book} [about] {it} [in]
{the library} &
       (Ronald [is moving to] {the library} then
   >> There is a book about French wines on the
first shelf.
(9) If (Ronald) [can] find {book} [about] {it} [in]
{the library} &
         (Ronald [is moving to] {the computer})
then
     >> Open a web search website and put the
keyword in the search field
```

When identifying the person's plan from speech, we basically rely on the linguistic semantic analysis as described in Section III B, but we certainly take advantage of the obvious lexical

signs of plan and intentions expression. For example, such verbs and phrases like "want", "wish", "plan", "going to", etc. (we defined 19 expressions like this in total) in the certain syntactic context and in the present or future tense clearly point at the person's intention to do something. And vice versa, negative forms of verbs like "I don't want", "I have no wish to…", "You don't want to…" as well as verbs in the past tense serve as stop-words, and signal that this plan should be discarded and not taken into account, because no system response is needed.

V. Conclusions and Future Work

During this challenging project we have posed and solved several difficult problems. We have:

- managed to deal with spatial relationships (based on the fixed "anchor" objects in the room)
- made semantic fusion of events not coinciding in time
- achieved good results in speaker identification - synchronisation between image and speech identification
- created an open framework to manage fusion between two (in our case) or more modalities (in enhanced future work)
- designed the system so that each component can run in a separated machine thanks to the distribution mechanism interchanging data through a TCP/IP network.

However, we are not going to stop at this point, but we have even more problems to solve in our future work. To name just a few, we should:

- implement effective learning mechanism
- perform efficient decision making even from information fragments
- handle spatial relationships relatively to moving people
- perform 3D video analysis
- identify detection of orientation of the people in the scene
- add at least one more modality - eye gaze tracking
- recognize various types of gestures
- learn to deal with natural language redundancy (repeating the same idea in different words).

VI. Acknowledgement

The research described herein was started under the European FP6 SIMILAR Network of Excellence (www.similar.cc) and continues by the European FP6-35182 OpenInterface project (www.oi-project.org).
We thank Diego Ruiz and Ronald Moncarey (UCL-TELE, Belgium) for precious help in the project preparation.

VII. References

[1] A Gentle Introduction to Soar: 2006 update, http://ai.eecs.umich.edu/soar/sitemaker/docs/misc/GentleIntroduction-2006.pdf

[2] Bolt R. A., "Put-that-there: Voice and gesture at the graphics interface," in *International Conference on Computer Graphics and Interactive Techniques*, July 1980, pp. 262–270

[3] Bos J. et al., "Wide-coverage semantic representations from a CCG-parser" Proc. of Int. Conf. COLING, 2004

[4] Bradski G., Davis J. *"Motion Segmentation and Pose Recognition with Motion History Gradients"* IEEE WACV'00, 2000

[5] Caetano T.S., Olabarriaga S.D., Barone B.A.C., *"Performance evaluation of single and multiple-Gaussian models for skin color modeling"*, in: Proc. XV Brazilian Symposium on Computer Graphics and Image Processing, 2002, pp. 275-282

[6] Chai J., Pan S. and Zhou M., MIND: A Context-based Multimodal Interpretation Framework, Kluwer Academic Publishers, 2005

[7] Cole L., Austin D., Cole L., "*Visual Object Recognition using Template Matching"*, Proceedings of Australasian Conference on Robotics and Automation, 2004

[8] Curran J., Clark S. and Bos J. (2007): *Linguistically Motivated Large-Scale NLP with C&C and Boxer*. Proceedings of the ACL 2007 Demonstrations Session (ACL-07 demo), pp.29-32.

[9] Davis J., Bobick A., *"A Robust Human-Silhouette Extraction Technique for Interactive Virtual Environments"*, IFIP Workshop on Modeling and Motion Capture Techniques for Virtual Environments (CAPTECH98), November 1998

[10] Davis J., Bobick A., "The Representation and Recognition of Action *Using Temporal Templates*" MIT Media Lab Technical Report 402, 1997

[11] Heijmans H.J.A.M., "*Morphological image operator*s" Advances in Electronics and Electron Physics, P. Hawkes (Ed.), Ac. Press: Boston, suppl. 24, Vol. 50, 1994

[12] Intel Corporation, Open Source Computer Vision Library – OpenCV, http://www.intel.com/technology/computing/opencv/index.htm, 2008

[13] Kamp H. and Reyle U., "From Discourse to Logic. Introduction to Model-theoretic Semantics of Natural Language, Formal Logic and Discourse Representation Theory", Kluwer, Dordrecht, 1993.

[14] Kolsch M., Turk M. "*Fast 2D Hand Tracking with Flocks of Features and Multi-Cue Integration*" Proceeding of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshop (CVPR'04)

[15] Lucas B.D., Kanade T., *"An Iterative Image RegistrationTechnique with an Application to Stereo Vision"* In Proc. Imaging Understanding Workshop, pages 121–130, 1981

[16] Oviatt S. et al., "Toward a theory of organized multimodal integration patterns during human-computer interaction", Proc. of the Int. Conf. ICMI, 2003

[17] Oviatt S. and Cohen P., "Multimodal interface research: A science without borders," *Proc. of 6th Int. Conference on Spoken Language Processing*, vol. 43, no. 3, pp. 45–53, 2000

[18] Pfleger N., "Context based multimodal fusion", Proc. of the Int. Conf. ICMI, 2004

[19] Pfleger N. and Alexandersson J. "Towards Resolving Referring Expressions by Implicitly Activated Referents in Practical Dialogue Systems" In: Proc. of the Workshop on the Semantics and Pragmatics of Dialogue, 2006.

[20] Protégé http://protege.stanford.edu/

[21] Reynolds D., Rose R. Robust text-independent speaker identification using Gaussian mixture speaker models // Speech and Audio Processing, IEEE Transactions on, Vol. 3, No. 1. (1995), pp. 72-83

[22] Ruiz D. and Macq B., A master-slaves volumetric framework for 3D reconstruction from images, // Proc. SPIE, Volume 6491. Videometrics IX, J.-Angelo Beraldin, Fabio Remondino, Mark R. Shortis, Editors, 64910G (San Jose, CA, USA, 2007)

[23] Soar: A Functional Approach to General Intelligence, http://ai.eecs.umich.edu/soar/sitemaker/docs/misc/SoarFunctionalOverview.pdf
[24] Soar http://sitemaker.umich.edu/soar/home
[25] Vybornova O., Gaitanis K. and Macq B., "Plan recognition using multimodal integration," 2006, Proc. of Int. Conf. CogSci, 2006.
[26] Vybornova O., Gemo M., Macq B.: Multimodal Multi-Level Fusion Using Contextual Information // ERCIM News, No. 70, July 2007.
[27] Vybornova O., Gemo M., Moncarey R., Macq B.: Ontology-Based Multimodal High Level Fusion Involving Natural Language Analysis for Aged People Home Care Application // In: Proceedings of INTERSPEECH 2007, August 27-31, Antwerpen, Belgium.
[28] Vybornova O., Gemo M., Moncarey R., Macq B.: Contextual Information Sharing in Natural Language and Gesture Crossmodal Integration for Aged People Assistive Home Care Application // In: Proceedings of the 29th Annual Conference of the Cognitive Science Society (CogSci 2007), August 1-4, Nashville, USA.
[29] W. Walker, et. al. "Sphinx-4: A Flexible Open Source Framework for Speech Recognition" Technical Report, SMLI TR2004-0811, 2004 SUN MICROSYSTEMS INC.

Olga Vybornova received a PhD degree in applied and mathematical linguistics from Moscow State Linguistic University, Russia, in October, 2002. From May 2005 she was a post-doctoral researcher and from January 2008 to the present time a research assistant in Université Catholique de Louvain at the Communications and Remote Sensing Lab (UCL-TELE) in Louvain-la-Neuve, Belgium.
Her main research interests are in the fields of multimodal data integration, assistive technology supporting active ageing and social cohesion, semantics-driven multimedia presentation generation, context-based inference, meaning representation, goals, intentions and commitments in communication, dialog modeling; cognitive linguistics; memory and attention, computational semantics.

Hildeberto Mendonça has a bachelor's degree in Computer Science at University of Fortaleza, Brazil. During his undergraduate studies he worked in the Laboratory of Human Computer Interaction with distance learning and digital TV projects. Afterwards, he worked in the Laboratory of Knowledge Management with public security and knowledge bases projects. Currently, he has finished his Diploma of Extended Studies in Human Computer Interaction Meta Representation as a requirement for the PhD in Applied Science in the field of Multimodal User Interaction at Catholic University of Louvain, Belgium. He has a rich professional experience in web application development, collaborative systems, and software engineering and software architecture projects. His technical skills are in the Java platform and its main extensions, C++, XML, relational databases, distributed systems, web services, web and desktop user interfaces, and creation and maintenance of ontology.

Daniel Neiberg received an Master of Science in field of electrical engineering, KTH (Royal Institute of Technology), Stockholm, Sweden, in 2003. From 2002 to 2006 he worked as a Research engineer at Centre of Speech Technology (CTT) at KTH, were he was doing work in Speaker Verification and Emotion Recognition. From 2006 - 2007 he attended a Trainee Internship at Asahi Kasei, where the working tasks included development and evaluation of Automatic Speech Recognition of Swedish. In 2007, he enrolled as PhD student at Centre of Speech Technology (CTT) at KTH, where the main field of research is methods for automatic speech recognition and speech production.
David Antonio Gómez Jáuregui was born in Mexico City on June 13, 1980. He received a BSc in Computer Systems Engineering in 2002 and a Msc. in Computer Science in 2004 with a specialization in Intelligent Systems and Computer Vision. He obtained these grades from Tec de Monterrey Campus Cuernavaca, Mexico. In 2003 he participated in the Robotic Soccer tournament (FIRA 2003) celebrated in Vienna, Austria in representation of Tec de Monterrey. During 2004 - 2007 he worked as a software developer in several companies (Real Time Services, Infomedia, Banco Azteca, Cognosite) learning and using different software programming technologies.
In 2007, he obtained a scholarship from the Mexican Government to study a PhD in Computer Vision at Télécom SudParis (Evry, France). He is currently studying his second year of his PhD with the thesis subject ?3D Human Gesture Acquisition by Computer Vision and Virtual Rendering?. He has big research interest in Computer Vision, Artificial Intelligence and Computer Graphics. Other interests include videogames, learning new languages, eating pizza, listening music and meeting people.

Ao Shen has a bachelor's degree in Communication Science & Engineering at Fudan University, Shanghai, China and a bachelor's degree in Electronic Communication Engineering at Birmingham University, UK. He is experienced in coding mobile application and won the second prize in Vodafone Betavine Student Competition. Now he has finished his first year of PhD study in Electronic Electrical & Computer Engineering Department, University of Birmingham.
His main interests multimodal systems and semantic level applications, knowledge base and ontology creation.

.

Sign-language-enabled information kiosk

Pavel Campr, Marek Hrúz, Alexey Karpov, Pınar Santemiz, Miloš Železný, and Oya Aran

***Abstract*—This paper presents design and creation of multimodal sign-language-enabled information kiosk which was developed during eNTERFACE'08 workshop. The kiosk uses automatic computer-vision-based sign language (SL) recognition, automatic speech recognition (ASR) and touchscreen as input modalities. The outputs are presented on a screen displaying 3D signing avatar and on a touchscreen showing special graphical user interface designed for the deaf users. The kiosk was tested on a dialogue providing information about train connections, but the scenario can be easily changed. The kiosk can be used both by hearing (ASR can be used) and deaf users (even who cannot read), in several languages. The human-computer interaction is controlled by a computer-driven dialogue system. A prototype of the kiosk was build and the work experience raised some new usability questions.**

***Index Terms*—multimodal information kiosk, sign language recognition, sign language synthesis, dialogue system, automatic speech recognition**

I. Introduction

THE aim of this eNTERFACE 2008 project is to design an information kiosk for deaf people that will use sign language (SL) as a main communication mean. Background of the idea is based on current research on sign language processing carried out at UWB (University of West Bohemia) and BU (Boğaziçi University). Methods for synthesis (visual animation) and recognition of sign language are under research. The project also follows the work done during previous eNTERFACE workshops (project No. 3 *Sign Language Tutoring Tool* at eNTERFACE 2006 in Dubrovnik [1] and project No. 3 *A Multimodal Framework for the Communication of Disabled* at eNTERFACE 2007 in Istanbul).

Deaf or hearing-impaired users have limited possibility of communication with hearing people, which can be a problem especially in the case of communication with authorities or information providers (train connection, sales, etc.) These people also cannot use speech-based automatic information services. In these cases dialogue systems shoud be designed to be accessible by deaf users to solve this problem. This project aims to design a dialoge system (information kiosk) in this way.

The idea is to combine sign language recognition and synthesis tasks to develop a simple information kiosk [2] for providing information such as train connections for deaf people who use sign language. The kiosk will be a standalone machine equipped with a standard PC, a touch screen display, a big screen for an avatar a microphone, and one or several cameras. The cameras will capture body/facial gestures of a signing person, while the big screen will present the rendered sign language output and the touch screen will allow touch commands as an alternative input. The system (kiosk) will wait for an input from user. When the user occurs in front of the kiosk, it will start its SL recognizer. It will decode the information user is looking for, detect important signs for obtaining data needed (time, destination, etc.) and requests from the user. As an output, information based on this data will be produced.

Pavel Campr, Marek Hrúz and Miloš Železný are with the Department of Cybernetics, Faculty of Applied Sciences, University of West Bohemia, Pilsen, Czech Republic.

Oya Aran and Pınar Santemiz are with the Computer Engineering Department, Boğaziçi University, İstanbul, Turkey.

Alexey Karpov is with the Speech and Multimodal Interfaces Laboratory, St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences, St. Petersburg, Russia.

II. System Overview

A. The information kiosk setup

Similarly as in case of speech-based information systems, we need to design a concept for sign language-based information system. For such a system special hardware will be required to be able to acquire input data. To be able to recognize sign language (both manual gestures and oral articulation) we need to collect both detailed (facial) and upper body visual data. It is putting restrictions to the recording conditions. We present the design of hardware setup for such information system as an information kiosk allowing to be used in public places, such as train station and at the same time allowing to capture data allowing best possible recognition rates.

The information system can be divided into several basic parts: central control unit, input sensors, and output part. Input sensors comprise a visual part, an acoustic part, and a touch screen part. Output is presented to the user using graphical

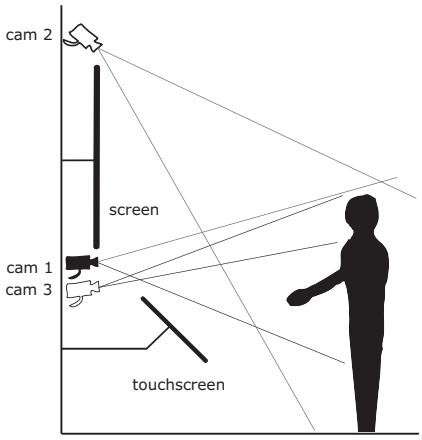


Fig. 1. Setup of cameras proposed for information kiosk for deaf users. This setup was used for recording SLR-A and SLR-B databases.

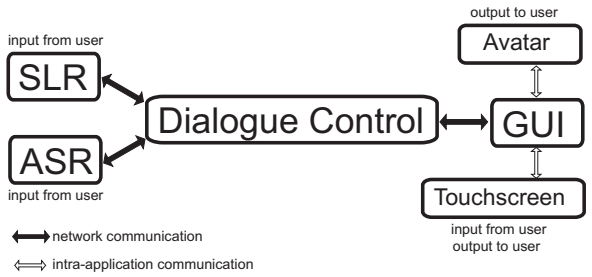


Fig. 2. Modules overview

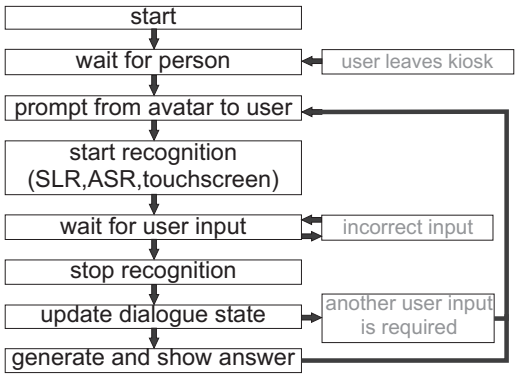


Fig. 3. Dialogue flow

screen. Central control unit will be typically a multimedia PC equipped with special hardware for recording and processing of the input data. Central control unit will also take care of the connection to external data sources, such as various types of databases (about train connections etc.)

Visual input sensors are made up by cameras situated in the rear part of the kiosk. Camera 1 looks from horizontal view at the user (the upper part of the body) standing in front of the kiosk. Camera 2 is situated at the top of the rear part of the kiosk. It looks downwards at the user's whole body. It enables to gather 3D information about the position of hands. Camera 3 is situated near Camera 1, but looks at the user's face. Data from Camera 3 will be used for automatic lip-reading, facial expressions and head gestures. Whole hardware setup is depicted in Figure 1.

Acoustic input sensor will be a microphone situated at the top of the screen, closest to the user. It should capture voice of the user with highest possible sensitivity. Acoustic part is an optional part of the system and can be used in case when communication of hearing (and speaking) people with the system is expected. However in this case, the expected recognition rate of such speech will be very low and can be used only as an accompaniment.

Haptic input sensor will be realized as a touch screen. For some easy tasks such as selection from several choices it will be more efficient to present graphically several choices and let user select by touching the screen at the position of one of the choices instead of forcing him to use sign language, as the expression in sign language can be complicated (i.e. city names that has to be spelled etc.)

Graphical screen output will be used to guide the user through the whole information providing process, allowing him/her to quickly enter most frequently used modes, letting him to choose from several choices when appropriate, and giving him feedback for his/her actions. Design of the graphical user interface will be done with respect to the special needs of users and will be as intuitive as possible to efficiently help the user to reach his/her goal.

III. Dialogue Control Module

A. Computer-driven dialogue

The interaction between the user and the kiosk is managed by a computer-driven dialogue system (Fig. 3). The dialogue consists of several scenarios, which are defined as a set of questions, their possible answers and answer generator. When the user selects a scenario (e.g. train connection information), the dialogue control asks the user for an answer of the first unanswered question. The answer can be entered as one- or multiple-word sentence (supported by speech recognizer, planned for SL recognizer in the future). All entered sentences must satisfy a context-free grammar defined by Backus–Naur Form (BNF). The grammar allows to generate a list of all sentences, their parts or words. All possible words, which can follow in the current sentence, are displayed in GUI (fig. 4, right). Later it allows to validate the user's answer. When all required questions are answered the dialogue system generates an answer for the user's query.

Scenarios are defined in one configuration file (YAML format) which can be easily modified. Here is an example with a part which define "departure information" scenario:

```
dialogue:
...
  screens:
    layout:
      template: layout.html
    ...
    departures:
      template: body_departures.html
      prompt: here you can find information about
              departures from this train station
      legend: departures
      form:
        departure:
          prompt: select departure
          label: departure
          type: grammar
          grammar: towns
          default: "town_prague."
        destination:
          prompt: select destination
          label: destination
          type: grammar
          grammar: towns
        date:
          prompt: select date of departure
          label: date of departure
          type: grammar
          grammar: date
        ...
```

The default language used in the system is English. Other languages are translated using internationalization configuration file which contains all necessary translations from English to other languages (Czech, Turkish, Russian, etc.) both spoken and signed (for sign synthesis).

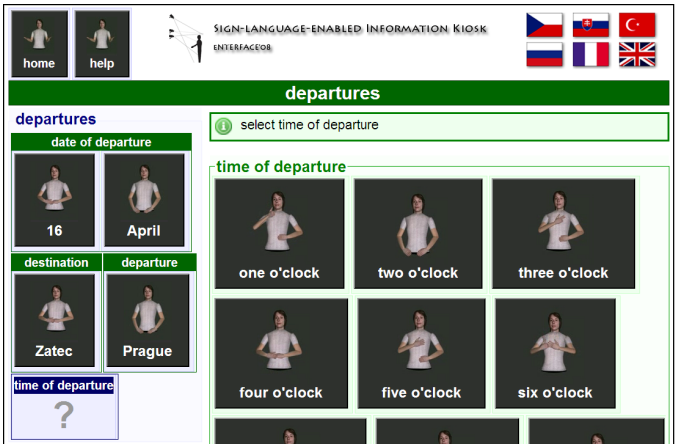


Fig. 4. Sample screen of touchscreen graphical user interface. Gray buttons and flags are clickable. Top: main menu. Left: dialogue status with answered and unanswered questions. Right: current question and all possible answers which can be selected by haptic, speech or sign-language modalities.

IV. Graphical User Interface Module

Graphical user interface (GUI) consists of two screens. The first shows a signing avatar which guides the user through the dialogue by asking questions and presenting an answer when all questions are answered. To increase interactivity the avatar turns in the direction to the user which is achieved by face detection.

The second touchscreen shows current dialogue status. There is a list of both answered and unanswered questions. By clicking on an answered question this answer can be reset. This is an advantage of using graphical representation of information which is able to present more information at one time in comparison to speech dialogue systems. Next part of the screen is used to present the list of all possible answers for current question. If the answers are multiple-word sentences then the list contains only words on the first places in the sentences and next words are listed later after the first is selected.

The first screen with signing avatar is rendered online and is described in section ”Sign language synthesis”. The second screen contains graphical user interface which is generated as a XHTML document. Particular screens are generated from XHTML templates which are filled by data generated by the dialogue control module. The signing avatars displayed on the buttons are pre-generated and showed in the XHTML document as flash animations.

V. Sign Language Recognition Module

A. Database

The sign language recognition module is intended for recognizing isolated signs. For this purpose a database was created using an application *sign_capture* developed at this workshop. This application displays a sign that should be performed by the user on the screen. This approach was chosen since the users were not familiar with Signed Czech Language. After the user is ready to perform the sign the system automatically detects the movement and starts recording from the camera

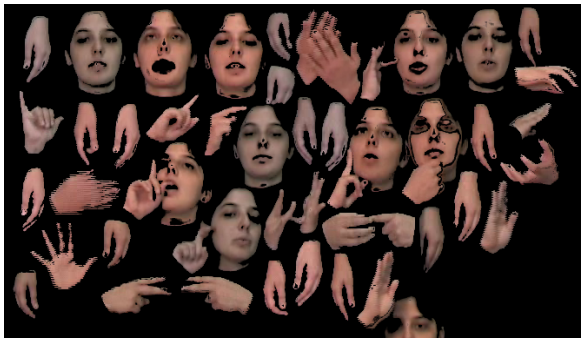


Fig. 5. Skin-color examples of one signer.

to a video file. After the hands are at the lower body and no movement is detected the capturing stops and the file is saved.

In total 338 files were recorded with one male and one female signer. The database contains 50 signs from Czech signed language such as Czech towns, days, and miscellaneous signs (i.e. yes, no, information). Every sign was repeated three times by each user. While recording the conditions were long sleeves, non-skin-colored clothes, uniform background, and constant illumination.

B. Skin Color Segmentation

Skin color is widely used to aid segmentation in finding parts of the human body [3]. We learn skin colors from a training set and then create their model 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. We prepared the set of training data by extracting images from our input video sequences and manually selected the skin colored pixels. We used images of both speakers under slightly different lighting conditions. In total, we processed 50 video segments. An example of training images is shown in Fig 5.

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 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,

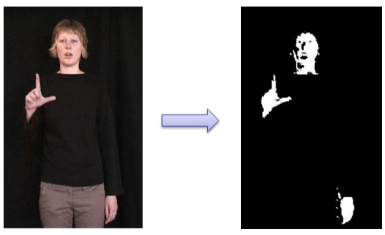


Fig. 6. Skin color segmentation example

provided the information we have is the mean, variance and gain of each Gaussian in the mixture. We have precomputed the likelihoods and stored them in a look-up table. The range of the likelihood is from 0 to 255. A likelihood of greater than or equal to 128 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 obtain the likelihood of the color of each pixel from the look-up table. According to the likelihood of the color, we decide whether it belongs to a skin segment or not. For each frame, we create a mask of skin color segments by thresholding the likelihood image. The result is as shown in Fig 6.

C. Tracking

1) The Joint Particle Filter (PF): We use a joint PF that calculates a combined likelihood for all objects by modeling the likelihood of each object with respect to others. Figure 7 shows the pseudo-code of the joint PF. $\mathbf{x}_t^n$ is the joint object state, as defined in Eq.1. The first step is to determine the initial states and the weights of the particles, with respect to the prior distribution. The initial distribution of the particles are determined with respect to an explicit detection step based on connected component labeling. The particles at time t are determined by the re-sampling, prediction and weight setting steps. At the re-sampling step, new particles are sampled with replacement from the weighted particles at time $t-1$. At this step the weights of the new particles are equally assigned. The states of the re-sampled particles at time t are determined by the object dynamics and by an additional mean-shift step. The weights are determined by normalizing the joint likelihood.

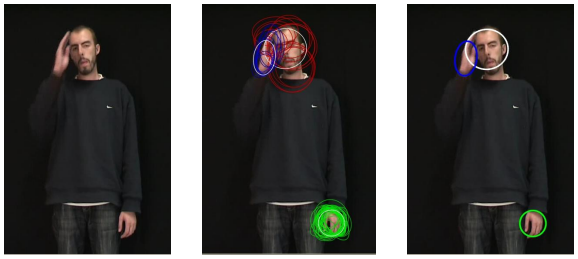
1) Initialization: $\{(\mathbf{x}_0^n, \pi_0^n)\}_{n=1}^N$
2) For $t > 0$
   a) Re-sampling: $\{(\mathbf{x}_{t-1}^n, \pi_{t-1}^n)\} \rightarrow \{(\mathbf{x'}_{t-1}^n, 1/N)\}$
   b) Prediction: $\mathbf{x''}_t^n = f(\mathbf{x'}_{t-1}^n)$ $\{(\mathbf{x'}_{t-1}^n, 1/N)\} \rightarrow \{(\mathbf{x''}_t^n, 1/N)\}$
   c) Mean-shift iterations: $\mathbf{x}_t^n = MS(\mathbf{x''}_t^n)$ $\{(\mathbf{x''}_t^n, 1/N)\} \rightarrow \{(\mathbf{x}_t^n, 1/N)\}$
   d) Weight setting: $\pi_t^n \propto \mathbf{z}_t^n = h(\mathbf{x}_t^n)$ $\{(\mathbf{x}_t^n, 1/N)\} \rightarrow \{(\mathbf{x}_t^n, \pi_t^n)\}$, $\sum_{n=1}^N \pi_t^n = 1$
   e) Estimation: $\hat{\mathbf{x}}_t = E[\{(\mathbf{x}_t^n, \pi_t^n)\}]$

Fig. 7. Joint PF algorithm

2) Object Description: The state vector for a single object consists of the position, the velocity and the shape parameters. The shape parameters are selected as the width, the height and the angle of an ellipse surrounding the object. Thus, for a single object we have a seven dimensional state vector. Then, the joint particle is a single vector containing all the objects in the scene:

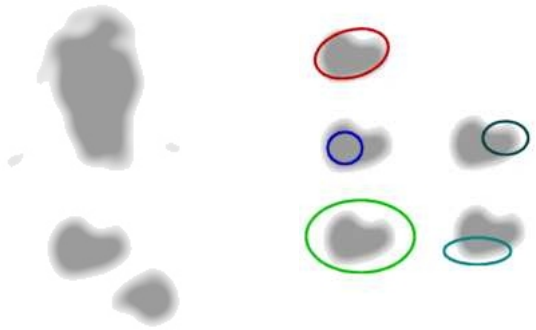
$$\mathbf{x}_t^n = \{\mathbf{x}_t^{n,f}, \mathbf{x}_t^{n,r}, \mathbf{x}_t^{n,l}\}^T \quad (1)$$

where f, r, l are indexes to the face, right and left hands. In the rest of the paper, we will refer to $\mathbf{x}_t^n$ as the joint particle and $\mathbf{x}_t^{n,i}$ as the particle or as the sub-particle, alternatively.



(a) (b) (c)

Fig. 8. (a) Original image, (b) Particle distribution with joint PF, (c) Estimated hand positions



(a) (b)

Fig. 9. (a) The thresholded image that is used in likelihood calculation, (b) hand region and different particles. The likelihood function gives the highest likelihood for the particle at the top.

3) Dynamic Model: For each object, the position and the velocity parameters are modeled by a damped velocity model and the shape parameters are modeled by a random walk model.

For multiple objects in the joint PF, the dynamic model is applied to each object. We additionally apply mean shift [4] to the sub-particles of each object independently. The MS algorithm moves the particle centers to the areas with high skin color probability. This allows us to use particles effectively, since the particles with low weights will be less likely. As a result, a PF with MS needs fewer particles than a standard PF.

4) Appearance Model: We use the skin color probability image, which has a positive probability for skin color pixels and zero probability for other colors.

To calculate the likelihood of a single object, we make two measurements based on the ellipse that is defined by the state vector of the particle:

- A: The ratio of the skin color pixels to the total number of pixels inside the ellipse.
- B: The ratio of the skin color pixels to the total number of pixels at the ellipse boundary.

These two ratios are considered jointly in order to make sure that our measurement function gives high likelihood to particles that contains the whole hand without containing many non-hand pixels [5]. If we do not take the ellipse boundary into account, smaller ellipses are favored and particles tend to get smaller. We design our measurement function Eq.2 to return

a high likelihood when A is as high as possible and B is as low as possible:

$$z_t^{n,i} = \begin{cases} 0 & , \; if \; A < \Phi_p \\ 0.5 \cdot A + 0.5 \cdot (1 - B) & , \; otherwise \end{cases} \quad (2)$$

where $z_t^{n,i}$ denotes the likelihood of a single object, i, for the particle n.

The first line in Eq.2 is required to assign low likelihood to particles that have zero or very few skin color pixels. Otherwise these particles will receive 0.5 likelihood value even if they do not contain any skin color pixels. The equation takes its highest value when there are no skin colored pixels at the boundary ($B = 0$), and when all the inner pixels are skin colored. Figure 9b shows the grey-level hand image and possible particles. The particle at the top receives the highest likelihood by Eq.2 where as the other particles receive lower likelihoods.

5) Joint Likelihood Calculation: A joint particle is a combination of sub-particles which refer to the objects we want to track, i.e. two hands and the face (Eq.1). We calculate the joint likelihood with respect to the following:

1) The likelihood of a single object based on the appearance model;
2) The distance of each object to the other objects to handle interactions. Assign low likelihood if the sub-particles are close to each other;
3) Additional constraints on respective object locations. This criterion is needed to prevent wrong object assignments especially after occlusions.

We define partial likelihoods for each criterion above and calculate the joint likelihood by the multiplication of the partial likelihoods.

The joint likelihood is calculated for the objects that stay in the scene. If any of the objects disappears, it is excluded form the likelihood calculations. We assume an object has disappeared if all of the sub-particles of that object have zero weight.

D. Handshape Feature Extraction

In order to recognize the signs we also need shape information in addition to the tracking features. In order to describe the shape we implemented five algorithms and tested their performances over the set of finger alphabet in Czech Sign Language, which can be seen in Fig 10.

After segmenting and tracking of the hands, we obtain the segmented hand image for each sign. Using this image, we find the contour and the gray scale image of the hand. Then we find DCT coefficients from the gray scale image [6], Hu moments from the segmented hand image [7], and Fourier descriptors from the contour points of the hand shape [8]. The feature extraction methods can be seen in Fig 11.

From the points on the contour, we get the following:

- Complex coordinates
- Distances of the points from the centroid of the hand
- Angles between two consecutive points

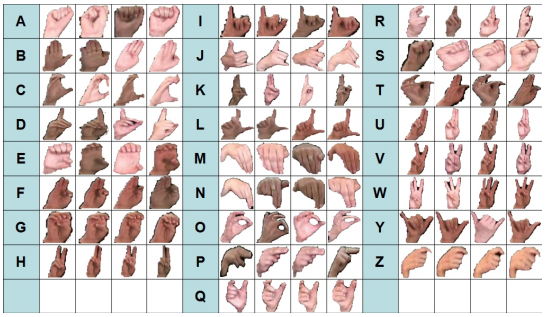


Fig. 10. Finger alphabet of the Czech sign language

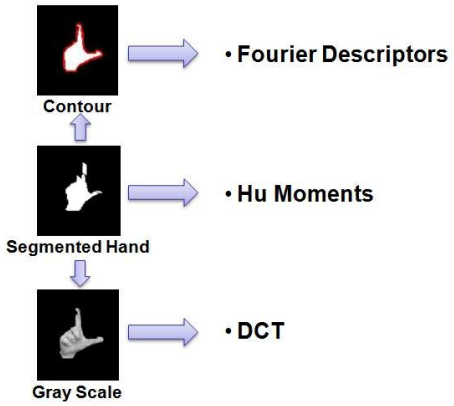


Fig. 11. Feature extraction methods

Then we calculate Fourier descriptors using these features. The procedure of obtaining Fourier descriptor features can be seen in Fig 12.

In the first method, to eliminate the effect of translation, we first subtract from each coordinate the center, compute the Fourier transform of the complex coordinate and find its absolute value. So, if the point is $P_i = (x_i, y_i)$ and the center of the hand is $P_c = (x_c, y_c)$, then the first feature is

$$f_i^1 = \text{abs}\Big[\text{fft}\big((x_i - x_c) + i(y_i - y_c)\big)\Big] \quad (3)$$

In the second method, we find the distance of the boundary points from the centroid of the shape. So,

$$f_i^2 = \text{abs}\Big[\text{fft}\big(\sqrt{(x_i - x_c)^2 + (y_i - y_c)^2}\big)\Big] \quad (4)$$

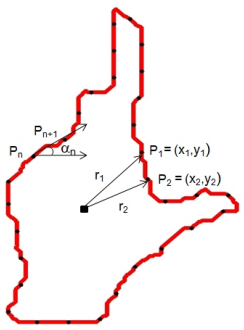


Fig. 12. The procedure to obtain the complex coordinates, central distances and the angle between consecutive points

In the third method, we first compute the angles between two consecutive points and compute the fft features using the formula

$$f_i^3 = \text{abs}\Big[\text{fft}\big(\arctan\frac{y_{i+1}-y_i}{x_{i+1}-x_i}\big)\Big] \quad (5)$$

In the fourth method, we compute the DCT features from the gray scale image of the hand [6]. Then we trace the upper-left corner of the DCT image diagonally and choose the first 15 features and eliminate the first feature, which is the DC component.

Finally in the last method, we compute the Hu moments from the segmented hand image [7]. In order to find the Hu moments, we first compute the internal moments of order p+q using the formula

$$m_{pq} = \iint (x-x_c)^p (y-y_c)^p dxdy \quad (6)$$

where x, y are the pixel coordinates belonging to the hand. Then we normalize the moments for eliminating the scale factor

$$n_{pq} = \frac{m_{pq}}{m_{00}^{\frac{p+q}{2}+1}} \quad (7)$$

We use seven Hu moments which are invariant to rotation. The first six are also reflection invariant whereas the seventh moment is skew orthogonal invariant, which is useful in distinguishing mirror images.

$$\begin{aligned}
S_1 &= n_{20}+n_{02}\\
S_2 &= (n_{20}+n_{02})^2+4n_{11}^2\\
S_3 &= (n_{30}-3n_{12})^2+(n_{03}-3n_{21})^2\\
S_4 &= (n_{30}+n_{12})^2+(n_{03}+n_{21})^2\\
S_5 &= (n_{30}-3n_{12})\cdot(n_{30}+n_{12})\cdot\\
&\quad \big((n_{30}+n_{12})^2-3(n_{03}+n_{21})^2\big)\\
&\quad -(n_{03}-3n_{21})\cdot(n_{03}+n_21)\\
&\quad \cdot\big(3(n_{30}+n_{12})^2-(n_{03}+n_{21})^2\big)\\
S_6 &= (n_{20}-n_{02})\cdot\big((n_{30}+n_{12})^2-(n_{03}+n_{21})^2\big)\\
&\quad +4n_{11}^2\cdot(n_{30}+n_{12})\cdot(n_{03}+n_{21})\\
S_7 &= (3n_{21}-n_{03})\cdot(n_{30}+n_{12})\cdot\\
&\quad \big((n_{30}+n_{12})^2-3(n_{03}+n_{21})^2\big)\\
&\quad -(n_{30}-3n_{12})\cdot(n_{03}+n_{21})\\
&\quad \cdot\big(3(n_{30}+n_{12})^2-(n_{03}+n_{21})^2\big)
\end{aligned}$$

In our experiment, we use a database containing 241 images from the Czech sign language finger alphabet. In this database there are 23 classes which are some of the letters in the alphabet. Each class contains six to 18 samples. To create our training set, we select five samples from each class randomly and put the remaining samples to the test set. We form six folds in total, where we guarantee that each sample is put at least once to the training set. So, in each fold training set includes 115 samples and test set includes 126 samples. We use 1-nearest neighbor method as a classifier. The recognition results can be seen in Table V-D.

According to our results, DCT obtained significantly better results than the others. The reason for that is DCT also takes into account the gray scale texture information in the hand region whereas the others only use shape information.

Method	Accuracy	Std Deviation
FFT-Complex	22%	2,05%
FFT-Centroid	23%	2,51%
FFT-Angle	17%	3,70%
DCT	**75%**	**3,92%**
Hu Moments	30%	2,54%

TABLE I
RECOGNITION RESULTS USING THE SET OF CZECH SIGN LANGUAGE FINGER ALPHABET

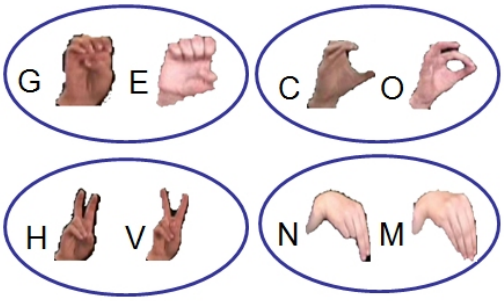


Fig. 13. Examples of images that are misclassified by the DCT method

Some signs have a similar shape, so using only shape information it is difficult to differentiate between them. Some examples can be seen in Fig 13.

E. Sign Recognition

For the purpose of recognition we use Hidden Markov Model (HMM). The signs are modeled as an 8-state HMM (two of these states are non-emitting). Each state is modeled with a single Gaussian. This was due to the relative low amount of data. The structure of the HMM can be seen in Fig. 14.

One model of a sign was trained on 4 out of 6 video sequences. PCA and ICA were tested to reduce the dimensionality of the data and to align the data according to the feature space coordinate system. Both methods failed to improve the recognition rate since there were too few samples available. The remaining two video sequences were used to test the system. Results can be seen in Table II.

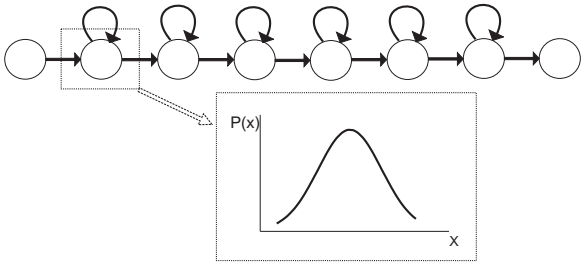


Fig. 14. The structure of HMM used for modeling the signs.

No. parameters	Method	Recognition rate
49	x	81.63%
49	PCA	78.57%
39	PCA	78.57%
29	PCA	76.53%
19	PCA	70.41%
9	PCA	47.96%

TABLE II
RECOGNITION RATE OF SLR

The hands of a signer were approximated as ellipses. Following features were extracted :

- Center of the ellipse
- Size of the ellipse (width and height)
- Angle between the image x-axis and the major axis of the ellipse
- Velocity of the ellipse
- DCT coefficients of images of both hands

Out of 338 video sequences 15 were not tracked correctly. DCT coefficients improved the recognition rate significantly.

VI. SIGN LANGUAGE SYNTHESIS MODULE

A. Feedback Animation

1) Animation Model: The feedback animation module can be divided to a module for rendering the animation model and to a module for forming the animation trajectories (a trajectory generator). 3D geometric animation model of the avatar is in compliance with the H-Anim standard. Currently the animation model covers 38 joints and body segments. Each segment is represented as textured triangular surface, 16 segments is used for fingers and the palm, one for the arm and one for the forearm, totally 1372 vertices and 1772 triangles per hand. The thorax and the stomach are together represented by one segment, 182 vertices and 360 triangles. The talking head is composed from seven segments, totally 4692 vertices and 9243 triangles.

The body segments are connected by the avatar skeleton. One joint per segment is sufficient for this purpose. A controlling of the skeleton is carried out thought the rotation of segments (3 DOF per joint). The rotation of the shoulder, elbow, and wrist joints are completed from 3D positions of the wrist joints by the inverse kinematics.

The animation of the talking head is performed by a local deformation of the triangular surfaces. It is primarily used for the animation of the avatar's face and the tongue [9]. The triangular surfaces are deformed according to influence zones defined on the triangular surface by spline functions constructed from several 3D control points. The collection of these points is currently taken by 9 animation parameters. The rendering of the animation model is implemented in C++ and OpenGL code.

2) Trajectory Generator: Firstly for the manual component of signed speech, the trajectory generator performs the syntactic analysis of the symbolic string on the input HamNoSys string and creates parse tree structure. Currently the trajectory generator uses 374 parse rules. However, it is difficult to define rules and actions for all symbol combinations to cover the entire notation variability. We made a few restrictions in order to preserve maximum degree of freedom. In this assumption, the annotation of the sign have a good meaning for the user familiar with HamNoSys as well as signs are obvious enough for the transformation to the avatar animation. Next, the structurally correct string is decomposed to nodes determined by parsing rules. Two key frames data structure to distinguish the dominant and non dominant hand are used. The data structure of the key frame is composed from items specially designed for trajectory generation purpose (Figure 15 on the left). Next step of trajectory forming is filling of the leaf nodes from the symbol descriptors stored in the definition file (Figure 15 on the right). The definition file covers 138 HamNoSys symbols.

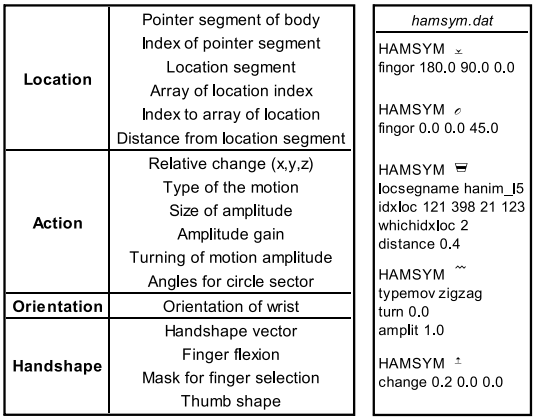


Fig. 15. Left panel: The list of all items. Right panel: An example of the items stored in the definition file.

Block	HamNoSys Symbols		Parser	
	#Base	#Auxiliary	#Rules	#Actions
Symmetry	4	0	8	8
Handshape	12	11	35	6
Finger and palm orientation	26	3	34	5
Location	30	17	85	14
Action	40	42	204	15
Link of blocks	0	0	8	1

Fig. 16. The statistic of symbols, rules, and actions used by the HamNoSys parser.

39 rule actions were added in a manner that one rule action is connected with each parse rule. The parse tree is processed by several tree walks. The initial tree walk put together the items of the key frames according to the type of the rule actions. The reduced tree is joined and transformed to the trajectories accordance with the timing of the particular nodes. Finally, the final trajectories for both hands are contained in the root node. The number of used symbols, parse rules, and actions used in the system is summarized in Figure 16. The final step is the conversion of the trajectories into the avatar animation and synchronization with articulatory trajectories generated by the selection of articulatory targets [9].

3) Evaluation of Synthesized Signs by Deaf: A subject evaluation of quality of synthesized signs has been performed in advance with the manual and non-manual components of a sign. The perception test by deaf children from primary school

was scored on the isolated signs. Two experiments both composed of two tests were designed. The second test followed the first one after three weeks with identical procedure. Five deaf pupils were chosen from the preliminary class and the first class (5-6 years) as participants for the first experiment and six deaf pupils from the sixth and seventh class (11-13 years) for the second experiment.

Fig. 17. An example of two choices from the sheet used for filling answers in multiple-choice test. Three choices are offered in form of three illustrations.

Test material contains the synthesized animation of isolated sign recorded to the video files. 15 signs from videotapes used in the curriculum of the preliminary class were collected, thus the signs would be known by all participated pupils. The sign editor has been used for HamNoSys notation of the sings. Next, we have prepared sheets for the multiple-choice test with three response options (one correct response) composed from randomly arranged pictures of the tested words, Figure 17.

Procedure The tested signs were presented on the wall in the classroom. At the beginning of the experiments, five extra non-scored words were presented to demonstrate various options of the study. The size of the signing avatar was approximately 30 cm on the wall. The pupils were not familiar with the tested words before the experiments and the scribing was prevented.

Results Tests have been scored as follows. A pupil got one point for the correct answer and no points for the wrong answer. The similarity of some signs was not taken into account. We tested the hypothesis that pupils filled the multiple-choice test by a chance. For this purpose, we have used one-sample and one-sided t-test. The planned comparisons are carried out for both tests of the first experiment and for the first test of the second experiment, ($\alpha = 0.01$). The results show significantly better understanding of the signed speech than a chance (three options, the chance level 33.3%, p <0.01). The average 80% success was achieved in the first test of first experiment (t (4) = 6.6, p = 0.0014) and 85% for the second test (t (3) = 17.91, p $<$ 0.001). Better results were achieved in the second experiment with older pupils. There were for the first test on average 95% correct answers (t (5) = 27.59, p $<$ 0.001) and the retesting without the possibility of choosing, on average 80% correct answers. The means of achieved scores are summarized in Table III.

