Multimodal Integration of Sensor Network

Joachim Neumann, Josep R. Casas, Dušan Macho, Javier Ruiz Hidalgo
Signal Theory and Communications Department
UPC – Technical University of Catalonia
Campus Nord edifici D5 Jordi Girona 1-3, 08034 Barcelona, SPAIN
{joachim,josep,dusan,jrh}@gps.tsc.upc.edu

Abstract. At the Universitat Politècnica de Catalunya (UPC), a Smart Room has been equipped with 85 microphones and 8 cameras. This paper describes the setup of the sensors, gives an overview of the underlying hardware and software infrastructure and indicates possibilities for high- and low-level multi-modal interaction. An example of usage of the information collected from the distributed sensor network is explained in detail: the system supports a group of students that have to solve a lab assignment related problem.

1 Introduction and Motivation

The smart room at UPC has been designed to hold group meetings, presentations and undergraduate courses in small groups. The multimodal integration of the sensors in distributed sensor network aims at providing services to the participants in the smart room, which go beyond the computing capabilities of non-integrated computer and sensor-networks.

The UPC smart room permits implementation and testing of a large variety of audio technologies, such as Automatic Speech Recognition (ASR), Speaker Identification (SID), Speech Activity Detection (SAD), Speaker Localization & Tracking (SLT), Acoustic Event Detection (AED), etc. At UPC we are currently active in SID, SAD, SLT, and AED audio technologies.

For video technologies, the multicamera setup in the smart room allows experimenting with visual analysis technologies that strongly rely in exploiting the available redundancy when the same scene is seen from up to 8 different cameras. Not only 3D visual analysis is possible in the smart room, but also any 2D visual analysis approach can be improved by selecting at any time the best camera for a given analysis task. The list of video technologies currently being developed in the smart room are Person Localization and Tracking (PLT), Face Detection (FD), Face

ID (FID), Body Analysis (BA), Gesture Recognition (GR), Object Detection (OD) and Analysis (ODA) and Text Detection (TD).

In addition, multi-modal approaches (audio + video) are being currently investigated for the Person Identification and Person Localization & Tracking technologies.

2 Sensor setup

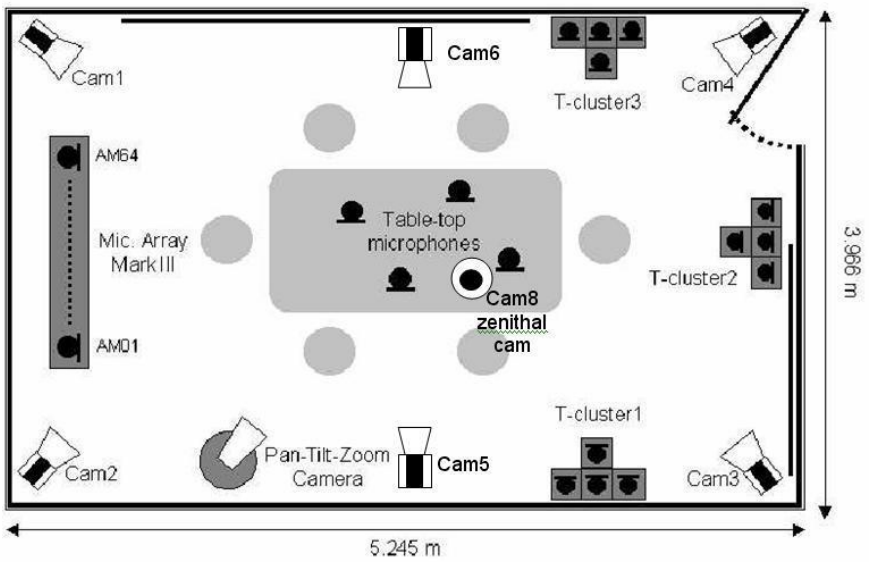


Fig. 1. Sensor set-up of the Smart Room at UPC: the multi-sensor system consists of various audio and video sensors

In order to provide the services to the group of students, the distributed sensor network needs to identify the participants in the room, track their positions over time as well as detect speech and identify voices. The system is capable of continuous monitoring of the UPC smart room [1]. It provides the necessary infrastructure to perform an audio-visual scene analysis as well as a basic modeling of room scenarios. The multi-sensor system is also be used for data collection during technology development.

Audio Sensors

The multi-microphone network should provide audio data for analysis of the acoustic scene in the smart-room by allowing detection and localization of multiple acoustic events, speech activity detection and speech recognition, speaker localization and tracking, etc.

A NIST Mark III 64 microphone array provides audio signal sampled at 44.1 kHz with 24-bit sample representation with all channels sample synchronized. The array is connected to the acquiring computer via Ethernet cable and it is placed close to the wall and approximately 4 m from the main talker area (see Figure1).