TABLE III
SIGN LANGUAGE SYNTHESIS UNDERSTANDING EXPERIMENTS

Experiment	1	2
Test 1	80%	95%
Test 2	85%	80%

VII. AUTOMATIC SPEECH RECOGNITION MODULE

A speaker-dependent automatic speech recognition (ASR) system was developed and embedded into the information kiosk as an optional alternative to the interactive GUI and the interface based on sign language recognition. ASR system is multilingual one and able to recognize voice commands both in English and Czech. The lexicon of ASR contains 101 diverse words for each language (Czech towns, digits, names of months and weekdays, etc.)

A. Signal processing and feature extraction

The audio signal is captured by a microphone of a headset and sampled at 16 KHz with 16 bits on each sample using a linear scale. The system is intended for the distant talking and human-computer interaction, so the microphone was selected to be able to capture speech signal at the distance of 2 meters from a speaker with an acceptable SNR. The signal is divided into the frames and cepstral coefficients are computed for the 25 ms overlapping frames with 10 ms shift between adjacent frames applying the bank of triangular filters calculated according to the Mel-scale frequencies 8:

$$Mel(f) = 2595 \log_{10} 1 + \frac{f}{700} \tag{8}$$

Mel-frequency cepstral coefficients (MFCCs) are calculated from the log filterbank amplitudes using the discrete cosine transform. So the audio speech recognizer system calculates 13 MFCCs (including 0-th coefficient) as well as estimates the first and second order derivatives that forms an observation vector of 39 components. The acoustical modeling is based on left-right continuous Hidden Markov Models (HMMs) [10], applying mixtures of Gaussian probability density functions that are defined according to the equation 9:

$$\mathcal{N}(o; \mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} e^{-\frac{1}{2}(o-\mu)^T \Sigma (o-\mu)} \tag{9}$$

where $\mathcal{N}$ is a Gaussian with mean vector μ and covariance matrix Σ, and n is the dimensionality of the observable vector o. HMMs of phonemes have three meaningful states and two "hollow" states intended for concatenation of the models of phonemes in the models of words. Each word of the vocabulary is obtained by concatenation of context-independent phonemes (triphones). As the base technology for realization of speech recognizer Hidden Markov Models Toolkit (HTK) [11], developed by the Cambridge University Engineering Department, was used. HTK 3.4 toolkit is employed on all the levels of audio signal processing. Modeling of speech by HTK includes two main stages:

- Training HMMs of acoustical items using a phonetically labeled speech corpus.
- Speech recognition in on-line or off-line modes.

B. System's training and evaluation

In order to train the speech recognizer a speech corpus was recorded in office conditions using the distant talking directed microphone. About 1000 utterances of two users were recorded and used for training HMMs of phonemes. Totally we have recorded above 10 minutes of speech data from each speaker, these data were labeled semi-automatically in the terms of phoneme sets. It is required to notice that the phonemic alphabets, for instance SAMPA alphabets [12] are rather different for English and Czech, the latter contains more phonemes. Totally 41 different phonemes are used in transcriptions of the lexicon, some words of which have several variants of pronunciations taking into account peculiarities of the speakers. 20 % training utterances were manually labeled by the WaveSurfer software, the rest of the data were automatically segmented by the Viterbi forced alignment method with the flat start [11]. In general the stage of acoustical models training includes the following steps:

- Manual transcription of a lexicon of an applied domain
- Creation of a grammar or a statistical language model (for instance, bi-gram or tri-gram model)
- Preparation of a training speech corpus
- Coding the speech data (feature extraction)
- Definition of topology of HMMs (prototypes)
- Creation of initial HMMs by flat start
- Re-estimation of HMMs parameters of monophones using a labeled speech corpus and Baum-Welch algorithm
- Re-estimation of HMMs of triphones by Baum-Welch algorithm
- Mixture splitting

The speech decoder uses Viterbi-based token passing algorithm [11]. The input phrase syntax is described in a simple grammar that allows to recognize one command in a hypothesis. The audio speech recognizer operates quite fast (less than 0.5xRT) so the result of speech recognition is available almost immediately after detection of speech end by an energy-based voice activity detector. The performance of ASR was evaluated by another speech data, collected in the same office conditions as the training part. The word recognition rate (WRR) was more than 90% on testing phrases pronounced by two male speakers. This rate is acceptable for our task since ordinary single microphone is used for distant speech capturing providing quite low SNR. It was found that SNR for audio signal is under 20 db because of far position (about 2 meters) of the speaker in front of the kiosk and the microphone. In further research a microphone array and corresponding digital signal processing methods for speaker localization and noise elimination are supposed to be applied [13].

VIII. Evaluation and Conclusion

Because the kiosk should be usable by all hearing-impaired people (even who cannot read) we designed the GUI in a way that all important text labels are accompanied by an animation of signing avatar with corresponding sign. This new graphical user interface component can be used as a clickable button. This component and whole GUI layout is designed as a XHTML web page, so that the whole application could be used online, but without ASR and SLR components, and controlled only by a mouse.

The biggest usability problem of the kiosk design is the SL recognition where the user have to start and finish the performed sign in the initial position. This expectation must be explained to the user in the early beginning of the dialogue.

The other parts of the kiosk proved that they can be used without any major usability difficulties. This is achieved by using similar concept of the dialogue as is used in web page browsing which is well-known and the users don't have to think how to control the kiosk.

All modules and scenario configuration are designed in a way that the kiosk can be easily modified to provide another information service or even another type of application (e.g. sign language tutoring tools [1], sign language games or sign language dictionaries [14]).

Acknowledgment

This project is supported by The Scientific and Technological Research Council of Turkey (TUBITAK) project 107E021, Bogazici University project BAP-03S106.

This research was supported by the Grant Agency of Academy of Sciences of the Czech Republic, project No. 1ET101470416 and by the Ministry of Education of the Czech Republic, project No. ME08106.

We would like to express our great thanks to Zdeněk Krňoul (University of West Bohemia, Pilsen) for contribution to this project by providing 3D avatar for sign language synthesis.

REFERENCES

[1] O. Aran, I. Ari, A. Benoit, A. H. Carrillo, F.-X. Fanard, P. Campr, L. Akarun, A. Caplier, M. Rombaut, and B. Sankur, "Sign Language Tutoring Tool," *Proceedings of eNTERFACE 2006, Summer Workshop on Multimodal Interfaces, Dubrovnik, Croatia*, 2006.

[2] M. Železný, P. Campr, Z. Krňoul, and M. Hrúz, "Design of a multimodal information kiosk for aurally handicapped people," in *SPECOM 2007 proceedings*, Moscow, Russia, 2005, pp. 751–755.

[3] V. S. V. Vezhnevets and A. Andreeva, "A survey on pixel-based skin color detection techniques," in *Graphicon*, 2003, pp. 85–92.

[4] E. Maggio and A. Cavallaro., "Hybrid particle filter and mean shift tracker with adaptive transition model," in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2005.

[5] O. Aran and L. Akarun, "A particle filter based algorithm for robust tracking of hands and face under occlusion," in *IEEE 16th Signal Processing and Communications Applications (SIU 2008)*, 2008.

[6] R. C. Gonzales and R. E. Woods, *Digital Image Processing*. Prentice-Hall, 2001.

[7] M. Hu, "Visual pattern recognition by moment invariants," *IRE Trans. Inf. Theory*, vol. 8, p. 179–187, 1962.

[8] D. S. Zhang and G. Lu, "A comparative study of three region shape descriptors," in *Proc. of the Sixth Digital Image Computing Techniques and Applications (DICTA02)*, 2002, pp. 86–91.

[9] Z. Krňoul and M. Železný, "A development of Czech talking head," in *Proceedings of ICSPL 2008*, in press, 2008.

[10] R. Rabiner and B. Juang, "Fundamentals of speech recognition," in *New Jersey: Prentice-Hall, Englewood Cliffs, USA*, 1993.

[11] S. Y. et al., "The htk book," in *HTK Version 3.4, Cambridge University Engineering Department*, 2006.

[12] "Sampa alphabet, http://www.phon.ucl.ac.uk/home/sampa/."

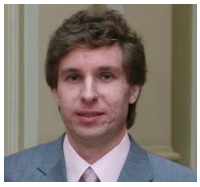
[13] S. Brandstein and D. Ward, "Microphone arrays," in *Springer Verlag*, 2000.

[14] J. Langer, "Přínos elektronických výukových pomůcek a slovníků znakového jazyka," *In Vzdělávání sluchově postižných. Praha: MŠMT, 2006.*, 2006.

Oya Aran Oya Aran received the BS and MS degrees in Computer Engineering from Boğaziçi University, Istanbul, Turkey in 2000 and 2002, respectively. She is currently a PhD candidate at Boğaziçi University working on dynamic hand gesture and sign language recognition under the supervision of Prof. Lale Akarun. Her research interests include computer vision, pattern recognition and machine learning. She is a student member of the IEEE.

Pavel Campr Pavel Campr was born in 1981 in the Czech Republic. He graduated in cybernetics from the University of West Bohemia (UWB) in 2005. As a Ph.D. candidate at the Department of Cybernetics, UWB, his research interests focus on hand gesture and sign language recognition, computer vision, machine learning and multimodal human-computer interaction. He is participating in the research project MUSSLAP. He is also teaching assistant and maintainer of the departmental website.

Marek Hrúz was born in 1983 in Slovakia. He received his M.S. degree in cybernetics from the University of West Bohemia (UWB) in 2006. As a Ph.D. candidate at the Department of Cybernetics, UWB, his research interests focus on hand gesture and sign language recognition, particularly tracking and image parametrization, computer vision, machine learning and multimodal human-computer interaction. He is participating in the research project MUSSLAP and is a teaching assistant at the Department of Cybernetics, UWB.



Alexey Karpov 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. His main research interests include automatic speech and speaker recognition, text-to-speech, multimodal interfaces based on speech and gestures, audio-visual speech processing. Currently he is a senior researcher of Speech and Multimodal Interfaces Laboratory of SPIIRAS. He has been the (co)author of more than 60 papers in refereed journals and International conferences, for instance, Interspeech, Eusipco, TSD, etc. His main research results are published by the Journal of Multimodal User Interfaces and by the Pattern Recognition and Image Analysis (Springer). He is a coauthor of the book "Speech and Multimodal Interfaces" (2006, in Russian), and a chapter in the book "Multimodal User Interfaces: From Signals to Interaction" (2008, Springer). He has also been involved in EU SIMILAR Network of Excellence as well as several research projects funded by EU INTAS association and Russian scientific foundations. He was a winner of the 2-nd Low Cost Multimodal Interfaces Software (Loco Mummy) Contest (2006, Brussels). Dr. Karpov is a member of organizing committee of series of International conferences "Speech and Computer" SPECOM, as well as member of the EURASIP, ISCA and OpenInterface associations.

Pınar Santemiz Pınar Santemiz received the B.S. degree in mathematics from Boğaziçi University, Istanbul, Turkey, in 2006. She is currently an M.Sc. student in the Department of Computer Engineering, Boğaziçi University. Her research interests are in the areas of computer vision, sign language analysis, machine learning, and pattern recognition.

Miloš Železný was born in Plzeň, Czech Republic, in 1971. He received his Ing. (=M.S.) and Ph.D. degrees in Cybernetics from the University of West Bohemia, Plzeň, Czech Republic (UWB) in 1994 and in 2002 respectively.

He is currently a lecturer at the UWB. He has been delivering lectures on Digital Image Processing, Structural Pattern Recognition and Remote Sensing since 1996 at UWB. He is working in projects on multi-modal speech interfaces (audio-visual speech, gestures, emotions, sign language). He is a member of ISCA, AVISA, and CPRS societies. He is a reviewer of the INTERSPEECH conference series.

Design and Evaluation of an Audio-Haptic Interface

Emma Murphy (1), Camille Moussette (2), Charles Verron (3), Catherine Guastavino (1)

(1) McGill University, Canada (2) Umeå Institute of Design, Sweden (3) Orange Labs, France

Abstract

A non-visual 3D virtual environment, composed of a number of parallel planes has been developed to explore how auditory cues can be enhanced using haptic feedback for navigation. 23 users were asked to locate a target in the virtual structure across both horizontal and vertical orientations of the planes, with and without haptic feedback. In this report, the design of auditory and haptic cues are described and a perceptual evaluation is presented in terms of experimental design, results and analysis.

***Index Terms*—multimodal, haptic feedback, audio feedback, perceptual evaluation.**

I. Introduction

The main aim of this project was to create a non-visual virtual environment in order to investigate issues of integration and effectiveness of information delivery between audio and haptic modalities. It is proposed that an interface involving a target finding task with audio and haptic (touch) feedback is relevant to this aim. Various studies have indicated that the use of non-speech audio and haptics can help improve access to graphical user interfaces [8, 10], by reducing the burden on other senses, such as vision and speech. Furthermore a mixture of sound and haptic feedback has been found to assist visually impaired users by aiding navigation through virtual environments [11, 16].

There is a substantial body of literature from the field of interface design for musical expression that can inform multimodal interactions for user interface design. Although this discipline is concerned with designing interactions for creative purposes, the issues of mapping between audio and haptic modalities are relevant to user interface design. For example, both gesture-based control of sound synthesis [14] and natural interactions between audio and haptic modalities [1] are also central issues for interaction design in HCI. Wanderly and Orio [13] have made direct parallels between musical tasks and tasks in HCI: "In particular, target acquisition may be similar to the performance of single tones (acquiring a given pitch as well as a given loudness or timbre), while constrained motion may be similar to the performance of specific phrase contours" [13]

Furthermore Marentakis and Brewster [5] have investigated the process of using gesture to control 3D sound for purposes of information display. In this experimental study, time and accuracy ratings indicate that deictic gesture interaction with a spatial audio display is a robust and efficient interaction technique. Studies have also investigated the combination of audio and haptics to convey object location in non-visual spatial structures [15, 4, 6]. In terms of target acquisition tasks previous research has investigated the addition of haptics to visual interfaces for target-finding [9,12]. More recently Kim and Kwon [3] implemented a haptic and audio grid in order to enhance recognition for ambiguous visual depth cues. The authors implemented a haptic vertical grid and pitch variation to convey a target location to users and subsequent evaluations revealed that the multimodal cues increased precision, particularly in the z-axis [3].

The design of both audio and haptic cues for the virtual interface presented in this study have been informed by previous literature and research. In the following sections, the design of auditory and haptic cues are described and the user evaluation is presented in terms of experimental design, results and analysis.

II. Audio-Haptic Interface

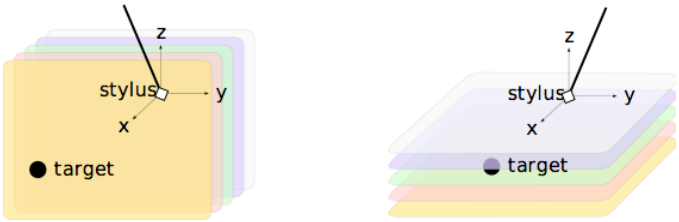


Figure 1: Virtual environment composed of a number of parallel planes positioned either horizontal or vertical to the phantom stylus.

A virtual environment was created composed of a number of parallel planes of the same dimensions that could be navigated by moving between planes or remaining on the one plane. The virtual structure could be rotated in space so that the planes were either horizontal or vertical (see figure 1). A virtual target was located at a random location on one of the planes. In order to find the target the user first has to navigate through the planes and identify which plane contains the target and then use the auditory cues to locate it. The design idea behind this virtual structure was to create a structure that could be navigated in 3D space. Practical applications of the plane structure include the design of a menu structure or browser for a restricted visual space or non-visual interface.

A. Audio

Auditory cues were integrated with the haptic environment using MAX/MSP. While the design and evaluation for this interface was not based on a musical instrument, the sounds were informed by a musical design. The idea was realised as a way to complement the haptic movement through the planes as a plucked sound and on a plane as a smooth bowing sound.

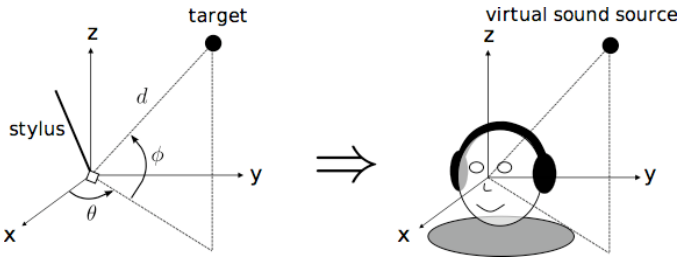


Figure 2: 3D Sound Mappings

The sound samples were cello recordings (retrieved from www.freesound.org/). Each plane was assigned a plucked pitch corresponding to the first 5 notes of a major scale mapped to the ascending plane number (A3, B3, C#3, D3, E3). The plucked sound was intended to complement the haptic gesture as the users pushed through magnetic feedback on the planes. The bowed sound location cue was intended to complement to the smooth haptic texture while the user navigated the planes. The pitch of the bowed location cue was the same as the plucked sound for that plane level. When the user successfully located a target a distinctive bowed chord auditory cue was played.

B. 3D Audio Rendering

The bowed sound location cue was spatialised using binaural synthesis with the "ears in hand" technique similar to that reported in [7]. Let (O,X,Y,Z) be the fixed reference coordinate system for both configurations (horizontal and vertical planes). The position of the target is calculated in a system (x,y,z) which translates in (O,X,Y,Z) with its origin attached to the stylus (see figure 2). Let (θφd) be the spherical coordinates of the target in (x,y,z). The virtual sound source (the bowed cello sound) is spatialised in the same direction in a coordinate system attached to the head of the listener.

The spat object from IRCAM for MAX/MSP was used for binaural sound rendering over headphones [2] Directional cues (θφ) are simulated with Head Related Transfer Function (HRTF) filtering based on KEMAR measurements. Reverberation is also used in the object to enhance the externalisation of the virtual source (Jot, 1999). The distance cue is simulated by a 6 dB attenuation of the direct sound when the target-stylus distance doubles. Let d0 be a reference distance of 4 cm, for which the virtual sound source level is calibrated to approximately 50 dB-SPL, the gain g of the direct sound is given by:

if d>d0 then g = d0/d else g=1

The virtual sound source is played only when the target and the stylus are located on the same plane (horizontal or vertical, according to the configuration). Consequently, in the horizontal configuration, the elevation of the virtual sound source is 0 degree and the azimuth varies continuously between -180 to +180 degrees; in the vertical configuration, the azimuth of the virtual source is either -90 or +90 degrees, and the elevation varies continuously between -90 and +90 degrees.

C. Haptics

The haptic effects were designed using H3D, an open source haptic graphic API based on X3D. The experiment was conducted using a Phantom Omni haptic device from Sensable Technologies Inc. This unit offers 3 degrees of freedom with a stylus-type grip and provides a workspace area of 160 W x 120 H x 70 D mm. The Phantom Omni can produce a maximum exertable force of 3.3 Newtons.

The virtual scene created for this interface contained five support planes equally spaced across the device's working area plus a target, a 3D spherical object positioned on one of the plane. A magnetic effect was used to create rigid surfaces to hold users to the planes while they were navigating. The magnetic effect consisted of forces generated in order to keep the proxy (device) on the surface. If the proxy is pulled outside a specific delta distance from the surface it will be freed from the magnetic attraction. The target also had a magnetic force applied, but it should be noted that the forces were weak enough to not attract or guide the device towards the target.

The forces were limited to the planes and the target. The scene and planes were large enough to totally filled the workspace area. Naturally the users were limited also to the mechanical limits of the device in all three axes.

III. Perceptual Evaluation

A. Experimental Design

The aim of the evaluation of the experiment was to compare the differences for usability and navigation using the two different orientations of the virtual structure; horizontal and vertical, and the two different types of feedback: audio and audio-haptic. Specifically, we investigated whether it would be possible to navigate the interface without the support of the haptic planes. Due to the fact that the interface was not designed with redundant information between modalities we hypothesised that it would be more difficult for users to navigate the environment without haptic feedback.

The experimental design was based on the following independent variables;

- Feedback: audio only, audio-haptic
- Orientation of the virtual planes: vertical, horizontal

This resulted in four experimental conditions:

1. Audio-Haptic Vertical
2. Audio-Haptic Horizontal
3. Audio-Only Vertical
4. Audio-Only Horizontal

The following dependent variables were integrated into the experimental design:

- Completion times; time taken for the user to navigate through the planes and find the target
- Trajectories; path travelled by user from the starting point of the stylus to the target
- Perceived effectiveness and ease of use of both audio and haptic cues
- Cognitive strategies reported by participants to navigate the virtual structure and located the target.

B. Procedure

23 participants volunteered for the study, 20 males and 3 females between the ages of 24 and 55. Participants were either involved in the eNTERFACE '08 project or part of the wider research community at LIMSI.

Participants were provided with both a short training introduction and a practical trial session before beginning the main experiment. The initial training introduction involved presenting a visual description of the virtual structure. While the entire experiment was non-visual it was considered useful to show the users a visual representation of the virtual structure so that they could visualise the task mentally. Furthermore for many of the participants, English was not a first language and therefore the provision of a visual representation ensured that users were not only relying on the experiment protocol instructions delivered through English. During this introductory session users became familiar with some of the audio cues using a visual representation of the planes. It was considered relevant to train the users on the sounds to convey changing planes, locating the target and the confirmation sound when the target was successfully hit. However, users were not give any information about the 3D audio mappings or the haptic feedback.

In a practical training session, users were presented with 8 trials, 2 of each condition. Users were asked to navigate the planes, firstly find the plane with the target and then locate that target. For the main experiment participants were presented with 44 Trials (11 per condition)

The condition and target position within and across planes were randomised. In order to move onto the next trial participants were asked to press a button on the phantom stylus when they were ready to move on. Timings were recorded from this point until the user located the target. User trajectories across and within the virtual planes were also recorded. Participants were also asked to complete a post-task questionnaire concerning perceived effectiveness and ease of use of the audio-haptic cues and the participant's cognitive strategies for finding the target.

C. Participants

23 participants volunteered for the study, 20 males and 3 females between the ages of 24 and 55 (mean age: 32, SD: 8.2). Participants were either involved in the eNTERFACE '08 workshop or part of the wider research community at LIMSI, CNRS, France, involved in audio, computer science or engineering related research.

D. Procedure

Participants were provided with both a short training introduction and a practical trial session before beginning the main experiment. The initial training introduction involved presenting a visual description of the virtual structure. While the entire experiment was non-visual it was considered useful to show the users a visual representation of the virtual structure so that they could visualise the task mentally. Furthermore for many of the participants, English was not a first language and therefore the provision of a visual representation ensured that users were not only relying on the experiment protocol instructions delivered through English. During this introductory session users became familiar with some of the audio cues using a visual representation of the planes. It was considered relevant to train the users on the sounds to convey changing planes, locating the target and the confirmation sound when the target was successfully hit. However, users were not given any information about the 3D audio mappings or the haptic feedback.

In a practical training session, users were presented with 8 trials, 2 of each condition. Users were asked to navigate the planes, firstly find the plane with the target and then locate that target. For the main experiment participants were presented with 44 trials (11 per condition), except for the first 3 participants who completed 40 trials (10 per condition). The condition and target position within and across planes were randomised. In order to move onto the next trial participants were asked to press a button on the phantom stylus when they were ready to move on. Timings were recorded from this point until the user located the target. User trajectories across and within the virtual planes were also recorded. Participants were also asked to complete a post-task questionnaire concerning perceived effectiveness and ease of use of the audio-haptic cues and the participant's cognitive strategies for finding the target.

IV. INITIAL RESULTS

A. Completion Times

Users successfully located the target across all trials for the audio-haptic condition. In the audio-only condition, 4 users failed to locate the target for a combined total of 10 trials (6 horizontal, 4 vertical). Their timings were recorded for the unsuccessful trials (AV 167s) but not included in the data for completion times.

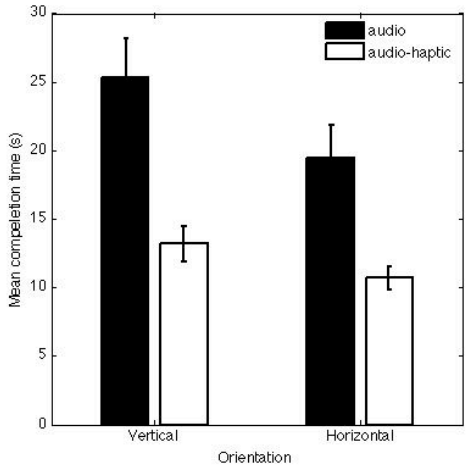


Figure 3: Mean completion times for vertical and horizontal orientations across both audio only and audio-haptic conditions

Furthermore, 62 trials out of 1000 trials were excluded from the data analysis due to technical difficulties. A repeated measures factorial ANOVA for completion times revealed significant effects of orientation and feedback. The time taken for users to complete the task in the horizontal condition was significantly less than that of vertical condition ($F(1, 936) = 6.1$, $p=0.02$) (see figure 3).

In addition, as illustrated in figure 3, completion times for the audio-only condition were significantly longer than that of the audio-haptic condition ($F(1, 936) = 49.3$, $p<0.001$). Furthermore, post-hoc tests revealed a significant difference between the vertical and horizontal orientations for both audio-only ($t(243) = -2.94$, p two-tailed =0.004) and audio-haptic conditions ($t(228) = -3.07$, p two-tailed =0.002). There were no interaction effects between feedback and orientation.

B. Qualitative Post task Feedback

In the post-task questionnaires users were asked to describe their strategies for finding the target using the multimodal cues. Most users described a process of first navigating the virtual space to determine the orientation of the planes, then using the auditory cues to first determine the correct plane and concentrate on the location cue to find the target. A further analysis of qualitative user comments will be conducted in conjunction with an analysis of recorded user trajectories.

V. Conclusion And Future Work

The focus of this project was to enhance an auditory navigation task in a non-visual virtual environment though haptic cues. Evaluation experiments revealed that the audio-haptic feedback condition was significantly less difficult for users to navigate. Furthermore, the horizontal orientation was easier to navigate for both audio-haptic modalities. Following a more detailed analysis of results from the experiment presented in this report, we intend to develop this study in terms of the audio-haptic cues, and also using other haptic devices.

VI. Acknowledgements

This study was supported by an NSERC-SRO team grant on ENACTIVE Interfaces (P.I.: M. Wanderley) and took place during the eNTERFACE'08 workshop at LIMSI, CNRS, France. We would like to thank Orange Labs for providing the haptic device, all of the Enterface and LIMSI staff and participants that volunteered for the study, Dr. Antoine Gonot for helpful discussion on navigation with auditory cues, and Dr. Ilja Frissen for his help and advice in the analysis of quantitative results.

References

[1] Essl, G. & O'Modhrain, S. 2006. An enactive approach to the design of new tangible musical instruments. Organised Sound 11(3), pp. 285-296.

[2] Jot, J. 1999, "Real-time Spatial processing of sounds for music, multimedia and interactive human-computer interfaces," ACM Multimedia Systems J. 7(1).

[3] Kim, S., Kwon, D. Haptic and Sound Grid for Enhanced Positioning in a 3-D Virtual Environment, Haptic and Audio Interaction Design, Second International Workshop, HAID 2007, Seoul, South Korea, November, pp. 29-30.

[4] Lahav, O. & Mioduser, D., 2004. Blind persons' acquisition of spatial cognitive mapping. Sharkey P., McCrindle R. & Brown D. (Eds) Proceedings of 2004 International Conference on Disability, Virtual Reality and Associated Technologies; September 20-22; Oxford, UK, pp. 131-138.

[5] Marentakis, G. & Brewster, S.A.(2005.. Gesture interaction with spatial audio displays: Effects of target size and inter-target separation. Proceedings of the 2005 International Conference on Auditory Display; July 6-9; Limerick, Ireland, pp. 77-84.

[6] Murphy, E., Kuber, R., Strain, P., McAllister, G. & Yu, W., 2007. Developing Sounds for a Multimodal Interface: Conveying Spatial Information to Visually Impaired Web Users. Proceedings of the 2007 International Conference for Auditory Display; June 26-29; Montreal, Canada.

[7] Magnusson, C., Danielsson, H & Rassmus-Gröhn, K., 2006. Non Visual Haptic Audio Tools for Virtual Environments, Haptic and Audio Interaction Design, First International Workshop, Glasgow, Scotland, 31st August - 1st September.

[8] Mynatt, E. & Weber, G., 1994. Nonvisual presentation of graphical user interfaces: contrasting two approaches. Adelson, B., Dumais, S. T. & Olson J. S. (Eds) Proceedings of the SIGCHI Conference on Human Factors in Computing Systems; April 24-28; Boston, Massachusetts, United States. New York: ACM Press, pp. 166-172.

[9] Oakley, I., McGee, M. R., Brewster, S. & Gray, P., 2000. Putting the feel in look and feel. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '00), pp. 415–422.

[10] Ramstein, C., Martial, O., Dufresne, A., Carignan, M., Chasse, P. & Mabilleau, P., 1996. Touching and hearing GUIs: Design issues for the PC-access system. Proceedings of the 1996 Annual ACM Conference on Assistive Technologies; April 11-12; Vancouver, British Colombia, Canada. New York, ACM Press, pp. 2-9.

[11] Sjöström, C., 2001. Designing haptic computer interfaces for blind people. Proceedings of the 2001 International Symposium on Signal Processing and its Applications; August 13-16; Kuala Lumpur, Malaysia, pp. 68-71.

[12] Wall, S. A. , Paynter, K., Shillito, A. M., Wright, M. & Scali, S., 2002. The effect of haptic feedback and stereo graphics in a 3d target acquisition. In Proceedings of Eurohaptics, pp. 23–29.

[13] Wanderley, M. & Orio, N., 2002. Evaluation of input devices for musical expression: Borrowing tools from HCI. Computer Music Journal, 26(3) pp. 62-76.

[14] Wanderley, M. M. & Depalle. P., 2004. Gestural control of sound synthesis. Proceedings of the IEEE, Johannsen, G. (Ed.) Special Issue on Engineering and Music - Supervisory Control and Auditory Communication, 92(4), pp. 632-644.

[15] Wood, J., Magennis, M., Francisca, E., Arias, C., Graupp, H., Gutierrez, T. & Bergamasco, M., 2003. The design and evaluation of a computer

game for the blind in the GRAB haptic audio virtual environment. Proceeding of EuroHaptics 2003; July 6-9; Dublin, Ireland.

[16] Yu, W. & Brewster, S. (2002). Multimodal virtual reality versus printed medium in visualization for blind people. Proceedings of the 2002 International ACM conference on Assistive technologies; July 8-10; Edinburgh, Scotland. 57-6

Emma Murphy is currently a post-doctoral research fellow at the Multimodal Interaction Lab, McGill University, Montreal. She is an honours graduate of Trinity College, Dublin, where she studied Music and Philosophy. Since graduating from TCD in 2000, her career path has followed from full-time involvement in research and development in industry to academic study. Emma completed an MSc in Music Technology in 2003 at the Centre for Computational Musicology and Computer Music, University of Limerick. In 2007, she completed and successfully defended her doctoral thesis entitled "Designing Auditory Cues for a Multimodal Web Interface: A Semiotic Approach" at the Sonic Arts Research Centre, Queen's University Belfast.

Camille Moussette Born in Montreal, Canada. Hyperactive by nature, Camille Moussette likes the blurry connections between humans, atoms and bits. He holds a Physics degree, a bachelor in Industrial Design and a Masters in Interaction Design. His work experience includes microelectronics R&D (IBM and NRC Canada), web consulting and various involvement in projects ranging from building architectural snow structures in Scandinavia and Italy, to developing new mobile systems in collaboration with Nokia. He is currently pursuing a Ph.D. researching mobile haptic interfaces and teaching classes in Experience Prototyping, Hardware Sketching and Design Ethnography at the Umeå Institute of Design in Sweden.

Chareles Verron Charles Verron was born in Caen, France, in 1980. He received an engineering degree from ENSIETA (Brest, France) in 2003, and a master's degree in acoustics, signal processing and computer sciences applied to music from IRCAM (Paris, France) in 2004. In 2006 he was a research assistant in the Computer and Audio Research Laboratory at Sydney University, Australia. He is currently doing a Ph.D. at Orange Labs, France. His research interests include sound synthesis and sound spatialization.

Catherine Guastavino Dr. Guastavino is a Professor at McGill University (Canada). She holds a Ph.D. in Psychoacoustics (Université Paris 6, France), and received post-doctoral training in the fields of memory and cognition at McGill, before joining the McGill School of Information Studies in 2005. Dr. Guastavino also serves as Associate Director of the Centre for Interdisciplinary Research on Music Media and Technology (CIRMMT), an inter-institutional research center based in Montreal, and is an associate member of the McGill Schulich School of Music. She investigates what information can be best conveyed using sound or touch as opposed to relying on text and graphics.

Capture and machine learning of physiological signals

Benedicte Adessi(1), Rami Ajaj(2), Carole Arrat(3), Ivan Chabanaud(4), Matthieu Courgeon(2), Nicolas Déflache (3), Nesli Erdogmus(4), Fabienne Gotusso(5), Hikmet Gökhan Himmetoğlu(6), Christian Jacquemin(2), Loïc Kessous(2), Jean-Claude Martin(2), Michal Osowski(7), Thomas Pachoud(8), Jana Trojanova(9)

(1) Independent Arttist, Marseilles, (2) LIMSI-CNRS, Orsay, France, (3) ESAV, Toulouse, (4) Independent Arttist, Marseilles, (3) Independent Sound Engineer, Paris, (4)METU - Electrical and Electronics Engineering, Ankara, Turkey, (5) Independent Arttist, Paris, (6)Koç University, Istanbul, Turkey, (7) Independent Arttist, Paris, (7) Independent Arttist, Amsterdam, (8) Independent Arttist, Paris, (9) Department of Cybernetics, University of West Bohemi, Czech Republic

Abstract

The topic of the eNTERFACE project #5 concerned emotion recognition. Its purpose was to design and experiment a set of components that would relate input signals to emotions, and use it as clue to generate correlated sound or graphics.

***Index Terms*—affective computing, physiological signals, emotions**

I. Overview

The topic of the eNTERFACE project #5 concerned emotion recognition. Its purpose was to design and experiment a set of components that would relate input signals to emotions, and use it as clue to generate correlated sound or graphics (Figure 1). Two types of input signals have been considered: signals captured by physiological sensors (heart beat and breath) and live images. As for the analysis, several tools have been used: statistical measures, neural networks, image processing, and clustering. The interfaces between the modules have been a focus of attention for a better connection when the developments are completed. Several output modalities have been considered: live music, live graphics, and particles systems.

Because of the diversity of the participants, we found it useful to have regular presentations so that we could discover the member's scientific interests. 11 presentations have been made in the morning and have covered a wide variety of topics: real-time video processing, live 3D graphics, facial feature extraction, emotions and facial animation, sonification from sensor data, 3D audio through WFS, facial animation, audio signal analysis, and tactile interface.

Due to the importance of the task, the whole processing sequence has not be completed, but several modules have however been realized and could be demonstrated autonomously by simulating the input they were supposed to receive. Since participants had so diverse background such as performance or laboratory, some being students, others professionals, artists or scientists, since some were discovering sensors and others were experienced, we found it useful to have regular presentations so that we could discover the member's interests. The participants have however gained a rich experience from the scientific exchanges through the collaborative design of components, through the regular morning presentations, and through informal talks with participants from the other groups.

The anticipated inputs for the correlated sound and graphics generation were 'breath signal', 'heart beat', 'voice' and 'live images of face'. All of those modules were realized except for the 'voice' due to the lack of a specialized person in the team on this subject. Detailed information on the feature extraction from the physiological signals and face images is given in the next section.

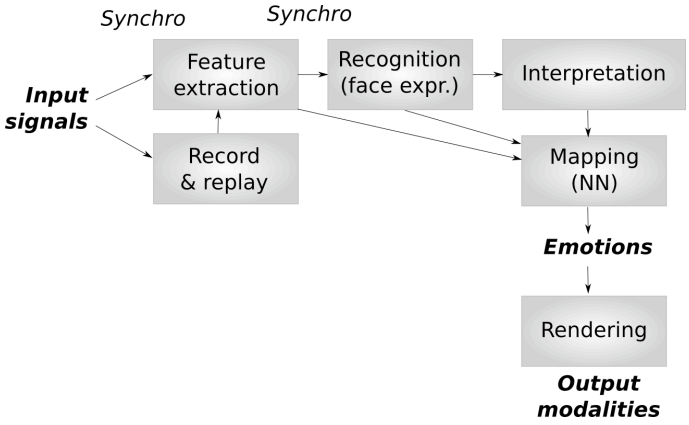


Figure 1 Overview of the system for emotion recognition

II. Feature Extraction

Extraction of features from signals has been done to obtain meaningful parameters for the control of the audio-visual generators.

A. Breath Signal

On the signal of the nasal respirometer, we extracted different statistical features. Then using a real-time algorithm for local maxima extraction we compute both explicit features and statistical features. By hypothesis, as the sensor measure the temperature at the output of the nose, we consider that the expiration phase corresponds to increasing of the sensor value (temperature) and decreasing value to the inspiration phase. Following this principle, we can extract several features that are relative to range and duration of the inspiration, expiration and full respiration cycles. In a second step, we extracted 5 statistical features from each of the previous features. These 5 features are computed on N successive values of each previous

feature. Using a 'window' of N values, we extracted the minimum, the maximum, the mean, the standard deviation, the skewness (measure of the asymmetry of the data around the sample mean) and the kurtosis (measure of how outlier-prone a distribution is) of each feature.

B.Heart Beat

The ECG signal is the manifestation of contractile activity of the heart. It can be used to measure heart rate (HR), inter-beat intervals (IBI) and to determine the heart rate variability (HRV). In our project, we just used the heart rate that was extracted from the ECG signal as a component for emotion recognition.

C.Live Images of Face

Live face image analysis has also been considered as a complementary means for emotion recognition. There are two main approaches for feature extraction from images: holistic methods and geometric methods. We are using a geometric method proposed by Cootes [3]. The method is mostly known under acronym ASM (Active Shape Model). Its basic idea is as follows: ASM is a statistical model which is able to deform to fit to a similar shape as the ones who were presented during training period (see Figure 2).

The main problem in ASM is the localization of the points of the model. There were several enhancement since the ASM was introduced. We would like to implement the LBP (Local binary patterns) to improve the fitting of the model.

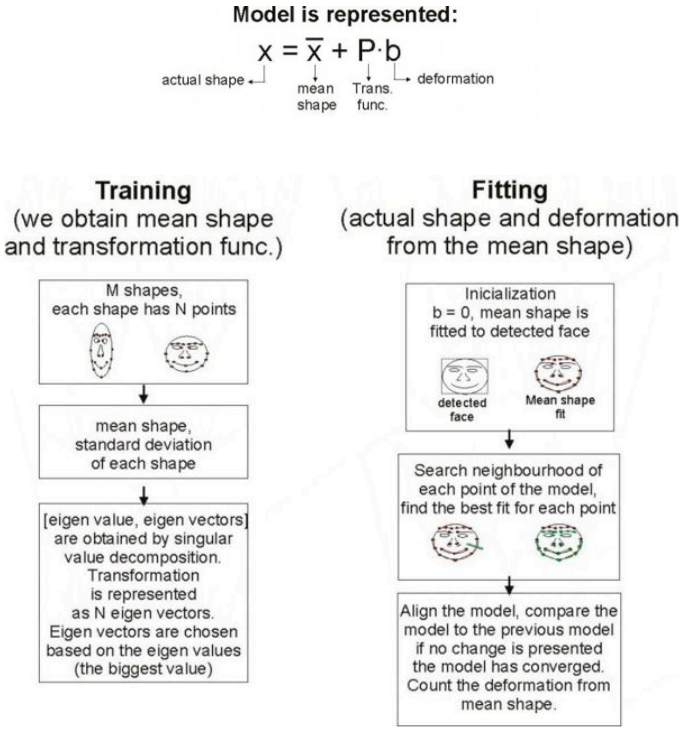


Figure 2 Live face emotion analysis

III.Real-Time Synchronization of the Features

For the synchronization and sensor management, a kind of server will receive the information from the all sensors. This server will able to calibrate and normalize the values in different feature vectors and record them. Afterwards, it resends the normalized and also the raw value, if requested to each feature extraction process. It is better to send as many values as possible, to put more information into emotion recognition unit (see Figure 3).

In order to synchronize the process, a time mark is sent each *t* milliseconds. When this time mark is received within each extracted feature, two actions can be proposed:

- if the process works in real time, only the current response value is sent
- if the process is short enough, first the process is completed then the response value is sent

The server will wait for each process to respond (The mark time is sent back to check the synchronization), then send the data to the neural network in order to do the emotions classification.

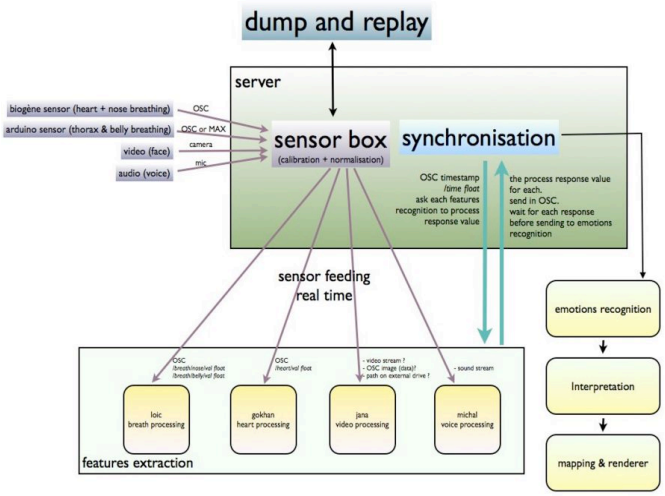
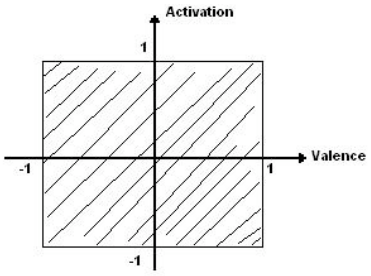


Figure 3 Feature extraction

IV.Classification of the Features into Emotions

After the extraction and the synchronization of the feature vectors from heart beat, breath and the facial images, classification among the emotions is made. It is not like labeling the feature vectors with an emotion but it is more like mapping a feature vector to the following 2 dimensional emotion space:



In order to map our N-dimensional feature vector to the two dimensional emotion space, Stuttgart Neural Network Simulator modified for Java (JavaNNS) is used. JavaNNS is a simulator for artificial neural networks, i.e. computational models inspired by biological neural networks. It enables the user to use predefined networks or create your own, to train and to analyze them.

The designed network that takes physiological features as input and gives the corresponding emotion point (A,V), is first trained with a labeled training set. After the training is completed, the model is exported out as a C++ function that can easily be mounted into the system.

Unfortunately, there are many difficulties in the process. One of the challenging points here is to extract right features that contain the distinguishing data on emotions and to label these feature vectors correctly. In addition to that, the right structure for the network has to be built. Back-propagation method for training is proven to be fast and accurate enough in most cases and thus it can be used in real time processing. For activation and output functions of the input and output nodes and the hidden nodes, many comprehensive tests and trials have to be conducted with the real data.

V. Interpretation of Sensor Data

Emotion is a complex phenomenon that no conceptual theory describes at a sufficient level at the moment. It involves percepts as well as cognition (concepts), conation (impulse) & affects (feelings, physiological systems). Everyday language provides a manageable simplification of concepts & experiences.

Emotional life is a large and complex space that we can't represent completely, because of technical, conceptual and experimental limitations.

- Technical: we need to be aware of what the multimodal inputs we have are capable of describing. We realize we have few parameters and they are superficial.
- Conceptual: are we interested in basic emotions that statistically occur more often (scientific approach), or are we going to fit subjective needs (artistic approach) ?
- Experimental: there is no obvious link, and we don't have knowledge to link measurements to the inner emotion of the human being sensed.

A subject found in a measured state equivalent to a state in which he was before does not mean that he is going through the same particular emotion. Inversely, being in a particular emotional state could result in a comparable measure, even though it is not necessarily the case. Anyway, it would require one step more of analysis to link the representation to an inner emotion.

So, what is represented and what is not? Are we recognizing emotions, are we recognizing concepts, or are we recognizing face expressions + sensor features?

With an artistic approach, it is possible to rely on the feedback of the multimedia result on the performer's perception to create a link between capitation and actual inner emotion. If expressivity is placed at a sufficient level to allow this loop to exist, the performer addresses directly to the multimedia content that includes things we wouldn't normally accede, because the performer isn't even conscious about. Because it is explicit, this is the only space where we know where we are.

Thus, we need a system of representation with emphasis on experiential description.

Intuitively, we proposed to use the Activity / Valence representation, which we assume to be a broad enough concept to fit the view of several artists, or at least to be of interest for them. This representation is not describing the context (environment, cognition, affect...) but taking it into account. Emotional life is divided into 4 poles, 4 poetic worlds that the artists describe with general sentences. An example of a poetic world mapped to sound generators is given in the following figure 4.

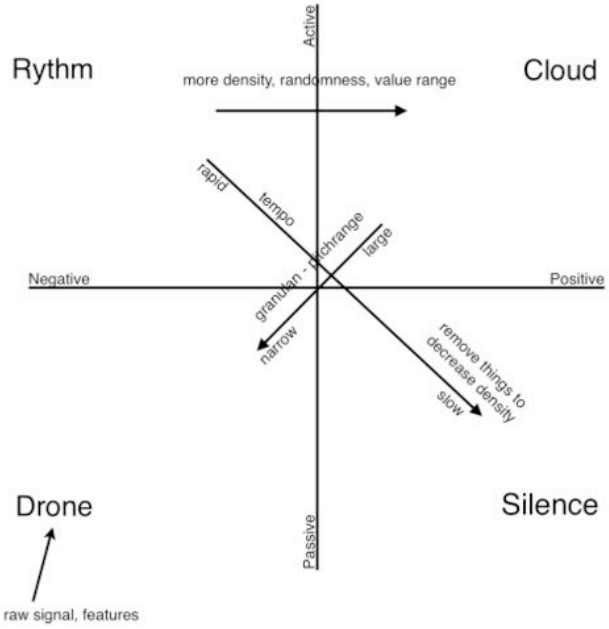


Figure 4 Poles of emotional life

VI. Artistic work

We now present a few artistic applications that we made during the course of the workshop.

A.Sound Generation

Breathing and Thoracic Volume: For breathing we used the sensor signal to modulate a filtered noise. We used a white noise and a band-pass filter with a moderate coefficient of resonance. The signal from the breath temperature sensor was connected to both the amplitude of the output signal and the center frequency of the filter. We apply a flanging effect at the output of the filter in order to give coloration to the resulting sound. For the thoracic volume we used a similar method but with a voiced source as signal input to feed the filter.

A toolbox for Heart beat sonification: We designed a toolbox for Heartbeat sonification using two different methods. These methods have different levels of flexibility, in particular, concerning spectral transformation. The first system uses complex resonances. Using parallel resonance filter bank based on the CNMAT tools, we made a module which allows us to synthesize an arbitrary resonance model and to control and modify several high-level parameters. The sonic interaction designer can control and modify the number of resonances and the lower resonance frequency. This model can be modified by changing other parameters at a high-level control as the global resonance and the spectral amplitude slope. A second module was build to use sampling techniques. This module use a polyphonic player that allows playing recorded sounds by triggering it and avoids clicks due to a re-triggering when it's already playing a sound, by allocating a new playing voice to it. Modifiable parameters are speed (pitch modified), time of beginning for reading into the sound file and parameters of a linear amplitude envelope defined by fade in time and fade out time.

B.Live performances based on input signals

The point is to create a sound-space worked by the variation of the emotions felt during the act of performing. Michal Osowski wishes to build an « emotional curve » from analyzing the video of the performance and then to program automatic sounds on the computer that happen in relation to the variations worked by the curve.

The performance is shot by the camera of Ivan Chabanaud. Fabienne Gotusso creates the performance choosing as a point of departure to the improvisation to play with the sound of the German word « Holzwege » (meaning forest path) and the English word « to die ». The aim is to create some estrangement from the meaning of both words and then to liberate the games and associations of meaning produced by the pronunciation of the words relating to improvised acts. She chooses to improvise in a sandy path with a tree standing by. Then Michal and Fabienne manage to extract some range of emotions by watching the video of the performance to finally order them as a map.

Fabienne dresses in a black suit on which several captors are fixed. She begins to move after several necessary preliminary measures in order to identify her skeleton. Skeleton that will appear at the time on the computer screen as she is acting the performance. The skeleton reproduces in real time the motions composed by Fabienne. The inscription and architecture of the motions are plainly faithful to the quality involved in the dancing. It is impossible not to think about E J Marey's works.

VII. Conclusion

Even though this workshop has not allowed the participants to design a full working system from signal capture to emotion recognition and reactive digital art, it has however had numerous positive achievements:

- through regular seminars, participants have discovered new methods and new techniques in signal analysis, machine learning, computer graphics and sound, live performance, and affective computing,
- through joint developments, the researchers have gained a better insight on the requirements of physiological signal analysis and interpretation,
- through human experience and joint collaboration, the participants have discovered shared scientific interests and defined new perspectives for future collaborative works.

Acknowledgement:

Part of this research was supported by the Grant Agency of Academy of Sciences of the Czech Republic, project No. 1ET101470416 .

References

[1] T. Jehan, A. Freed, and R. Dudas. "Musical applications of new lter extensions to Max/MSP", Proceedings of the International Conference on Computer Music, pages 504-507, 1999.

[2] R Cowie, "Emotional life: Terminological and conceptual clarifications", HUMAINE - D3i, 2006 http://emotion-research.net/projects/humaine/deliverables/D3i%20final.pdf

[3] Cootes, T. F., Edwards, G. J., and Taylor, C. J. 2001. Active Appearance Models. IEEE Trans. Pattern Anal. Mach. Intell. 23, 6 (Jun. 2001), 681-685. DOI= http://dx.doi.org/10.1109/34.927467

Multimodal Feedback from Robots and Agents in a Storytelling Experiment

S. Al Moubayed (#), M. Baklouti (^), M. Chetouani (*), T. Dutoit (+), A. Mahdhaoui (*), J.-C. Martin (~), S. Ondas (@), C. Pelachaud (°), J. Urbain (+), M. Yilmas (&)

(#)Center for Speech Technology, Royal Institute of Technology, KTH, SWEDEN, (^) Thalès, FRANCE, (*) University of Paris VI – FRANCE, (+) *Faculté Polytechnique de Mons – BELGIUM, (~) LIMSI – FRANCE,* (@) Technical University of Košice – *SLOVAKIA,(°) INRIA – FRANCE, (&) Koc University – TURKEY*

***Abstract* — In this project, which lies at the intersection between Human-Robot Interaction (HRI) and Human-Computer Interaction (HCI), we have examined the design of an open-source, real-time software platform for controlling the feedback provided by an AIBO robot and/or by the GRETA Embodied Conversational Agent, when listening to a story told by a human narrator. Based on ground truth data obtained from the recording and annotation of an audio-visual storytelling database, and containing various examples of human-human storytelling, we have implemented a proof-of-concept ECA/Robot listening system. As a narrator input, our system uses face and head movement analysis, as well as speech analysis and speech recognition; it then triggers listening behaviors from the listener, using probabilistic rules based on the co-occurrence of the same input and output behaviors in the database. We have finally assessed our system in terms of the homogeneity of the database annotation, as well as regarding the perceived quality of the feedback provided by the ECA/robot.**

I. Introduction

THIS project lies at the intersection between Human-Computer Interaction (HCI) and Human-Robot Interaction (HRI).

Human-Robot Interaction (HRI) is a multi-disciplinary field involving research on robot control (planning, sensor…), speech processing (recognition, synthesis), vision (human localization, environment characterization), artificial intelligence, cognitive science and other fields [1]. Various robots are now available for such studies, and are provided with specific programming tools. In this project, we have focused of the Sony AIBO dog, and the URBI (Universal Real-time Behaviours Interface) language [2].

Human-Computer Interaction is restricted here to Embodied Conversational Agents (ECAs). The term ECA has been coined in Cassell et al. [3] and refers to human-like virtual characters that typically engage in face-to-face communication with the human user. In this project, we have used GRETA [4], an ECA , whose interface obeys the SAIBA (Situation, Agent, Intention, Behavior, Animation) architecture [5].

Several methods have been proposed for the improvement of the interaction between humans and agents or robots. The key idea of their design is to develop agents/robots with various capabilities: establish/maintain interaction, show /perceive emotions, dialog, display communicative gesture and gaze, exhibit distinctive personality or learn/develop social capabilities [6, 7, 8]. Theses social agents or robots aim at naturally interacting with humans by the exploitation of these capabilities.

We have investigated one aspect of this social interaction: the engagement in the conversation [9]. The engagement process makes it possible to regulate the interaction between the human and the agent or the robot. This process is obviously multi-modal (verbal and non-verbal) and requires an involvement of both the partners. Some mechanisms as motivation, curiosity can be useful for this purpose [8].

Our project more specifically aims at exploring multimodal interaction between a human speaker telling a story (typically a cartoon) to (i) an ECA or (ii) an AIBO robot. More particularly, we focused on the design of an open-source, real-time software platform for designing the feedbacks provided by the robot and the humanoid during the interaction. The multimodal feedback signals we consider here are limited to facial and neck movements by the agent, while the AIBO robot uses all possible body movements, given its poor facial expressivity. We did not pay attention to arms or body gestures.

This paper is organized as follows. In Section II, we formalize SAIBA, a common architecture for embodied agents, and introduce its application to feedback modeling. This leads us, in Section III, to exposing the contents of the eNTERFACE08_STEAD database developed for this project and containing various annotated examples of

This report, as well as the accompanying database and the source code for the software developed during the project, are available online from the eNTERFACE'08 web site: www.enterface.net/enterface08.

human-human storytelling. This database was used for designing several feedback components in subsequent Sections. Sections IV and V respectively focus on the speech and video analysis modules we have used. This is followed in Section VI by details on possible control of the agent state via ASR. We then give, in Section VII, a description of the feedback rules we have established for triggering feedback from our software and hardware rendering engines, namely AIBO and GRETA, and show the behaviors we have been able to synthesize with them. Finally, Section VIII gives details on the compared performances of our HCI and HRI systems.

II. SAIBA AND FEEDBACK

Although we are all perfectly able to provide natural feedback to a speaker telling us a story, explaining how and when you do it is a complex problem. ECAs are increasingly used in this context, to study and model human-human communication as well as for performing specific automatic communication tasks with humans).

Examples are REA [10], an early system that realizes the full action-reaction cycle of communication by interpreting multimodal user input and generating multimodal agent behaviour, the pedagogical agent Steve [11] which functions as a tutor in training situations, MAX [12] a virtual character geared towards simulating multimodal behaviour, Gandalf [13] provides real-time feedback to a human user based on acoustical and visual analysis. Carmen [14] a system that supports humans in emotionally critical situations such as advising parents of infant cancer patients. Other systems realize presentation agents [15], i.e. one or more virtual agents present some information to the user. They can adopt several roles, such as being a teacher [15; 17], a museum guide [18, 19, 12] or a companion [20, 21]. In robotics, various models have been proposed for the integration of feedbacks during interaction [7]. Recently, the importance of feedbacks for discourse adaptation has been highlighted during an interaction with BIRON [22].

In a conversation, all interactants are active. Listeners provide information to the speaker on how they view the conversation goes on. By sending acoustic or visual feedback signals, listeners show if they are paying attention, understanding or agreeing with what is being said. Taxonomies of feedbacks, based on the meaning these signals convey, have been proposed [23, 24]. The key idea of this project is to automatically detect the communicative signals in order to produce a feedback. Contrary to the approach proposed in [LOH08], we focus on non-linguistic features (prosody, prominence) but also on head features (activity, shake, nod).

Our system is based on the architecture proposed by [4], but progressively adapted to the context of a storytelling (figure 1). We developed several modules for the detection and the fusion of the communicative signals from both audio and video analysis. If these communicative signals match our pre-defined rules, a feedback is triggered by the Realtime BackChannelling module, resulting in two different behaviors conveying the same intention.

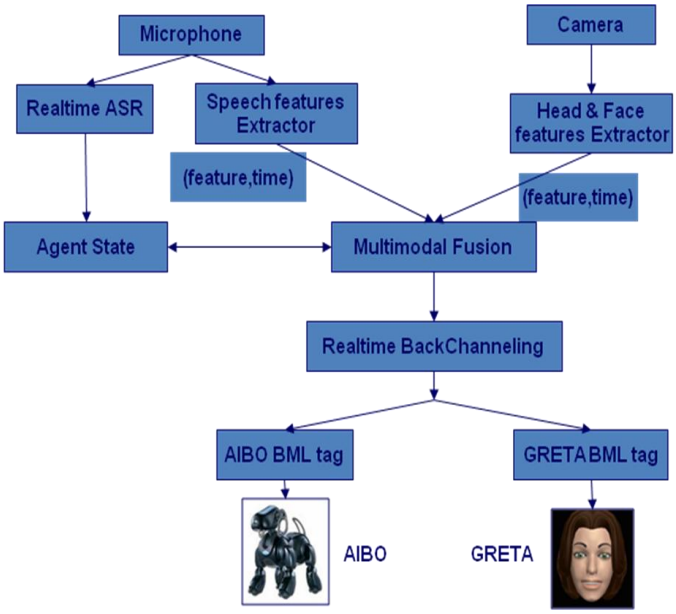


Fig. 1. Architecture of our interaction feedback model

III. THE ENTERFACE08_STEAD DATABASE

In order to model the interaction between the speaker and the listener during a storytelling experiment, we first recorded and annotated a database of human-human interaction: the eNTERFACE08_STEAD database. This database was used for extracting feedback rules (section VII, but also for testing the multi-modal feature extraction system (section VIII).

We followed the McNeill lab framework [25]: one participant (the speaker), who has previously observed an animated cartoon (Sylvester and Tweety), tells the story to a listener immediately after viewing it. The narration is accompanied by spontaneous communicative signals (filled pauses, gestures, facial expressions, etc.). In contract, instructions are given to the listener to express his/her engagement in the story by giving non-verbal audio-visual gestures in response to the story told by the speaker.

eNTERFACE_STEAD Contents

Twenty-two storytelling sessions telling the "Tweety and Sylvester - Canary row" cartoon story were recorded.

Thirteen recording sessions were done by a French listener and a French speaker. The last two recordings have exaggerated non-verbal activity (closer to acting than to real-life storytelling).

Four recording sessions were done by an Arabic listener and an Arabic speaker.

Five recording sessions were done by a speaker and a listener who do not speak or understand each other's languages; these recordings can be used to study the isolated effect of prosody on the engagement in a storytelling context. The languages used in these sessions were Arabic, Slovak, Turkish, and French.

Annotation Schema

A part of the MUMIN [26] multimodal coding scheme was used for annotating the database. MUMIN was originally created to experiment with annotation of multimodal communication in video clips of interviews taken from Swedish, Finnish and Danish television broadcasting and in short clips from movies. However, the coding scheme is also intended to be a general instrument for the study of gestures and facial displays in interpersonal communication, in particular the role played by multimodal expressions for feedback, turn management and sequencing.

The videos were annotated (with at least two annotators per session) for describing simple communicative signals of both speaker and listener: smile, head nod, head, shake, eye brow and acoustic prominence. These annotations were done using the ANVIL [27] annotation program (Fig. 2) and are summarized in Table 1.

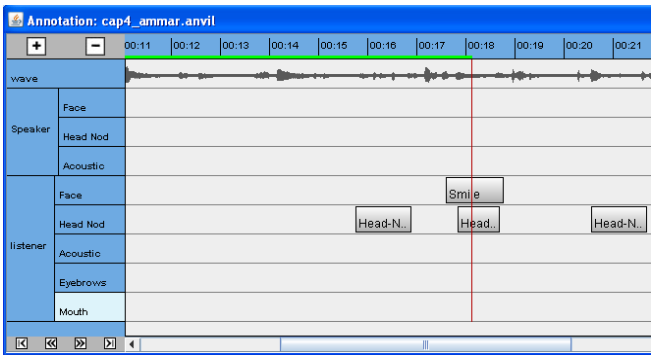


Fig. 2 Anvil, as used for our database annotation

Facial display feature	**Form of expression/Movement values**	
Facial display feature	**Value**	**Short tag**
General face	smile	smile
Mouth(opening)	Open mouth	Open-M
	Closed mouth	Close-M
Head	Nod	Head-Nod
	Shake	Head-shake
Eyebrows	Frowning	Frown
	Raising	Raise
Acoustic	Prominence	Prominence
	laughter	laughter

Table 1: Coding scheme used for eNTERFACE_STEAD annotations.

Annotators had instructions to annotate prominence in the audio recording of the speaker by only listening to the audio signal without looking to the video recording; some annotators spoke French, others did not.

Annotation evaluation

Manual annotations of videos were evaluated by computing agreements using corrected kappa [28] computed in the Anvil tool [27], shown in Fig. 2:

$$kappa = (P0 - 1/z) / (1 - 1/z)$$

where z is the number of categories and $P0$ is like in Cohen's kappa.

Table 2 presents the agreements among annotators for each track. We can see that the best agreement is obtained for the Listener.Acoustic track which is expected since the listener is not assumed to speak and when he/she does simple sounds are produced (filled pauses). Other tracks have a lower agreement such as Speaker.Acoustic. The speaker always speaks during the session and prominent events are less identifiable. However, the agreements measures are high enough to allow us to assume that selected communicative signals might be reliably detected.

Track Name	**Cohen's Kappa**	**Corrected Kappa**	**Agreement(%)**
Speaker.Face	0.473	0.786	89.306
Speaker.Acoustic	0.099	0.786	84.500
Listener.Face	0.436	0.559	77.960
Listener.HeadNod	0.464	0.694	84.622
Listener.Acoustic	0.408	0.929	95.972

Table 2 Agreement between annotators of our database (eNTERFACE_STEAD)

The eNTERFACE_STEAD License

The eNTERFACE_STEAD contents, and all the annotations are released under an MIT-like free software license and is available from the eNTERFACE'08 website (www.enterface.net/enterface08).

IV. Speech Analysis

The main goal of the speech analysis component is to extract features from the speech signal that have been previously identified as key moments for triggering feedbacks (cf. section VII). In this study, we do not use any linguistic information to analyze the meaning of the utterances being told by the speaker, but we focus on the prosodic cross-language features which may participate in the generation of the feedback by the listener.

Previous studies have shown that pitch movements, especially at the end of the utterances, play an important

role in turn taking and backchannelling during human dialogue [29]. In this work, we propose in this work to use the following features extracted from the speaker's speech signal: Utterance beginning, Utterance end, Raising pitch, Falling pitch, Connection pitch, and Pitch prominence.

To extract these important features from the speech stream, we decided to work in Pure Data [30], a graphical programming environment for real-time audio processing. The patch we developed for speech feature extraction is shown in figure 3; it provides the following features: Utterance beginning, Utterance end, Raising pitch, Falling pitch, Stable pitch, and Acoustic prominence.

Fig. 3. Pure Data patch for Speech Feature Extraction

Audio acquisition is performed by the `adc~` object. It provides 64-samples blocks (at a sampling frequency set to 16kHz) to objects `vad_framing~` and `pitch_framing~`, which are responsible for voice activity detection and pitch estimation for prominence estimation. These algorithms are written in C, using audio processing functions from a C library developed by the Center of Speech Technology at KTH, Sweden. They were compiled as externals for Pure Data, so that they can be used as ordinary Pure Data objects.

Since we wanted to compute some features on overlapping audio segments longer than 64 samples, we developed a specific framing routine. The objects `vad_framing~` and `pitch_framing~` take two arguments *X* and *Y*, which impose blocks of *X* samples with a shift of *Y* samples between successive blocks: a buffer of *X* samples is filled with the input blocks of 64 samples; when this buffer is full, it is sent to the analysis algorithm, the samples in the buffer are then shifted by *Y* samples, and the buffer is filled again, etc.

For every input speech frame, `vad_framing~` and `pitch_framing~` output an integer, which is set to 0 if there is no event (feature) is detected in the speech, and to a number (index) indicating the id of the detected feature otherwise. These indices are sent to the "route" PureData object, which triggers a string depending on its input; this string is later sent through a tcp/ip connection to the Multimodal Fusion module.

The vad_framing~ object is a Voice Activity Detection object, which contains an adaptation of the SPHINX Vader functionality [31]. This object sends "1" if there is a detected change in the audio stream from Silence to Speech, and "2" when there is a detected change from Speech to Silence, otherwise the output of this object is always "0".

The pitch_framing~ object is used to extract the rest of the speech features. This object contains an implementation of the realtime fundamental frequency tracking algorithm YIN [32]. For cleaning the output of the YIN algorithm, a median filter of size "5" (60 msec) is applied on the extracted F0 to compensate for outliers and octave jumps.
This object sends "1" when a Raising Pitch is detected, "2" for Falling pitch, "3" for Stable Pitch, and "4" for Acoustic prominence.

The TILT model:
The TILT model [33] is used to extract Raising pitch, Falling pitch, Connection pitch. We tried in this implementation to not compensate for the unvoiced segments by using any type of interpolation; nevertheless, the movements of the pitch are detected only at the end of the voiced segments, and no movements are detected when the voiced segment duration is shorter than 125 msec.

Audio prominence estimation
In the literature, several definitions of acoustical prominent events can be found showing the diversity of this notion [34, 35]. Terken [35] defines prominence as words or syllables that are perceived as standing out from their environment. Most of the proposed definitions are based on linguistic and/or phonetic units.

We propose in this project another approach using statistical models for the detection of prominence. The key idea is to assume that a prominent sound stands out from the previous message. For instance, during our storytelling experiment, speakers emphasize words or syllables when they want to focus the attention of the listener on important information. These emphasized segments are assumed to stand out from the other ones, which makes them become salient.

Prominence detectors are usually based acoustic parameters (fundamental frequency, energy, duration, spectral intensity) and machine learning techniques (Gaussian Mixture Models, Conditional Random Fields) [36, 37]. Unsupervised methods have been also investigated such as the use of Kullback-Leibler (KL) divergence as a measure of discrimination between prominent and non-prominent classes [38]. These statistical methods provide an unsupervised framework adapted to our task. The KL divergence needs the estimation of two covariance matrices (Gaussian assumption):

$$KL_{ij} = \frac{1}{2}\left[\log \frac{\Sigma_j}{\Sigma_i} + tr\left(\Sigma_i \Sigma_j^{-1}\right) + \left(\mu_i - \mu_j\right)^T \Sigma_j^{-1} \left(\mu_i - \mu_j\right) - d \right]$$

where μ_i, μ_j and Σ_i, Σ_j denote the means and the covariance matrices of *i*-th (past) and *j*-th (new event) speech segments respectively. *d* is the dimension of the speech feature vector. An event *j* is defined as prominent if the distance from the past segments (represented by the segment *i*) is larger than a pre-defined threshold.

One major drawback of the KL divergence approach is that since the new event is usually shorter than the past events, the estimation of their covariance matrices is less reliable. In addition, it is well-known that duration is an important perceptual effect for the discrimination between sounds. Taking these points into account, we propose to use another statistical test namely the T^2 Hotteling distance defined by:

$$H_{ij} = \frac{L_i L_j}{L_i + L_j}\left[\left(\mu_i - \mu_j\right)^T \Sigma_{i \cup j}^{-1} \left(\mu_i - \mu_j\right) \right]$$

where $i \cup j$ is the union of *i*-th (past) and *j*-th (new event) segments. L_i and L_j denote the length of the segments. The T^2 Hotteling divergence is closely related to the Mahalanobis distance.

In this work only the fundamental frequency (*F0*) is used as a feature to calculate the Hotteling distance between two successive voiced segments. In this sense, a prominence is detected when the Hotelling distance between the current and the preceding Gaussian distributions of *F0* is higher than a threshold. We have used a decaying distance threshold over time, where the initial value of this threshold is the highest distance during the first utterance of the speaker; whenever this threshold is reached by a following segment, a Pitch Prominence event is triggered, and the new distance becomes the distance threshold. Since we estimate a Gaussian distribution of the pitch for a voiced segment, we only estimate it when there are enough pitch samples during the voiced segment, (we set this duration threshold to 175 msec).

V. Face Analysis

The main goal of the face analysis component (Fig. 4) is to provide the feedback system with some knowledge of communicative signals conveyed by the head of the speaker. More specifically, detecting if the speaker is shaking the head, smiling or showing neutral expression are the main activity features we are interested in. The components of this module (Fig. 5) are responsible for face detection, head shake and nod detection, mouth extraction, and head activity analysis. They are detailed below.

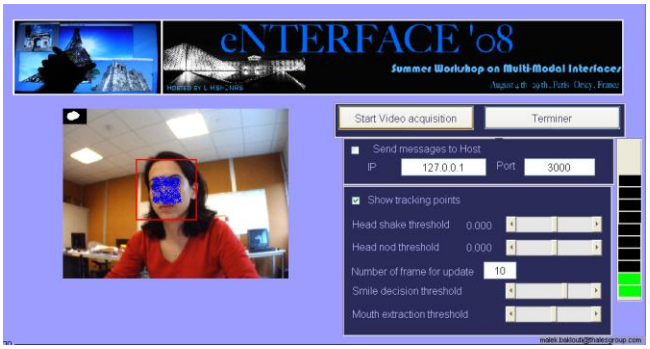


Fig. 4. Screenshot of the face analysis component, which runs on a PC and shows the analysis results through a MATLAB-based user interface It sends its results to the multimodal fusion module via TCP-IP.

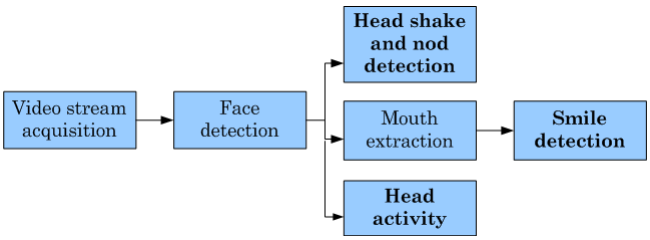


Fig. 5. Overview of the face analysis module

Face detection

The face detection algorithm that we used exploits Haar-like features that have been initially proposed by Viola & Jones [39]. It is based on a cascade of boosted classifiers working with Haar-like features and trained with a few hundreds of sample views of faces. We used the trained classifier available in OpenCV.

The face fetection module outputs the coordinates of existing faces in the incoming images.

Smile detection

Smile detection is performed in two steps: mouth extraction followed by smile detection. We use a colorimetric approach for mouth extraction. A thresholding technique is used after a colour space conversion to the YIQ space (Fig. 6).

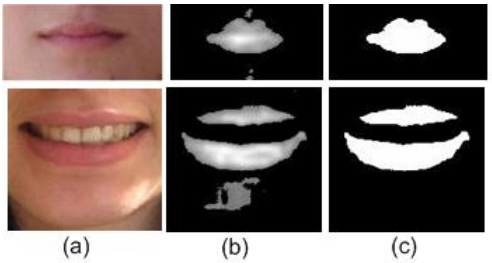


Fig. 6. Mouth extraction. (a) original images (b) color conversion and thresholding (c) elimination of small regions

Once the mouth is extracted, we examine the ratio between the two characteristic mouth dimensions, P1P3 and P2P4 (see Fig. 7), for smile detection. We assume that when smiling, this ratio increases. The decision is obtained by thresholding.

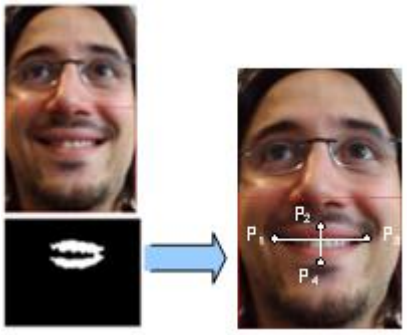


Fig. 7. smile detection

Head shake and nod detection

The purpose of this component is to detect if the person is shaking his/her head or doing a nod. The idea is to analyze the motion of some feature points extracted from the face along the vertical and horizontal axes.

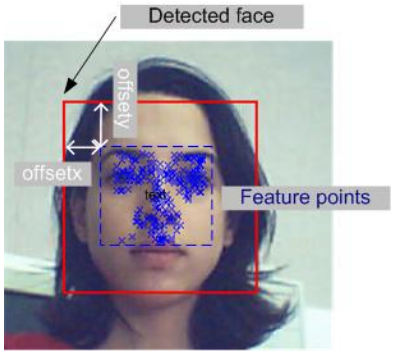


Fig. 8. Feature points extraction

Once the face has been detected in the image, we extract 100 feature points using a combined corner and edge detector defined by Harris [40]. Feature points are extracted in the central area of the face rectangle using offsets (Fig 8).

These points are then tracked by calculating the optical flow between a set of corresponding points in two successive frames. We make use of the Lucas Kanade [41] algorithm implementation available from in the OpenCV library (http://sourceforge.net/ projects/ opencvlibrary/).

Let *n* be the number of feature points and $Pt_i(x_i, y_i)$ the *i*th feature point defined by its 2D screen coordinates (x_i, y_i). We then define then the overall velocity of the head as:

$$V = \begin{cases} V_x = \frac{1}{n}\sum_{i=1}^{n}(x_i - x_{i-1}) \\ V_y = \frac{1}{n}\sum_{i=1}^{n}(y_i - y_{i-1}) \end{cases}$$

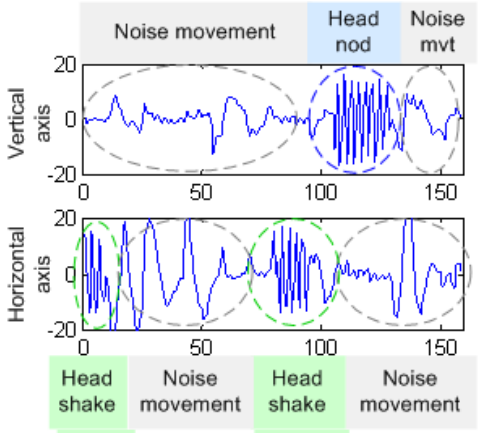


Fig. 9. Feature point velocity analysis

Fig. 9 shows the velocity curves along the vertical and horizontal axes. The sequence of movements represented is composed by one nod and two head shakes. We notice that the velocity curves are the sum of two signals: (1) a noise movement which is a low frequency signal representing the global head motion and (2) a high frequency signal representing the head nods and head shakes.

The idea is then to use wavelet decomposition to remove the low frequency signals. More precisely, we decomposed the signal using symlet-6 wavelet. Fig. 10 shows the reconstruction of the detail at the first level of the signal shown in Fig. 8. The head nod and shake events can be reliably identified by this process.

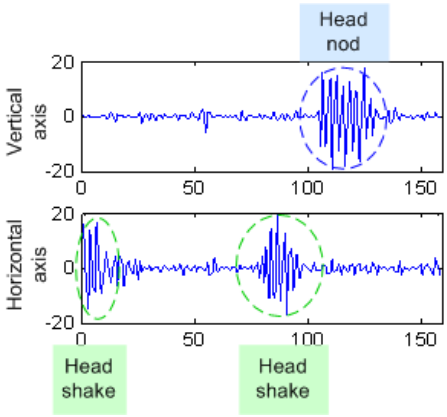


Fig. 10 Signal denoising via wavelets.

Head activity analysis

Analysis of recordings of the storytelling experience has shown a correlation between the head activity of both speaker and listener. To characterize the head activity, we use the velocity of the feature points defined in (1), to quantify the overall activity A:

$$A = \sum_{t \in timewindow} V_{x,t}^{\ 2} + V_{y,t}^{\ 2} \qquad (2)$$

where *timewindow* is set to 60 frames (30 frames/s).

This measure provides information about the head activity levels. In order to quantize head activity into levels (high, medium or low), we analyzed the head activity of all the speakers of the eNTERFACE08_STEAD corpus. Assuming that the activity of one given speaker is Gaussian, we set up various thresholds defined in Table 3. By using these thresholds, the algorithm becomes more sensitive to any head movement of a stationary speaker, while it raises the thresholds for an active speaker, thus resulting in a flexible adaptive modeling.

Amplitude	**Interpretation**
< mean	LOW ACTIVITY
< mean + standard deviation	MEDIUM ACTIVITY
Otherwise	HIGH ACTIVITY

Table 3. Segments are categorized according to the amplitude of their maxima. Mean and standard deviation statistics are related to head activity.

VI. AGENT/ROBOT STATE CONTROL

In the feedback model proposed in [4], the state of the agent/robot is characterized by the following features: *disagreement, agreement, acceptance, refusal, belief, disbelief, liking, disliking, interest, no_interest, understanding, no_understanding,* and *mimicry*. For this project, we have reduced this set to: *interest, understanding,* and *liking*. Our goal was then the design means of modifying these features through some analysis of the audio and video streams from the speaker.

To achieve this goal, we have used an English speaking ASR system based on keyword spotting, which makes it possible to modify the agent state according to the recognized words. The ASR system is thus integrated into an *Agent State Manager* (ASM) module (Fig 11), which consists in three main parts: the ASR engine, the State Planner and the Message Generator. These components are detailed in the next paragraphs.

ASR engine

The speech engine we have used is based on ATK/HTK [42] and is available as a dynamic-link library (dll), whith a simple API.

It uses freely available British English triphone acoustic models, which are part of the ATK distribution, and were trained on the WSJCAM0 speech corpus [43] recorded at the Cambridge University and composed of readings of the Wall Street Journal.

As a language model we have used a speech grammar which enables the recognition of keywords in phrases. Other words are modeled as "filler words", and not recognized. The keyword spotting grammar actually puts all keywords in parallel (with no specific syntactic constraints), together with a filler model. It is written in BNF (Backus Naur Form) format and then it is translated to HTK-compatible SLF format.

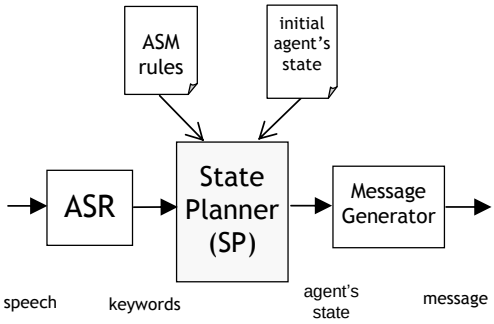


Fig. 11. Architecture of the Agent State Manager

For the purpose of storytelling we needed to define appropriate keywords. The best way to obtain this information was to look at our storytelling recordings, available in the from eNTERFACE08_STEAD database (see Section III). We used two recordings by native English speakers, as well as recordings in Slovak, and transcribed them into English words. The transcriptions were then analyzed in terms of word counts via a simple Python script. After eliminating the articles (*the, a*), conjunctions (*and*) and pronouns (*she, he, it*), which were naturally the most frequent words, we identified a group of keywords for our storytelling experiment, containing: *hotel, reception, door, room, stairs, luggage, baggage, bags, birdcage, umbrella, clerk, tomcat, cat, silvester, woman, lady, tweety, bird, knock, carry, call, run, trick, discover, hit, hide, ring, uncover, pick up, cartoon, story, treatment, outside*.

Notice that the ASR engine also uses a simple pronunciation dictionary to specify the expected pronunciation of keywords. We modified the default pronunciation of words in the dictionary by removing the expected short pauses (sp) after each word. (As a matter of fact, we use continuous speech in this project, in which there are no short pauses between words.)

State Planner

The State Planner is the main part of the Agent state Manager. Its task is to modify the state of the agent according to spoken input (keywords). For this purpose it uses a rule-based approach. This component is initialized with the initial agent state as well as with a set of rules loaded from an initialization file. It takes spoken keywords as input, and changes the value of the *interest, understanding,* and *liking* as output.

Each rule consists of the following fileds: *feature*, *keyword*, *step*, *max_value* and *opposite_feature* (which is optional). When the speech recognizer recognizes a *keyword*, State Planner looks for an appropriate rule. If it finds it, it increases the value of the related *feature* by a given *step* value, while *max_value* is not reached. If some *opposite_feature* is defined in the rule, it is decreased by the same step.

Message Generator

Every three words which have triggered the application of rules, the Message Generator prepares an XML file containing the feature values, which represent the new agent

state. After converting special characters into hexadecimal, it sends the XML file through a TCP socket to the ECA (Greta).

VII. Multimodal fusion and Feedback Behaviors for AIBO and GRETA

Extracting rules from data

Based on the selected communicative signals, we have defined some rules to trigger feedbacks. The rules are based on [44, 45], which involved mainly only mono-modal signals. The structure of such rules is as follows:

"If some *signal* (eg. head-nod, pause, pitch accent) is received, then the listener sends some *feedback signal* with probability X."

We have extended these rules by analyzing the data annotated from our eNTERFACE08_STEAD storytelling database. We looked at the correlation of occurrence between each speaker mono-modal and multi-modal signal and each listener feedbacks (where we understand *multi-modal signal* as any set of overlapping signals that are emitted by the speaker within a time window, defined as the time interval of any speaker signal plus 2 seconds). This gave us a correlation matrix between speaker and listener signals, whose elements give, for each speaker signal, the probability that the listener would send a given feedback signal. In our system we use this matrix to select listener's feedback signals. When a speaker's signal is detected, we choose from the correlation matrix, the signal (ie feedback) with the higher probability.

From this process, we identified a set of rules (which can be found in the repository of the project) such as:

- Mono-modal signal ⇒ mono-modal feedback: *head_nod* is received, then the listener sends *head_nod_medium*.
- Mono-modal signal ⇒ multi-modal feedback: *smile* is received, then the listener sends *head_nod and smile.*
- Multi-modal signal ⇒ mono-modal feedback: *head_activity_high* and *pitch_prominence* are received, then the listener sends *head_nod_fast.*
- Multi-modal signal ⇒ multi-modal feedback: *pitch_prominence* and *smile* are received, then the listener sends *head_nod* and *smile*.

Rules can be made probabilistic via associated probabilities: in case there is more than one rule with the same input, every rule will have a probability of execution.

Multi-modal fusion

The multi-modal fusion module is responsible for activating the rules mentioned above, when input signals are detected, and will eventually trigger feedbacks from the agent/robot.

For realtime consideration, the rule contains a response time variable, which defines when the output of the rule should be executed after the reception of the last input signal; the last variable is rule duration, rule duration defines how long this rule can be active, so in case not all the input signals are received, the rule will be deactivated after this specified period.

Reactive behaviors

In our architecture, we aim to drive different types of virtual and/or physical agents: the GRETA ECA (Fig 12), and The AIBO ERS-7 Robot (Fig 13). To ensure high flexibility we are using the same control language to drive all the agents, the Behavior Markup Language BML [5]. BML encodes multimodal behaviors independently from the animation parameters of the agents.

Through a mapping we transform BML tags into MPEG-4 parameters for the GRETA agent and into mechanical movements for the AIBO robot. Various feedbacks are already available for GRETA such as acceptance (*head_nod*), non-acceptance (*head_shake*) or smile. Concerning AIBO, we developed similar feedbacks conveying the same meaning but in a different way. To develop the reactive behavior of AIBO, we used the URBI (Real-Time Behavior Interface) library [2] allowing a high-level control of the robot.

Fig. 12 The GRETA ECA

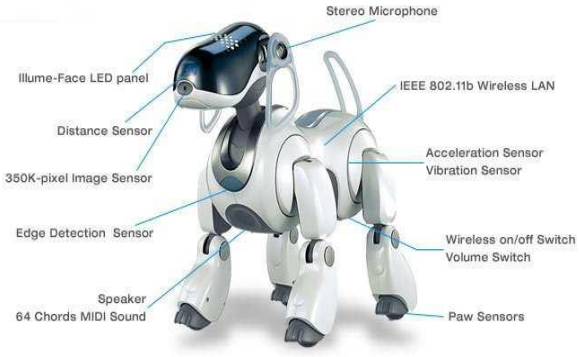


Fig. 13 The ERS-7 Sony Aibo Robot. (from http://www.sony.net/Products/aibo/)

VIII. Assessment and Discussion

Evaluation research is still underway for virtual characters [46, 47, 48] and for human-robot interaction [49, 50].

Since the goal of the project was to compare feedback provided by two types of embodiments (a virtual character and a robot) rather than to evaluate the multimodal feedback rules implemented in each of these systems, we decided to have users tell a story to both Greta and Aibo at the same time. The feedback system tested was the one described in the previous Sections, with the exception of the ASR system, which was not used here.

An instruction form was provided to the subject before the session. Then users watched the cartoon sequence, and were asked to tell the story to both Aibo and Greta (Fig 14). Finally, users had to answer a questionnaire. The questionnaire was designed to compare both systems with respect to the realization of feedback (general comparison between the two listeners, evaluation of feedback quality, perception of feedback signals and general comments). The evaluation form is provided in appendix. Sessions were videotaped using a Canon XM1 3CCD digital camcorder.

Fig 14. The assessment set-up

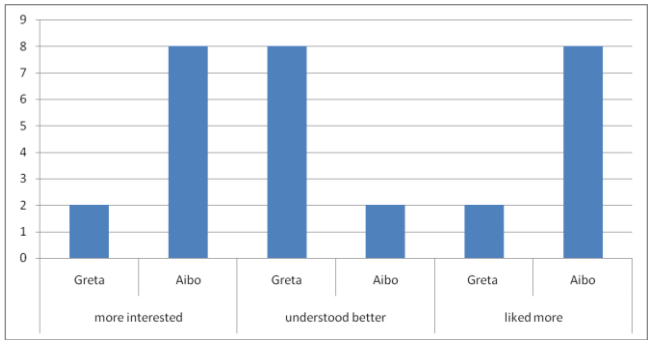


Table 4 Comparing the feedback provided by the virtual character and the robot

As illustrated by table 4, 8 out of 10 users estimated that GRETA understood better the story than AIBO. Yet, 8 out of 10 users felt that AIBO looked more interested and liked the story more than GRETA did.

Further evaluations could be investigated with such a system. Another possibility would be to have the speaker tell two different stories one to Greta, and then another one to Aibo. The order of the listeners should be counterbalanced across subjects. This would avoid having the speaker to switch his attention between Aibo and Greta. Perceptive tests on videos combining speakers and Aibo/Greta listeners could also be designed to have subjects 1) compare random feedback with feedback generated by analyzing user's behavior , or 2) rate if the listener has been designed to listen to this speaker or not.

IX. Conclusion and Further Developments

We presented a multi-modal framework to extract and identify Human communicative signals for the generation robot/agent feedbacks during storytelling. We exploited face-to-face interaction analysis by highlighting communicative rules. A real-time feature extraction module has been presented allowing the characterization of communicative events. These events are then interpreted by a fusion process for the generation of backchannel messages for both AIBO and GRETA. A simple evaluation was established, and results show that there is an obvious difference in the interpretation and realization of the communicative behavior between humans and agents/robots.

Our future works are devoted to the characterization of other communicative signals using the same modalities (speech and head). Prominence detection can be improved by the use of syllable-based analysis, which can be computed without linguistic information. Another important issue is to deal with the direction of gaze. This communicative signal conveys useful information during interaction and automatic analysis (human) and generation (robot/agent) should be investigated.

Acknowledgments

We are grateful to Elisabetta Bevacqua for her advice in the organization of our work and her help on interfacing our software with GRETA.

We also want to acknowledge Yannis Stylianou for the feedback he gave during discussions on our project.

Some members of this were partly funded by Région Wallonne, in the framework of the NUMEDIART research programme.

FP6 IP CALLAS is also gratefully acknowledged for having funded participants of this project.

REFERENCES

[1] K. Dautenhahn (2007) Methodology and Themes of Human-Robot Interaction: A GrowingResearch Field. International Journal of Advanced Robotic Systems 4(1) pp. 103-108.

[2] J.C. Baillie (2006) URBI tutorial [online] http://www.gostai.com/doc/en/urbi-tutorial/

[3] Cassell, J., Sullivan, J., Prevost, S., and Churchill, E. (eds). Embodied Conversational Agents. MIT Press, 2000.

[4] E. Bevacqua, M. Mancini, and C. Pelachaud, A listening agent exhibiting variable behaviour, Intelligent Virtual Agents, IVA'08, Tokyo, September 2008.

[5] H. Vilhjalmsson, N. Cantelmo, J. Cassell, N. E. Chafai, M. Kipp, S. Kopp, M. Mancini, S. Marsella, A. N. Marshall, C. Pelachaud, Z. Ruttkay, K. R. Thorisson, H. van Welbergen, R. van der Werf, The Behavior Markup Language: Recent Developments and Challenges, Intelligent Virtual Agents, IVA'07, Paris, September 2007.

[6] T. Fong, I. Nourbakhsh and K. Dautenhahn (2003): A Survey of Socially Interactive Robots, Robotics and Autonomous Systems 42(3-4), 143-166.

[7] C. Breazeal (2004). "Social Interactions in HRI: The Robot View," R. Murphy and E. Rogers (eds.), in IEEE SMC Transactions, Part C.

[8] Oudeyer P-Y, Kaplan , F. and Hafner, V. (2007) Intrinsic Motivation Systems for Autonomous Mental Development, IEEE Transactions on Evolutionary Computation, 11(2), pp. 265--286.

[9] Sidner, C.L.; Lee, C.; Kidd, C.D.; Lesh, N.; Rich, C., "Explorations in Engagement for Humans and Robots", Artificial Intelligence, May 2005

[10] Cassell, J., Bickmore, T., Billinghurst, M., Campbell, L., Chang, K., Vilhjálmsson, H. and Yan, H. (1999). "Embodiment in Conversational Interfaces: Rea." Proceedings of the CHI'99 Conference, pp. 520-527. Pittsburgh, PA.

[11] J. Rickel and W.L. Johnson, Animated Agents for Procedural Training in Virtual Reality: Perception, Cognition, and Motor Control, Applied Artificial Intelligence, 13, pp. 343-382, 1999.

[12] S. Kopp and I. Wachsmuth, Synthesizing Multimodal Utterances for Conversational Agents, The Journal Computer Animation and Virtual Worlds, 15(1), PP 39-52, 2004.

[13] J. Cassell and K. Thórisson, The Power of a Nod and a Glance: Envelope vs. Emotional Feedback in Animated Conversational Agents, Applied Artificial Intelligence, 13(3), 1999.

[14] S. Marsella, Interactive Pedagogical Drama: Carmen's Bright IDEAS Assessed. Intelligent Virtual Agents, pp 1-4, 2003.

[15] E. André, T. Rist, S. van Mulken, M. Klesen and S. Baldes, The automated design of believable dialogues for animated presentation teams, in J. Cassell, J. Sullivan, S. Prevost and E. Churchill (Eds) Embodied Conversational Characters, MITpress Cambridge, MA, 2000.

[16] W.L. Johnson, H. Vilhjalmsson, S. Marsella, Serious Games for Language Learning: How Much Game, How Much AI?, 12th International Conference on Artificial Intelligence in Education, Amsterdam, The Netherlands, July 18-22, 2005.

[17] A.L. Baylor, S. Kim, C. Son and M. Lee, Designing effective nonverbal communication for pedagogical agents, Proceedings of AI-ED (Artificial Intelligence in Education), Amsterdam, July 2005.

[18] J. Gustafson, N. Lindberg, M. Lundeberg, The August sopken dialog system, Proceedings of Eurospeech'99, Budapest, Hungary, 1999.

[19] Martin, J.C., Buisine, S., Pitel, G., and Bernsen, N. O.: Fusion of Children's Speech and 2D Gestures when Conversing with 3D Characters. In T. Dutoit, L. Nigay and M. Schnaider (Eds.): Multimodal Human-Computer Interfaces. Special Issue of Signal Processing (Elsevier) Vol. 86, Issue 12, December 2006

[20] T. Bickmore, J. Cassell, Social Dialogue with Embodied Conversational Agents", in J. van Kuppevelt, L. Dybkjaer, N. Bernsen (Eds) Advances in Natural, Multimodal Dialogue Systems, Kluwer Academic, New-York, 2005.

[21] L. Hall, M. Vala, M. Hall, M. Webster, S. Woods, A. Gordon and R. Aylett, FearNot's appearance: Reflecting Children's Expectations and Perspectives, Intelligent Virtual Agents IVA, Marina del Rey, USA, pp. 407-419, 2006.

[22] M. Lohse, K. J. Rohlfing, B. Wrede; G. Sagerer, "Try something else!" - When users change their discursive behavior in human-robot interaction, IEEE Conference on Robotics and Automation, Pasadena, CA, USA, 3481-3486, 2008.

[23] J. Allwood, J. Nivre, and E. Ahlsen. On the semantics and pragmatics of linguistic feedback. Semantics, 9(1), 1993.

[24] I. Poggi. Backchannel: from humans to embodied agents. In AISB. University of Hertfordshire, Hatfield, UK, 2005.

[25] D. McNeil, Hand and mind: What gestures reveal about thought, Chicago IL, The University, 1992.

[26] Jens Allwood, Loredana Cerrato, Kristiina Jokinen, Costanza Navarretta and Patrizia Paggio. The MUMIN coding scheme for the annotation of feedback, turn management and sequencing phenomena. Journal of Language Resources and Evaluation, Springer. Volume 41, 2007. pages 273-287

[27] Michael Kipp (2001) Anvil - A Generic Annotation Tool for Multimodal Dialogue. *Proceedings of the 7th European Conference on Speech Communication and Technology (Eurospeech)*, pp. 1367-1370.

[28] R. L. Brennan, D. J. Prediger: Coefficient κ: Some uses, misuses, and alternatives. In: Educational and Psychological Measurement. 41, 1981, 687–699.

[29] R. M. Maatman, Jonathan Gratch, Stacy Marsella: Natural Behavior of a Listening Agent. IVA 2005: 25-36.
[30] Pure Data: http://www.puredata.org
[31] The CMU Sphinx open source speech recognizer http://cmusphinx.sourceforge.net
[32] de Cheveigne, A., Kawahara, H.: YIN, a fundamental frequency estimator for speech and music. The Journal of the Acoustic Society of the America 111. 2002.
[33] Paul Taylor. The Tilt Intonation model, ICSLP 98, Sydney, Australia. 1982.
[34] B.M. Streefkerk, L. C. W. Pols, L. ten Bosch, Acoustical features as predictors for prominence in read aloud Dutch sentences used in ANNs, Proc. Eurospeech'99, Vol. 1, Budapest, 551-554, 1999.
[35] J.M.B. Terken, Fundamental frequency and perceived prominence of accented syllables. Journal of the Acoustical Society of America, 95(6), 3662-3665, 1994.
[36] N. Obin, X. Rodet, A. Lacheret-Dujour, "French prominence: a probabilistic framework", in International Conference on Acoustics, Speech, and Signal Processing (ICASSP'08), Las Vegas, U.S.A, 2008.
[37] V. K. R. Sridhar, A. Nenkova, S. Narayanan, D. Jurafsky, Detecting prominence in conversational speech: pitch accent, givenness and focus. In Proceedings of Speech Prosody, Campinas, Brazil. 380-388, 2008.
[38] D. Wang, S. Narayanan, An Acoustic Measure for Word Prominence in Spontaneous Speech. IEEE Transactions on Audio, Speech, and Language Processing, Volume 15, Issue 2, 690-701, 2007.
[39] P.Viola and M.J.Jones, "Robust Real-Time Face Detection", *InternationalJournal of Computer Vision*, 2004, pp 137-154.
[40] C.G. Harris and M.J. Stephens, "A combined corner and edge detector", Proc. Fourth Alvey Vision Conf., Manchester, pp 147-151, 1988
[41] Lucas, B., and Kanade, T. An Iterative Image Registration Technique with an Application to Stereo Vision, Proc. of 7th International Joint Conference on Artificial Intelligence (IJCAI), pp. 674-679.
[42] S. Young: "ATK: An application Toolkit for HTK, version 1.4, Cambridge University, May 2004
[43] T. Robinson, J. Fransen, D. Pye, J. Foote, and S. Renals. WSJCAM0: A British English speech corpus for large vocabulary continuous speech recognition. In Proc IEEE ICASSP, pages 81-84, Detroit, 1995.
[44] R. M. Maatman, Jonathan Gratch, Stacy Marsella, Natural Behavior of a Listening Agent. Intelligent Virtual Agents, IVA'05, 25-36, 2005.
[45] N. Ward, W. Tsukahara, Prosodic features which cue back-channel responses in English and Japanese. Journal of Pragmatics, 23, 1177-1207, 2000.
[46] Dehn, D. M. and van Mulken, S. (2000). "The impact of animated interface agents: a review of empirical research." International Journal of Human-Computer Studies(52): 1-22.
[47] Ruttkay, Z. and Pelachaud, C. (2004). From Brows to Trust - Evaluating Embodied Conversational Agents, Kluwer. http://wwwhome.cs.utwente.nl/~zsofi/KluwerBook.htm
[48] Buisine, S., Martin, J.-C. (2007) The effects of speech-gesture co-operation in animated agents' behaviour in multimedia presentations. International Journal "Interacting with Computers: The interdisciplinary journal of Human-Computer Interaction". Elsevier, 19: 484-493.
[49] Dan R. Olsen, Michael A. Goodrich (2003) Metrics for Evaluating Human-Robot Interactions. Performance Metrics for Intelligent Systems Workshop held in Gaithersburg, MD on September 16-18, 2003. http://www.isd.mel.nist.gov/research_areas/research_engineering/Performance_Metrics/PerMIS_2003/Proceedings/Olsen.pdf
[50] Mohan Rajesh Elara, Carlos A. Acosta Calderon, Changjiu Zhou, Pik Kong Yue, Lingyun Hu, Using Heuristic Evaluation for Human-Humanoid Robot Interaction in the Soccer Robotics Domain. Second Workshop on Humanoid Soccer Robots @ 2007 IEEE-RAS International Conference on Humanoid Robots Pittsburgh (USA), November 29, 2007 http://www.informatik.uni-freiburg.de/~rc06hl/ws07/papers/HSR-2-110.pdf

X. Appendix

Instructions

You are going to watch a short cartoon sequence.
Then you will have to tell this story.
Your story will be listened to and watched by the Greta virtual agent displayed on a screen and an Aibo robot at the same time.
Both will react to the story that you are telling.
Please tell the story as you would tell a story to two people listening to you at the same time.

Evaluation form

Subject number :
Date of session :

INFORMATION ABOUT SUBJECT

Last name :
First name :
Age :
Male / Female
e-mail :

GENERAL COMPARISON BETWEEN THE TWO LISTENERS

Did you like telling your story to Greta?
I liked very much / I liked / I did not like
Did you like telling your story to Aibo?

I like very much / I liked / I did not like
Which one did you prefer?
Greta / Aibo
Why?
Which listener was mostly interested in your story?
Greta / Aibo
Which listener understood better your story?
Greta / Aibo
Which listener liked better your story?
Greta / Aibo
What did you like in Greta that was not present in Aibo?
What did you like in Aibo that was not present in Greta?

EVALUATION OF FEEDBACK QUALITY

Who most clearly displayed when it was interested in your story?
Aibo / Greta
Who most clearly displayed when it understood what you said?
Greta / Aibo
Who most clearly displayed when it liked what you said?
Aibo / Greta
How would you qualify the behavior displayed by Greta when you told your story ?
How would you qualify the behavior displayed by Aibo when you told your story ?

PERCEPTION OF FEEDBACK SIGNALS

How did Greta displayed that she was interested in your story?
How did Greta displayed that she was understanding in your story?
How did Greta displayed that she liked in your story?
How did Aibo displayed that it was interested in your story?
How did Aibo displayed that it was understanding in your story?
How did Aibo displayed that it liked in your story?

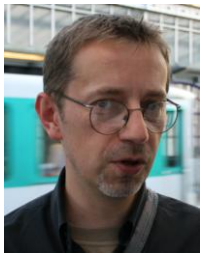
GENERAL COMMENTS

How did you evaluate the fact to interact with both at the same time?
Please feel free to provide any additional comments:

Sàmer Al Moubayed is a PhD student since 2008 at the Center of Speech Technology CTT, Royal Institute of Technology KTH, Sweden, and at the Graduate School of Language Technology GSLT, Gothenburg, Sweden. He obtained his MSc degree in Artificial Intelligence and Speech and Language Technology from KULeuven, Belgium.
His main research is carried out in the field of speech communication, and is doing his thesis in Multimodal Speech Synthesis, and is involved in many National and European projects. Samer main research interest is in Talking Agents and behavior prediction, machine learning and pattern recognition in speech, and psycholinguistics

Malek Baklouti graduated as engineer in from the Tunisian Polytechnic School and received the M.S. degree in Applied Mathematics in 2006. She is currently a PhD student in Robotics and Signal processing at Thalès Security System and Services, France and University of Versailles.
She will be in PAMI lab at the University of Waterloo (Canada) as a Visiting Scholar.

Mohamed Chetouani received the M.S. degree in Robotics and Intelligent Systems from the University Pierre and Marie Curie (UPMC), Paris, 2001. He received the PhD degree in Speech Signal Processing from the same university in 2004. In 2005, he was an invited Visiting Research Fellow at the Department of Computer Science and Mathematics of the University of Stirling (UK). Dr. Chetouani was also an invited researcher at the Signal Processing Group of Escola Universitaria Politecnica de Mataro, Barcelona (Spain).
He is currently an Associate Professor in Signal Processing and Pattern Recognition at the University Pierre et Marie Curie. His research activities, carried out at the Institute of Intelligent Systems and Robotics, cover the areas of non-linear speech processing, feature extraction and pattern classification for speech, speaker and language recognition.
He is a member of different scientific societies (ISCA, AFCP, ISIS). He has also served as chairman, reviewer and member of scientific committees of several journals, conferences and workshops.



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 full 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, 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 a member of the INTERSPEECH'07 organization committee. He was the initiator of eNTERFACE workshops and the organizer of eNTERFACE'05.

Ammar Mahdhaoui was born in Tunisia, on january 31, 1983. He received the M.S. degree in Engineering of Person-System communication from the university of Grenoble, Grenoble, 2007. Since October 2007, he is PHD Student in signal processing and pattern recognition at the University Pierre and Marie Curie Paris-6, his research activities carried out at the Institute of Intelligent Systems and Robotics.

Jean-Claude Martin is Associate Professor at CNRS-LIMSI, France.
His research topic is multimodal communication, both in human-human and human-computer contexts ; the study of individual differences, the multimodal expression and perception of social behaviors and the evaluation of user's multimodal interaction with embodied conversational agents.
He received the PhD degree in Computer Science in 1995 from the Ecole Nationale Supérieure des Télécommunications (ENST, Paris). He passed his habilitation to direct research in 2006 on Multimodal Human-Computer Interfaces and Individual Differences.
Annotation, perception,
representation and generation of situated multimodal behaviors. He is the head of the Conversational Agents topic of research within the Architecture and Models for Interaction Group (AMI). He co-organised a series of three international workshops on multimodal corpora at LREC 2002, 2004, 2006 and he is a guest editor of a special issue of the international Journal on Language Resources and Evaluation to appear in 2007 on multimodal corpora.

Stanislav Ondas graduated from the Technical University of Košice in 1999. In 2007 he finished your PhD. study at the Department of Electronics and Multimedia Communications at the same university and he is currently an assistent at mentioned department.
His research interests in spoken dialogue systems, dialogue management, voice service designing and natural language processing**.** He has been involved in several national projects related to spoken and multimodel dialogue systems (ISCI, MOBILTEL).

From October 2008, **Catherine Pelachaud** will be a director of research at CNRS in LTCI, TELECOM ParisTech**.**She received her PhD in Computer Graphics at the University of Pennsylvania, Philadelphia, USA in 1991. Her research interest includes representation language for agent, embodied conversational agent, nonverbal communication (face, gaze, and gesture), expressive behaviors and multimodal interfaces. She has been involved in several European and national projects related to multimodal communication (EAGLES, IST-ISLE), emotion (FP5 NoE Humaine, FP6 IP CALLAS, FP7 STREP SEMAINE), to believable embodied conversational agents (IST-MagiCster, FP5 PF-STAR, RIAM ACE, ANR My-Blog-3D), and to handicap (CNRS-Robea HuGeX, RIAM LABIAO).

Jérôme Urbain graduated as an electrical engineer from the Faculté Polytechnique de Mons (FPMs), Belgium, in 2006. He is currently PhD student at the Signal Processing and Circuit Theory (TCTS) Lab of the same University, working on speech processing in the framework of FP6 IP CALLAS.
He is focusing on laughter modeling, synthesis and recognition.

Mehmet Yilmaz is a senior undergraduate student at Koç University, Electrical and Electronics Department. His research interests are visual tracking, and interactive multimodal systems.

Activity-related Biometric Authentication

G. Ananthakrishnan, H. Dibeklioğlu, M. Lojka, A. Lopez, S. Perdikis, U. Saeed, A.A. Salah, D. Tzovaras, A. Vogiannou

***Abstract*—This project aims at developing a biometric authentication system exploiting new features extracted by analysing the dynamic nature of various modalities, including motion analysis during ordinary tasks performed in front of a computer, analysis of speech, continous face and facial movement analysis, and even patterns for grasping objects. We test the potential and contribution of each of these modalities for biometric authentication in the face of natural, uncontrolled environments, as well as their fusion.**

***Index Terms*—Biometric authentication, activity recognition, face recognition, motion analysis, speaker recognition, audio based event recognition**

I. Introduction

THIS project attempts to address the limitations of unimodal biometrics by deploying activity-related multimodal biometric systems that integrate the evidence presented by multiple sources of information. Therefore, the combination of a number of independent modalities is explored to overcome the possible restrictions set by each modality. With a simple sensor setup, we aim at more robust biometric identification through the fusion of physiological, behavioral and soft biometric modalities, keeping also in mind the unobtrusiveness and comfort of the subject.

The term behavioral biometrics refers to Person Recognition using shape based activity signals (gestures, gait, full body and limb motion) or face dynamics. Activity-specific signals [6], [23] provide the potential of continuous authentication, but state-of-the-art solutions show inferior performance compared to static biometrics (fingerprints, iris). This drawback could hopefully be eliminated by the inferential integration of different modalities.

Behavioral information from face videos for person recognition may also be investigated in order to exploit the underlying temporal information in comparison to image-based recognition [39]. Methods for person recognition from face dynamics can be classified into holistic methods (head displacements and pose evolution [30]), feature-based methods (exploitation of individual facial features [12]) and hybrid methods [14]. Various probabilistic frameworks have been proposed in recent works, usually employing a Bayesian Network (Hidden Markov Models, Coupled and Adaptive HMMs, etc.) as the mathematical model for recognition [35].

Soft biometrics (gender, height, age, weight etc.) are believed to be able to significantly improve the performance of a biometric system in conjunction with conventional static biometrics [22], yet their exploitation remains an open issue. Microphones for voice recognition, sound based sensors for monitoring activities or other modalities could also be considered.

This report, as well as the source code for the software developed during the project, is available online from the eNTERFACE'08 web site: www.enterface08.limsi.fr.

G. Ananthakrishnan is with Royal Institute of Technology, SWEDEN. E-mail: agopal@kth.se.

H. Dibeklioğlu is with Perceptual Intelligence Laboratory, Department of Computer Engineering, Boğaziçi University, 34342 Bebek, Istanbul, TURKEY. E-mail: hamdi.dibeklioglu@cmpe.boun.edu.tr.

M. Lojka is with Technická Univerzita v Košiciach, SLOVAKIA. E-mail: martin.lojka@tuke.sk.

A. Lopez is with Technical University of Catalonia, Barcelona, SPAIN. E-mail: alopez@gps.tsc.upc.edu.

S. Perdikis, A. Vogiannou and D. Tzovaras are with CERTH/ITI, GREECE. E-mail: {perdik,tvog }@iti.gr.

U. Saeed is with EURECOM, FRANCE. E-mail: usman.saeed@eurecom.fr.

A.A. Salah is with Centrum Wiskunde & Informatica, 1090 GB Amsterdam, THE NETHERLANDS. E-mail: a.a.salah@cwi.nl.

In this report, we look at some of these modalities in a specific fixed-seat pilot. Our experimental setup is described in Section II, including the details of the collected database. The individual modalities are investigated in separate sections, starting with model-based motion analysis in Section III, which tracks the user via calibrated cameras during ordinary activities. The sounds that ensue during these activities are analysed for robust activity classification. This part is exposed in Section IV. Once speech is detected among the sound events, it can be further used for authentication. Section V deals with speaker authentication. Our model flexibly integrates data coming from seemingly unrelated modalities. Section VI exemplifies this by making use of an advanced interface for recognizing activity, namely a Cyberglove, which is used to collect and analyse grasping patterns. The more common face modality is used to serve as a benchmark. Continuous authentication from captured static face images is explained in Section VII, and the optical-flow based analysis of facial motion for authentication is detailed in Section VIII. Section IX builds on the motion analysis to recognize types of activities, and evaluates the authentication potential of each of these activities.

The mathematical framework we establish here is employed to seamlessly integrate an arbitrary number of sources that provide partial authentication information. Our experimental results are given in Section XI. The report concludes with a discussion of these results and on possible future directions in Section XII.

II. The Experimental Setup

The proposed biometric system is evaluated in a fixed - seat office pilot, where the user is able to move his arms, head and torso and manipulate objects on a desk while seated. This experimental setup selection serves multiple aspects of the problem of activity-related biometric authentication:

- **It is portable and easy to setup**
- **It can be part of a normal authentication system scenario (e.g. secured indoor premises)**
- **It can easily incorporate all the equipment for selected modalities**
- **An office environment is involved in many work - related activities, which makes the pilot ideal for testing the activity - related authentication module**
- **It is fully unobtrusive to the user**

The selected pilot consists of a desk upon which a number of objects is placed, in stable predefined positions. This constraint implies a static environment, which slightly affects the generality of the setup, but significantly facilitates the activity recognition task. The objects are: a) desk phone, b) glass (on a pad), c) keyboard, d) mouse, e) computer screen, f) pencil (in a pencil case) g) a piece of paper for writing. The sensorial equipment is as inexpensive and unobtrusive as possible. It comprises of three Logitech QuickCam webcams (two for body motion tracking and one for continuous face authentication and facial motion analysis) and a regular low - budget microphone. Two cameras are mounted on the desktop screen facing the user (these are the frontal motion tracking camera and the face camera, which is zoomed on the user's head area), while the third camera (lateral motion tracking camera) is placed on a tripod on the left side of the desk. The microphone is mounted on the desk, next to the keyboard. Fig. 1 illustrates the actual pilot setup.

A. Recording Scenario and Data Gathering

Within the project a database of 15 persons performing a number of actions has been recorded. Each person was asked to execute six actions in a particular order, responding to the environmental stimuli (phone ringing, instructions on the screen or on a writing form). A recording scenario has been prepared so as to enhance the database's consistency, to meet the requirements and constraints of every modality and to ensure the user's concentration and relaxation, so that he performs the required

Fig. 1. The pilot setup, shown during one of the recordings. The frontal cameras are mounted on the display, and the side camera is mounted on a pod to the left of the subject.

actions in his natural (and therefore consistent) way. The six recorded actions were:

- **Mouse manipulation (playing a computer game)**
- **Phone Conversation (real dialogue with a team member)**
- **Typing in the keyboard (filling in a given questionnaire)**
- **Writing (filling in a questionnaire in a writing form)**
- **Drinking (taking the cup, and leaving it back to its place)**
- **Reading (specific texts provided in the screen).**

Every session consisted of one repetition of the six actions and 10 **sessions were recorded for each user in order to provide enough training and testing data for all the modalities. The database size was limited to** 15 **persons due to limited time available for recordings.**

During the data gathering users were asked to act in their natural way, without any further instructions or constraints. The selected activities are common work - related activities involving usual office objects, there was no previous knowledge about their suitability for authentication. The evaluation of their discriminative power is among the objectives of this project.

III. Model-Based Motion Analysis

Markerless human motion capture is a challenging problem that involves the estimation of a high-dimensional configuration of a three-dimensional non-rigid and self-occluding object. Since a wide range of applications are derived from the unobtrusive characterization of human activity, this research area has recently undergone several advances due to the yielded interest.

A common approach is to consider an articulated body model with several degrees of freedom per joint, depending on the complexity of the possible poses and the quality of the available data. This representation implies the use of kinematic constraints on the motion. Additional assumptions and motion constraints can be adopted at the cost of generality of the solution which we intend to preserve. To this end, Particle Filters [2] have become a relevant technique due to their ability to handle multi-modal non-linear and non-Gaussian distributions. Several approaches such as partitioned sampling [37], hierarchical sampling [41] and annealing particle filter [15] have been developed to cope with high-dimensional limitations of the classical Condensation algorithm [21].

We present a particular implementation of the annealing particle filter for a simplified body model in order to retrieve the human body poses of a subject performing different actions in a multi-view scenario. We propose simplifications of the body tracking problem without almost no loss of generality in the given pilot and with the capability of coping with realistic scenarios.

TABLE I
Articulated Body Model Joints

Angle	Joint	Rotation Axis	Range
θ_1	Base of the Neck	**y**	$\left[-\frac{\pi}{4}, \frac{\pi}{4}\right]$
θ_2	Right Shoulder	**x**	$\left[-\frac{\pi}{4}, \pi\right]$
θ_3	Left Shoulder	**x**	$\left[-\frac{\pi}{4}, \pi\right]$
θ_4	Right Shoulder	**y**	$\left[-\frac{\pi}{4}, \pi\right]$
θ_5	Left Shoulder	**y**	$\left[-\frac{\pi}{4}, \pi\right]$
θ_6	Right Shoulder	**z**	$\left[-\frac{\pi}{4}, \frac{\pi}{2}\right]$
θ_7	Left Shoulder	**z**	$\left[-\frac{\pi}{4}, \frac{\pi}{2}\right]$
θ_8	Right Elbow	**y**	$[0, \pi]$
θ_9	Left Elbow	**y**	$[0, \pi]$

A. Body Model

A simplistic articulated body model will fulfill the requirements of the scenario presented in section II. This model is based on the kinematic chain framework and comprises a set of joints. In our case, this set of joints are the base of the neck, shoulders and elbows. Every joint has a maximum of three degrees of freedom according to the complexity of the motions that we want to capture. Each degree of freedom is represented by an axis of rotation defined in a default body configuration, where all the angles are set to zero (see fig. 2). The range of joint angles is also defined according to this default body pose. In our model, a total of nine degrees of freedom are defined (see table I). In order to set the model in a world position, a three-dimensional coordinate system built with the base of the neck as origin and a body orientation are defined. Our model reference point is set to be the base of the neck. Therefore, our body model defines a thirteen-dimensional state vector:

$$\mathbf{x_t} = \{x_0, y_0, z_0, \theta_0, ..., \theta_9\} \tag{1}$$

Angle θ_0 **is the orientation of the whole body model while all the other angles are designed following basic kinematic constraints. The use of angles ensures a compact representation in front of a state defined by only 3D coordinates. Knowing the limbs' dimensions we can go from a set of angles to Cartesian coordinates by means of exponential twists formulation [7]; every point of interest can be computed from its initial location with respect to the reference point in the default body configuration and the product of the exponential maps affecting the motion of this point:**

$$p(\mathbf{x_t}) = \prod_i \mathbf{M}_i(\mathbf{x_t}) p_0 \tag{2}$$

$$\mathbf{M}_i = \begin{bmatrix} \mathbf{R_i}(\mathbf{x_t}) & \mathbf{t_i}(\mathbf{x_t}) \\ 0 & 1 \end{bmatrix} \tag{3}$$

where $p(\mathbf{x_t})$ **represents a point of interest as a function of the state vector, that encodes model position, model orientation and joint angles, and** $\mathbf{M}_i(\mathbf{x}_t)$ **is the exponential map in the chain where** p **is found. The exponential map comprises the rotation matrix** R **and the translation vector** t**. The whole notation is being presented in homogeneous coordinates due to its compactness.**

B. Particle Filter

Particle Filters (PF) [2] are recursive Bayesian estimators derived from Monte Carlo sampling techniques which can handle non-linear and non-Gaussian processes. Commonly used in tracking problems, they are used to estimate the posterior density $p(\mathbf{x}_t|\mathbf{z}_t)$ **by means of a set of**

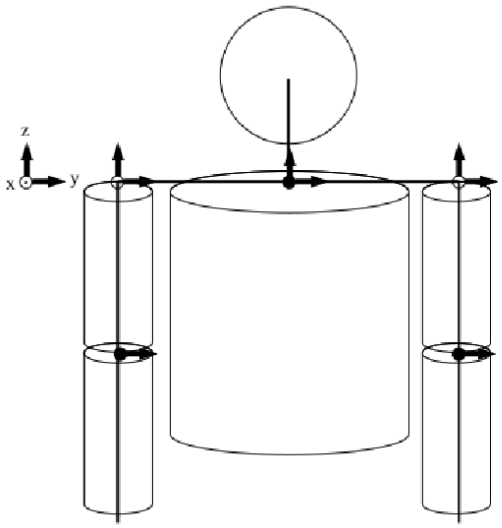


Fig. 2. Simple articulated model for body tracking

N_s **weighted samples or particles. Given a Bayesian recursive estimation problem:**

$$p(\mathbf{x}_{0:t}|\mathbf{z}_{1:t}) = \frac{p(\mathbf{z}_{1:t}|\mathbf{x}_t)p(\mathbf{x}_t|\mathbf{x}_{t-1})}{p(\mathbf{z}_t|\mathbf{z}_{1:t-1})} p(\mathbf{x}_{0:t-1}|\mathbf{z}_{1:t-1}) \quad (4)$$

we want to draw samples from the posterior such that:

$$p(\mathbf{x}_{0:t}|\mathbf{z}_{1:t}) \approx \sum_i^{N_s} w_t^i \delta(\mathbf{x}_t - \mathbf{x}_t^i) \quad (5)$$

where w_t^i is the weight associated to the i-th particle. This discrete approximation of the posterior requires the weights evaluation. This is done by means of the importance sampling principle [16], with a probability density function (pdf) $q(\mathbf{x}_{0:t}|\mathbf{z}_{1:t})$ from which we can generate samples that can be evaluated with the posterior (up to proportionality). Applying the importance sampling principle in Eq. 4:

$$w_t^i \propto \frac{p(\mathbf{x}_{0:t}|\mathbf{z}_{1:t})}{q(\mathbf{x}_{0:t}|\mathbf{z}_{1:t})} \quad (6)$$

$$w_t^i \propto \frac{p(\mathbf{z}_{1:t}|\mathbf{x}_t)p(\mathbf{x}_t|\mathbf{x}_{t-1})}{p(\mathbf{z}_t|\mathbf{z}_{1:t-1})q(\mathbf{x}_{0:t}|\mathbf{z}_{1:t})} p(\mathbf{x}_{0:t-1}|\mathbf{z}_{1:t-1}) \quad (7)$$

and choosing this importance distribution in a way that factors appropiately we have:

$$w_t^i \propto \frac{p(\mathbf{z}_{1:t}|\mathbf{x}_t)p(\mathbf{x}_t|\mathbf{x}_{t-1})p(\mathbf{x}_{0:t-1}|\mathbf{z}_{1:t-1})}{p(\mathbf{z}_t|\mathbf{z}_{1:t-1})q(\mathbf{x}_t|\mathbf{x}_{0:t-1},\mathbf{z}_t)q(\mathbf{x}_{0:t-1}|\mathbf{z}_{1:t-1})} \quad (8)$$

$$w_t^i \propto w_{t-1}^i \frac{p(\mathbf{z}_{1:t}|\mathbf{x}_t)p(\mathbf{x}_t|\mathbf{x}_{t-1})}{p(\mathbf{z}_t|\mathbf{z}_{1:t-1})} \quad (9)$$

Moreover, if we apply the Markov assumption the expression is simplified regarding the fact that observations and current state only depend on the previous time instant. Therefore, the PF is a sequential propagation of the importance weights.

Two major problems affect the PF design. The first is the choice of the importance distribution. This is crucial since the samples drawn from $q()$ must hit the posterior's typical set in order to produce a good set of importance weights. It has been shown in [16] that $q(\mathbf{x}_t|\mathbf{x}_{t-1},\mathbf{z}_t) = p(\mathbf{x}_t|\mathbf{x}_{t-1},\mathbf{z}_t)$ is optimal in terms of variance of the weights. The second problem deals with particle degeneracy in terms of variance of the weights. After several iterations the majority of the particles have negligible weights and as a consequence of this the estimation efficiency decays. An effective measure for the particle degeneracy is the survival rate [34] given by:

$$\alpha = \frac{1}{N_s \sum_{i=1}^{N_s} (w_t^i)^2} \quad (10)$$

In order to avoid the estimator degradation the particle set is resampled. After likelihood evaluation a new particle set must be drawn from the posterior estimation, hence particles with higher weights are reproduced with higher probability. Once the new set has been drawn all the weights are set to $\frac{1}{N_s}$, leading to a uniformly weighted sample set concentrated around the higher probability zones of the estimated posterior.

The Sampling Importance Resampling (SIR) Particle Filter proposed by Gordon et. al [18] is a method commonly used in computer vision problems. It's characterized by applying resampling at every iteration and by defining the importance distribution as the prior density $p(x_t|x_{t-1})$. By substituting this importance density in 8, it's easy to realize that weight computation only depends on the likelihood. Consequently, the design of the particle filter is basically a problem of finding an appropiate likelihood function.

C. Likelihood Evaluation

In computer vision problems probability density functions usually are not directly accessible, thus an observation model is required to approximate the likelihood function. It is necessary to determine which image features are more correlated with the true body configuration. Therefore, finding the appropiate likelihood approximation involves both image and body model. Deutscher et al. [15] proposed a matching of the model projection with foreground segmentation and edges. Their flesh model consists of conic sections with elliptical cross-sections surronding virtual skeleton segments. Raskin et al. [50] add the body part histogram as an additional feature. Other authors use Visual Hull approaches [27] to work with voxel data. In that case, they can use three-dimensional flesh models, like ellipsoids [40] or three-dimensional Gaussian mixtures [9].

Our challenge is to produce a likelihood approximation able to deal with moving objects, clothing, limited number of views and low frame rate. In our approach we should not rely on a 3D reconstruction because only a few views are available, thus a projection of the model onto the images is required. Our proposal is to avoid the computational cost of projecting the whole set of sampling points of a 3D flesh model by projecting a reduced set of points per body part. Our flesh model will be set of cylinders around all the skeleton segments except the head, which will be modelled by a sphere (see Fig. 2). Therefore, our reduced set of projected points will be defined by the vertices of the trapezoidal section resulting from the intersection of a plane, approximately parallel to the image plane, with the cylindric shapes modelling the limb (or spherical shape in the case of the head).

To define an intersecting plane for a given cylinder, we compute the vectors going from the camera center towards each one of the limit points of the limb. Then the cross product of these vectors with the one defined by the limb itself is computed to determine two normal vectors that lie on the intersecting plane and along which we will find the key points to project. The head template is handled with a similar procedure using as limb vector the one going from the body model reference point to the head center. The norm of the cross product, as well as the area of the projected trapezoid, can be used as a quality measure in order to determine whether the limb is properly aligned with the view (this does not apply for the head). If this quality measure is above a certain threshold, we can change the trapezoidal projected shape by a circle or an ellipse. However, in our scenario the views are set so that they capture good limb alignments in most of the frames, thus we can obviate the computation of this measure.

Regarding the image features, we have seen that common likelihood approximations like [15] do not perform well in our scenario with the described body model. We propose modifications on this approximation while keeping common features that are easy to extract, like foreground silhouettes, contours and detected skin. We extract foreground silhouettes by means of a background learning technique based on Stauffer and Grimson's method [56]. A single multivariate Gaussian $\mathcal{N}(\mu_t, \mathbf{\Sigma}_t)$ with diagonal covariance in the RGB space is used to model every pixel value $\mathbf{I_t}$. The algorithm learns the background model for every pixel using a set of background images and then, for the rest of the sequence, evaluates the likelihood of a pixel color value to belong to the background. With every pixel that matches the background the pixel model is updated, adaptively learning smooth illumination changes:

$$\mu_t = (1-\rho)\mu_{t-1} + \rho\mathbf{I_t} \quad (11)$$

$$\mathbf{\Sigma}_t = (1-\rho)\mathbf{\Sigma}_{t-1} + \rho(\mathbf{I_t} - \mu_{t-1})^T(\mathbf{I_t} - \mu_{t-1}) \quad (12)$$

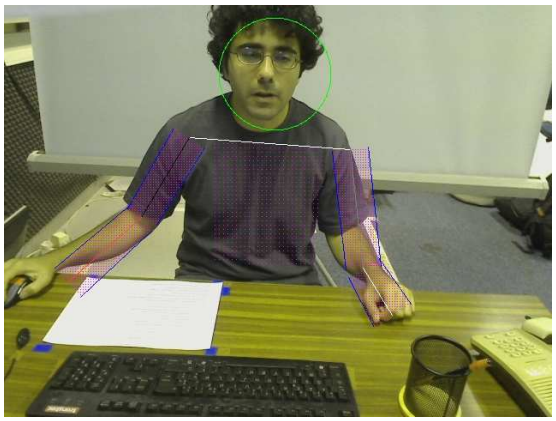
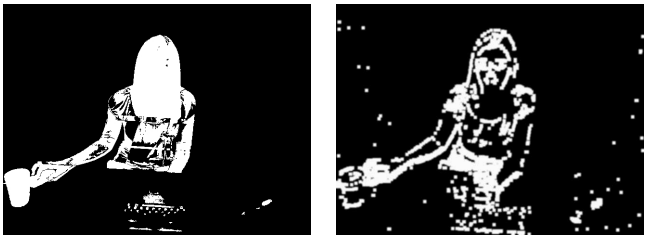


Fig. 3. Projection of the flesh model associated to a given particle

A shadow removal algorithm [65], based on the color and brightness distortion, is used to enhance the segmentation.



(a) Foreground Mask (b) Contours Mask

Fig. 4. Extracted Image features

Contour detection is performed by means of the Canny edge detector [10]. The result is dilated with a 8-connectivity and 5x5 structuring element, and smoothed with a Gaussian mask. In order to avoid spurious contours, we subtract the background contours. This implies also deleting some pixels in the edges of interest but the body structure it's in general preserved. Finally, a simple skin detection method based on evaluating the likelihood ratio between skin and non-skin hypothesis is performed. The likelihood functions are estimated by 8-bins color histograms of several skin and non-skin samples.

The likelihood evaluation procedure involves the projection of the flesh model of every particle onto the image coordinate system. The resulting shape is scanned and matched with the foreground segmentation. The weight is computed as follows:

$$\omega^{fl} = \frac{1}{N}\sum_{n=1}^{N}(1 - I_n^f) \tag{13}$$

Since pixel intensities in the foreground masks (I_t^f) have 0 or 1 as possible values, the weighting function is obtained by a normalized sum of the background pixels falling inside the projected flesh model. In the head model case, we add skin detection information:

$$\omega^{fh} = \frac{1}{N}\sum_{n=1}^{N}(1 - I_n^f I_n^s) \tag{14}$$

Therefore, the final foreground weight ω^f is the averaged sum of all the limbs and head weights. Foreground segmentation provides data that are generally invariant to clothing and most of the background conditions.

Since many configurations can be explained via this feature, foreground information is used to penalize false poses rather than to single out the correct one. Moreover, the proposed measure shows how well the model fits the observation, but doesn't evaluate how well the observations are being explained by the model. Suppose the likelihood $p(\mathbf{z}_t|\mathbf{x}_t)$ is available and that a given pose generates a pdf. A measure that can be used to assess the similarity of the likelihood and the generated pdf is the Kullback-Leibler divergence. At this point is important to remark that the KL divergence will provide different results depending on the factor order (except if both pdfs are identical). We can establish an analogy with our likelihood approximation. We are trying to determine the mutual information of the model and the observations. Therefore, we propose to include an additional divergence measure between the projection of the flesh model and the foreground masks to see how well a particle explains the observations.

$$\omega^d = \frac{1}{N_f}\sum_{n=1}^{N_f}(I_n^f(1 - B_n)) \tag{15}$$

This divergence basically aims for projecting a given particle and computing the overlap between the pixels B_n of this projection and the N_f foreground pixels of the observation.

Contours found in the body usually provide good information on the location of the arms and the legs. However, in some cases, clothing and background can introduce spurious contours that reduce the reliability of this feature. As mentioned above, we try to minimize the background impact by subtracting the background contours. The proposed weighting function for this feature is a sum of squared differences between the contour pixels and the edges of the flesh model aligned with the axis of the limb:

$$\omega^e = \frac{1}{N}\sum_{n=1}^{N}(1 - I_n^e)^2 \tag{16}$$

Finally, all these weights are combined for every camera:

$$\omega = exp\left(\sum_{c=1}^{C}(\lambda_c^f\omega^f + \lambda_c^e\omega^e + \lambda_c^d\omega^d)\right) \tag{17}$$

We use a set of weights for every camera and measure to adjust the importance of every feature according to its importance and visibility. Since in our scenario the subject stays in his seat, we assume that the visibility component can be determined beforehand.

D. Annealing Particle Filter

It has been shown in several works that SIR Particle Filters are a good approach for tracking in low dimensional spaces, but they become inefficient in high-dimensional problems. Deutscher et. al [15] proposed a variation of the SIR framework by introducing the concept of Annealing PF. In body pose tracking problems, the likelihood approximation often is a function which has several peaked local maxima. Annealing PF deals with this problem by evaluating the particles in several smoothed versions of the likelihood approximation. After the weights are computed via the modified likelihood, particles are resampled and propagated with Gaussian noise with zero mean and a covariance that decreases at every step. Each one of this steps (weighting with a smoothed function, resampling and propagation) is called an annealing run. In the last annealing run the estimation is given by means of the Monte-Carlo approximation of the posterior mean:

$$\hat{\mathbf{x}}_\mathbf{t} = \sum_{i=1}^{N_s} w_t^i \mathbf{x}_t^i \tag{18}$$

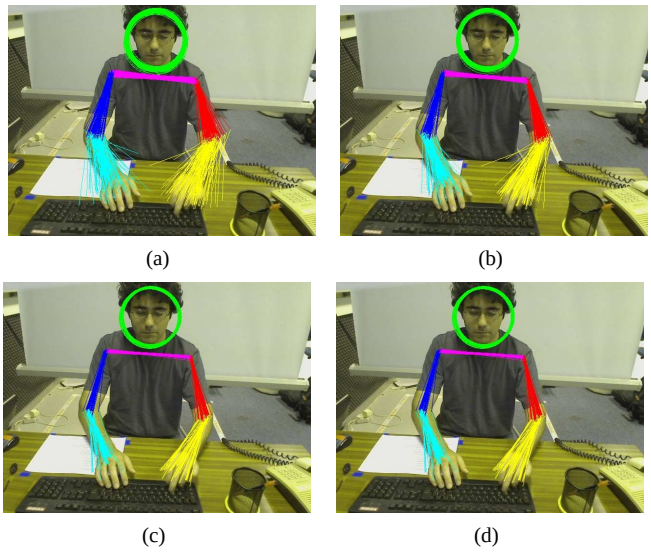
The most usual way to smooth the weighting function is by means of an annealing rate, an exponent $\beta < 1$. In the first layer β is minimum but it progressively increases with each layer, sharpening the likelihood approximation. In [15] a method for tuning β with the survival rate after each annealing run is proposed.

The sharpness of the likelihood function is due to the high dimensional space in which is defined, the use of a hierarchical model [11] is another possible strategy in order to have annealing layers. Since our model is quite simple, a hierarchical approach is not justified. We have implemented an annealing particle filter in which the smoothing is done by means of an exponent β. In our case, the annealing rate is updated according to the survival rate of the preceding layer $\alpha(\beta_{t-1})$. Given a desired survival rate α_T:

$$\beta_t = \beta_{t-1} - \lambda(\alpha_T - \alpha(\beta_{t-1})) \tag{19}$$

Due to the image feature characteristics, we also introduce β in (17), giving higher importance to the foreground-based measures in the first layers and to the contour-based measures in the last layers.

$$\omega = exp\left(\sum_{c=1}^{C}(\lambda_c^f(\frac{1}{\beta})\omega^f + \lambda_c^e(\beta)\omega^e + \lambda_c^d(\frac{1}{\beta})\omega^d)\right) \tag{20}$$



(a) (b) (c) (d)

Fig. 5. Annealing Layers. The covariance used in the propagation step is progressively reduced through four annealing layers while the estimator gets closer to the true pose

Therefore, we propose to work with overall smoothing and feature-based smoothing. However, more work needs to be done in this area in order to show that this approach can help to efficiently reach the true pose.

IV. Sound-Based Event Detection

This section deals with detection of sound activity and classification of sounds into the typical events that would be encountered. In the first step, any sort of sound activity is detected and in the second step it is classified. The details of each step are explained below.

A. Sound Activity Detection

The field of Sound Activity Detection has been researched for several years. Most of the research has been in the field of Voice activity detection in noisy conditions. This is essentially different from the current experiment in which all sound activity needs to be detected. This makes it a slightly difficult problem in a way, because a threshold on the length of the activity cannot be provided. The detection has to be made on short bursts of sounds like clicks of mouse as well as continuous speech. So a dynamic threshold needs to be provided, based on the current noise level.

Previous work done in voice activity detection was mainly by Mak *et al.* [38] and Nemer *et al.* [45]. Nemer *et al.* proposed a method based on the residual of the signal, and used higher order statistics of the noise in order to set the threshold to detect sound activities. Renevey and Drygajlo [52], proposed an Entropy based threshold for activity detection. The method used in this experiment uses the entropy found on the residual as a measure to detect activity.

The following steps are taken to detect sound activity

- **The signal is windowed with a window size of 40 ms and a shift of 20 ms.**
- **The signal within one window is approximated by 2 Linear Prediction Coefficients (LPC). This is done to grossly approximate the frequency spectrum and calculate the bias.**
- **The residual of the signal, which is the error between the estimated LPC and the true signal is calculated. Fig. 6 shows the spectrum of the signal and Fig. 7 shows the corresponding residual. One can observe that the bias has been canceled and the spectrum has been whitened.**
- **The Entropy is calculated for the residual, assuming a gaussian distribution, since whitening has been performed. The evidence of activity is given by the entropy. A higher entropy indicates a higher level of activity.**
- **A dynamic threshold is calculated, which decides whether the entropy is high enough to be classified as noise.**

The biggest problem with sound activity detection is the hysteresis associated with detection. After detecting a certain sound, we cannot hear other less louder sounds occurring after it. Hence a dynamic threshold has to be calculated based on the statistics of the past. Since the distribution of the sound activity entropy is unknown, a histogram is calculated, for the entropy over a history of around 10 seconds. If the entropy level is in the highest L% range of the histogram, then it considered as activity. However, the entropy level has to go below the 50% range of the past activity to be classified as background noise. Fig. 8 shows the Entropy variation of a short segment of the signal. The two dynamic thresholds are also indicated along with the decision.

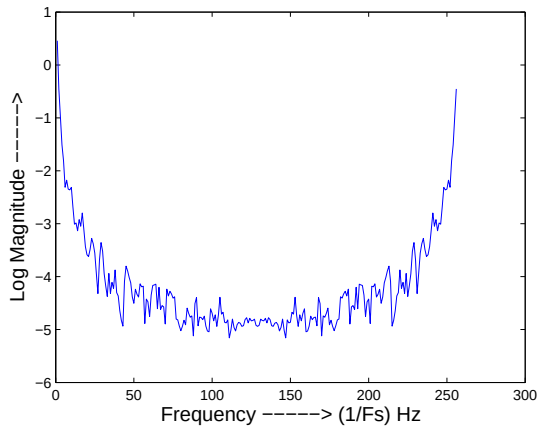


Fig. 6. The log-frequency spectrum of a typical signal

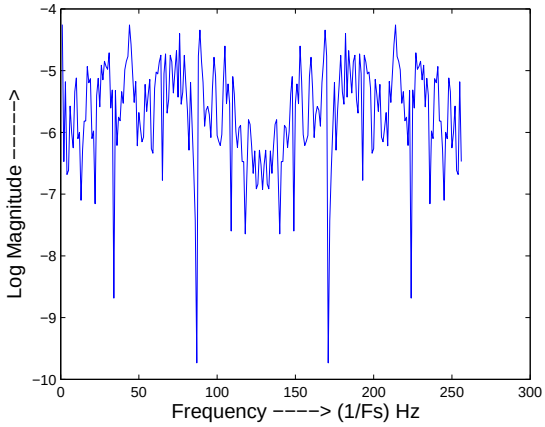


Fig. 7. The log-frequency spectrum of the residual

The value of L decides the region in the Detection Error Trade-off (DET) curve as shown in Fig. 9. Most of the errors that occur are due to the fact that length of the detected activity is either shorter or longer than the annotated activity. Often what is annotated as a contiguous activity is split into several activities or what is annotated as different activities, is detected as a single activity. The DET curve for length independent detection is shown in Fig. 10.

B. Sound Event Classification

Sound event classification has been commonly called auditory scene analysis in literature. The most seminal work on auditory scene analysis is discussed by Bregman [8]. Several methods and several features have been tried for this purpose. Among the most common features used are Bark-filter coefficients, wavelet coefficients, Linear Prediction Coefficients etc. Similarly, Support Vector Machines (SVM), Self Organizing Maps (SOM), Artificial Neural Networks (ANN) and Gaussian Mixture Models (GMM) and their combinations have been used for this purpose.

In our experiments Bark filter coefficients are used as features for classification, because the Bark filters mimic the subjective measurements

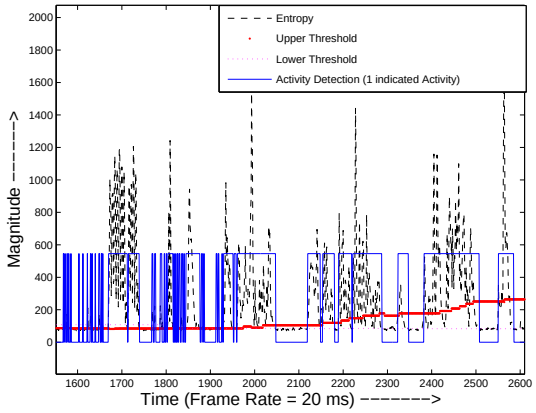


Fig. 8. The Entropy variation for a short segment of the signal. The dynamic thresholds are also shown

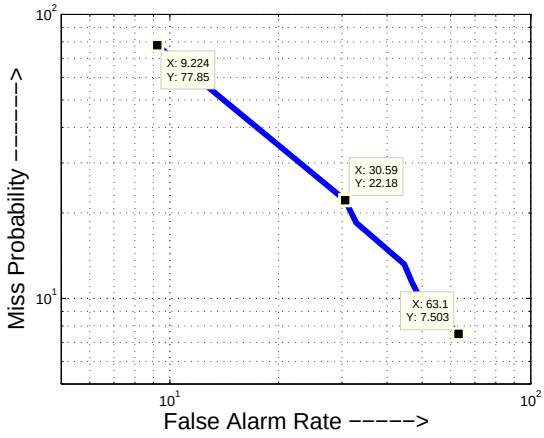


Fig. 9. The log scale plot of the Detection Error Trade-off Curve

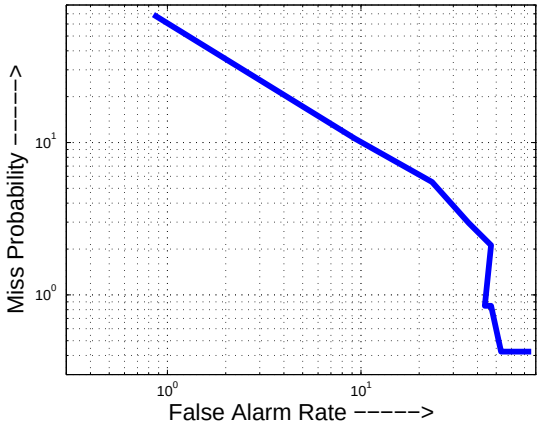


Fig. 10. The log scale plot of the Detection Error Trade-off Curve, independent of length

TABLE II
SOUND EVENT RECOGNITION RESULTS

Sound	Accuracy	False Alarm	Most confused
Voice	67.3%	12.3%	Pencil keep
Telephone ringing	54.3%	0%	Voice
Writing sound	5.2%	0%	Silence
Keyboard typing	45.0%	15.3%	Mouse click
Glass use	89.3%	56.3%	-
Mouse click	43.3%	19.5%	Typing
Phone receiver	63.3%	32.4%	Keyboard Typing
Pencil use	54.3%	12.8%	Voice
Overall	53.8%	23.7%	-

of loudness of the human ear. Since we have sound events with different durations, and since we are classifying contiguous blocks of signals, one-state HMMs are used for event classification where the observation probability distribution is expressed with a GMM. This helps in coupling the likelihoods of each of the frames of the signal to give a single likelihood value.

The most important question in these models is to decide how many mixture components will be employed. This is a difficult problem, especially because there are only a few available sounds, with varying length and duration. The number of Gaussians for each sound class is decided by maximizing the Bayesian Information Criterion (BIC). The sound classes that we used are as follows:

1) **Voice**
2) **Telephone Ringing**
3) **Typing Sound**
4) **Writing sound (with a pencil)**
5) **Placing the glass on the table**
6) **Clicking of the Mouse**
7) **Picking up the phone receiver or putting it back**
8) **Picking up or placing the pencil**

One can see that, a few of these sounds are quite similar and difficult to distinguish even for human beings. However, since the experiment is set-up in a controlled environment, one can expect a decent performance. Table II denotes the results of the recognition of each of the sounds in the list.

As we can see, the accuracy is higher for detection of voice and glass use, but the false alarm is also high for the same two sounds. There is a very high confusion rate between mouse click and typing for example. It is expected that we assume higher priors for more probable events and lower priors for less probable events. However in that case most of the sounds would be classified as voice, because the voice includes sounds similar to each of these mentioned sounds. So the classification is done assuming an equal prior. The overall accuracy may be boosted if the priors were selected according to their probability of occurrence, but then the overall accuracy evaluation would be biased. It does not make sense to use a weighted average for calculation of accuracy, because one wrongly classified event with low probability would affect the overall accuracy greatly.

More work can be done in the direction of a better classifier, using combination of GMM with classifiers like ANNs or SVM. More evaluation is necessary to deal with different time lengths of each of these sound events. Different number of models and modeling the dynamics of the sounds could be other options. A varying length window in order to calculate the Bark coefficients maybe another direction of research.

V. SPEAKER VERIFICATION

The speaker verification system provides a Boolean authentication decision based on the analysis of a speech fragment. Speech-based verification systems can be classified into two main types. In the first approach, the speaker utters a word or a sentence, which is fixed for all authentication attempts. This is called the text-dependent approach. In the more difficult text-independent approach, which is more appropriate for this scenario, the speaker can utter any sentence, and the textual content is not known a priori. For a good survey of speaker verification systems, the reader is referred to [47]. Suffice it to say that all such systems need a speaker model, and an impostor model to determine the decision for authentication. Frequently employed methods for modeling the speaker as well as the impostor include dynamic time warping (DTW),

vector quantization (VQ), Gaussian mixture models (GMM), and hidden Markov models (HMM).

The DTW is used for non-linear aligning of two time sequences and computing the minimum distance between them. Use of DTW in speaker verification system is based on assumption that every speaker is uttering the same word or sentence approximately in the same manner but differently from other speakers. Here the speaker is represented by a template of one or limited set of words or sentences. As such, this method is not adequate for text-independent verification. Vector quantization methods are based on the assumption that the acoustic space of a speaker's speech output can be divided into non-overlapping classes, representing different kinds of sounds, for example phonemes. Each class is defined by one vector, centroid and so each speaker is represented by a set of these classes, thus by his own codebook of vectors. In the GMM approach, the codebook vectors are the means of the Gaussian distributions. Here, the noise around each mean is assumed to be normally distributed. Each speaker is represented by a Gaussian mixture density, which is weighted linear combination of Gaussian distributions of each speaker's acoustic class. Thus speaker is represented by a set of weigths, means and variances. In the HMM approach, the speech dynamic is modeled by a Markov model, where the states are modeled by codebooks of the VQ (discrete HMM) or by Gaussian mixture densities (continuous and semi-continuous HMM). In the particular case of text-independent verification systems, ergodic models are preferred where all interstate-transitions have non-zero probabilities.

In this work, we follow the GMM approach based on the results reported in [4], [51]. First, a number of features are extracted from the input signal. Following [20], we use Mel-filter cepstral coefficients (MFCC) by applying the following transformations:

- **Preemphasis filter**
- **Division of signal into frames**
- **Fast Fourier transformation for obtaining frequency spectrum**
- **Logarithmic transform**
- **Application of Mel-filter banks to the spectrum**
- **Discrete cosine transform**

In speech recognition, usually 13 coefficients are selected from the MFCC. The first and second derivatives (i.e. velocity and acceleration) are added to these coefficients to indicate the history and evolution of the signal, resulting in 39-dimensional feature vectors. N-dimensional feature vector implies using N-dimensional Gaussian distributions, thus the N-dimensional mean and NxN covariance matrix. Because of sufficient effectiveness in modeling the components are restricted to have diagonal covariances.

Once a speaker model is learned, there are two ways of authenticating a particular speaker [47]. In the first approach, a threshold is selected for the probability $P(\lambda^t|O)$, where λ denotes the model parameters of the target speaker, and O is the observed signal. In the second approach is a treshold selected to the ratio of proability of the genuine speaker and probability of the impostor model, which is trained on all speakers in the system except the genuine speaker class. This implies that for every person in the system, two models will be trained. In the case of sufficiently many subjects, a single and generic impostor model can be employed. The implementation of the GMM approach is done by using the Hidden Markov Models Toolkit (HTK) [66]. The GMM was built as an HMM with just one state, as shown in Fig. 11.

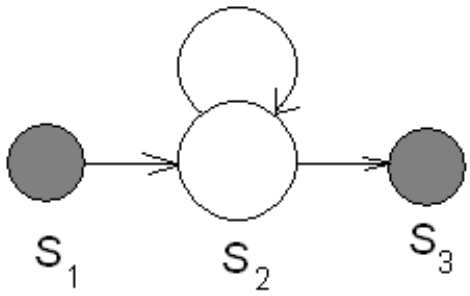


Fig. 11. One-state HMM

VI. Contact-Based Biometrics

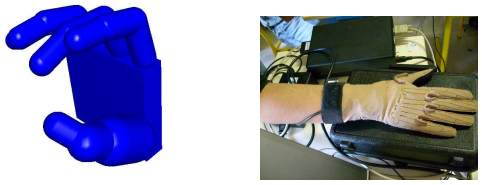
The concept of Contact-Based Biometrics derives from the simple observation that every person handles the objects of the surrounding environment quite differently. For example, the action of picking up a glass or holding a knife depends on the physiological characteristics of each person and the way that this person is used to manipulate objects. Contact-Based Biometrics can also been thought as a specialized part of Activity-Related Biometrics for every activity which involves an object.

In the context of this project we intend to investigate the feasibility of such biometric features in user authentication applications. The proposed approach exploits methods from different scientific fields, such as collision detection and pattern classification, to solve the problem of authentication. The major parts of the final implementation scheme are the setup of a 3D virtual environment, the registration of the user and the objects in this environment, the extraction of collision features during an action between the user and an object and the classification procedure.

A. 3D Environment Setup and Model Registration

Collision detection algorithms can only be used in a 3D environment with full knowledge of the geometry of each object. The virtual environment of the presented pilot requires only the 3D representation of the user's hand and each object that is of interest. The user's hand is modeled as a set of five fingers connected to the palm, which is modeled as a simple rectangle (Figure 12(a)). Each finger has four degrees of freedom (DOF) and consists of three phalanxes which are modeled as simple capsules.

For the registration of the hand we used the CyberGlove® (http://www.immersion.com/3d/). The CyberGlove® (Figure 12(b)) provides the angles between the phalanxes of the hand, so it is possible to reconstruct the 3D representation of the hand. Note, that the virtual representation of the hand is not perfectly accurate because the size of the fingers and the phalanxes are not known. In order to satisfy the requirements of a realistic pilot we cannot make any assumptions or measures on the user so this inaccuracy is considered as noise.



(a) (b)

Fig. 12. (a) The 3D representation of the hand. (b) The CyberGlove®.

The objects of the environment can be registered using computer vision techniques for object tracking. However, it is not absolutely necessary to have an accurate representation of the object in the virtual environment. In particular for rigid objects, which are typically encountered in an office environment, we can simplify the geometry of the object using a priori information. This simplification is possible as the real shape of each object is mostly related to the specific action that is used and not to the way it is handled. For example, a glass can be represented by a cylinder since the user grabs only the outer surface of the glass.

B. Collision Feature Extraction

The classification features consist of any information that can be acquired by employing state-of-the-art algorithms for proximity queries. These include penetration depth [24], closest distance [26], [32], contact points etc. The literature in the field is vast and there are numerous algorithms to accurately perform queries in real-time. The interested reader is directed to [17], [33], [57], [58], [60] for further details. For our purposes we used the algorithms for rigid convex objects [59], [60] of the software package SOLID (http://www.dtecta.com/).

Proximity queries are performed between the object and every finger of the user's hand. Each query refers to either of two states, collision or no collision between the two virtual shapes. For example penetration depth can only be calculated when two objects intersect since it is always zero otherwise. However, in a user-object interaction scheme it is necessary to continuously produce discriminant feature samples. Thus, any proximity query as a single feature would not provide adequate information to a classifier.

In the proposed method we employ the combination of the penetration depth and the closest distance, depending on the collision state, to define the feature space. The penetration depth and the closest distance are usually described as 3D vectors in virtual simulations. However, in our case we prefer to describe them as the pair of points (p_{finger}, p_{object}), one on the finger and the other one on the object, that define the respective vector $\mathbf{v} = p_{finger} - p_{object}$. This way the 3D position of each finger affects the values of the feature vector, while $\mathbf{v}$ would only describe the relative direction which is most probable to be similar even for different fingers. Let pd_k and cd_k denote the points of the penetration depth and the closest distance respectively for either the finger or the object k. The feature sample $f_e(i, O)$ for the finger e and the object O on the i-th frame is

$$f_e(i,O) = \begin{cases} (pd_e, pd_O), & \text{e and O collide} \\ (cd_e, cd_O), & \text{e and O do not collide} \end{cases}$$

The final feature vector $F = \bigcup_e \{f_e\}$ is formed using the collision information from all the five fingers and is a 30-dimensional vector.

VII. Continuous Face Authentication

With the rapid increase of video surveillance equipment and webcam usage, it became necessary to develop robust recognition algorithms that are able to recognize people using video sequences, which not only provide abundant data for pixel-based techniques, but also record the temporal information. This project inspects two complementary approaches to face biometrics from continuous video, detailed in this section and the next.

The processing for the face and facial motion analysis modules starts with detecting the face. We use the OpenCV face detection module that relies on the adaboosted cascade of Haar features, i.e. the Viola-Jones algorithm for this purpose [61]. The face camera is positioned so that the face image roughly covers a 150×150 pixel area, which changes greatly as the subject moves around.

One of the assumptions we have in the face authentication module is that the statistical models that incorporate general face information are trained offline, prior to the actual experimental setup. This means that the bulk of the training database should consist of external data. For this purpose, we have used the world model of 300 face images that accompany the BANCA database [3], enriched with one gallery image per enrolled person. This is a realistic assumption, and since the gallery is acquired with different illumination conditions as well, the actual experimental environment presents a formidable challenge, with completely uncontrolled illumination under ordinary (and poor) office lighting.

For continuous face authentication, we take a straightforward approach. The detected faces are cropped, rescaled to a fixed size, projected to a previously computed subspace, and compared to the templates residing in the gallery. For controlling the illumination, we apply a image enhancement procedure proposed by Savvides and Kumar [54]. In this procedure, the pixel intensities are mapped to a logarithmic range, which nonlinearly allocates a broader range to dark intensity levels, increasing the visibility.

The subspace is found by applying the Karhunen-Loeve transform to the enhanced training set. The matching of a claim with a gallery image can be achieved by thresholding a Mahalanobis-cosine distance between projected vectors. If the subspace-projected query is denoted by $\boldsymbol{u} = [u_1 u_2 \dots u_p]'$ and the subspace-projected gallery template is denoted by $\boldsymbol{v} = [v_1 v_2 \dots v_p]'$, denote their corresponding vectors in the Mahalanobis space with unit variance along each dimension as:

$$m_i = \frac{u_i}{\sigma_i} \tag{21}$$

and

$$n_i = \frac{v_i}{\sigma_i} \tag{22}$$

where σ_i is the standard deviation for the i^{th} dimension of the p-dimensional eigenspace. Then the Mahalanobis cosine distance is given by [49]:

$$d_{MC}(\boldsymbol{u}, \boldsymbol{v}) = cos(\theta_{mn}) = \frac{mn}{|m||n|} \tag{23}$$

A. Adaptive Cropping

The preprocessing of the external database is not replicated in our acquisition conditions. This means that the eigenspace projection that models the variation in aligned face images is not necessarily the ideal projection for a given query image. To remedy this situation, we apply an adaptive cropping algorithm that fine-tunes the face detection result so as to minimize the reprojection error e. Assume the eigenspace is denoted with $[\lambda, \boldsymbol{e}]$, where λ stands for the sorted eigenvalues and $\boldsymbol{e}$ are the corresponding eigenvectors. The projection of query $\boldsymbol{x}$ to the eigenspace is:

$$\boldsymbol{u}_{(p\times 1)} = \boldsymbol{e}'_{(p\times d)}(\boldsymbol{x}_{(d\times 1)} - \mu_{(d\times 1)}), \tag{24}$$

where μ denotes the data mean, and the subscripts indicate dimensionality. The reprojection error is given by:

$$e = ||\boldsymbol{x}_{(d\times 1)} - \boldsymbol{e}_{(d\times p)}\boldsymbol{u}_{(p\times 1)} + \mu_{(d\times 1)}||. \tag{25}$$

The pseudocode of the algorithm is given in Fig. 13. Fig. 14 shows the cumulative effect of illumination correction and adaptive cropping on a sample frame.

```
algorithm Adaptive_Cropping(faceImg)
    cropping ← [0,0,0,0]
    oldError ← Infinity
    found = False
    cropDir = 1
    while NOT found
        oldError ← newError
    /*Crop the image in one of four directions*/
        cropping(cropDir) ← cropping(cropDir) + 1
        croppedImg ← crop(faceImg,cropping)
    /*Scale to fixed size*/
        scaledImg ← scale(croppedImg)
    /*Illumination normalization*/
        normalizedImg ← logTransform(scaledImg)
    /*Projection*/
        projImg ← eigenVectors'*(normalizedImg-meanImg)
    /*Re-projection into the original space*/
        reprojImg ← (eigenVectors*projImg)+meanImg
    /*Update the error*/
        reprojError = norm(reprojImg - normalizedImg)
        if reprojError < oldError
            newError ← reprojError
        else
    /*Reverse the cropping*/
            cropping(cropDir) ← cropping(cropDir) - 1
        end
    /*Update the next cropping direction*/
        cropDirection ← mod(cropDirection+1,4)
        found ← (updated in the last cycle of four directions)
    end
    return cropping
end
```

Fig. 13. Adaptive Cropping Algorithm

Fig. 14. a) The original captured frame. b) The illumination compensated image. c) The result of the adaptive cropping

B. Probabilistic Matching

The activity model necessitates a short video sequence to be recorded for training purposes. This allows us to use a larger training set for the face authentication module as well. For each subject in the gallery, one sequence of recordings is processed with the face detection and adaptive

cropping modules. The ensuing cropped images are projected to the Mahalanobis space, and modeled with a mixture distribution.

The general expression for a *mixture model* is written as

$$p(\boldsymbol{x}) = \sum_{j=1}^{J} p(\boldsymbol{x}|\mathcal{G}_j)P(\mathcal{G}_j) \tag{26}$$

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. 26 are Gaussian distributions:

$$p(x|\mathcal{G}_j) \sim \mathcal{N}(\mu_j, \boldsymbol{\Sigma}_j) \tag{27}$$

Typically, the covariance expression is restricted in MoG models to control the complexity of the model, as a diagonal covariance scales linearly with dimensionality, whereas a full covariance scales quadratically. In this work we use the factor analysis approach to model the covariance, where 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 $\boldsymbol{\epsilon}$:

$$\boldsymbol{x} - \boldsymbol{\mu}_j = \boldsymbol{\Lambda}_j \boldsymbol{z} + \boldsymbol{\epsilon}_j \tag{28}$$

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. 26 with its FA formulation.

To learn the distribution of training faces of a single class, we use the incremental mixtures of factor analysers (IMoFA) algorithm, which automatically determines the number of components in the mixture, and tunes the latent variable dimensionality for each mixture component separately. For more details, the reader is referred to [53]. The ensuing model for the subject is $(\boldsymbol{\Lambda}_j, \boldsymbol{\mu}_j, \boldsymbol{\epsilon}_j, \pi_j)$, with π_j being the component prior, and j is the index for mixture components. The authentication of a normalized and projected image $\boldsymbol{x}_t$ is effected by checking a pre-fixed threshold:

$$p(x_t|\mathcal{G}) \geq \tau \tag{29}$$

At any point in time, the continuous face authentication module evaluates the most recent frame, and returns a Boolean decision. The threshold τ depends on the Mahalanobis space dimensionality, and scales approximately linearly with it. For a 300-dimensional Mahalanobis space, we have used a threshold of -400 for the log-likelihood, a higher value will reject more frames and ensure a more secure system, whereas a lower value will favour user convenience over security. It is also possible to base the decision on all the frames up to time t, by using any classifier combination method.

VIII. Behavioral Face Biometrics

The previous section dealt with the static facial appearance, ignoring the behavioral cues that can be potentially useful for discriminating identities. Recently there is much attention to biometric systems that exploit temporal information in videos, and most of the proposed approaches involve a heterogeneous mixture of techniques. These approaches can roughly be classified into the following categories:

- Holistic approach: This family of techniques analyze the head as a whole, by extracting the head displacements or the pose evolution. In [30] Li et al. propose a model-based approach for dynamic object verification and identification using videos. In 2002, Li and Chellappa were the first to develop a generic approach for simultaneous object tracking and verification in video data, using posterior probability density estimation through sequential Monte Carlo methods [29]. Huang and Trivedi in [19] describe a multi-camera system for intelligent rooms, combining PCA based subspace feature analysis with Hidden Markov Models (HMM). Liu and Cheng proposed a recognition system based on adaptive HMMs [35]. They first compute low-dimensional feature vectors from the individual video frames by applying a Principal Component Analysis (PCA); next they model the statistics of the sequences and the temporal dynamics using a HMM for each subject. In [1] Aggarwal et al. have modeled the moving face as a linear dynamical system using an autoregressive and moving average (ARMA) model. The parameters of the ARMA model are estimated for the entire database using the closed form solution. Recently, Lee et al. developed a unified framework for tracking and recognition, based on the concept of appearance manifold [28]. In this approach, the tracking and recognition components are tightly coupled: they share the same appearance model.
- Feature based approach: The second group of methods exploits the individual facial features, like the eyes, nose, mouth and eyebrows. One of the first attempts to exploit facial motion for identifying people is presented by Chen et al. in [12]. In their work, they propose to use the optical flow extracted from the motion of the face for creating a feature vector used for identification.
- Hybrid approach: These techniques use both holistic and local features. Colmenarez et al. in [14] have proposed a Bayesian framework which combines face recognition and facial expression recognition to improve results; it finds the face model and expression that maximizes the likelihood of the test image.

This section proposes a new person recognition system based on temporal features from facial video. As in the previous section the face area is first detected in each frame of the video. The registration, or the alignment problem, however, has different criteria to satisfy. Since we will track the features, the alignment is not absolute, but relative to the previous frame, minimizing a mean square error measure. For aligned faces, the optical flow is calculated from consecutive frames, and used as feature vectors for person recognition.

Once the faces are detected with the Viola-Jones method, a representation called the "integral image" is created using Haar-like features.

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.

Next the resulting image is cropped as shown in Fig. 15 based on anthropological measures to limit the image to facial features that exhibit more motion.

Fig. 15. Detected and cropped face images in two frames.

Face alignment was required due to the simple fact that we wanted to focus our attention on motion of local features from the face such as the lips and the eyes. If this step is not performed before feature extraction, global motion of the head significantly affects the results. Alignment of the faces detected in two different frames was carried out by minimizing the mean square error of the integral image difference:

$$\arg\min \frac{1}{M \times N} \sum_{i=1}^{M} \sum_{j=1}^{N} (I_1(i,j) - I_2(i,j))^2 \tag{30}$$

where,... Fig. 16 shows two facial images found in consecutive frames aligned with this method.

Fig. 16. Two facial images aligned and superimposed.

We have decided to use optical flow vectors for person recognition, calculated by the Lucas-Kanade technique [36], which uses the spatial intensity gradient of the images to guide the search for matching locations, thus requiring much less comparisons with respect to algorithms that use a predefined search pattern or search exhaustively. Then block means are

taken to reduce the size of the feature vector to standard dimensionality of 200. Fig. 17 shows the optical flow computed from the images aligned in Fig. 16.

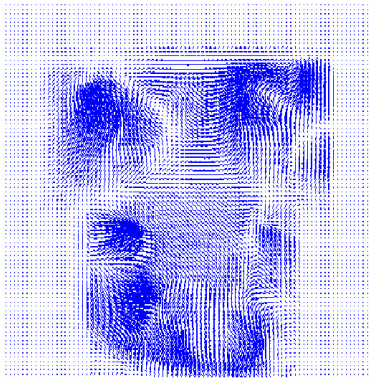


Fig. 17. Optical flow from consecutive frames.

IX. Continuous Activity - Related Biometric Authentication

Among the project's prominent objectives is to investigate the effectiveness and applicability of activity - related biometric technologies. Activity - related biometrics is a novel and innovative concept in biometric user authentication and refers to biometric signatures extracted by analyzing the response of the user to specific stimuli, while performing predefined but natural work - related activities. The novelty of the approach lies in the employment of dynamic features extracted by the moving human model as biometric signal, as well as in the fact that the biometric measurements will correspond to the user's response to specific events, being, however, fully unobtrusive and fully integrated in the user's workspace. The activity - related biometric authentication module evaluates the fundamental assumption that each user's dynamic behavioral profile contains unique intrinsic characteristics that can be used for authentication. Furthermore, a reliable implementation of an activity - related biometric authentication system is ideal for continuous user authentication, thus alleviating the main limitation of some successful state-of-the-art approaches (fingerprint, iris etc.) which cannot be recovered once forged.

In the following, the modules and methods that were implemented to perform activity - related authentication will be described. In addition to that, the pilot setup and the experimental procedures followed in order to evaluate activity - related biometrics will be presented.

A. Activity Detection and Recognition Module

As stated above, the user's dynamic profile extraction is based on the response to specific environment - generated stimuli. Any human behavior is associated to some action or activity. The aim of stimuli generation is to trigger the execution of specific actions by the user, upon which his behavioral profile can be then calculated. It is therefore clear that the extraction of the activity - related features must be preceded by an action detection, segmentation and recognition procedure. This goal is achieved by means of a multimodal approach that uses the output of the sound event recognition Module, the Object Occlusion Tracking Module and the body motion tracking module along with a Coupled Hidden Markov Model formulation in order to detect the generation of the stimuli and segment the user's response (action). The segmentation output of the Activity Recognition Module can be then fed to the Activity - Related Biometric Authentication Module. Fig. 18 illustrates the above inter - module relationships.

Numerous relevant approaches for activity recognition have been reported in the literature using object manipulation context information [43], [46], [64] and/or object trajectory information in the given scene [5], [31]. Sound event detection has also been previously employed to assist inference of ongoing activities [55], [63].

The proposed method for Activity Recognition is based on the detection of three different kinds of Scene Events occurring in the scene: Sound Events (e.g. Phone Ringing), detected by the sound event recognition Module, Proximity Events (e.g. "Hand close to Glass"), detected by the Human Body Tracking Module along with predefined knowledge of the object positions on the controlled workspace and Object Occlusion Events detected by the respective tracker. An Object Occlusion Event is emitted when some object in the scene is missing from its "normal" position.

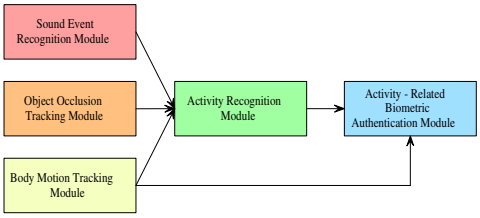


Fig. 18. Module cooperation for Activity Recognition

In order to achieve action recognition, a two - stream Coupled HMM is associated to every action class and trained on two sets of discrete observation symbols (one for each stream) extracted by the primitive events described above (i.e. second layer events). The first set of second layer symbols is a subset of the Sound Event set that can be associated to a particular action. For example, the Phone Conversation Coupled HMM only handles relative sound events (Ringing, Speech, Silence etc.) and disregards the rest (e.g. Writing sound). The observation symbols of the second stream are formed as meaningful (for the particular action class) combinations of the Object Occlusion and Proximity Events of the first layer. For instance, the state "Phone receiver missing" AND "Left Hand close to Head" forms a single second layer event that is used as observation symbol of the second stream of the Phone Conversation CHMM to represent the state of "talking on the phone".

At every timestamp of some activity sequence, first and second layer events are detected and form N double - stream discrete observation sequences, where N the number of actions to be recognized and segmented. Each CHMM uses an overlapping sliding window that goes through its own observation sequence. The size of the sliding windows and the size of overlapping are experimentally defined. The CHMM of each action is trained on manually annotated sequences and a probability threshold is defined, above which the respective action is recognized and a portion in the size of the sliding window is segmented and fed to the Activity Biometrics Module. Fig. 20 graphically depicts the Activity Recognition Module. The reason for performing the mapping from first layer events to second layer events is to impose a smaller size on the final observation sets and process the three initial streams of events into only two - stream Coupled HMMs, which results in making training more efficient.

B. Coupled Hidden Markov Models

The need of a Coupled Hidden Markov Model formulation is justified by the fact that Scene Event detection is often erroneous, producing many false alarms, wrong inferences and multiple occlusions over time. Consequently, detected event symbols would better be thought of as the probabilistic output of some underlying process, rather than as deterministic events. Furthermore, Coupled HMMs offer a robust mathematical background for integrating multimodal observations and fusing different but correlated processes (sound events + human activity based events).

Our Coupled HMM implementation is based on the formulation presented by Ara V. Nefian et al. [44], where the hidden nodes of each stream interact and at the same time have their own observations (Fig. 21). The elements of the CHMM (Initial, Transition and Observation probabilities) are described as:

$$\pi(\mathbf{i}) = \prod_s \pi^s(i_s) = \prod_s P(q_1^s = i_s) \tag{31}$$

$$b_t(\mathbf{i}) = \prod_s b_t^s(i_s) = \prod_s P(\mathbf{O}_t^s | q_t^s = i_s) \tag{32}$$

$$\alpha(\mathbf{i}|\mathbf{j}) = \prod_s \alpha^s(i_s|\mathbf{j}) = \prod_s P(q_t^s = i_s | \mathbf{q_{t-1}} = \mathbf{j}) \tag{33}$$

The CHMMs are trained using an EM algorithm, based on the calculation of the forward and backward variables, $a_t(\mathbf{i}) = P(\mathbf{O_1}, ..., \mathbf{O_t}, \mathbf{q_t} = \mathbf{i})$ and $\beta_t(\mathbf{i}) = P(\mathbf{O_{t+1}}, ..., \mathbf{O_T}, \mathbf{q_t} = \mathbf{i})$ respectively, where T the length of the observation sequence:

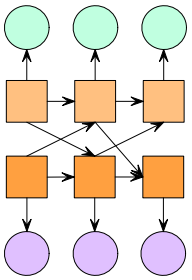


Fig. 19. Coupled Hidden Markov Model structure. Squares denote the hidden nodes of each interacting process and circles the associated observable outputs

$$a_1(\mathbf{i}) = \pi(\mathbf{i})b_1(\mathbf{i}) \tag{34}$$

$$a_t(\mathbf{i}) = b_t(\mathbf{i}) \sum_j \alpha(\mathbf{i}|\mathbf{j})a_{t-1}(\mathbf{j}) \tag{35}$$

for $t = 2, 3, ..., T$

$$\beta_1(\mathbf{i}) = 1 \tag{36}$$

$$\beta_t(\mathbf{j}) = \sum_i b_{t+1}(\mathbf{i})\alpha(\mathbf{i}|\mathbf{j})\beta_{t+1}(\mathbf{i}) \tag{37}$$

for $t = T, T-1, ..., 2$

The probability of the r_{th} observation sequence O_r of length T_r is computed as $a_{r,T}(N_1, N_2, ...N_S) = \beta_{r,1}(1, ..., 1)$

The scaled version of the forward and backward variables ($\hat{a}$,$\hat{\beta}$) [48] obtained in the E step are used to re-estimate the transition and observation parameters as follows:

$$\tilde{\alpha}^s(i|\mathbf{j}) = \frac{\sum_r \sum_{\mathbf{i},s.t.i_s=i} \sum_t \hat{a}_{r,t}(\mathbf{j})\alpha(\mathbf{i}|\mathbf{j})b_{r,t+1}(\mathbf{i})\hat{\beta}_{r,t+1}(\mathbf{i})}{\sum_r \sum_t \hat{a}_{r,t}(\mathbf{j})\hat{\beta}_{r,t}(\mathbf{j})\frac{1}{c_t}} \tag{38}$$

$$\tilde{b}_i^s(k) = \frac{\sum_r \sum_{\mathbf{i},s.t.i_s=i} \sum_{t,s.t.O_t^s=k} \hat{a}_t(\mathbf{i})\hat{\beta}_t(\mathbf{i})\frac{1}{c_t}}{\sum_r \sum_t \sum_{\mathbf{i},s.t.i_s=i} \hat{a}_t(\mathbf{i})\hat{\beta}_t(\mathbf{i})\frac{1}{c_t}} \tag{39}$$

where c_t the scaling coefficient for time t.

The number of states has been defined taking into consideration the inherent structure of each action. For instance, the Phone Conversation action consists of the "natural" states "Ringing" - "Reach Phone"- "Bring close to Head" - "Speech" - "Hang Up", upon which various second layer events can be defined.

C. Activity - Related Biometric Authentication Module

The aim of the activity - related biometric authentication module is to receive the dynamics of the human posture produced by the body motion tracking module on some user action segmented by the Activity Recognition Module and output some authentication results (Fig 18). Within this project we would like to evaluate the assumption that behavior can be employed as biometric signal as well as the hypothesis that our belief measure on the user's identity increases with time. Furthermore, various work - related motions should be tested with regard to their discriminative power.

Related work includes several model - based and feature - based methods for human gait identification and authentication [62], [13]. Key stroke dynamics have been also employed for activity - related person authentication [42]. To our knowledge, activity - related person authentication based on environment generated stimuli and work - related activities is a completely novel concept and has never been implemented before.

The output of this module for a particular action could either be a strict authentication result (Accepted/Rejected) or a belief measure that can be integrated with future partial inferences of the same modality and/or inferences of other modalities to converge in a final authentication result at later time stamps (Continuous Authentication). The latter approach seems more promising, as the user's "natural" behavior can be more reliably confirmed on multiple action instances. In general, a user's way

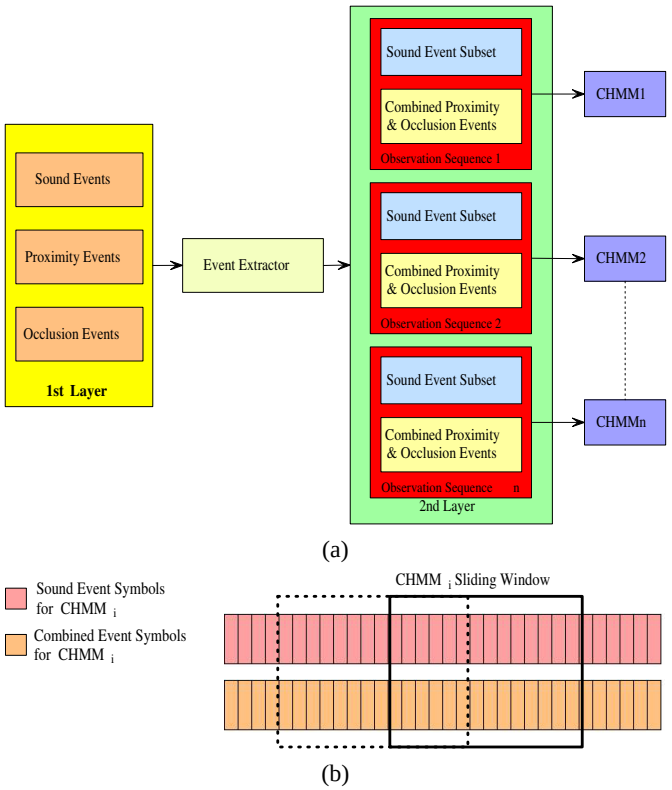


Fig. 20. a) Event Extraction b) Sliding Window for CHMM

of execution of some motion can diverge from its usual dynamics on single instances depending on various factors (psychological condition, unusual environmental conditions etc.). Despite that, it can be assumed that over longer periods of times where multiple instances of many actions take place, the user's identity could be reliably inferred.

The Activity - Related Biometric Authentication Module assumes a mapping of a user's behavior to his identity, therefore tools, methods and features that have been used for action and gesture recognition can be applied. In this implementation the body joint angles and position of the central point of the human model (III) and their derivatives are used as features for modeling the user's natural way of executing some action, since those features can powerfully represent the human model posture and its dynamics. Principal Component Analysis for each action class is used to reduce the dimensionality of the feature vector.

For biometric authentication Hidden Markov Models with Multivariate Gaussian outputs are used to capture the spatio - temporal dynamics of the human behavior. Standard HMM classification is performed by assigning one model to every individual enrolled in the authentication system. Given some extracted observation sequence $O_{1:T}$ of length T associated to a segmented action, and the set of HMMs $\lambda_i, i = 1, \ldots, N$ where N the number of enrolled users, the probability $P(O|\lambda_i)$ is calculated for all HMMs. By assigning an authentication threshold to each user's HMM, direct authentication results based on single actions can be obtained. A more promising option is propagating all the above probabilities to an integration module that emits authentication results on longer periods of activity. Fig. 21 graphically represents the Activity - Related Biometric Authentication Module.

X. Integration of Decisions

A typical authentication system presents a DET (Detection Error Trade-off) curve which enables a system to select a point on the curve to trade off between security and ease of use of a system. However, a continuous authentication system needs to traverse this DET curve based on the current situation. If the system is confident based on past inferences, temporary drops in the probability of the target class should not cause the rejection of the user. However if there is an elongated period of diffidence about the authenticity of the target person, then the system should be able to reject the person eventually.

The second problem is the integration of the inferences from the different modalities. Each mode produces different inferences with a

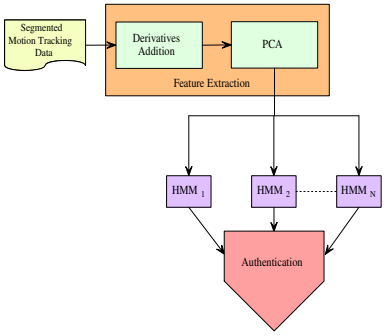


Fig. 21. Activity - related biometric feature extraction and authentication

different probability and these inferences are available at different points in time. There is the additional complication of assessing the reliability (and consequently the relative weight) of each modality. This problem is termed in the literature as "Holistic Fusion".

Among previous work done on holistic fusion, the most significant are Zhang *et al.* [67] and Kittler *et al.* [25]. Zhang *et al.* suggested a two state Hidden Markov Model, where the two states are "safe" and "attacked". A decay factor was proposed, which exponentially weighted over the previous observations, as well as weighted sums to integrate over modalities, where the weights were the assessed reliabilities of the modalities. The area under the Receiver Operating Characteristics (ROC) curve for each modality is used to quantify reliability. The approach we will present now is similar in some respects to this method, but it does not use HMMs.

Let λ_Ω be the model of 'target' person, the person whom we want to authenticate. Let λ_m be one among the M impostor models. Let O_t^n be the t^{th} observation in time, among Γ observations from the n^{th} modality among N modalities. Each module produces the likelihood of λ_Ω given O_t^n, i.e. $p(O_t^n|\lambda_\Omega)$. Since the likelihoods from different modalities have the inherent problem of being in different scales, it becomes difficult to find suitable weights. So the posterior is calculated as follows

$$P(\lambda_\Omega|O_t^n) = \frac{p(O_t^n|\lambda_\Omega) * P(\lambda_\Omega)}{p(O_t^n)} \tag{40}$$

The next question is about calculating the prior $P(\lambda_\Omega)$ and the observation probability $p(O_t^n)$. The observation probability can be given by

$$p(O_t^n) = p(O_t^n|\lambda_\Omega) * P(\lambda_\Omega) + \sum_{m=1}^{M} p(O_t^n|\lambda_m) * P(\lambda_m) \tag{41}$$

How to estimate $P(\lambda_\Omega)$ is an interesting problem. This value is tunable and different points on the DET curve can be achieved by changing this value. Increasing this value makes the system more confident about the authenticity of the subject and thereby increases the false acceptance rate (FAR). Reducing this value increases the false rejection rate (FRR) while decreasing the FAR.

A continuous authentication system is typically used after the authenticity is verified by an independent system. The initial estimate of the prior, $P_0(\lambda_\Omega)$, can be received from this entry system or taken to be an arbitrarily high value. The subsequent values of this prior are calculated as shown below

$$P_t(\lambda_\Omega) = \frac{\sum_{n=1}^{N} P_t(\lambda_\Omega|O_t^n) * P(O_{1:t}^n)}{\sum_{n=1}^{N} P(O_{1:t}^n)} \tag{42}$$

where

$$P(O_{1:t}^n) = \frac{\left(\frac{p(O_i^n|\lambda_\Omega) + \sum_{m=1}^{M} p(O_i^n|\lambda_m)}{M+1}\right)}{\frac{1}{t}\sum_{i=1}^{t}\left(\frac{p(O_i^n|\lambda_\Omega) + \sum_{m=1}^{M} p(O_i^n|\lambda_m)}{M+1}\right)} \tag{43}$$

Now with a time-varying estimate of prior available equation 40 can be combined with 41 and written as shown below.

$$P_t(\lambda_\Omega|O_t^n) = \frac{p(O_t^n|\lambda_\Omega) * P_{t-1}(\lambda_\Omega)}{p(O_t^n|\lambda_\Omega) * P_{t-1}(\lambda_\Omega) + \sum_{m=1}^{M} p(O_t^n|\lambda_m) * P_{t-1}(\lambda_m)} \tag{44}$$

where $\forall m$

$$P_{t-1}(\lambda_m) = \frac{1 - P_{t-1}(\lambda_\Omega)}{M} \tag{45}$$

The prior is updated at every calculation and the confidence of the system depends on all the different modalities. In a system such as the one described in the experiment, it may not be possible to get a new inference from each modality at each instance of time. So the latest inference from each modality is used for re-computing the estimate of the prior, λ_Ω.

The strategy proposed builds a confidence value about the identity of a person. This confidence is in terms of the updated posterior probability. If the different modalities ascribe low confidence to the authenticity of the person, then the overall confidence drops down. But if the modalities provide high confidence to the authenticity, then the overall confidence in the person builds up. At some point, if one of the modalities ascribes low confidence to the authenticity of the target, then it is weighed by how probable the occurrence of such an observation is. So if an observation is not very probable in the model of the entire system, then a lower weight is given in the overall confidence calculation.

If at any point, the user is switched with an impostor, it will take some time for the system to bring down the confidence levels due to the high confidence levels initially built on the user, and the impostor is likely to be authenticated for some time. But the overall confidence will drop eventually, with a speed that depends on the confidence scores of each modality. Using a window-approach that takes into account the last *k* frames in assessing probabilities may be useful in providing a fast decrease under switched persons.

Further testing needs to be done in the case of impostor switching and hysteresis of the system under these circumstances.

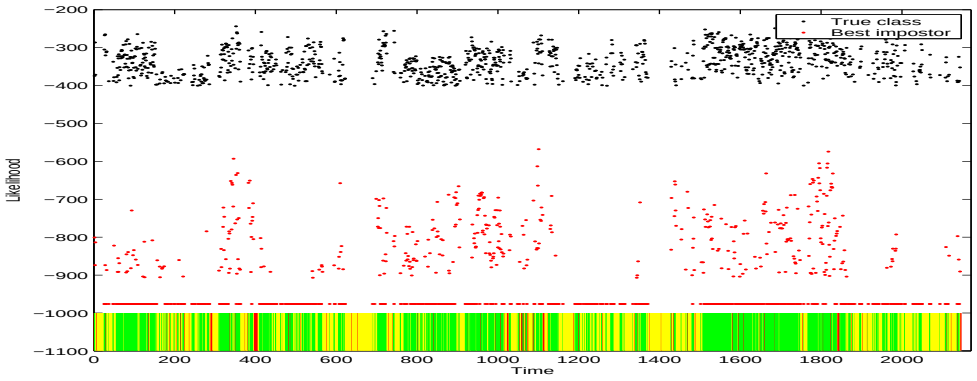
XI. Experimental Results

A. Continuous Face Authentication

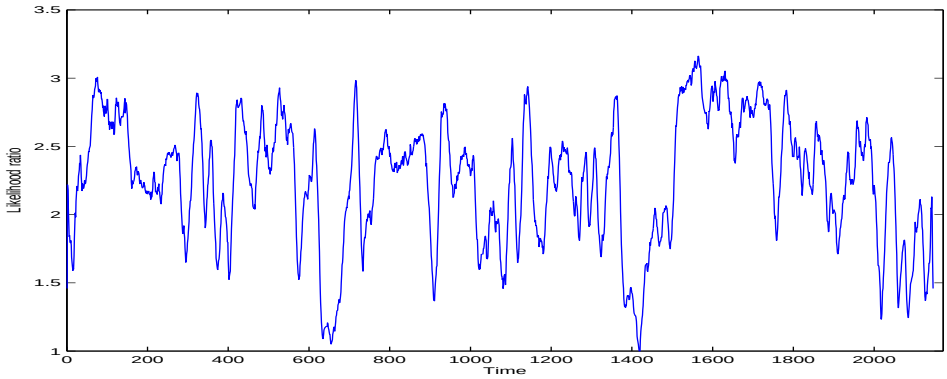
The face authentication module is tested with the recordings of 11 individuals. The first session is used to construct the statistical models for each person. The remaining nine sessions are used for reporting the success of the algorithm. For 99 sessions, the face detection module locates faces 92.3 per cent of the total recording time, with a standard deviation equal to 7.4 per cent. This means that for a 1000 frame session, about 923 face images are processed for authentication. Some of these faces are false alarms, caused by the failure of the Viola-Jones face detector.

In general, the face detection module is robust enough to correctly localize faces during activities like phone conversations. This implies that for these frames, the cropped face area contains the hand and the phone itself. We have observed that the face authentication module frequently stays below authentication threshold for these cases. Fig. 22 shows the authentication result for a single session. The horizontal axis is the time, and the vertical axis is the likelihood value obtained by the class models. Each face is shown as a dot on this plot. We only report the likelihood from the genuine class and the best impostor claim for that frame. The threshold is selected as -400, and the shown sequence justifies this choice nicely. In fact the threshold is optimized on a separate set, but since it strictly depends on the subspace dimensionality, it produces uniformly good results across the test sessions, as shown by the low variance of the results. At the bottom of the figure, a coloured band indicates when faces are not detected in the video (with red), when they are detected but the true class authentication does not follow (with yellow) and correct authentications (with green). The parts with longer bands of yellow are the activities where the face is not isolated or completely frontal.

The complete testing data consists of 91250 frames, recorded from nine sessions per subject, and 11 subjects. For each frame, the best impostor access is selected by evaluating the remaining 10 models. We demostrate the effect of selecting different thresholds in Fig. 23, where the false accept rate and the false rejection rate of the system are plotted for a range of threshold values. For the selected threshold of -400, the system has 0.3 per cent false acceptance rate and 30.1 false rejection rate. This means that for a video sequence with 1000 detected faces, roughly 3 frames would admit impostors, and 700 frames would indicate the presence of the true user. At this level, there is no interpretation of these results. In practice, a session of continuous authentication can



(a)



(b)

Fig. 22. The output of the continuous face authentication module during one session. (a) The likelihood of the genuine and best impostor claims. The band shows correct authentications (green), no authentication (yellow), and no detection (red) cases. (b) The likelihood ratio of the genuine class to the best impostor class for the same session.

operate on a sliding window of frames, where the genuine and impostor likelihoods are compared, and the system outputs a decision at every time slot. Under these controlled conditions (i.e. difficult but similar illumination conditions in training and test sessions), it is obvious that the face modality provides very robust authentication.

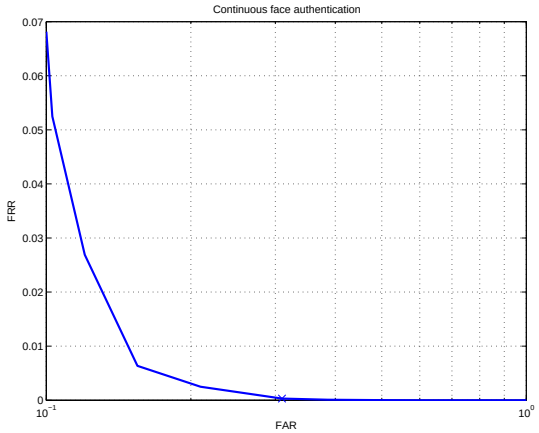


Fig. 23. The receiver operator characteristic curve for a range of thresholds of authentication. The genuine class is evaluated against the best impostor model for each frame. The average values for 99 sessions are reported. The cross indicates the selected threshold for the operating point of the system.

B. Speaker verification

For purposes of training and testing, approximately 20 seconds of speech is recorded during each session in form of a telephone conversation in addition to 40 seconds of speech in form of reading a paragraph

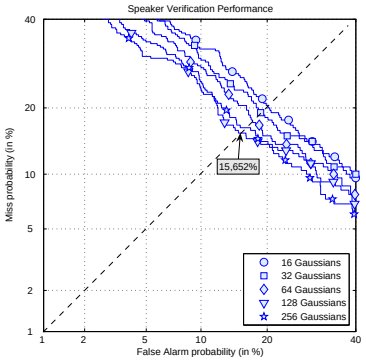


Fig. 24. DET curve for speaker verification module

of written text. 15 subjects have contributed to the database, with 10 recording sessions per subject. The results reported in this section are obtained by training with sessions one to five, and testing with the session six and seven, for 10 subjects. We have evaluated GMMs with different numbers of components.

Fig. 24 shows that the best results are achieved using 128 components for Gaussian mixture densities.

C. Contact-based Biometrics

The experimental setup includes one testing action and eight subjects. In particular, the right hand of each user and the glass of the office were registered in the virtual environment for the action notated as "grabbing the glass". For the classification we implemented standard techniques of pattern recognition. PCA was used to reduce the dimensionality of the feature space while neural networks were trained for the final classification. Each person performed the action 10 times which produced

1000 sample frames on average for each subject due to the high sampling frequency of the CyberGlove®. From these samples 70% were used to train the network and 30% for testing. Fig. 25 displays the final ROC curve of the FAR and FRR rates for the testing data of the eight subjects.

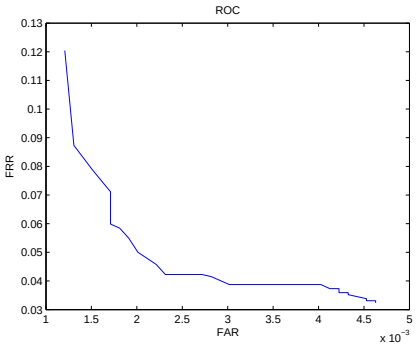


Fig. 25. ROC curve for the action "grabbing the glass" and eight subjects.

The results show that collision features are comparable to other Activity-Related biometrics and therefore comprise a very interesting approach for user authentication.

D. Body Motion Tracking

The body tracker was tested in the office pilot. Two webcams, one frontal and one lateral, recording at 9.5 fps provided the frames onto which the 3D articulated model was projected. 3D body part locations (head, shoulders, elbows and wrists) have been manually annotated in one subject sequence in order to test the tracker performance. The error is expressed as the mean distance between the annotated and the estimated joints. Comparative results between the APF with the common likelihood approach (comprising edges and foreground matching) and our proposal are shown in Fig. 26. In both cases we used the body model and the projection procedure explained in section III-C. Final mean error obtained by our approach for this sequence was 85 mm. Common likelihood evaluation makes the tracker vulnerable to track loss, leading to higher mean error. On the other hand, the divergence measure and the feature-based smoothing of the likelihood approximation make the tracker more robust under our experimental conditions.

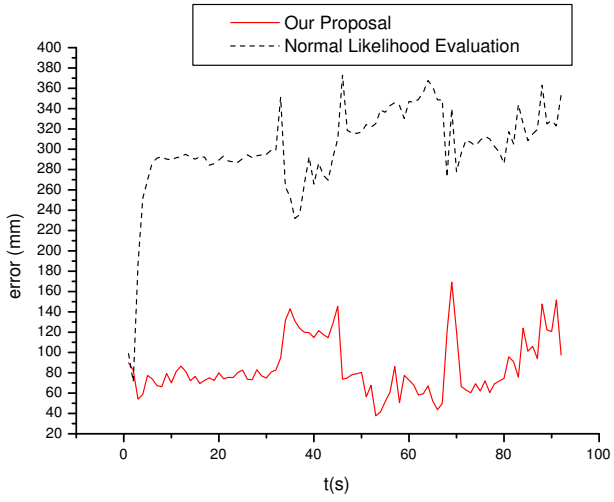


Fig. 26. Comparative results using 3 layers and 200 particles per layer with the normal likelihood aproximation and our proposal.

We found out that some spurious contours due to clothing and objects caused our tracker to fail in its estimation. The aparent motion recorded in the images was very fast in some of the actions required for activity-based recognition. These aparent fast motions caused blurs in the image and abrupt translation of body parts. Since the implemented annealing PF works with contours as most determinant feature, the algorithm was not able to track several of these fast motions. However, it was able to recover some poses after a tracking error. Similarly, we detected that some poses couldn't be retrieved due to self-occlusions caused by the lack of additional views. Therefore, for some of the actions and poses, the problem becomes ill-posed and, as a consequence, more information is needed.

After testing several sequences, it was found that for several non-fast motions good results can be obtained with 3 layers and between 100 and 200 particles per layer. However, a more exhaustive study with ground truth angles must be done under similar conditions in order to refine the likelihood approximation, the annealing parameters and the number of particles.

E. Other Modules

The results of the Sound - based Event Detection module are illustrated in the respective section IV. Testing of the Activity Recognition module and the Behavioral Face Biometrics module remain as future work.

Preliminary results for the Activity - related biometric module reveal a potential of using work - related activities as biometric signals. Experimenting on 7 manually segmented sequences (5 for training and 2 for testing) of the action classes Writing and Phone Conversation, we found out that the true person receives good HMM log - likelihood ranking. Despite that, the need for more accurate and stable 3D Motion Tracking was obvious, as it is the case for most state of the art model - based techniques. Future work includes testing on larger sets and more action classes, with improved motion tracking data. A feature - based approach (direct feature extraction on segmented human blobs) will also be implemented.

XII. Conclusions and Future Directions

In this project we have evaluated several activity related biometric modalities for their relative success in continuously determining and verifying the identity of a user in a typical and non-obtrusive work environment scenario. Apart from more traditional face and speech based verification, facial actions and movement patterns were assessed for authentication. A pilot setup with different action scenarios is defined, and a large database is collected from 15 subjects. Each subject contributed 10 sessions, which are manually annotated by the project group for further evaluation.

The experimental evaluation of all the modalities is not achieved exhaustively, and their possible integration remains to be a future endeavor. The latter is partly due to the success of individual modalities on the restricted pilot setup, which suggests that under closely resembling training and testing conditions there will be no marked benefit under fusion scenarios. However, the results demonstrate that activity-based biometrics is a promising venue for further study.

XIII. Acknowledgements

The authors thank Christophe D'Alessandro and the organization team of the eNTERFACE'08 for all their efforts. Albert Ali Salah is supported by the Dutch BRICKS/BSIK project, and a scientific mission grant from EU COST 2101 Action. Martin Lojka is supported by Ministry of education of Slovak Republic under research project VEGA 1/4054/07 and Slovak Research and Development Agency under research project APVV-0369-07. This work was supported in part by the EC under contract FP7-215372 ACTIBIO.

References

[1] G. Aggarwal, A. Chowdhury, and R. Chellappa. A system identification approach for video-based face recognition. *Proc. Int. Conf. on Pattern Recognition*, 4:176–178, 2004.

[2] M. Arulampalam, S. Maskell, N. Gordon, T. Clapp, D. Sci, T. Organ, and S. Adelaide. A tutorial on particle filters for online nonlinear/non-GaussianBayesian tracking. *Signal Processing, IEEE Transactions on [see also Acoustics, Speech, and Signal Processing, IEEE Transactions on]*, 50(2):174–188, 2002.

[3] E. Bailly-Bailliere, S. Bengio, F. Bimbot, M. Hamouz, J. Kittler, J. Mariethoz, J. Matas, K. Messer, V. Popovici, F. Poree, et al. The BANCA Database and Evaluation Protocol. *LNCS*, pages 625–638, 2003.

[4] P. Balucha. Automatic speaker recognition with gaussian mixture models. Master's thesis, Technical University of Košice, 2006.

[5] F. Bashir, A. Khokhar, and D. Schonfeld. Object trajectory-based activity classification and recognition using hidden markov models. *Image Processing, IEEE Transactions on*, 16(7):1912–1919, July 2007.

[6] N. Boulgouris and Z. Chi. Gait Recognition Using Radon Transform and Linear Discriminant Analysis. *IEEE Transactions ON Image Processing*, 16(3):731, 2007.

[7] C. Bregler and J. Malik. Tracking People with Twists and Exponential Maps. *IEEE Computer Society Conference On Computer Vision And Pattern Recognition*, 1998.

[8] A. Bregman. *Auditory scene analysis*. MIT Press Cambridge, Mass, 1990.

[9] F. Caillette, A. Galata, and T. Howard. Real-Time 3-D Human Body Tracking using Variable Length Markov Models. *British Machine Vision Conference*, 1:469–478, 2005.

[10] J. Canny. A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 8(6):679–698, 1986.

[11] C. Canton-Ferrer, J. Casas, and M. Pardas. Exploiting Structural Hierarchy in Articulated Objects Towards Robust Motion Capture. *Lecture Notes in Computer Science*, pages 82–91, 2008.

[12] L. Chen, H. Liao, and J. Lin. Person identification using facial motion. *Image Processing, 2001. Proceedings. 2001 International Conference on*, 2, 2001.

[13] M.-H. Cheng, M.-F. Ho, and C.-L. Huang. Gait analysis for human identification through manifold learning and hmm. *Pattern Recogn.*, 41(8):2541–2553, 2008.

[14] A. Colmenarez, B. Frey, and T. Huang. A Probabilistic Framework for Embedded Face and Facial Expression Recognition. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1:592–597, 1999.

[15] J. Deutscher, A. Blake, and I. Reid. Articulated body motion capture by annealed particle filtering. *PROC IEEE Comput Soc Conf Comput Vision Pattern Recognit*, 2:126–133, 2000.

[16] A. Doucet, S. Godsill, and C. Andrieu. On sequential Monte Carlo sampling methods for Bayesian filtering. *Statistics and Computing*, 10(3):197–208, July 2000.

[17] C. Ericson. *Real-Time Collision Detection*. Morgan Kaufmann, 2004.

[18] N. Gordon, D. Salmond, and A. Smith. Novel approach to nonlinear/non-Gaussian Bayesian state estimation. *Radar and Signal Processing, IEE Proceedings F*, 140(2):107–113, 1993.

[19] K. Huang and M. Trivedi. Streaming face recognition using multicamera video arrays. *Proc. Int. Conf. on Pattern Recognition*, 4:213–216, 2002.

[20] X. Huang, A. Acero, and H. Hon. *Spoken Language Processing*. Prentice Hall, 2001.

[21] M. Isard and A. Blake. CONDENSATION-Conditional density propagation for visual tracking. *Int. Journal of Computer Vision*, 29(1):5–28, 1998.

[22] A. Jain, S. Dass, and K. Nandakumar. Soft biometric traits for personal recognition systems. *Lecture notes in computer science*, pages 731–738.

[23] A. Kale, N. Cuntoor, and R. Chellappa. A framework for activity-specific human recognition. *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (Orlando, FL)*, 706, 2002.

[24] Y. J. Kim, M. C. Lin, and D. Manocha. Incremental penetration depth estimation between convex polytopes using dual-space expansion. *IEEE Transactions on Visualization and Computer Graphics*, 10(2):152–163, 2004.

[25] J. Kittler, M. Hatef, R. Duin, and J. Matas. On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20:226–239, 1998.

[26] E. Larsen, S. Gottschalk, M. Lin, and D. Manocha. Fast distance queries with rectangular swept sphere volumes. volume 4, pages 3719–3726, 2000.

[27] A. Laurentini. The visual hull concept for silhouette-based image understanding. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 16(2):150–162, 1994.

[28] K. Lee, J. Ho, M. Yang, and D. Kriegman. Visual tracking and recognition using probabilistic appearance manifolds. *Computer Vision and Image Understanding*, 99:303–331, 2005.

[29] B. Li and R. Chellappa. A generic approach to simultaneous tracking and verification in video. *IEEE Transactions on Image Processing*, 11:530–544, 2002.

[30] B. Li, R. Chellappa, Q. Zheng, and S. Der. Model-based temporal object verification using video. *IEEE Transactions on Image Processing*, 10(6):897–908, 2001.

[31] Z. Li, S. Wachsmuth, J. Fritsch, and G. Sagerer. View-adaptive manipulative action recognition for robot companions. *Intelligent Robots and Systems, 2007. IROS 2007. IEEE/RSJ International Conference on*, pages 1028–1033, 29 2007-Nov. 2 2007.

[32] M. C. Lin and J. F. Canny. A fast algorithm for incremental distance calculation. pages 1008–1014, 1991.

[33] M. C. Lin and S. Gottschalk. Collision detection between geometric models: A survey. In *In Proc. of IMA Conference on Mathematics of Surfaces*, pages 37–56, 1998.

[34] J. Liu and R. Chen. Sequential Monte Carlo methods for dynamical systems. *Journal of the American Statistical Association*, 93(5):1032–1044, 1998.

[35] X. Liu and T. Chen. Video-Based Face Recognition Using Adaptive Hidden Markov Models. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1, 2003.

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

[37] J. MacCormick and M. Isard. Partitioned Sampling, Articulated Objects, and Interface-Quality Hand Tracking. *Lecture Notes in Computer Science*, pages 3–19, 2000.

[38] B. Mak, J.-C. Junqua, and B. Reaves. A robust speech/non-speech detection algorithm using time and frequency-based features. *Proc. ICASSP*, 1:269–272, 1992.

[39] F. Matta and J. Dugelay. A behavioural approach to person recognition. *Proceedings of IEEE International Conference on Multimedia and Expo (ICME2006)*, pages 9–12, 2006.

[40] I. Mikic. Human Body Model Acquisition and tracking using multi-camera voxel Data. *PhD. Thesis, University of California, San Diego*, 2003.

[41] J. Mitchelson and A. Hilton. Simultaneous pose estimation of multiple people using multiple-view cues with hierarchical sampling. *In Proc. of BMVC, September*, 2003.

[42] F. Monrose and A. Rubin. Authentication via keystroke dynamics. In *CCS '97: Proceedings of the 4th ACM conference on Computer and communications security*, pages 48–56, New York, NY, USA, 1997. ACM.

[43] D. J. Moorey, I. A. Essaz, and M. H. H. Iiiy. Objectspaces: Context management for human activity recognition. In *Second International Conference on Audio- Vision-based Person Authentication*, 1999.

[44] A. V. Nefian, L. Liang, X. Pi, X. Liu, and K. Murphy. Dynamic bayesian networks for audio-visual speech recognition. *EURASIP J. Appl. Signal Process.*, 2002(1):1274–1288, 2002.

[45] E. Nemer, R. Goubran, and S. Mahmoud. Robust voice activity detection using higher-order statistics in the LPC residual domain. *IEEE Transactions on Speech and Audio Processing*, 9:217–231, 2001.

[46] D. Patterson, D. Fox, H. Kautz, and M. Philipose. Fine-grained activity recognition by aggregating abstract object usage. *Wearable Computers, 2005. Proceedings. Ninth IEEE International Symposium on*, pages 44–51, Oct. 2005.

[47] P. Psutka, L. Müller, J. Matoušek, and V. Radová. *Mluvíme s počítačem česky*. Academia, 2004.

[48] L. R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. pages 267–296, 1990.

[49] N. Ramanathan, R. Chellappa, and A. Roy Chowdhury. Facial similarity across age, disguise, illumination and pose. *Proc. Int. Conf. on Image Processing*, 3, 2004.

[50] L. Raskin, E. Rivlin, and M. Rudzsky. Using Gaussian Process Annealing Particle Filter for 3D Human Tracking-Volume 2008, Article ID 592081, 13 pages. *EURASIP Journal on Advances in Signal Processing*, 2008.

[51] A. Raynolds. Speaker identification and verification using Gaussian mixture speaker models.

[52] P. Renevey and A. Drygajlo. Entropy Based Voice Activity Detection in Very Noisy Conditions. In *Seventh European Conference on Speech Communication and Technology*. ISCA, 2001.

[53] A. Salah and E. Alpaydın. Incremental mixtures of factor analyzers. *Int. Conf. on Pattern Recognition*, 1:276–279, 2004.

[54] M. Savvides and B. Vijaya Kumar. Illumination normalization using Logarithm transforms for face authentication. *Lecture notes in computer science*, pages 549–556.

[55] M. Stager, P. Lukowicz, and G. Troster. Implementation and evaluation of a low-power sound-based user activity recognition system. In *ISWC '04: Proceedings of the Eighth International*

Symposium on Wearable Computers, pages 138–141, Washington, DC, USA, 2004. IEEE Computer Society.

[56] C. Stauffer and W. Grimson. Learning Patterns of Activity Using Real-Time Tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 747–757, 2000.

[57] M. Teschner, S. Kimmerle, G. Zachmann, B. Heidelberger, L. Raghupathi, A. Fuhrmann, M.-P. Cani, F. Faure, N. Magnetat-Thalmann, and W. Strasser. Collision detection for deformable objects. In *Eurographics State-of-the-Art Report (EG-STAR)*, pages 119–139. Eurographics Association, 2004.

[58] F. Thomas and C. Torras. 3d collision detection: A survey. *Computers and Graphics*, 25:269–285, 2001.

[59] G. Van and D. Bergen. Efficient collision detection of complex deformable models using aabb trees. *J. Graphics Tools*, 2, 1997.

[60] G. van den Bergen. *Collision Detection in Interactive 3D Environments*. Morgan Kaufmann, 2003.

[61] P. Viola and M. Jones. Rapid Object Detection Using a Boosted Cascade of Simple Features. *IEEE Computer Society Conference On Computer Vision And Pattern Recognition*, 1, 2001.

[62] L. Wang, T. Tan, H. Ning, and W. Hu. Silhouette analysis-based gait recognition for human identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(12):1505–1518, 2003.

[63] J. A. Ward, P. Lukowicz, G. Troster, and T. E. Starner. Activity recognition of assembly tasks using body-worn microphones and accelerometers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(10):1553–1567, 2006.

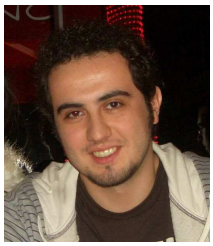
[64] J. Wu, A. Osuntogun, T. Choudhury, M. Philipose, and J. Rehg. A scalable approach to activity recognition based on object use. *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pages 1–8, Oct. 2007.

[65] L. Xu, J. Landabaso, and M. Pardas. Shadow Removal with Blob-Based Morphological Reconstruction for Error Correction. *Acoustics, Speech, and Signal Processing, 2005. Proceedings.(ICASSP'05). IEEE International Conference on*, 2, 2005.

[66] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland. *The HTK Book*. Cambridge University Engineering Department, 2006.

[67] S. Zhang, R. Janakiraman, T. Sim, and S. Kumar. Continuous Verification Using Multimodal Biometrics. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29:687–700, 2007.

Gopal Ananthakrishnan received his Master of Science (Engg) from the Indian Institute of Science, Bangalore, India, in 2007. Currently, he is a PhD candidate in the Royal Institute of Technology, Stockholm, Sweden. His main research interests include Auditory to Articulatory Inversion, Speech production, Analysis of Speech and Audio signals, Perceptual and Neuro-Physiological analysis of hearing, Modeling and Simulation of Perceptual properties of human hearing, Pattern Recognition, On-line Handwriting Recognition.



Hamdi Dibeklioğlu was born in Denizli, Turkey in 1983. He received his B.Sc. degree from Yeditepe University Computer Engineering Department, in June 2006, and his M.Sc. degree from Boğaziçi University Computer Engineering Department, in July 2008. He is currently a research assistant and a Ph.D. student at Boğaziçi University Computer Engineering Department. His research interests include 3D face recognition, computer vision, pattern recognition and intelligent human-computer interfaces. He works with Professor Lale Akarun on 3D Face Recognition.

Martin Lojka was born in Snina, Slovakia, in 1984. He received engineer's degree in 2007. He is currently a PhD student in department of Electronics & Multimedia Communications of Technical University of Košice under the supervision of Doc. Ing. Jozef Juhar, PhD. His research interests focus on algorithms of audio processing in embedded systems.

Adolfo López was in born in Barcelona in 1982. He received his BS in Telecommunication Engineering from Universitat Politècnica de Catalunya (UPC) in Barcelona in 2007. He is currently a PhD student in the Image and Video Processing Group of the UPC under the supervision of professor Josep Ramon Casas. His research interests include Markerless Body Pose and Motion Estimation, Gesture and Motion Recognition and Particle Filter based Visual Tracking.

Serafeim Perdikis received his Diploma in Electrical and Computer Engineering from Aristotle University of Thessaloniki, Thessaloniki, Greece, in 2007. Currently, he is a PhD candidate in the Aristotle University of Thessaloniki and a research fellow with the Informatics and Telematics Institute, Centre for Research and Technology Hellas, Thessaloniki. His main research interests include activity recognition, gesture recognition, human - computer interaction, dynamic biometrics and signal processing. He is also a member of the Technical Chamber of Greece.

Usman Saaed 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.

Albert Ali Salah received his PhD in 2007 from the Dept. of Computer Engineering of Boğaziçi University, with a dissertation on biologically inspired 3D face recognition. This work was supported by two FP6 networks of excellence: BIOSECURE on multimodal biometrics, and SIMILAR on human-computer interaction. His research areas are pattern recognition, biometrics, and multimodal information processing. He received the inaugural EBF Biometrics Research Award in 2006, and joined with the Signals and Images group at CWI, Amsterdam as a BRICKS scholar. Recent scientific assignments include program committee memberships for BIOID'08, biometrics track of ICPR'08, and ICB'09.

Dr. Dimitrios Tzovaras is a Senior Researcher (Grade B) at the Informatics and Telematics Institute. He received the Diploma in Electrical Engineering and the Ph.D. in 2D and 3D Image Compression from the Aristotle University of Thessaloniki, Greece in 1992 and 1997, respectively. Prior to his current position, he was a senior researcher on the Information Processing Laboratory at the Electrical and Computer Engineering Department of the Aristotle University of Thessaloniki. His main research interests include information management, multimodal data fusion and knowledge management. His involvement with those research areas has led to the coauthoring of over thirty articles in refereed journals and more than eighty papers in international conferences. He has served as a regular reviewer for a number of international journals and conferences. Dr. Tzovaras is an Associate Editor of the Journal of Applied Signal Processing (JASP).

Athanasios Vogiannou received his Diploma degree in electrical and computer engineering from Aristotle University of Thessaloniki, Thessaloniki, Greece, in 2007. Currently, he is a PhD candidate in the Aristotle University of Thessaloniki and a research fellow with the Informatics and Telematics Institute, Centre for Research and Technology Hellas, Thessaloniki. His main research interests include virtual reality, collision detection, physically based modeling, real time simulations and computer vision. He is also a member of the Technical Chamber of Greece.

Tracking-dependent and interactive video projection

Matei Mancas (1), Donald Glowinski (2), Pierre Bretéché (3), Jonathan Demeyer (1), Thierry Ravet (1), Gualtiero Volpe (2), Antonio Camurri (2), Paolo Coletta (2)
(1) FPMS, Mons, Belgium (2) Casa Paganini/InfoMus Lab, Genova, Italy
(3) Laseldi Lab, Montbéliard, France

Abstract

Gestures' expressivity, as perceived by humans, may be related to the amount of attention they attract. In this paper, we present three experiments that quantify behavior saliency by the rarity of selected motion and gestural features in a given context. The first two ones deal with the current quantity of motion of a person's silhouette compared to a brief history of his quantity of motion values and with the current speed compared to a brief history of the person's speed. The third one focuses on the motion speed of a person compared to the motion speed of other persons around him. Considering both features (speed and quantity of motion) and contexts (space and time), we compute an attention index providing cues on the behavior novelty. This can be considered as a preliminary step to an expressive gesture analysis based on behavioral. In order to achieve accurate tracking, a fusion between color and IR camera streams is achieved. This fusion let us have a robust tracking system with respect to illumination and partial occlusions issues.

***Index Terms*—computational attention, saliency, rarity, data fusion, tracking, gestures.**

I. PRESENTATION OF THE PROJECT

A. Introduction: towards a context-based gestural analysis

A lot of research effort has been devoted to robustly track humans in a scene and to analyze their gestures in order to individuate and characterize their behavior. Gestural analysis, often applies in situations where either the human on which the analysis is carried on is previously selected or the same kind of analysis is performed to all the subjects that can be distinguished in the scene. A recent field of research aims at investigating collective behaviors [1]. Still, the object of the analysis is already defined and the work mainly focuses on characterizing collective displacements. The possibility of dynamically selecting the person to carry analysis on or to adapt analysis to the current behavior of a person in a context-dependent way would open new directions for gesture research. Human beings naturally show the capacity to dedicate limited perceptual resources to what is of particular interest. However, computers capacity to exhibit a behavior worthy of attention remains very limited.

B. Computational attention interest in Human-Computer Interfaces (HCI)

In many real situations, where people interact freely together, it can be difficult to select the participants that exhibit a behavior worthy of attention.
The design of (expressive) gestures interface can gain from a better understanding of individual- and context-dependent human behavior. It can ensure their usability in a more naturalistic environment. Automatic attention cues are also able to simplify information access in those complex-situations which leads, in the HCI domain at foster interaction, anticipating focus of attention as automatic zoom on the Region Of Interest (ROI).

C. Work overview

The project consisted in two main steps:

- *A robust tracking system*

This system uses both color video cameras and infra-red cameras to be robust enough to illumination changes and partial occlusions. Infra-red (IR) camera is much less sensitive to light changes if those lights are not directly in the camera field of view (FOV). The color camera FOV is larger and it is able to keep on the tracking process if infra-red markers are occluded or out of the infra-red camera FOV.

- *Motion attention: human-like reactions*

Once participant tracking is robust enough to handle naturalistic scenarios, an automatic attention index can be computed which highlights which movements should be the most "interesting" for a human observer. Attention is computed both in a spatial and in a temporal context on several features: speed and quantity of motion.

In section II, after a hardware and software overview, we will describe the video data acquisition and processing (blob segmentation) for each one of the two video modalities. In section III, the fusion mechanism which leaded to a robust tracking system will be explained. Section IV deals with attention computation both in spatial and temporal context. Finally, we will conclude by a discussion in section V. The source code of this project and some video demos can be found on the eNTERFACE 2008 workshop website [2].

II. SIGNAL ACQUISITION

A. *Material and system overview*

We developed a setup that analyzes human behavior in a flexible environment with regard to illumination changes. Three cameras were used to capture video: two "Eneo VKC-1354" color analogical cameras with a 752*582 pixel resolution at 25 frames per second (fps) and one "Imaging Source DMK 31BF03" monochrome digital camera delivering a 1024*768 pixel resolution at 30 fps. They were equally placed on the side of a 3 x 3 meters area, at a height of 2.5 meters, downward looking above the participants and recording with constant shutter, manual gain and focus. The participants used each one a red hat (for color segmentation) and a halogen light which emits visible but also infra-red light (for IR segmentation) in all space directions. Figure 1 shows the setup configuration.

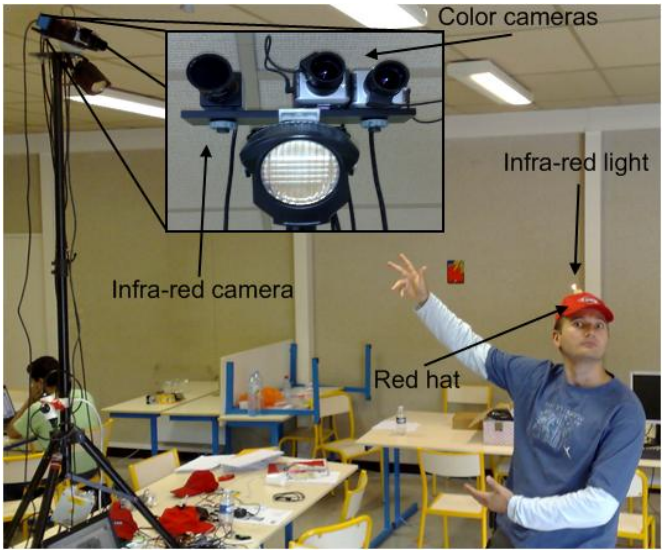


Fig. 1: Experimental setup

As described on Figure 2, two computers were used to perform the IR video and color video stream acquisition and processing (IR-and Color-based blob detection and tracking). The third PC was used to achieve data fusion between the IR and color stream data and to further higher-level processing and rendering.

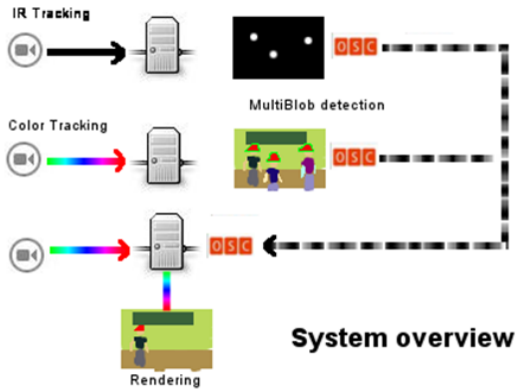


Fig. 2: System overview

A non-invasive video-based approach was adopted based on the EyesWeb XMI free software platform. We were interested in automatically extracting the displacement of people moving in front of the camera and computing their motion features. Head detection and tracking solutions were privileged to fully exploit the reduced available space, obtain depth information and avoid occlusions due to the interactions between participants.

The EyesWeb XMI (www.eyesweb.org) is a free software platform [3]. It consists of two main components: a kernel and a graphical user interface (GUI). The GUI manages interaction with the user and provides all features needed to design patches. It allows hand fast development of custom interfaces for use in artistic performances and interactive multimedia installations. The kernel manages real-time data processing, synchronization of multimodal signals. It supports the integration of user-developed plugins; an SDK (Software Development Kit) is provided to simplify the creation of such plugins by means of Microsoft Visual C++. The user-developed plugins, together with the ones provided with EyesWeb are the building blocks that the end user can interconnect to create an application (*patch*).

B. *Blob Detection*

The present system's analysis of human activity starts from foreground segmentation based on the analysis of the color and infra-red video streams. This analysis provides a binary mask of the spatial extension of the region of interest through time (blob detection).

- *Color video stream*

A skin color detection algorithm was used on the signal coming from the color camera. We developed a modified version of the Continuously Adaptive Mean Shift Algorithm (CamShift) which is itself an adaptation of the Mean Shift algorithm for object tracking. Our method consisted in manually selecting the color of interest (COI) which is converted into a HSV colorimetric system. The set of colored pixel is quantized in a one-dimensional histogram to create the COI's model. Furthermore, a bandwidth of acceptable Hue and Saturation values are defined to allow the tracker to compute the probability that any given pixel value corresponds to the selected color. In order to enhance the system robustness to illumination changes, the color model was updated in several areas of the scene which have different illumination. In that way the color model was resistant to moderate illumination changes as those present on our scenario visible on Figure 3 between the left-side and the right-side of the scene.

Fig. 3: Color-based red hat blob detection

- *Infra-red video stream*

The signal coming from the IR camera is not affected by illumination variations but might be artefacted by light reflections or some static infrared sources. We processed the video stream with a background subtraction to eliminate static elements. Then, we binarized the signal with an empirically-tested threshold value to extract the moving regions of interest (blobs). Figure 4 shows that the IR lights located on the top of the red hats are very clearly detected.

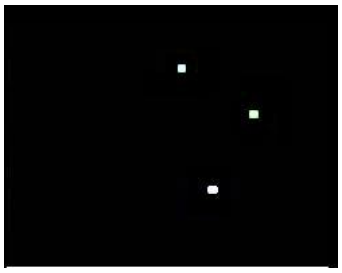
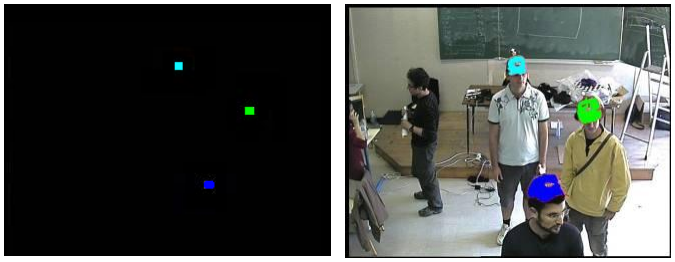


Fig. 4: Infra-red lights located on the top of the red hats detection

C. Single and Multi-Blob Tracking

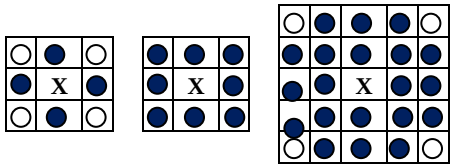
Resulting from the pre-processing step (color or IR based), we obtained a binary image where white represents the foreground objects. The next step is to assign a label that identifies the different white blobs, and to track them. The tracking result can be seen on Figure 5.



(1) **(2)**

Fig. 5: Multi-blob tracking : from left to right the labeling (identification) of (1) the three blobs of the IR video stream (2) the three blobs of the color video stream

To achieve the image tracking, we defined an adjacency measure based on the n-connectivity of two pixels (Figure 6).



(1) **(2)** **(3)**

Fig. 6. From left to right, illustration of (1) a 4-connectivity: the four filled circles (pixels) are connected to the one of interest (the cross); we defined this case as adjacency 1. (2) 8-connectivity, defined as adjacency = 2; (3) 20-connectivity defined as adjacency = 3. This adjacency measure can be generalized to n-connectivity

Basing on this definition, we have used the following algorithm for image segmentation and tracking:

- *Definitions*

Image segmentation *image_sgm* (video frame *f*) returns the set of distinct connected components *cc* (*f*) such that each pixel of an item in *cc* (*f*) is at distance x ≥ threshold from any other pixel belonging to other items in *cc* (*f*)
Valid region *valid_reg* is user defined as the maximum Euclidean distance between two blobs' baricentres in two consecutive frames.
Minimum difference *min_diff* returns the following:

$$\text{min\_diff} = \alpha \times \text{dist}(b_1, b_2) + \beta \times |\, \text{area}(b_1) - \text{area}(b_2) \,| \qquad \textbf{(1)}$$

where *dist* (b1, b2) is the Euclidean distance between the baricentres of blobs b1 and b2 and α and β are the weights of area and position, comprised between:
A = [0,1] and β = [0,1]
These values are manually decided according to the cam's FOV and the foreground objects to track (e.g. humans)

- *Procedure*

Initialization: $t = 0$
  cc(*t*=0) = *image_sgm(frame(0));*
  store *cc*(0) and assign a new label to each item in *cc*(0)
For $t = 1, 2, \ldots$
  cc(*t*) = *image_ sgm(frame(t));*; i.e. the set of distinct blobs in current frame (frame(t));
  store *cc*(t)
  for each item *x* in *cc*(*t*)
    find *x* in *cc*(*t*-1) which minimizes *min_diff*
    if *x* is in *valid_reg*, then
      {
      assign *y* the same label as *x*
        if *x* is recognized in N consecutive frames,
          then the item is considered to be trackable (i.e., we can measure its velocity, direction, etc.)
        else
          the item is unstable (recognized but not tracked)
      }
    else
      y is assigned a new distinct label

III. DATA FUSION

A. A common reference

Prior to any fusion algorithm, we need to provide a common reference to the signals to be fused. As we worked with different cameras, their fields of view (FOV) were different. A robust transformation was necessary to match the position of a point computed on a frame from one video camera, to the position of the same point computed on a frame from another video camera. The projective transformation meets our need because the perspective is subject to change with respect to the camera focal distance. A correction of the radial distortion was not necessary as the distortion due to wide-angle lenses was not very important.

The projective transformation conserves the proportions between two sets of two points: a quadrilateral is projected onto another one. Since it is not a linear transformation, we used homogeneous coordinates to compute the transformation with a matrix product.

The transformation matrix H_{ab} performing the projection of the points P_a in image "a" to the points P_b in image "b" can be written in homogeneous coordinates as follows (points are located on the same subjective plane):

$$P_a = \begin{pmatrix} x_a \\ y_a \\ 1 \end{pmatrix}, P'_b = \begin{pmatrix} x'_b \\ y'_b \\ w_b \end{pmatrix}, \mathbf{H}_{ab} = \begin{bmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & h_{33} \end{bmatrix}$$

Then:

$$P'_b = \mathbf{H}_{ab} \cdot P_a \quad \text{and} \quad P_b = {}^{P'_b}/_{w'} = \begin{pmatrix} x_b \\ y_b \\ 1 \end{pmatrix} \qquad \textbf{(2)}$$

In order to visually test the accuracy of our transformation, we developed EyesWeb blocks which perform projective image transformation according to an input matrix H_{ab}. We also developed a block which composes this projective matrix from the correspondence of four points in the original and the referent image.

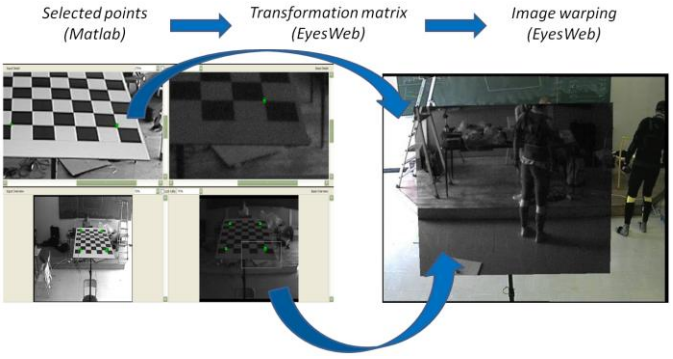


Fig. 7. IR and color video stream registration process. From left to right: selection of the four corresponding points in a color and IR snapshot in Matlab, computation of the transformation matrix in EyesWeb, image warping in EyesWeb

Figure 7 displays the entire transformation process. We first selected four points in a snapshot of the IR and color video streams using Matlab. Then, the coordinates of those points are used by EyesWeb to compute the transformation matrix. Finally, this matrix is used to warp the initial image: the superposition of the color and IR image show a good registration in the plane where the points in the two images were chosen. Figure 8 shows how the initial video stream is transformed into a new video stream according to a projection matrix H_{ab}.

Fig. 8. Left: initial video frames, right: real-time projective warping using OpenCv-based EyesWeb block

Once the registration was visually validated, the matrix H_{ab} was used to compute the projective transformation for the blob baricenters and not for the whole image. The use of a newly implemented EyesWeb block which performs matrix multiplication greatly reduced the computational cost: this solution avoided the computation of the projective transformation for all pixels in each frame of the video stream. Figure 9 shows the rendering of the final video stream with the surimposition of IR (green markers) and color video tracking (red markers) which converges onto the participants hats.

Fig. 9. IR tracking of the lights on top of the red hats (green), color tracking of the top of the red hats (red

The projective transform blocks were achieved by using some functions already implemented in OpenCV (Open Computer Vision) [4]. OpenCV is a library developed by Intel with a BSD license. EyesWeb can easily wrap OpenCV functions to handle them as blocks in the software platform.

B. Confidence Level

The second step performed before the fusion was to compute a confidence level on each modality (IR and color).
The weight (confidence level) changes according to the participants' visibility with respect to the camera's field of view (FOV) and obstruction occurrences. Data range from 0 when the participants are not visible by the camera to 1 when they are visible. For the color modality, the confidence level also gradually changes between 0 and 1 depending on blob area variations: abrupt variations due for example to sudden illumination changes decrease the confidence level whereas stable blob areas over time increase it.

- *Color video stream confidence level*

Blob tracking with skin color detection can have a lack of accuracy due to illumination variation or even a loss of coordinates due to visibility issues (obstruction or disappearance from the camera's FOV).
This confidence level is built on two hypotheses. The first one is to consider location information obsolete after a short delay (mobile objects tracking). The second hypothesis is that blob's area abrupt variation can be related either to some undetected surface or to unwanted detections. According to these assumptions, the confidence level CL_{VID} is computed as the conjunctions of two terms:

$$CL_{VID} = (\textit{blob in FOV}) \wedge (\textit{stable blob's area}) \quad \textbf{(3)}$$

where $CL_{VID} \in [0, 1]$.
For the "*blob in FOV*" indicator computation, we used a clock generator to check that the elapsed time since the last coordinates' acquisition keeps below an acceptable delay. The "*stable blob's area*" indicator is computed on a temporal sliding window. We compared the current blob's area value with its mobile average and returned an inverse variation rate (minimum value between the ratio and its inverse).

- *Infra-red video stream confidence level*

The IR light tracking might be interrupted in two situations: the blobs might be obstructed by the occlusion of one participant with respect to the others or they can come out from the camera FOV. The blob detection could also be corrupted when the tracked participants move closely. A binary confidence level was developed to handle these tracking issues. We measured the elapsed time since the last valuable detection. If this delay exceeded an empirically tested threshold, the confidence degree was set to 0 until a new detection occurred. The confidence level CL_{IR} is computed as:

$$CL_{IR} = (\textit{blob in FOV}) \quad \textbf{(4)}$$

where $CL_{IR} \in \{0, 1\}$.
We measured the elapsed time since the last valuable detection. If this delay is over a threshold that we fixed, "*blob in FOV*" is fixed at 0 until a new detection occurs.

C. Fusion algorithm

- *Single blob fusion*

In order to fuse the 2D coordinates coming from the IR and the color video stream, a weighted mean rule was applied. The weights are the respective confidence levels previously computed for both modalities.

- *Extension to multi-blob fusion*

In the case of multi-blob tracking, each blob coordinates set was extracted and computed in each modality (color and IR) together with their confidence level.
Nevertheless, the fact that the blobs from both modalities will remain linked to the same target during the experiment is not obvious. To prevent this issue, our method needed to test continuously the relation between the coordinates of the same blobs located in the two modalities. We created a new fusion EyesWeb block where we considered only the non null confidence level elements and we matched each one of them in a modality with the nearest neighbor point (following a Euclidian distance) in the other modality. We fused these couples with the same weighted mean rule described in the previous section. Figure 10 shows the fusion results.

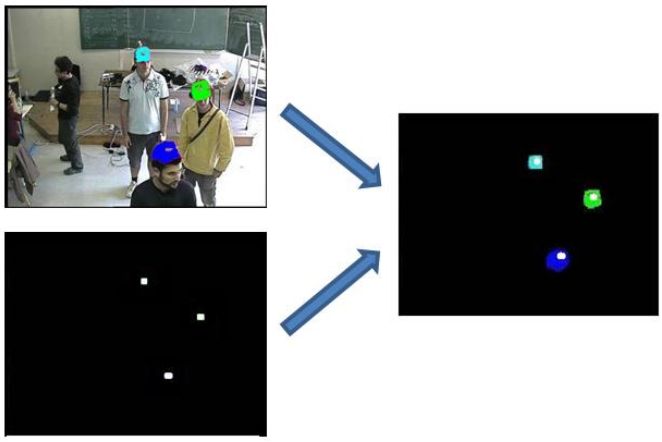


Fig. 10: Top-left: color image tracking, bottom-left: IR image tracking, right: modality fusion

IV. Salient Gestures: an attention filter

A. Computational Attention

The aim of computational attention is to automatically predict human attention on different kinds of data such as sounds, images, video sequences, smell or taste, etc… This domain is of a crucial importance in artificial intelligence and its applications are numberless from signal coding to object recognition. Intelligence is not due only to attention, but there is no intelligence without attention.
Attention is also very closely related to memory through a continuous competition between a bottom-up or unsupervised approach which uses the features of the acquired signal and a top-down or supervised approach which uses observer's a priori knowledge about the observed signal. We focused here only on bottom-up attention due to motion.

While numerous models were provided for attention on still images, time-evolving two-dimensional signals as videos have been much less investigated.
Nevertheless, some of the authors providing static attention approaches generalized their models to the time dimension: Dhale and Itti [5], Yee and Pattanaik [6], Parkhurst and Niebur [7], Itti and Baldi [8], Le Meur [9] and Liu [10]. Motion has a predominant place and the temporal contrast of its features is mainly used to highlight important movements. Zhang and Stentiford [11] provided a motion analysis model based on comparing image neighborhoods in time. The limited spatial comparison led to a "block-matching"-like approach providing information on motion alone more than on motion attention. Boiman and Irani [12] provided an outstanding model which is able to compare the current movements with others from the video history or video database. Attention is related to motion similarity. The major problem of this approach is in its high computational cost.
As we already stated in [13] and [14], a feature does not attract attention by itself: bright and dark, locally contrasted areas or not, red or blue can equally attract human attention depending on their context. In the same way, motion can be as interesting as the lack of motion depending on the scene configuration. The main cue which involves attention is the rarity or the contrast of a feature in a given context. A pre-attentive analysis is achieved by humans in less than 200 milliseconds. How to model rarity in a simple and fast manner?
The most basic operation is to count similar areas in the context. Within information theory, this simple approach based on the histogram is close to the so-called self-information. Let us note m_i a message containing an amount of information. This message is part of a message set M. A message self-information $I(m_i)$ is defined as:

$$I(m_i) = -\log(p(m_i)) \quad \textbf{(5)}$$

where $p(m_i)$ is the probability that a message m_i is chosen from all possible choices in the message set M or the occurrence likelihood. We obtain an attention map by replacing each message m_i by its corresponding self-information $I(m_i)$. The self-information is also known to describe the amount of surprise of a message inside its message set: rare messages are surprising, hence they attract our attention.
We estimate $p(m_i)$ as a two-terms combination:

$$p(m_i) = A(m_i) \times B(m_i) \quad \textbf{(6)}$$

The $A(m_i)$ term is the direct use of the histogram to compute the occurrence probability of the message m_i in the context M:

$$A(m_i) = \frac{H(m_i)}{Card(M)} \quad \textbf{(7)}$$

where $H(m_i)$ is the value of the histogram H for message m_i and $Card(M)$ the cardinality of M. The M set quantification provides the sensibility of $A(m_i)$: a smaller quantification value will let messages which are not the same but quite close to be seen as the same.
$B(m_i)$ quantifies the global contrast of a message m_i on the context M:

$$B(m_i) = 1 - \frac{\sum_{j=1}^{Card(M)} |m_i - m_j|}{(Card(M) - 1) \times Max(M)} \quad \textbf{(8)}$$

If a message is very different from all the others, $B(m_i)$ will be low so the occurrence likelihood $p(m_i)$ will be lower and the message attention will be higher. $B(m_i)$ was introduced to avoid the cases where two messages have the same occurrence value, hence the same attention value using $A(m_i)$ but in fact one of the two is very different from the others while the other one is just a little different.
In order to get a fast model of motion attention we have here a three-level rarity-based approach. While the first two approaches are bottom-up and use motion features context to attract computer's attention, the last one is mostly top-down and it learns a model of the scene which will be able to modify bottom-up attention by inhibiting some movements for example. The three levels of motion attention we propose here are:

- *Low-level instantaneous motion attention:*

Motion features are compared in the spatial context of the current perceived frame. Rare motion behaviors should immediately pop-out and attract attention. This low-level approach is pre-attentive (reflex) and it uses no memory capacities.

- *Middle-level short-term motion attention:*

Once a moving object was selected using low-level attention, its behavior into a short temporal context is than observed. Short-term memory (STM) is here used to save an object motion during 2 or 3 seconds (for longer time periods, motion details of an object are forgotten). Rare behaviors of an object through time will be quoted as interesting while repeating motion will be less important.

- *High-level long-term motion attention:*

This third top-down attention approach uses long-term memory capacities and it is a first step through motion and scene comprehension. The attention level of each pixel through time is accumulated which leads in areas of the scene which concentrate attention more than the others : a street accumulates more attention through time than a grassy area close to it, a tree which moves because of the wind or a flickering light will also accumulate attention through time. The scene can thus be segmented in several areas of attention accumulation and the motion in these areas can be summarized by only one motion vector per area. If a moving object passes through one of these areas and it has a motion vector similar to the one summarizing this area, its attention will be inhibited. If this object is outside those segmented attention areas or its motion vector is different from the one summarizing the area where it passes through, the moving object will be assigned a very high attention score. This third attention step builds an attention model learnt from instantaneous and short-term attention steps which is able to inhibit bottom-up attention if it corresponds to the model or to enhance it if the motion does not match with this model.

Within this project we developed an implementation of the two bottom-up motion attention comparing current motion features with a spatial and a short-term temporal context. The third top-down attention model is further discussed in section V of this article.

B. Instantaneous motion attention

An implementation of the spatial motion rarity was achieved as an EyesWeb XMI patch by using Equation (5) with $p(m_i) = B(m_i)$. In the scenario three people were tracked and their instantaneous speed was used. As only 3 motion vectors were available, the computation of the rarity $A(m_i)$ had not much sense from a statistical point of view.
Fig. 11 shows part of the tested scenario. Three people moving or not are present behind the cameras. Their instantaneous velocity vectors V_1, V_2, and V_3 are computed.

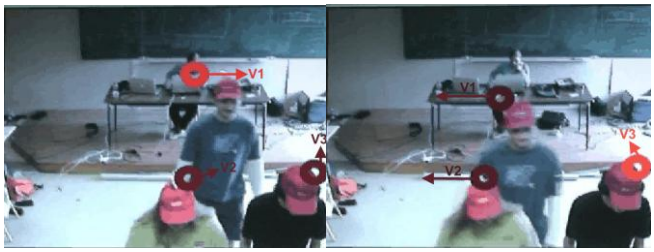


Fig. 11: Left: fastest moving object (V_1) is the most important (hot red), Right: slower moving object (V_3) is the most important.

On the left-side V_1 is very different from V_2 and V_3: V1 has high speed amplitude while V_2 and V_3 are very slow or stopped. In this case the faster object has a higher attention score (hot red) compared with the lower attention score (darker red) of V_2 and V_3. This situation can be compared with the one which often occurs if a classical motion detection module is used and where faster motion is well highlighted. That is not the case on the right image where the most different speed is V_3 (stopped) while V_1 and V_2 are quite the same (very fast). In this case, the attention score is higher (hot red) on V_3 which does not move while fast moving objects do not attract a large amount of attention. The result of this approach shows a different behavior compared to a simple motion detection algorithm. This first step enables the computer to choose the most "interesting" moving object very efficiently and then to apply to it short-term motion attention.

C. Short-term motion attention

- *Trajectory-related features*

The moving object speed was computed on a history of 3 seconds and the speed mean was computed on a 10 frames sliding window in order to avoid a too high variability of the speed due to segmentation and tracking noise.
The speed range was divided into 3 bins: static or very low speed, normal speed and high speed. The Equation (5) was applied with $p(m_i) = A(m_i)$ because there was enough data within the 3 seconds history to get a statistically reliable occurrence likelihood $p(m_i)$. Moreover, the contrast between the 3 speed bins is always very high so the $B(m_i)$ term is here less important.

Fig. 12: The brutal speed change is more important (red) then the speed value: the amplitudes of V(T1) and V(T4) are the same but they have a different attention score. Similar behavior can be seen with V(T2) and V(T3).

Fig. 12 shows the tested scenario. One participant moves his head from left to right very fast, then normally and finally he stops. When a change in the speed is detected (stop to normal speed, normal speed to high speed, high speed to stop, etc...), the attention score is very high, but it decreases exponentially when a stable speed occurs.

- *Silhouette-related features*

The same algorithm as in the previous point was applied but the feature used here was the Quantity of Motion (QoM). This measure is obtained by integrating in time the variations of the body silhouette (called Silhouette Motion Images - SMI).

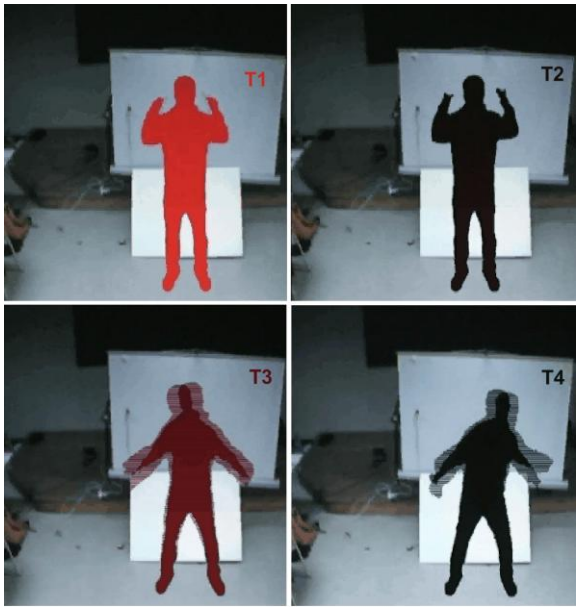


Fig. 13: The brutal QoM change is more important (red) than its value: the QoM at T1 and T2 are the same but they have a different attention score. The same behavior can be seen between T3 and T4.

Fig. 13 shows the test case. One participant is stopped, moves normally and moves a lot. If he moves fast while within his temporal context he is majorly stopped, this will be interesting. But if he moves fast while within his temporal context he also majorly moves fast the computer will classify this movement as uninteresting because it is repetitive and no new information is brought.
As in section IV.B, the short-term motion attention algorithm was implemented as an EyesWeb XMI patch.

D. Spatial and temporal motion attention fusion?

The previous two points (B and C) provide rarity-based attention indexes for moving objects based on spatial and on short-term temporal contexts. An interesting point is about a possible fusion of the results of those two approaches: how simultaneously take into account attention based on two contexts which do not have the same nature?
It is impossible to compare a given temporal context with a spatial one on the same basis: time and space are orthogonal which is also confirmed by the fact that time and space features are processed in two separate regions of the brain [15]. Instead of trying to fuse those two kinds of attention indexes, it seems more realistic to state that one of them (spatial context) is pre-attentive and it occurs first while the second one (short-term) is attentive and it focuses on the temporal context of an object previously selected by the spatial attention. Once this object is analyzed by the temporal attention, another one which may spatially pop-out can be tracked and its attention computed. Thus, spatial attention selects potential interesting moving targets while temporal attention verifies if those targets have also interesting behaviors through time.
In our opinion it is much more relevant to fuse the motion spatial attention map with a static image attention map based on color and gray level rarity [16]. As the context is in this case the same, it is realistic to compute rarity cues and compare static images (as background) and motion rarity (foreground) to get a final spatial attention which takes into account motion but also the other static pre-attentive features as colors or shape.

V. Results and Discussion

A. Results and rendering

We developed a solution to track human movement based on the fusion of two video modalities: a color and IR system. This solution showed robustness with respect to the light changing conditions and partial occlusions.
We analyzed spatio-temporal profiles of people activities at different levels of details. At gross level, human activity was analyzed in terms of the spatial trajectory of moving blobs corresponding to heads. However trajectory, by itself, hardly provided detailed information about the performed gestures [17]. A more-detailed level of person's activity was analyzed in terms of full-body quantity of motion (QoM) and found more relevant to characterize motion sequences and related gestures.
An attention index was developed to highlight motion saliency. Both the spatial and temporal contexts were used to compute rarity, thus attention indexes which provides additional information compared to a simple motion detection algorithm. We showed that motion perception can be radically different from simple motion detection: depending of the context, the lack of motion appeared to be more "interesting" than motion itself. Spatial context is used to select the participant which exhibits the highest saliency. Then, the selected participant can be tracked on a short period of time in order to see if he also has a salient behavior over time. This project is a first step towards systems which have more human-like reactions and perceive signals instead of only detecting them.

B. Further improvements

Several improvements could be achieved in addition to improve the already implemented system:

- *Video stream synchronization*

The blob detection data flows were not synchronized. A delay could be found between the tracking achieved on the IR video stream and the one achieved on the color video stream. This is due to the fact that the analog cameras and the digital camera had not the same time response. The output signal of the first ones must be digitalized with an extern FireWire/AD converter. We could observe a time lag between the video output and the filmed scene.
Moreover, the computation time for the color video processing is longer than the one needed for the IR video processing. Finally, the computers we used for the two video stream processing had not the same performances in terms of both CPU and graphical card.
The data flow synchronization could be improved if both cameras are digital and have the same characteristics. The synchronization could be even simpler if both cameras are acquired on the same computer using several FireWire ports. In this case, the EyesWeb XMI possibility to synchronize several processes on a single platform could be used and the two data flows would be perfectly synchronized. EyesWeb XMI highly improved the multimodal synchronization possibilities [3]: each block has two additional pins called "Sync-in" and "Sync-out" which can be used to propagate synchronization clock signals between blocks.

- *Confidence level improvement*

The present confidence level (CL) characterizes how much we should be confident in the tracked object position in each modality (IR and color video stream) in order to fuse them using a weighted mean. For the current implementation, we only used two assumptions: time since the last valuable detected position and smoothness of the detected blob variation over time. The first way to improve the modality confidence level could consist in adding other assumptions. For example, as we are tracking human heads, we could assume that abrupt speed variation is correlated with tracking inaccuracy and to decrease the positioning CL.
An additional improvement could consist in defining a confidence level for the fusion itself that we call fusion confidence level (FCL). A tracked object position could have a high CL in both modalities but the final fusion may not be

necessarily accurate if the two positions are very far one from the other for example. Other elements like speed and direction could help in defining if two positions from two different modalities can be fused. This FCL is able to point out if the data fusion should be considered or not.

- *Additional features: getting closer to human attention*

An interesting improvement could also be achieved by computing attention indexes for many other features. We can think first about motion direction which was not taken into account during this implementation: only the speed attention was computed. If the speed of a moving object is very low, the motion direction information is not very reliable because it can be due only to detection noise. Nevertheless, if the speed is high enough the information brought by motion direction is very important. From an attentional point of view, if many people have a common direction while one participant has a different direction, this is a very important cue on objects saliency. The same idea can be developed for temporal attention where brutal direction changes are very interesting.

A future work will consist in reaching a more complete description of human activities. The head tracking should be integrated by information about overall body motion. Camurri and all [18] revealed that bounding box variations or ellipse inclination, that approximate 2D translation of body, can account for expressive communication. A more detailed analysis of upper body-part should also be accomplished. Glowinski et al. [19] show that color-based tracking of head and hand can reveal expressive information related to emotion portrayals. On the basis of this refined description of human movement, we could consider new motor cues (e.g. symmetry, directivity, contraction, energy, smoothness…) accounting for the communication of an expressive content [18] that could be integrated, processed by the attention index for a more pertinent context-based analysis of expressivity.

If the fusion between spatial and temporal attention is not interesting (as discussed in section IV.D), it is crucial to fuse attention information coming from several features in the same context (space or time). This fusion can simply be done by using the maximum operator: if a feature highly attracts attention in a specific place, this area is very interesting at least from this feature point of view.

- *High-level motion attention*

In real life, when we observe a scene which does not change every time (fixed camera for example), we build knowledge models about those scenes. That model will have important influences on the final attention score and it will be able to modify attention coming from the bottom-up attention mechanisms described in this paper. A first simple implementation of this high level model is to use attention accumulation as a threshold [13]: only the objects which imply a bottom-up attention which is higher than the attention accumulation of the model are really interesting, all the other objects are inhibited.

Figure 14 displays an example of the results of this implementation: a frame of the video is extracted on the top-left image. The top-right image is an attention map of the same frame. The bottom-left image is a model of the scene: the trajectory of the moving person is visible and also is the tree which often moves because of the wind. The bottom-right image show the final attention map after the high-level attention inhibition: some noisy areas (grass which also moves because of the wind, moving person's trajectory) are inhibited. The moving tree area has also been inhibited and only few attention is focused on the tree area. The moving person remains the only area of high attention score.

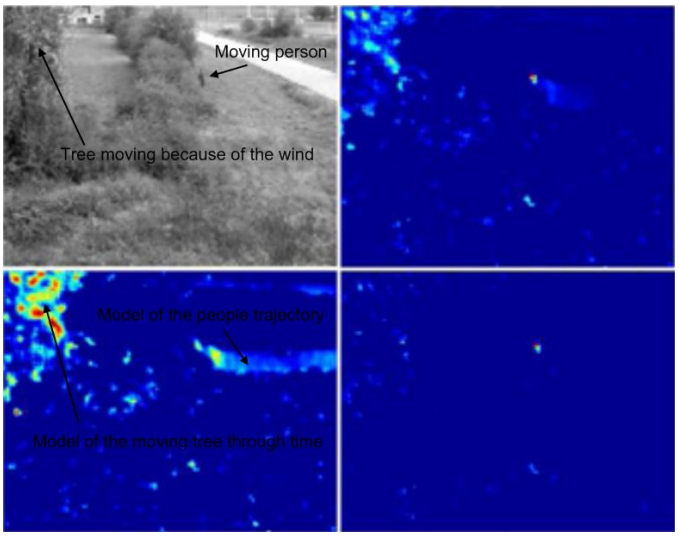


Fig. 14: Top-left: current video frame, top-right: bottom-up attention, bottom-left: scene model (top-down attention), bottom-right: final attention map after high-level attention inhibition

This simple approach already showed that it can inhibit some attention areas dues to noise or repetitive movements (trees which move because of the wind, flickering lights …).

A more complex implementation of the high-level attention model should segment the areas where a lot of attention accumulated and summarize them with a single motion vector. An object moving into these areas should be inhibited only if its motion vector is close to the motion vector which summarizes that area and its bottom-up attention score is bellow the model local amplitude. If the motion vector of this object is very different from the segmented attention area it passes through, the object will not be inhibited.

C. Discussion and Conclusion

We developed a context-based framework working in a non-controlled environment. It can deal with multiple heterogeneous situations caused by environmental disturbances and focus on relevant/rare events.

This system is more robust comparing to other applications based on video-tracking commonly developed for controlled environment. Stable environmental factors such as constant illumination and stable background greatly facilitate human activity monitoring. The present system can be a response to the growing demands, human monitoring systems in many application fields (e.g. video-surveillance, elder-people ambient assistant living, performing arts, museum spaces, etc.)

Moreover, our system open new research perspective for affecting computing and analysis of human expressivity.

Results from this study actually show that context-sensitive feature can help to better analyze gesture expressivity. The same expressive features are differently weighted depending on their spatial and temporal rarity. It put in evidence salient human motion either at the level of a single individual (rarity of behavior over time) or at the collective level (relative rarity of one member's behavior with respect to the others).

We plan to further investigate the potentialities of the rarity index in two directions: (i) by applying it to a more sophisticated set of expressive features (.e.g contraction, expansion, fluidity, impulsiveness) (ii) by analyzing how the visual feedback computed on the rarity index affect the subject behavior (e.g. whether it fosters expressive behavior).

ACKNOWLEDGMENTS

This work has been achieved in the framework of the eNTERFACE 2008 Workshop at the LIMSI Lab of the Orsay University (France). It was also included in the Numédiart excellence center (www.numediart.org) project 3.1 funded by the Walloon Region, Belgium. The authors thank to Johan DECHRISTOPHORIS who helped in IR sensor hardware components set-up. Finally, this work has been partially supported by the Walloon Region with projects BIRADAR, ECLIPSE, and DREAMS, and by EU-IST Project SAME (Sound And Music for Everyone Everyday Everywhere Every way).

REFERENCES

[1] Hongeng, S. and Nevatia, R., "Multi-agent event recognition" in ICCV, 2001, pp. II: 84-91

[2] eNTERFACE 2008 Workshop website: http://enterface08.limsi.fr/

[3] A. Camurri, A., Coletta, P., Varni, G., & Ghisio, S. "Developing multimodal interactive systems with EyesWeb XMI", Proceedings of the 2007 conference on new interfaces for musical expression (NIME07) (pp. 305-308). New York, USA, 2007

[4] OpenCV static wiki: http://opencvlibrary.sourceforge.net/wiki-static

[5] Dhavale, N., and Itti., L., "Saliency-based multi-foveated mpeg compression", IEEE Seventh International Symposium on Signal Processing and its Applications, 2003

[6] Yee, H., and Pattanaik, S., "Spatiotemporal sensitivity and visual attention for efficient rendering of dynamic environments", IACM, 2001

[7] Parkhurst, D.J., Niebur, E., "Texture contrast attracts overt visual attention in natural scenes", European Journal of Neuroscience, 19:783–789, 2004

[8] Itti, L., Baldi, P., "Bayesian Surprise Attracts Human Attention", Advances in Neural Information Processing Systems, Vol. 19 (NIPS 2005), pp. 1-8, Cambridge, MA:MIT Press, 2006

[9] Le Meur, O., Le Callet, P., Barba, D., and Thoreau D. "A Coherent Computational Approach to Model Bottom-Up Visual Attention", PAMI(28), No. 5, pp. 802-817, 2006

[10] Liu, F., and Gleicher, M., "Video Retargeting: Automating Pan-and-Scan", ACM Multimedia, 2006

[11] Zhang, V., and Stentiford, F.W.M., "Motion detection using a model of visual attention", IEEE ICIP, 2007

[12] Boiman, O., Irani, M., "Detecting Irregularities in Images and in Video", International Conference on Computer Vision (ICCV), 2005

[13] Mancas, M., "Computational Attention: Towards Attentive Computers", Similar edition, CIACO University Distributors, ISBN : 978-2-87463-099-6, 2007

[14] Mancas M., Gosselin B., Macq B., "A Three-Level Computational Attention Model", Proc. of ICVS Workshop on Computational Attention & Applications, 2007, Germany.

[15] Hubel, D.H., "Eye, brain and vision", New York: Scientific American Library, N°22, 1989

[16] Mancas, M., "Image perception: Relative influence of bottom-up and top-down attention", Proc. of the WAPCV workshop of the ICVS conference, Santorini, Greece, 2008

[17] Velastin, S., Boghossian, B., Lo, B., Sun, J., Vicencio-Silva, M.: Prismatica: toward ambient intelligence in public transport environments. IEEE Trans. Syst.Man Cybern. Part A 35(1), 164–182, 2005

[18] Camurri, A., De Poli, G., Leman, M., and Volpe, G., 2005. "Toward Communicating Expressiveness and Affect in Multimodal Interactive Systems for Performing Art and Cultural Applications", IEEE Multimedia, 12,1, 43-53, 2005

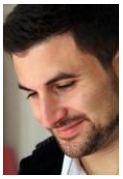
[19] Glowinski, D., Camurri, A., Coletta, P. , Bracco, F., Chiorri, C., Atkinson, A. An investigation of the minimal visual cues required to recognize emotions from human upper-body movements, (in press) Proceedings of the ACM 2008 International Conference Conference on Multimodal Interfaces (ICMI), Workshop on Affective Interaction in Natural Environments (AFFINE) Crete, 2008

Matei Mancas.

Matei Mancas was born in Bucarest in 1978. He holds an ESIGETEL (Ecole Supérieure d'Ingénieurs en informatique et TELecommunications, France) Audiovisual Systems and Networks engineering degree, and a Orsay University (France) MSc. degree in Information Processing. He also holds a PhD in applied sciences from the FPMs (Engineering Faculty of Mons, Belgium) on computational attention since 2007.

His past research interest is in signal and, in particular, image processing. After a study on nonstationary shock signals in industrial tests at MBDA (EADS group), he worked on medical image segmentation. He is now a Senior Researcher within the Information Processing research center of the Engineering Faculty of Mons, Belgium. His major research field concerns computational attention and its applications.

Donald Glowinski.



Donald Glowinski, Paris, 27-02-1977. He is doing a Phd in computing engineering at InfoMus Lab – Casa Paganini, in Genoa, Italy. (dir: Prof. Antonio Camurri). His background covers scientific and humanistic academic studies as well as high-level musical training.- EHESS (Ecole des Hautes Etudes en Sciences Sociales) MSc, in Cognitive Science, CNSMDP (Conservatoire National Supérieur de Musique et de Danse de Paris) MSc in Music and Acoustics, Sorbonne-Paris IV MSc. in Philosophy.

He was chairman of the Club NIME 2008 (New Interfaces for Musical Expression), Genoa, 2008. His research interests include multimodal and affective human-machine interactions. He works in particular on the modeling of automatic gesture-based recognition of emotions.

Pierre Bretéché.

Pierre Bretéché, Ivry sur Seine, 1981, received a MSc. degree in Computer Science in 2006 at University of Rouen (France). He is doing a PhD in Information and Communication Sciences at Laseldi Lab – University of Franche-Comté in Montbéliard (France). He previously worked in AI with a using massive multi-agent system to build semantic picturing of an environment.
He is now part of Organica research project. His research interest has moved to studying and designing new technology applications for public, cultural and artistic purpose.

Jonathan Demeyer.

Jonathan Demeyer received a MSc. in electrical engineering, Université Catholique de Louvain, Louvain-la-Neuve, Belgium in 2005 and a MSc. in applied sciences from the FPMs (Faculté Polytechnique de Mons, Belgium) in 2008.
He previously worked in the TCTS lab (Faculté Polytechnique de Mons, Belgium) developing a mobile reading assistant for visually impaired people.
He is now working on a project on automatic processing of high speed videoglottography for physicians. His main interests are medical image processing and computer vision.

Thierry Ravet.

Thierry Ravet was born in Brussels, Belgium on 31st Augustus 1976. He received an Electrical Engineering degree in 1999 in ULB (Université Libre de Bruxelles).
Afterwards, he worked as researcher in the Electronic – Microelectronic - Telecommunication department of ULB for 4 years with medical instrumentation projects.
Since February 2008, he has joined the TCTS Lab (Faculté Polytechnique de Mons). His main experiences are in non-invasive instrumentation, microprocessor system development, artefacts filtering and data fusion in the field of cardiorespiratory monitoring and polysomnographic research.

Gualtiero Volpe.

Gualtiero Volpe, Genova, 24-03-1974, PhD, computer engineer. He is assistant professor at University of Genova. His research interests include intelligent and affective human-machine interaction, modeling and real-time analysis and synthesis of expressive content in music and dance, multimodal interactive systems.

He was Chairman of V Intl Gesture Workshop and Guest Editor of a special issue of Journal of New Music Research on "Expressive Gesture in Performing Arts and New Media" in 2005. He was co-chair of NIME 2008 (New Interfaces for Musical Expression), Genova, 2008. Dr. Volpe is member of the Board of Directors of AIMI (Italian Association for Musical Informatics).

Antonio Camurri.

Antonio Camurri (born in Genova, 1959; '84 Master Degree in Electric Engineering; 1991 PhD in Computer Engineering) is Associate Professor at DIST-University of Genova (Faculty of Engineering), where he teaches "Software engineering" and "Multimedia Systems". His research interests include multimodal intelligent interfaces, non-verbal emotional and expressive communication, kansei information processing, multimedia interactive systems, musical informatics.
He was President of AIMI (Italian Association for Musical Informatics), is member of the Executive Committee (ExCom) of the IEEE CS Technical Committee on Computer Generated Music, Associate Editor of the international "Journal of New Music Research", a main contributor to the EU Roadmap on "Sound and Music Computing" (2007). He is responsible of EU IST Projects at DIST InfoMus Lab of University of Genova. He is author of more than 80 international scientific publications. He is founder and scientific director of the InfoMus Lab at DIST-University of Genova (www.infomus.org). Since 2005 he is scientific director of the Casa Paganini centre of excellence at University of Genoa on research in ICT integrating artistic research, music, performing arts, and museal productions (www.casapaganini.org).

Paolo Coletta.

Paolo Coletta born in Savona, Italy, in 1972. He received the "Laurea" degree in computer science engineering in 1997 and the Ph.D. degree in electronic engineering and computer science in 2001, both from the University of Genoa, Genoa, Italy.
From January 2002 to December 2002, he was a Research Associate at the Naval Automation Institute, CNR-IAN (now ISSIA) National Research Council, in Genoa.
In 2004 he was adjunct Professor of "Software Engineering and Programming Languages" at DIST (Dipartimento di Informatica, Sistemistica e Telematica), University of Genoa. Since 2004 he is a collaborator at DIST, University of Genoa.

Expressive Violin Synthesis A Study Case in Realtime Manipulation of Complex Data Structures

Nicolas d'Alessandro, Alexis Moinet, Thomas Dubuisson, *PhD students, Faculty of Engineering of Mons*[1], Jehan-Julien Filatriau, Wenyang Chu, *PhD students, Catholic University of Louvain*[2]

***Abstract*—This report gives the results of the project #10 of eNTERFACE 2008 summer workshop on mulimodal interfaces. This research highlights the problems of content-oriented instrumental sound synthesis, and an example for the violin is developed. The project has been planned as an extensive sound analysis processus, where multiple aspects of a given database have been studied. Then different resynthesis strategies have been implemented and some relevant choices in this field are discussed.**

***Index Terms*—violin, spectral analysis, granulation, viterbi**

I. Introduction

From the early days of electrical (even acoustical, through organs) modeling of instrumental sounds, a lot of work have been achieved. Today end-user solutions such as GarageBand [1] give an uncomparable access to high-quality musical sounds, allowing professionnal and non-professionnal musicians to compose nice soundtracks. With the same kind of temptation than in speech processing (thanks to high-quality unit selection synthesizers [2]) we could say that this topic reached a production level, where it is now just a matter of funding, in order to propose best collections of sounds.

However this is not really the point of view of performers. From their side, the possibility of producing high-quality musical sounds remains slightly schizophrenic. On the one hand, sensors and controllers have never been so user-friendly and exciting (e.g. WiiRemote), thanks to softwares like Max/MSP [3] and OSC protocol [4]. On the other hand, synthesis techniques often have to be re-designed from scratch (e.g. inside Max/MSP itself), as MIDI protocol is strangling the access to high-quality sample-based softwares. This is not just transmission conventions to be updated (e.g. OSC replacing MIDI), but also the architecture of these algorithms which is not compatible with more expressive setups.

Ircam's CataRT granulator [5] and its related softwares represents a significant break-through in this context. It brings creative ways of interacting with databases in the hands of the end-user. However we think that some important aspect of this "database-to-synthesis" pipeline could be re-discussed and more optimal solutions could be found. First because instrument sound synthesis is a complex and particular pitch-synchronous case of granulation where usual texture synthesis benchmarking [6] is no longer efficient. Then because libraries like FTM/Gabor/Mnm propose an "all-in-the-patch" view of the computer music langage which makes some simple things really unobvious to implement.

In this paper we propose to focus our investigation on a particular case study: the realtime synthesis of the violin. In section II we give an overview of what is the violin timbre from a perceptual point of view, thus allowing us to know what is the structure and dimensionality of the sound we want to control. In this project we use a monophonic violin database from FreeSound Project [7]. In section III we describe analysis algorithms that have been implemented offline in order to convert wavefiles into a structured cloud of frames. In section IV we finally explain some tracks that we followed in the field of data clustering and the possibility of generating a synthesis signal by browsing the cloud and selecting frames for OLA.

II. Background in Violin Timbre Perception

Timbre space is a feature space which is constructed by perceptual attributes selected based on human perceptual properties. The selection of perceptual attributes is one of the decisive factors of the success of granular synthesis of musical instruments. It decides not only the effectiveness of the segmentation of original sound tracks in the analysis phase but also that of unit selection from the granular database in the synthesis phase. The best selected attributes can construct a timbre space which allows to distinguish and classify all grains in the database.

To construct a timbre space which can be efficiently browsed and provides efficient mapping between feature trajectory and gesture trajectory, except pitch information, the signal processing timbral attributes of the violin has been seen the important clues to define the timbre space. However, only little research has been carried on the violins timbre space in a signal processing point of view. In the following survey, we try to summarize those important papers and make a list of the possible attributes.

A. Selection of the parameters

1) Categories of the attributes and their correlation: In the paper of James McDermott *et al.* [8], a set of forty timbral, perceptual, and statistical sound attributes are defined, studied,

[1] Information Technology Research Center, Faculty of Engineering of Mons, Belgium. Address: Boulevard Dolez, 31 - B7000 Mons (Belgium). Email: firstname.lastname@fpms.ac.be.

[2] Telecommunications and Teledetection Laboratory, Catholic University of Louvain, Belgium. Address: Place du Levant, 2 - B1348 Louvain-la-Neuve (Belgium). Email: firstname.lastname@uclouvain.be.

and categorized into the following domains: (1) Time-domain, (2) Fourier-transform, (3) Partial-domain, (4) Trajectory, (5) Periodic, and (6) Statistical, where the high correlation between attributes within each subset has been observed [8]. To avoid having high correlation among selected attributes and to have the minimum number of the attributes, we will not choose more than one attribute within each set in the early analysis stage.

Since the most well known perceptual properties are pitch, loudness, brightness, and attack time, it is reasonable to first find the corresponding statistical attributes which can represent those perceptual sound properties better. Then we further eliminate the less dominant attributes in each subset. First, we obtain the following attributes (description in the quotients represents the corresponding perceptual sound attributes): fundamental frequency (pitch), RMS energy (loudness), spectralcentreofgravity(CoG) (brightness) [8]. The change of them over time, which is perceptual, shows another information – trajectories inside the timbre space. Spectral flatness measure (SFM) can show the distinction between harmonic and non-harmonic sounds [8]. We avoidtopickupattributesdependingonVibratomeasures because it isdifculttobe denedor measured.

2) Attributes especially for the violin: The violin timbre and its perceptual attributes have been shown its dominant influence. Some interesting conclusion can be also drawn from the work of Jane Charles, etc [9]. Except the attributes mentioned in the last paragraph, SFM and the spectral contrast measure (SCM) are considered important because of their suitability as violin timbre features [9].

3) Other perceptual study: According to the research of Jan Stepanek [10] based on the human auditory experiment, levelsincriticalbands(barks), CoGand?overallspectrumlevelandlevelsofindividual play important roles in the perception. The experiment results show that the CoG and the first harmonic level have a dominant influence on the perception of violin tones [10].

III. Analysis of Expressive Violin Timbre

This section aims at describing the features that have been computed for the five pieces of the database. A strategy for the mapping of a source piece to a target piece is also presented.

A. Computation of the Fundamental Frequency

The fundamental frequency F_0 (and its temporal correspondance, the fundamental period T_0) is surely one of the most obvious parameters for the description of musical sound. In a tonal music context, it is indeed related to the perceived note, played by the performe, inside a given scale.

In speech and music processing domains, many pitch detection algorithms have been proposed. All these algorithms have to deal with strong difficulties of estimation when the fundamental frequency is high (which may be the case if sources like violin or soprano voice is considered). Among the proposed algorithms, the YIN algorithm [11] emerged for the last years and proved its ability to deal with high fundamental frequencies. For the purpose of this project, the fundamental frequency has been estimated in MAX/MSP environment, using the pitch detection tool included in the FTM/GABOR library. The values of fundamental frequency and the voiced/unvoiced decision have then been stored in files loadable in the MATLAB environment.

B. Pitch Contour Correction

Although YIN algorithm is known as efficient for musical sounds, it has been observed that the detected fundamental period suffers of several mistakes, essentially octave jumps. These jumps may be isolated or grouped into bursts. These artefacts have to be removed because there is a need of a reliable fundamental period, this value being the base of an efficient pitch-synchronous analysis. In order to correct the pitch contour, an algorithm consisting in two steps is proposed in this study.

First median filtering. This step aims at removing the isolated peaks in the pitch contour. It consists on computing, for every pitch occurence, the median of three values: the current pitch value and the two ones located on its left and right. Once the median filtered pitch contour is available, the difference between this contour and the original one is computed. For a given occurence, if the difference is higher than a given threshold (let us name it $Th1$), the pitch value is the one obtained after median filtering, otherwise the pitch value is the original one. The tests on different parts of the database showed that 100 seems to be a correct value for $Th1$.

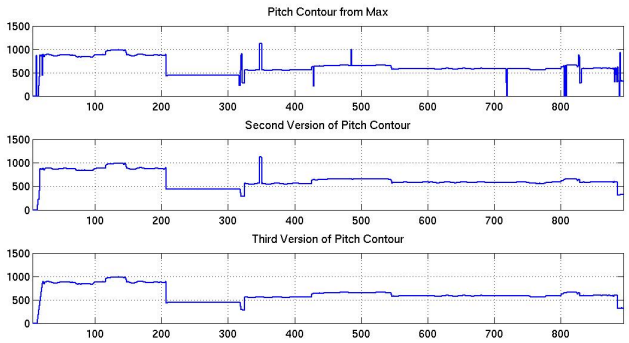


Fig. 1. Pitch contour correction

Then an anti-rebounce procedure. Although the previous step is able to deal with isolated peaks, it is not adapted to burst of peaks. In order to avoid these bursts, an anti-rebound procedure has been implemented. First the derivative of the corrected pitch contour from the first step is computed. It allows to highlight strong variations in the contour. Then, for each occurence of the derivative, if it is higher than a given threshold (let us name it $Th2$), a temporal interval is considered, beginning at the current pitch occurence and ending after a certain number of pitch occurences (let us name it $Th3$). If in this interval there is at least one value of derivative higher than $Th2$, the entire interval is tagged as a burst, which has to be removed. In the tagged interval, the corrected pitch values are obtained by computing the linear interpolation between the pitch values located at the boundaries of the interval. The tests on different parts of the

database showed that 50 and 5 seem to be a correct values for respectively $Th2$ and $Th3$. Fig. 1 shows the two steps of correction for an piece of pitch contour computed for the first piece of the database. For each occurence that has been corrected, its index is stored in order to avoid to use the corresponding frame in the synthesis operation.

C. Computation of Pitch Marks

As said before, the aim of this study is to perform a pitch-synchronous analysis, meaning that the descriptors are computed for frames centered on particular instants called pitch marks. At this point, an analogy with speech has been made. Indeed, particular instants during the phonation correspond to Glottal Closure Instant (GCI). These $GCIs$ occur every fundamental period and are particularly suited to sensitive analysis methods such as ZZT decomposition [12]. Besides these instants may be considered as particular pitch marks. For these reasons, it has been decided to compute analog instants for violin sounds, meaning that the same techniques than in speech are used. From now on, the computed pitch marks for violin sounds are called $GCIs$. Among the techniques proposed in litteratures, an interesting one use the temporal center of gravity (COG) [13] to locate $GCIs$:

$$Temporal\ COG = \frac{\sum_{i=1}^{N} n \times x(n)^2}{\sum_{i=1}^{N} x(n)^2} \qquad (1)$$

where $x(n)$ stands for a particular violin frame. The COG is computed for frames whose hopsize is one sample. A first estimation of $GCIs$ is obtained by finding the samples at which the COG crosses the zero-axis from positive to negative values. Fig. 2 shows the evolution of COG for successive periods of violin and Fig. 3 shows the detected $GCIs$ for a sequence of COG.

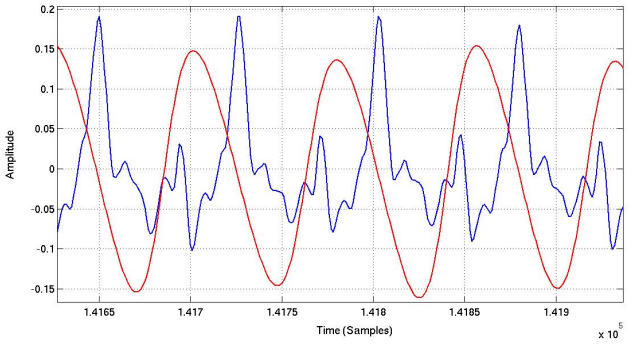


Fig. 2. Temporal Center of Gravity

Some $GCIs$ may be missed during this analysis step, which is problematic for the following analysis. As explained above, the $GCIs$ are spaced by the current fundamental period. This fact is exploited in order to detect the missing $GCIs$. For this purpose, the temporal interval between successive $GCIs$ and the difference between the pitch contour and the evolution of this interval are computed. For each pair of successive $GCIs$, if this difference is higher in absolute value than 80 percents of the corresponing value, it is decided that a GCI has been missed between the two successive $GCIs$. A new GCI is then inserted in the middle of the interval defined by the two existing $GCIs$. Fig. 4 shows an example of pitch marks insertion.

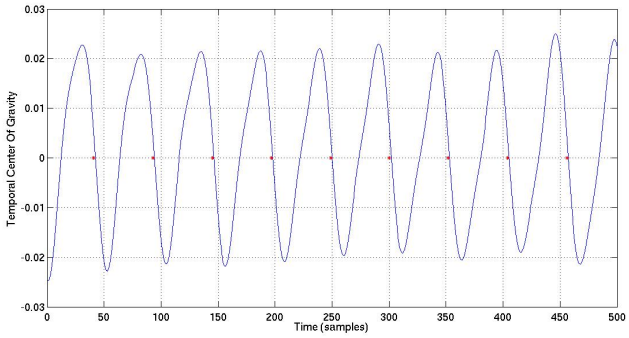


Fig. 3. Detected Pitch Marks from COG

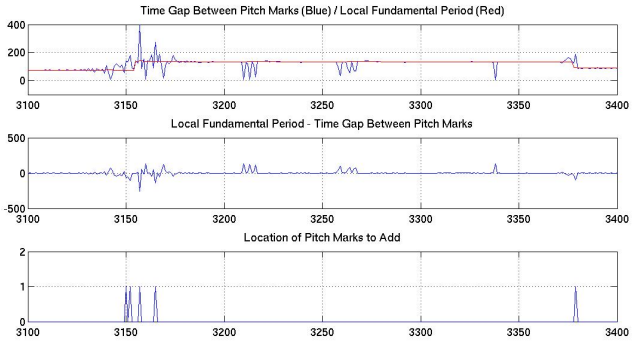


Fig. 4. Insertion of missing Pitch Marks

D. Pitch-Synchronous Analysis

Once the pitch marks are available, descriptors are computed for successive short time periods. In the case of pitch-synchronous analysis, these periods are frames centered on each pitch mark and whose length is twice the fundamental period associated to the considered pitch mark. In order to avoid strong discontinuities at the boundaries of the frame, each one is weighed by a Hanning window (very common in speech processing).

Many descriptors are proposed in signal processing litterature, some of them being particularly dedicated to speech or music analysis. For this study, it has been chosen to consider some descriptors implemented in the framework of CUIDADO project [14] aiming at developing a new chain of applications through the use of audio/music content descriptors. The chosen descriptors are presented in the following, where $x(n)$ stands for a frame and $X(f)$ for its magnitude spectrum.

1) Energy: The temporal energy of the frame $x(n)$ is defined as below (expressed in dB).

$$Temporal\ Energy = 10 \times log_{10} \sum_{i=1}^{N} x(n)^2 \qquad (2)$$

2) Total loudness: The loudness is defined as the energy in a band of the Bark scale [15]. This non linear scale has been defined so as two sinusoids of same amplitude and located in the same band are *perceived* in the same way while, if they are located in different bands, their perceived intensity is different. The Bark scale is used to mimic the behaviour of human ear in spectral domain. The loudness in a given frequency band is defined as below. The total loudness is the sum of all the loudness along the Bark scale. In this study, it has been chosen to consider 24 frequency bands.

$$Loudness = \sum_{Band} X(f)^{0.23} \tag{3}$$

3) Fundamental frequency (Hz): See Section III-A.

4) Spectral decrease: This descriptor aims at quantifying the amount of decrease of the amplitude spectrum. Coming from perceptual studies, it is supposed to be more correlated with human perception. It is computed as:

$$Decrease = \frac{\sum_{f=2}^{\frac{fe}{2}} \frac{X(f) - X(1)}{f-1}}{\sum_{f=2}^{\frac{fe}{2}} X(f)} \tag{4}$$

5) Spectral Center of Gravity: The spectral center of gravity (also known as spectral centroid) is a very common feature in MIR domain. Perceptually connected to the perception of brightness, it indicates where the "center of mass" of the spectrum is. The spectral center of gravity of the amplitude spectrum is known as an economic spectral descriptor giving an estimation of the major location of spectral energy. It is computed as:

$$COG = \frac{\sum_{f=1}^{\frac{fe}{2}} f \times X(f)}{\sum_{f=1}^{\frac{fe}{2}} X(f)} \tag{5}$$

6) Spectral Flatness Measure: The spectral flatness measure is defined as the ratio between the geometrical and arithemtic mean of the spectrum.

7) Tristimulus 1: According to the Bark scale, the first tristimulus is defined as a particular energy ratio:

$$Tristimulus_1 = \frac{\sum_{f_{Band_{[1]}}} X(f)}{\sum_{f_{Band_{[1\ldots24]}}} X(k)} \tag{6}$$

where $f_{Band_{[1]}}$ stands for the frequency bins included in the first Bark band and $f_{Band_{[1\ldots24]}}$ for the frequency bins included in the frequency band covered by Bark band from 1 to 24.

8) Tristimulus 2: According to the Bark scale, the second tristimulus is defined as a particular energy ratio:

$$Tristimulus_2 = \frac{\sum_{f_{Band_{[2,3,4]}}} X(f)}{\sum_{f_{Band_{[1\ldots24]}}} X(k)} \tag{7}$$

9) Tristimulus 3: According to the Bark scale, the third tristimulus is defined as a particular energy ratio:

$$Tristimulus_3 = \frac{\sum_{f_{Band_{[5\ldots24]}}} X(f)}{\sum_{f_{Band_{[1\ldots24]}}} X(k)} \tag{8}$$

Energy and fundamental frequency are base dimensions of control in a synthesis system from a performer point of view. For our database, these two features are distributed as shown in Fig. 5.

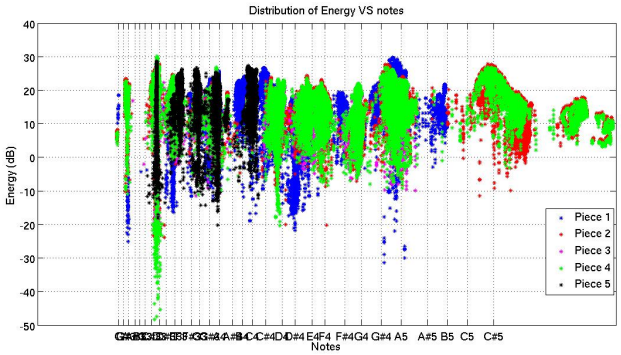


Fig. 5. Energy vs Notes for the five pieces of the database.

The total loudness has not been considered in the final synthesis system but it was computed for all the database. Fig. 6 shows the distribution of loudness vs notes for the five pieces of the database. It is interesting to note that it is highly correlated with the human perception (loudness curve).

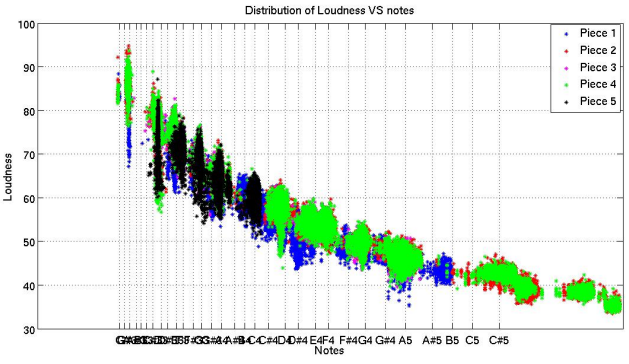


Fig. 6. Total Loudness vs. Notes for the five pieces of the database.

Considering the energy and the pitch in a synthesis system only allows to control the melodic part of the sound. Thus it would be interesting to include more dimensions in order

to reach other regions of expressivity (see [16] where more than 2 dimensions are used to synthesize singing voice). As 3 dimensions remains manageable and viewable by human, it has been chosen to select one spectral feature among the ones defined above to be the third dimension of our synthesis engine.

It has been observed that the COG and the $SpectralFlatnessMeasure$ are highly correlated. Conserving these two descriptors is useless since they bring similar information. It has been thus chosen to conserve the COG because it is widely used in the MIR domain. Concerning the three tristimuli, only the first is kept since the two others are highly correlated with the first one. Fig. 7, 8 and 9 show respectively the distribution of COG, $SpectralDecrease$ and $Tristimulus_1$ for the first piece of the database.

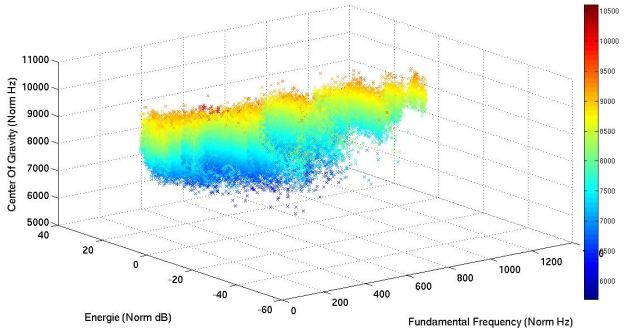


Fig. 7. Spectral CoG vs. Energy/Notes for the first piece of the database.

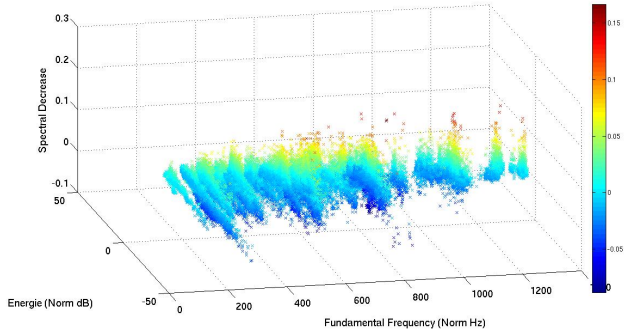


Fig. 8. Spectral decrease vs. Energy/Notes for the first piece of the database.

In order to select one feature among the three ones, the 3D space is projected on a 2D space by representing the distribution of each of these features according to the fundamental frequency. Fig. 10, 11 and 12 show respectively the distribution of COG, $SpectralDecrease$ and $Tristimulus_1$ for the five pieces of the database according to the notes.

Among the three features, one has to find the one which shows the highest variation for each zone of note and energy, but with a sufficient density of frames in order to have large possibilities of exploration. When looking at Fig. 10, 11 and 12, one can see that COG is the feature that meets the best these conditions.

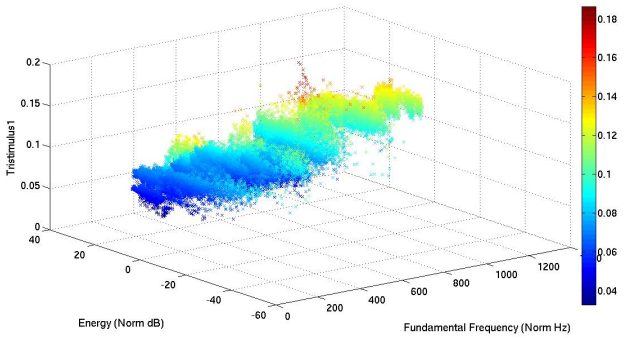


Fig. 9. First tristimulus vs. Energy/Notes for the first piece of the database.

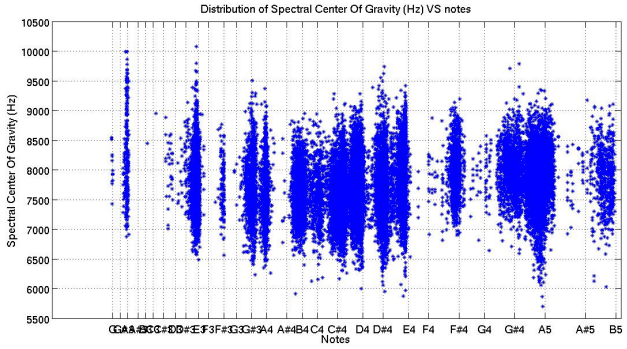


Fig. 10. Spectral CoG vs. Notes for the five pieces of the database.

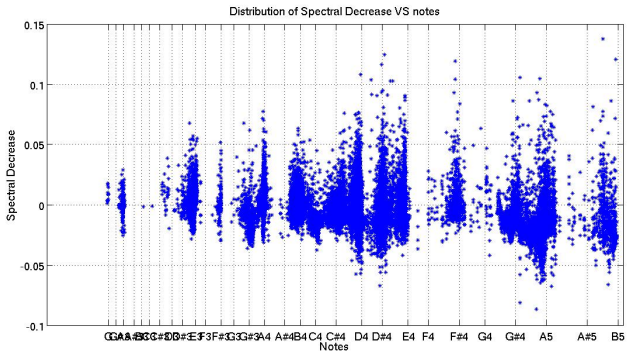


Fig. 11. Spectral decrease vs. Notes for the five pieces of the database.

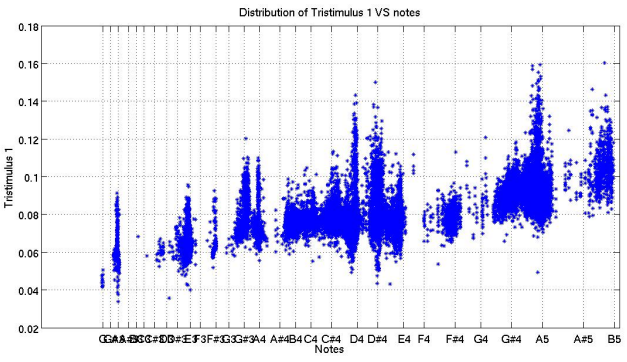


Fig. 12. First tristimulus vs. Notes for the five pieces of the database.

Three features are thus considered in our system to caracterize violin frames: the energy, the fundamental frequency

and the spectral center of gravity. In order to test strategies of synthesis, the database is split in two parts: the frame belonging to the first piece and the ones belonging to the four other pieces. The distribution of these two parts is shown in Fig. 13.

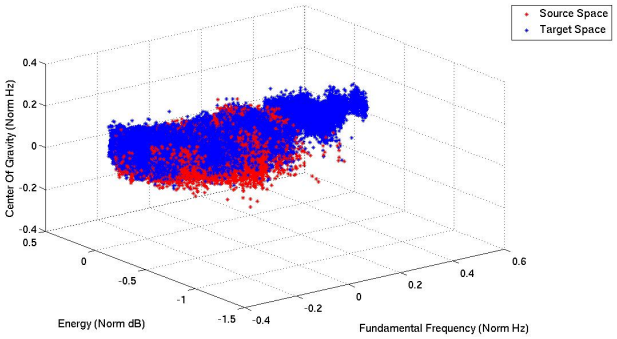


Fig. 13. The two parts of the database in the 3D space.

IV. Browsing the Cloud for Synthesis

Let us note by $x_{(i,j)}$, the j^{th} frame of the i^{th} violin piece of the database and by $X_{(i,j)}$ its corresponding vector of features. We also name x_t the frame $x_{(i,j)}$ selected at time t and X_t its vector of features.

We need to find a sequence of T database frames $(x_1, ..., x_T)$ whose features $(X_1, ..., X_T)$ best match a given sequence of features $(Y_1, ..., Y_T)$, whether these are extracted from an actual violin piece or produced by some gesture-based heuristic. The sequence of X_t needs to be consistent in many ways (realistic and continuous feature trajectories, overall acoustic and musical quality, ...). Moreover the selection process has to be fast enough to be used in a real-time environment such as Max or Pd.

A. Clustering and Pruning

First a standard k-means clustering is performed on the f_0-Energy-CoG normalized space of features so that its 150000 frames are spread around 128 centroids C_k $(k = 1, ..., 128)$. This k-means clustering is achieved thanks to Spider[1], a Matlab library of objects for machine learning purpose, released under the terms of the GPL. The result of the clustering in the f_0-Energy space is shown in Figure 14

Note that this clustering is an offline process that has to be done only once, when the database is created. Obviously a clustering for every X_t sequence generation would be everything but a real-time process.

Then for every vector Y_t that has to be matched by a vector X_t from the database we find the k-means centroid C_k that minimizes $\mathcal{M}(Y_t, C_k)$, the Mahalanobis distance,

$$\mathcal{M}(Y_t, C_k) = \sqrt{(Y_t - C_k)^T \Sigma^{-1} (Y_t - C_k)}, \qquad (9)$$

[1]http://www.kyb.tuebingen.mpg.de/bs/people/spider

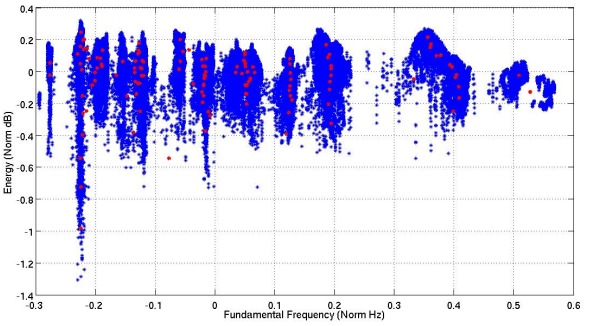


Fig. 14. Result of the k-mean clustering. Horizontal axis is f_0, verical axis is energy. Red dots marks the position of the cluster centroids.

where Σ is the covariance matrix of all the $X_{(i,j)}$ in the database.

X_t will be selected amid the subset of vectors $X_{(i,j)}$ attached to that centroid. Let us denote by X^{k_t} that subset.

In order to reduce the algorithm duration we can downsize that subset by discarding all the elements of X^{k_t} whose Mahalanobis distance to the target vector Y_t is beyond a certain threshold. Another approach is to limit the number of frames in the subset to the N closest to Y_t. We selected the latter in our implementation and fixed $N = 10$.

Finally the last step, detailed in section IV-B, is to add some continuity constraint when selecting one of the N remaining vectors in X^{k_t} to match Y_t .

B. Selection

The continuity constraint is added through a Viterbi-based method. Indeed the X_t sequence has to minimize a global cost which is a cumulative sum of target and concatenation costs.

The target costs at time t are computed as the Mahalanobis distance between each one of the N X^{k_t} and the target Y_t. These distances have already been computed during the pruning step.

As for the concatenation cost, it corresponds to "the price to pay" to go from one of the N vectors of $X^{k_{t-1}}$ at time $t-1$ to one of the N vectors of X^{k_t} at time t. A zero cost should correspond to a transition between two vectors that are in successive order in the original data used in the database.

For instance, if one of the $X^{k_{t-1}}$ is $X_{(2,64)}$ and one of the X^{k_t} is $X_{(2,65)}$ (i.e. respectively the 64^{th} and 65^{th} frames of the second violin piece), the concatenation cost between the two should be null since they can be concatenated seamlessly. Another way to phrase this is : "if the frame at time $t-1$ is $X_{(i,j)}$ then the optimal frame at time t would be frame $X_{(i,j+1)}$".

Therefore we measure the concatenation cost between a feature vector $X_{(i,j)}$ from $X^{k_{t-1}}$ and $X_{(i',j')}$ from X^{k_t} as the Mahalanobis distance between $X_{(i,j+1)}$ and $X_{(i',j')}$.

Finally the sequence $X_1, ..., X_T$ is obtained by using the Viterbi algorithm to minimize the overall cost δ_T.

$$\delta_t = \min(\delta_{t-1} + \mathcal{M}(X_{(i,j+1)}, X_{(i',j')})) + \mathcal{M}(X^{k_t}, Y_t), \quad (10)$$

where each δ_t is a N-dimension vector and

$$\delta_1 = \mathcal{M}(X^{k_1}, Y_1) \tag{11}$$

Then X_T is selected as

$$X_T = \text{argmin}(\delta_T), \tag{12}$$

and then the backtracking step of the algorithm selects every X_t, starting with $t = T - 1$, and so on.

$$X_t = \text{argmin}(\delta_t + \mathcal{M}(X_{(i,j+1)}, X_{t+1})), \tag{13}$$

that is to say X_t will be the $X_{(i,j)}$ whose following frame $X_{(i,j+1)}$ minimizes the overall cost to go to X_{t+1}.

V. Conclusion

In this project we had the opportunity of extensively study a sound space which was not in the track of what we did before. It gave us the possibility to develop our analysis tools (pitch analysis, pitch marking, spectral feature computation) in new directions. The problem of content-oriented sound synthesis with extra-small timing resolution has also been developed. In this context we can conclude that violin sound is particularly difficult to process by overlap-add and that new tracks have to be discussed in this field.

Acknowledgment

The authors would like to thank their respective PhD supervisors for the possibility of participating to a long workshop in another lab and country. We also thank the LIMSI/CNRS lab for the organisation and support. Finally we would like to thank the RW/Numediart project because of this opportunity of combining both regional and international human ressources.

References

[1] "Garage band," http://www.apple.com/ilife/garageband.
[2] "Loquendo," http://www.loquendo.com/en/technology/tts.htm.
[3] "Max/msp," http://www.cycling74.com/products/maxmsp.
[4] M. Wright, A. Feed, and A. Momeni, "Opensound control: State of the art 2003," in *Proceedings of NIME'03*, 2003, pp. 153–159.
[5] N. Schnell and D. Schwarz, "Gabor: Multi-representation real-time analysis/synthesis," in *Proceedings of DAFx'05*, 2005.
[6] M. Cardle and S. Brooks, "Directed sound synthesis with natural grains," in *Proceedings of the Cambridge Music Processing*, 2003.
[7] "Freesound project," http://www.freesound.org.
[8] J. McDermott, N. Grifth, and M. ONeill, "Timbral, perceptual, and statistical attributes for synthesized sound," in *Proceedings of the International Computer Music Conference*, 2006.
[9] J. Charles, D. Fitzgerald, and E. Coyle, "Quantifiing violin timbre," in *DMRN-06: DMRN Doctoral Research Conference*, 2006.
[10] J. Stepanek, "Spectralsourcesofbasicperceptual dimensionofviolintimbre," in *CFA/DAGA*, 2004.
[11] A. de Cheveigné and H. Kawahara, "Yin, a fundamental frequency estimator for speech and music," *Journal of the Acoustical Society of America*, vol. 111, pp. 1917–1930, 2002.
[12] B. Bozkurt, L. Couvreur, and T. Dutoit, "Chirp group delay analysis of speech signals," *Speech Communication*, vol. 49, pp. 159–176, 2007.
[13] H. Kawahara, Y. Atake, and P. Zolfaghari, "Auditory event detection based on a time domain fixed point analysis," in *ICSLP 2000, Proceedings of the International Conference on Spoken Language Processing*, ISCA, Ed., vol. 4, 2000, pp. 664–667.
[14] G. Peeters, "A large set of audio features for sound description (similarity and classification) in the cuidado project," IRCAM, Tech. Rep., 2004.
[15] E. Zwicker, "Subdivision of the audible frequency range into critical bands," *Journal of the Acoustical Society of America*, 1961.
[16] N. d'Alessandro and T. Dutoit, "Handsketch bi-manual controller," in *Proceedings of the NIME'07*, 2007, pp. 78–81.

Nicolas d'Alessandro holds an Electrical Engineering degree from the FPMs since 2004. He did his master's thesis in the Faculy of Music of the University of Montréal. That work gathered the development of an application based on perceptual analogies between guitar sounds and voice sounds, and a study of mapping possibilities between gestures and speech production models. He just started a PhD. thesis in the TCTS Lab of the FPMs related to the real-time control of unit-based synthetizers.

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/lter decomposition of speech and music, phase vocoder, voice activity detection, voice conversion and HMM synthesis.

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).

Jehan-Julien Filatriau 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. 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.

Wenyang Chu received the B.S. degree in electronic engineering from Fu Jen Catholic University, Taipei, Taiwan in 2004, and the M.S. degrees in information and communication technologies from Universitat Politecnica de Catalunya (Spain) and Universite Catholique de Louvain (Belgium) in 2008. He currently is pursuing the Ph.D degree in Laboratoire de Telecommunications et Teledetection, Ecole Polytechnique de Louvain (UCL). His research interests include multimodal signal processing and fusion, and human-computer interaction.

A Database for Stylistic Human Gait Modeling and Synthesis

Joëlle Tilmanne, Raphaël Sebbe, and Thierry Dutoit

***Abstract*—This project aimed at recording a database which can be used as training data for the development of three-dimensional statistical models of various human gait styles. The three-dimensional motion data was captured using the IGS-190 motion capture suit from Animazoo, an integrated wireless motion tracking system equipped with inertial sensors. Fourty participants were recorded, walking at several speeds and performing several direction changes.**

***Index Terms*—motion capture, gait modeling, database.**

I. Introduction

Synthesis of realistic human motion is one of the greatest challenges in computer graphics, as classical animation methods are very time-consuming. A new trend in motion synthesis consists in using Machine Learning techniques to model and reproduce humanlike motion data. When using this kind of techniques, there is a need for representative data. This project is inscribed in this context, as its main goal is to record a database of three-dimensional human gait data, which could later be used for the training of statistical models of human gait and provide mechanisms to generate stylistic variations of these models.

In the next section, we give a brief introduction to the context of walk synthesis and character animation. In section 3, we introduce our database, and on what purpose it was built. In Section 4, we will give some information about the IGS-190 motion capture suit which is used for the recordings. Section 5 will explain the recording setup and instructions, and thus explain the content of the database. Section 6 will explain the format under which the data is represented. In the last section, we will make some remarks on the recorded data, and give some perspectives about the use of the database.

II. Animation of virtual characters

As of today, the only application where motion synthesis is truly used and is not only studied for research purposes is the animation field. Animation consists of the rapid display of a sequence of images, which creates an optical illusion of movement due to the human phenomenon of retinal persistence. The animation process is applied to various domains, ranging from the coarse motions seen in video games to the precise humanlike motions of 3D movies, and including fields like virtual reality or character animations for human-computer interactions. There are currently three main kinds of techniques which are used for motion synthesis in the animation field:

- keyframe animations (hand-driven),
- model-driven techniques in which a software based on a set of expert-based rules (i.e. the model) helps the animator for motion production,
- motion capture (data-driven technique).

Not that long ago, those methods were the only ones existing. Unfortunately, all of them have strong drawbacks. The solutions that exist for human character animation or synthesis of human motion in general are either time and labor consuming, or they are unrealistic and of poor quality.

For a few years now, a new trend has appeared. It consists of using statistical learning techniques to automatically extract the underlying rules of human motion, without any prior knowledge, directly from training on 3D motion capture data. Starting from the statistical models trained that way, new motion sequences can be automatically generated, using only some high-level commands from the user. Two movements generated by the same command (for example executing two consecutive steps in a walk sequence) will never be exactly identical. The result presents indeed a random aspect as can be found in the human execution of each movement, and becomes potentially more realistic than the repetition of the same capture sequence over and over. The movements produced that way are thus visually different, but are all stochastically similar to the training movements.

Bipedal locomotion is by nature a periodic time sequence of movements, which involve all body parts, with an underlying sequence of repeated phases. Following [1], human gait can be viewed as a dynamical process with a simple temporal structure (i.e., repetitions of basic stance and swing phases) but with several stylistic variations (i.e., walking, running, trudging, ...). Besides, the gait observations depict high variability that can be explained by the incapacity of a walker to produce perfectly twice the same movements and by the varying nature of its environment. All these elements create the naturalness that is so difficult to achieve in walking avatar animation. The modelling of such a process can be addressed by using probabilistic graphical models [2], and more especially Dynamic Bayesian Networks (DBN) [3] and Hidden Markov Models (HMM) [4], which provide mathematically tractable and computationally efficient solutions for their estimation.

III. Need for a database

In all Machine Learning techniques, the first major issue is to obtain enough representative training data. As only motion capture can give realistic 3D human motion data, it is the only way to obtain representative training data for statistical modeling of human motion. The first option is of course to

use an existing database of captured motions. Several motion capture databases are available for free on the Internet, like the widely known Carnegie Mellon University (CMU) Graphics Lab Motion Capture Database [5], or motion provided for free by a society called Eyes, from Japan, referred to as "motion-data.com" [6], but in both cases the motions are only roughly described with a single definition word like "walk", "jump", etc. Some other sites sell the motions, like "Truebones" [7] but again, the description of those movements is very brief as they are aimed at a direct use in an animation sequence. There is thus almost no documentation on the recorded sequences of motion, on the people who performed them, or on the motion capture area. This makes it hard to compare walk sequences from different people as they are not all recorded in the same conditions, and the information about each recording is very sparse. The HDM05 database from Bonn University [8], [9] gives a little bit more information, as they have the same sequence played by five different subjects. The sequence is a series of different motions corresponding to instructions like walk 5 steps, turn right, walk 5 steps backwards, etc. But there are no speed variations nor extended study of direction changes, and the subdivision of the whole sequence into each separate class of motion is not made for style variations classes of walk.

For model training, we needed better documented data. Moreover, we are not interested in having as many different sequences as possible, but in acquiring enough examples of the same motion to train a model. If possible, we also want to take into account more precise parameters in the training, like the style of the motion, which is, along with other factors, linked to the subject performing it. This project gave the opportunity to massively collect human gait data with an accurate motion capture system. For each of the fourty subjects of the database, the recording conditions and instructions were the same, which enables us to truly compare the inter-people variations. One could then benefit from this valuable database to estimate complex statistical models. The data are collected for various gait conditions, with an inertial motion tracking system, the Inertial Gyroscopic System (IGS 190) from Animazoo [10].

IV. Motion capture equipment: the IGS 190

In most research, the three-dimensional nature of human gait is overlooked [11]. Observations are often reduced to measurements from the sagittal plane (e.g., video recordings from the side plane) and symmetry between left and right sides is assumed. From this limited perspective, much valuable information is lost. Observations can be improved by including measurements from other planes such as coronal (e.g., front plane) and transverse (e.g., top plane). However, human gait is ideally observed via three-dimensional measurements. To do so, various motion capture equipments can be proposed. All so called "motion capture" systems have the same goal, which is to record a real motion by transcribing it under mathematical form usable by a computer. This is achieved by tracking a number of key points through space across time, and combining them to obtain a tridimensional unified representation of the performance [12]. The subject's body is then most of the time represented under the form of a kinematic chain, whose root is situated at the hip level, and whose various segments represent a kind of simplified representation of the skeleton.

Several techniques can be used for motion capture. First, multiple camera views can be triangulated to estimate body marker positions. Such system requires fine calibration and robust computer vision algorithms to yield accurate data [13]. Marker-less optical solutions are more comfortable yet more complex [14]. Besides, mechanical solutions allow directly measuring joint angles but require intrusive sensors. In this project, we adopt the Inertial Gyroscopic System (IGS) developed by Animazoo [10]. The main advantage of this system is to deliver remotely accurate data in a unified manner with no extra solving and over large capture areas.

The IGS 190 motion capture suit contains 19 inertial sensors (InertiaCube3 from Intersense) consisting each of a 3 axis accelerometer, a 3 axis gyroscope and a 3 axis magnetometer. The data from those three sources are integrated and combined directly in the inertial sensor box, and angle data is thus provided straight from the sensor. Those angles which give the configuration of the inertial sensor are then postprocessed realtime trough the software provided with the suit. The output of the system is the set of angles between the body segments, at a rate of 30, 60 or 120 frames per second. In this database recording, the frame rate is chosen to be 60 fps (frames per second).

Although each of the sensors is wired to the main processing unit which is attached to the suit, the system is wireless as this main processing unit acts as emitter and communicates trough radio frequency with the receiver linked to the laptop trough USB connection. The wireless range of the suit is around 100 meter indoors and up to several hundred meters outdoors.

The AutoCal software enables the user to build a skeleton fitting any new actor wearing the suit with just two photographs taken in a given stance and within a 3D cube used for setting the scale of the picture. The actor will need to stand in the same stance, facing north, in the beginning of each motion capture session, in order to calibrate the system (which contains magnetometers).

There is no 3D position tracking system in the IGS suit, and the position of the actor is calculated by the software given a known initial position, using the length of the skeleton segments from the feet to the hip, the angles recorded between those segments for each frame, and always considering that the lower point of the skeleton is in contact with the ground.

Because of that, drifts can appear, but from our experiences with the equipment, if the calibration step is properly done, as well for the skeleton calibration as for the initial position of the actor at the beginning of each motion capture session, we obtain in most capture sequences very good quality of data.

V. Recording setup

This database contains 17 walk sequences for 40 subjects. Those 17 sequences correspond, for each subject, to rectilign walks at different speeds, and to segments with direction changes of various amplitudes in the trajectory. The subjects

were participants of other projects of the workshop who agreed to volunteer for the motion capture recording.

The motion capture area was a 6m by 4m space, in which subjects were asked to walk given some instructions. The first step was to take the calibration pictures and to calibrate a skeleton for each participant. Once this was done, each subject came for approximately one hour of recording. They were first asked to remove their shoes and put on the motion capture suit. The walk in the database is thus always barefoot, on a perfectly even floor.

The first part of the motion capture session was the recording of the trajectory changes. Several crosses were drawn on the floor, as shown in figure 1. For each one of the five direction change sequences, the performer began on the red cross on the left upper side of the figure, facing right. The instruction was to walk straight until they reached the red line (which was also drawn on the floor), and then they were free to change direction the way they wanted to, they just had to reach one of the other crosses drawn on the floor, with different angles from the initial straight trajectory. For the first sequence, the cross to reach was just straight ahead (0 degree), and for the following ones, the angles of the direction change were respectively 45, 90, 135 and 180 degrees.

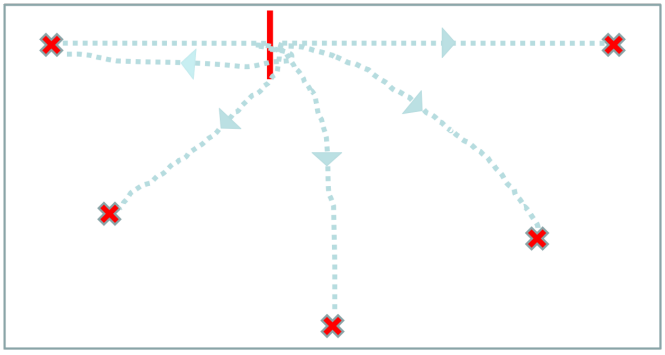


Fig. 1. Sketch of the motion capture area for the five direction change recordings. Only the red crosses and line were drawn on the floor.

For the second part of the motion capture, the aim was to study how walk changes when we change our velocity. This time, the performer was told to walk straight from one corner of the capture area to the opposite corner (approximately 7 meters). He was given four different instructions, and the sequence corresponding to each of these instructions had to be repeated three times. Those four instructions were the following:

- Walk (normally).
- Imagine you are in a park, the weather is nice, you take your time.
- You are going to work, but you are not late.
- You have to catch a train, and you are very late, but you are not allowed to run.

These instructions let to the performer the freedom to choose his speed for each scenario, even if the gradation from slower to faster was the same for everybody, with the first "walk normally" instruction as a reference for a standard self selected speed.

VI. Data representation

The output of our motion capture system are angles between body segments, and the calculated 3D position of the root (hips) which is evaluated from the angles and the segment lengths of the legs, as explained previously(section IV). That data is recorded under the form of a Biovision file format (.bvh), which is a "de facto" standard format for 3D motion data. A bvh file contains two parts. The first part is the hierarchical description of the skeleton, starting from the root (hip) to the extremities of each limb, and the geometric position of each joints, still standing. The second part of the file contains one line of data for each motion frame. The skeleton from the Animazoo software contains 20 body segments in the skeleton (see figure 2), and each data line contains thus 3 (3D position of the root) + 3×21 (3D angles for each segment and for the whole body) = 66 values. Three of the segments are in fact only added to make the skeleton look closer to the real skeleton but have no degree of freedom in the motion capture, and their value does thus not vary during motion capture. There are thus finally 57 values to analyze and model in our data.

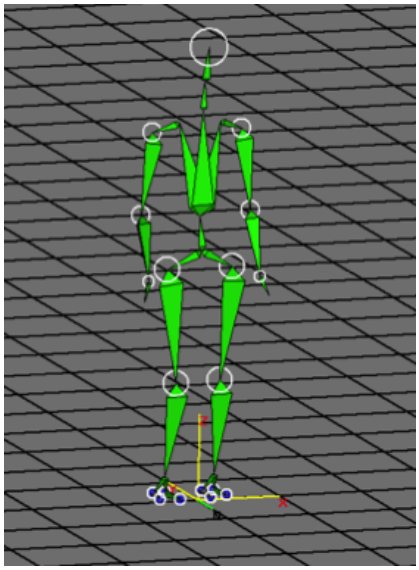


Fig. 2. Skeleton calibrated and displayed in Animazoo's software during motion capture.

As we have said in section IV, drifts can appear when using such an inertial motion capture system and it is the case for a few motions of the database, where the 3D position is not properly calculated. The appearance of the walk is then still good when looked at without 3D displacement, but the whole body looks like if instantly drifted backwards. Another small disturbance that can sometimes be observed is when one of the sensors moves because of the suit and not because of the subject's motion.

VII. Data visualization and segmentation

The project has a need for a tool to manually segment the motion recordings to keep only those parts needed by the training process. In this context, Motion Editor.app was

developed, which consists of a 3D view of the skeleton and an interactive timeline. Moreover, this application allows the labeling of time segments that can be used for specific training purposes.

The software parses the Biovision file (.bvh) and builds a memory representation of its content. This includes both the skeletal representation, i.e. the relative offsets of the joints, and the time sampled 3-axis rotations of each joint [1]. Then, it displays the skeleton at any point in time as an Open Scene Graph [2] 3D scene view.

A timeline is also provided, which enables time navigation with immediate display of the skeleton. Additionally, the timeline permits to select temporal segments, and associated motion frames can be deleted. Alternatively, the time segment can also be given a label for identification (and further processing). The edited, in-memory frame representation can then be saved as a new .bvh file, and the labels, if any, can be saved to a separate labels file.

The graphical user interface of Motion Editor is shown of figure 3.

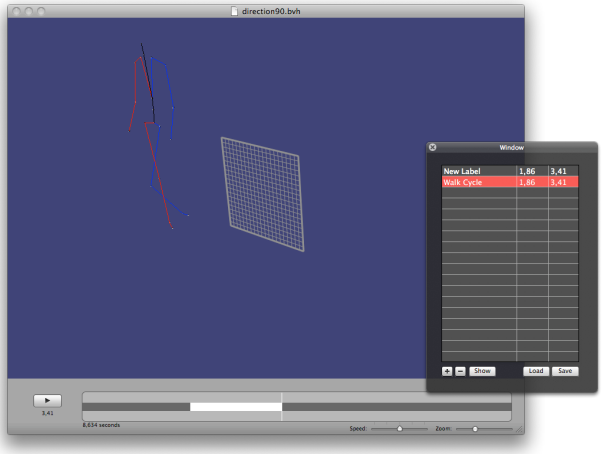


Fig. 3. Sketch of the motion capture area for the five direction change recordings. Only the red crosses and line were drawn on the floor.

VIII. Conclusion

The flexibility of the IGS 190 equipment enabled the recording of a large database of human walk and its variations. A preliminary visual analysis of the database already allowed us to observe the obvious style differences between walk from a woman and a man, the way that a person walks at different speeds, as well as the differences in walk parameters for two subjects with comparable morphology. From now on, this new database will allow us to study inter-subject variability in walk, walk variations linked to speed, and the way humans manage direction changes when they walk on an even ground without obstacles, and to train statistical models to model and to synthesize new motion.

[1] this 66 DOF record is called a frame

[2] Open Scene Graph is an open source library based on OpenGL use to display 3D content

Acknowledgment

The authors would like to thank all eNTERFACE'08 attendants who participated in the creation of the database.

This project was partly funded by the Ministry of Région Wallonne under the Numediart research program (grant $N^0$716631). J.Tilmanne receives a PhD grant of the Fonds de la Recherche pour l'Industrie et l'Agriculture (F.R.I.A.), Belgium.

References

[1] M. Brand and A. Hertzmann, "Style machines," in *Proc of SIGGRAPH 2000*, 2000, pp. 183–192. [Online]. Available: http://www.merl.com/reports/docs/TR2000-14.pdf

[2] S. L. Lauritzen, "Graphical models," 1996.

[3] Z. Ghahramani, "Learning dynamic bayesian networks," in *Adaptive Processing of Sequences and Data Structures*. Springer-Verlag, 1998, pp. 168–197.

[4] L. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," in *Proc. of IEEE*, vol. 77, no. 2, February 1989, pp. 257–286.

[5] "The Carnegie Mellon University Graphics Lab Motion Capture Database," http://mocap.cs.cmu.edu.

[6] "The "mocapdata" database," http://www.mocapdata.com.

[7] "Truebones website," http://www.truebones.com.

[8] "The HDM05 motion capture database," http://www.mpi-inf.mpg.de/resources/HDM05.

[9] M. Müller, T. Röder, M. Clausen, B. Eberhardt, B. Krüger, and A. Weber, "Documentation mocap database HDM05," Universität Bonn, Tech. Rep. No. CG-2007-2, June 2007. [Online]. Available: http://www.mpi-inf.mpg.de/resources/HDM05/07_MuRoClEbKrWe_HDM05.pdf

[10] "Animazoo inertial motion capture system IGS-190," http://www.animazoo.com.

[11] C. L. Vaughan, B. L. Davis, and J. C. OConnor, *Dynamics of Human Gait, 2nd ed.* Kiboho Publishers, 1999.

[12] A. Menache, *Understanding motion Capture for Computer Animation and Video Games*. San Francisco, CA, USA: Morgan Kauffman Publishers Inc., 1999.

[13] B. Rosenhahn, T. Brox, and H.-P. Seidel, "Scaled motion dynamics for markerless motion capture," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2007.

[14] A. Colahi, M. Hoviatalab, T. Rezaeian, M. Alizadeh, and M. Bostan, "Implementation of optical tracker system for marker-based human motion tracking," in *Proceedings of the IASTED International Conference on Applied Simulation and Modelling (ASM)*, June 2006.

Joëlle Tilmanne holds an Electrical Engineering degree from the Faculté Polytechnique de Mons (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 (Circuit Theory and Signal Porcessing) Lab of FPMs since september 2006, in the field of HMM based motion synthesis.

Raphaël Sebbe holds a PhD in the field of image processing as well as electrical engineering degrees from both FPMs and Supélec. Raphaël has worked on speech and image-related technologies and software development. He is now with TCTS Lab of FPMs, pursuing research on Numédiart project.

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 full 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, 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 a member of the INTERSPEECH07 organization committee. He was the initiator of eNTERFACE workshops and the organizer of eNTERFACE05.

eNTERFACE'08 Participants

- Bénédicte Adessi, France, benedicteadessi@hotmail.com
- Rami Ajaj, LIMSI-CNRS, Orsay, France rami.ajaj@limsi.fr
- Samer Al Moubayed, Royal Institute of Technology KTH - Speech Music and Hearing Dept., Sweden,sameram@kth.se
- Gopal Ananthakrishnan, Royal Institute of Technology KTH - Speech Music and Hearing Dept., Sweden, agopal@kth.se
- Oya Aran, Bogazici University - Computer Engineering, Turkey, aranoya@gmail.com
- Carole Arrat, France, acarole17@hotmail.com
- Malek Baklouti, Thales - D3S, France, malek.baklouti@thalesgroup.com
- Jean Bernard, Université Paris VIII, France, j.bernard@iut.univ-paris8.fr
- Tifanie Bouchara, LIMSI-CNRS, Orsay, France, tifanie.bouchara@hotmail.fr
- Pierre Breteche, LASELDI, Université de Franche-Comté, France, pierre.breteche@univ-fcomte.fr
- Pavel Campr, University of West Bohemia - Department of Cybernetics, Czech Republic, campr@kky.zcu.cz
- Antonio Camuri, University of Genova, Italia, antonio.camurri@unige.it
- Aleksandra Cerekovic, Croatia, aleksandra.cerekovic@fer.hr
- Mohamed Chetouani, Université Pierre et Marie Curie, France, mohamed.chetouani@upmc.fr
- Hee-Seung Choi, choimode@gmail.com
- Matthieu Courgeon, CNRS-LIMSI, Orsay, France, courgeon@limsi.fr
- Laurent Couvreur, Faculté Polytechnique de Mons - Théorie des Circuits et Traitement du Signal, Belgium, laurent.couvreur@fpms.ac.be
- Nicolas d'Alessandro, Polytech.Mons - Signal Processing Lab, Belgique, nicolas@dalessandro.be
- Christophe d'Alessandro, LIMSI-CNRS, Orsay, France, cda@limsi.fr
- Nicolas Déflache, Independant, France, contact@nicolasdeflache.fr
- Jonathan Demeyer, Faculté Polytechnique de Mons - TCTS Lab, Belgium, Jonathan.Demeyer@fpms.ac.be
- Hamdi Dibeklioglu, Bogazici University - Computer Engineering Department, Turkey, hamdimail@gmail.com
- Thomas Dubuisson, Faculté Polytechnique de Mons - TCTS Lab, Belgium, thomas.dubuisson@fpms.ac.be
- Thierry Dutoit, Faculté Polytechnique de Mons - TCTS Lab, Belgium thierry.dutoit@fpms.ac.be
- Nesli Erdogmus, Turkey, neslibozkurt@hotmail.com
- Jean-Julien Filatriau, Université catholique de Louvain (UCL) - Communication Lab (TELE), Belgium, jehan-julien.filatriau@uclouvain.be
- Takuya Furukawa, Kyoto Univ, Japan, furukawa@ii.ist.i.kyoto-u.ac.jp
- Simone Ghisio, INFOMUS, Genova, Italia, ghisio@infomus.org
- Donald Glowinski, University of Genova, Italia, donald.glowinski@unige.it
- David Antonio Gómez Jáuregui, Télécom SudParis - EPH, France, david.gomez@it-sudparis.eu

- Fabienne Gotusso, France, fabiennegot@orange.fr
- Catherine Guastavino, University McGill, Montreal, Quebec, Canada, catherine.guastavino@mcgill.ca
- Gökhan Himmetoglu, Turkey, hhimmetoglu@ku.edu.tr
- Marek Hruz, University of West Bohemia - Department of Cybernetics, Czech Republic,mhruz@kky.zcu.cz
- Hung-Hsuan Huang, Kyoto University - Graduate School of Informatics, Japan, huang@ii.ist.i.kyoto-u.ac.jp
- Christian Jacquemin, LIMSI-CNRS, Orsay, France , jacquemin@limsi.fr
- Alexey Karpov, SPIIRAS, Russia, karpov@iias.spb.su
- Loïc Kessous, LIMSI-CNRS, Orsay, France loic.kessous@limsi.fr
- Byung Jun Kwon, byungjun@gmail.com
- Ju-Hwan Lee, University of Oxford - Crossmodal Research Laboratory, UK (but Korean), juhwan.lee@psy.ox.ac.uk
- Martin Lojka, martin.lojka@tuke.sk
- Adolfo Lopez, UPC - Image & Video Processing Group, Spain, alopez@gps.tsc.upc.edu
- Benoît Macq, Université catholique de Louvain (UCL) - Communication Lab (TELE), Belgium,Benoit. Macq@uclouvain.be
- Ammar Mahdhaoui, ISIR - Percetion and Movement P&M , France, Ammar.Mahdhaoui@robot.jussieu.fr
- Matei Mancas, Faculté Polytechnique de Mons - TCTS Lab, Belgium Matei.Mancas@fpms.ac.be
- Jean-Claude Martin, LIMSI-CNRS, Orsay, France martin@limsi.fr
- Hildeberto Mendonca, Université catholique de Louvain - Electric Engineering, Belgium, hildeberto.mendonca@uclouvain.be
- Alexis Moinet, , Université catholique de Louvain (UCL) - Communication Lab (TELE), Belgium alexis.moinet@fpms.ac.be
- Camille Moussette, Umea Institute of Design, Sweden, camille.moussette@dh.umu.se
- Emma Murphy, University McGill, Montreal, Quebec, Canada, emma.murphy@mail.mcgill
- Yukiko Nakano, y_nakano@agate.plala.or.jp
- Daniel Neiberg, KTH - TMH, Sweden, daniel.neiberg@gmail.com
- Toyoaki Nishida, Kyoto University - Graduate School of Informatics, Japan, nishida@i.kyoto-u.ac.jp
- Stanislav Ondas, Faculty of Electrical Enginnering and Informatics, Technical University of Košice, Slovakia - Department of Electronics and Multimedia Communications, Slovakia, stanislav.ondas@gmail.com
- Michal Osowski, michalosowski@yahoo.com
- Thomas Pachoud, www.didascalie.net, France, thomas.pachoud@gmail.com
- Igor Pandzic, igor.pandzic@fer.hr
- Catherine Pélachaud, catherine.pelachaud@inria.fr
- Serafeim Perdikis, ITI/CERTH, Greece, perdik@iti.gr
- Laurent Pointal, CNRS - LIMSI, France, laurent.pointal@limsi.fr
- Thierry Ravet, Thierry.Ravet@fpms.ac.be
- Albert Rilliard, CNRS - LIMSI, albert.rilliard@limsi.fr
- Renaud Rubiano, didascalie.net, France, renaud@renaudrubiano.com
- Usman Saeed, Eurecom - Multimedia Dept, France, usman.saeed@eurecom.fr

- Albert Ali Salah, CWI - Signals and Images, The Netherlands, a.a.salah@cwi.nl
- Pinar Santemiz, Boğaziçi University - Computer Engineering, Turkey, psantemiz@gmail.com
- Ao Shen, University of Birmingham, UK, shenany@gmail.com
- Yannis Stylianou, yanni@csd.uoc.gr
- Shinya Takeda, Japan, shinya_ddr@yahoo.co.jp
- Joëlle Tilmanne, Faculté Polytechnique de Mons, Belgium, joelle.tilmanne@fpms.ac.be
- Jana Trojanova, University of West Bohemia - Department of Cybernetics, Czech Republic, trojana@kky.zcu.cz
- Dimitrios Tzovaras, Dimitrios.Tzovaras@iti.gr
- Jérôme Urbain, Faculty of Enginneering, Mons - Circuit Theory and Signal Processing Laboratory, Belgium, jerome.urbain@fpms.ac.be
- Charles Verron, charles.verron@orange-ftgroup.com
- Athanasios Vogiannou, CERTH/ITI, Greece, tvog@iti.gr
- Gualtiero Volpe, DIST - University of Genova - Casa Paganini - InfoMus Lab, Italy, gualtiero.volpe@unige.it
- Olga Vybornova, olga.vybornova@uclouvain.be
- Chu Wen-Yang, Taiwan, wenyang.chu@gmail.com
- Yuji Yamaoka, Japan, 50007646208@st.tuat.ac.jp
- Mehmet Mustafa Yilmaz, Turkey, mehyilmaz@ku.edu.tr
- Milos Zelezny, University of West Bohemia - Department of Cybernetics, Czech Republic, zelezny@kky.zcu.cz

Table of Contents