Three T-shaped microphone clusters consisting of 4 microphones are positioned on three walls except the wall with Mark III (see Figure 1) at the height of about 2 m. Similarly to Mark III, the clusters provide sample synchronized signals sampled at 44.1 kHz with 24-bit resolution.

Four omni-directional microphones placed on the table without having a fixed position. Additionally, fife close-talking/lapel microphones are very small, barely visible and their signal is wirelessly transferred to allow free movement.

Each of the sensors has its primary task, but it is not limited to it. For example, the Mark III will mostly be used for ASR of the beam formed signal, but its secondary task is acoustic source localization. On the other hand, the three T-shaped clusters are mostly used for audio localization, but they may also be used for ASR if e.g. the position of talker suggests it.

Video Sensors

Cameras placed in the room corners aim at covering the whole area of the smart room. These cameras are used for overall monitoring of the room, to detect and track the locations of people, for articulated body modeling and for classification of activities [2]. They might be also useful for identification of people, head pose estimation and gesture recognition. The requirement for camera placement is that a person in the room should always be in the field of view of at least two cameras. The zenithal camera provides valuable help for person localization and tracking and global activity detection [3]. Cameras in the long walls point at participants seating on the opposite side of the table [4,5]. They are mostly intended for face ID, head-pose and hand gesture recognition and multimodal speaker detection. Finally, the active pan-tilt-zoom camera is primarily aimed at the presenter or the person speaking. It also points at the door when idle, so that a shot of the face of any newcomer is obtained for face ID.

3 Applied Analysis Technologies

Table 1. List of technologies implemented in UPC's smartroom

Technology	Description
Tracker3D	Multi-camera perceptual component that detects foreground objects such as persons or chairs, laptops, etc. The module tracks these objects over time. Foreground (FG) detection is carried out separately over each one of the camera video sequences. The binary masks obtained from this process are input for a particular

Technology	Description
	application of the ShapeFrom Silhouette algorithm [6,7], which results in a) 3D voxelized representation of the foreground objects in the room, and b) improved robustness and consistency in the original 2D FG regions detected from the camera images. A 3D object tracker is then applied for the tracking of the 3D FG labeled objects/persons of interest.
Blob Analysis	Deeper analysis of the abovementioned 3D foreground objects: distinguishes chairs from people, detects body posture (standing, sitting, etc) and gestures (raising hand). A standard model of the human body is aligned to the detected 3D objects. The positions of joints and nodes of the human model are updated over time to track the object, always considering the restrictions of the human body model.
FaceDetector	Detects faces in a camera image and creates a mask containing the face. For 3D FG blobs identified as persons, we resort to the camera providing the best available image or to the PTZ camera (cf. Fig. 1). On this image, and taking into account the foreground region, the contour of the face is traced and input to the next module.
Face ID	Based on the analysis of the Face Detector, this module identifies a face from a face-database. A frontal face view is choosen whenever available. If not, the database considers also side and profile views. The ID of the face is assigned to the 3D object, which, now can be tracked over time until a confirmation/refresh of the ID can be performed.
Object detector	Certain 3D FG blobs not detected as persons are further analyzed by this module. A model-based classification algorithm, classifies them (e.g. "laptop") and analyses their state (lid open / close, on / off).
Teacher GUI	Graphical user interface (GUI) that provides a service to the teacher: it allows the teacher to browse through highlights (snapshot from one of the cameras and a text describing the situation). Also the teacher GUI informs the teacher if one of the students has called him (raising hand) and informs the Memory Jog service, if the teacher acknowledges the student's call.
Highlights	Service provided to the students: allows the memory Jog to give the students a hint, e.g. highlights of the work done by a previous group.
Question (cf. Fig 6)	A GUI that allows to directly interacting with the Memory Jog. It also controls the Selection of the assignment, the selection of the tower and the entry of the solution.
Answer[1]	Question and answering engine. This perceptual component

[1] This technology has been provided by the Natural language processing Group at UPC. However, since this paper is on multimodal integration, we do not go into details here.

Technology	Description
	consists of a front-end that knows about the topic (e.g. the selected tower) and a back-end that generates the answer from a database.
Recorder	Perceptual component that records highlights for the Teacher GUI, which are automatically generated from a snapshot of the scene from one of the cameras and an explaining text.
RoomStatus	A simple foreground pixel-counter that detects activity in predefined areas of the room (e.g. door open / door closed).
SpeechDetector	The output of speech activity detection contains the information about whether there is anybody speaking in the room or not. The UPC SAD system 9], similarly to the SID system, employs frequency filtering features. The original dimension of the feature vector (49) is reduced to 1 by applying linear discriminant analysis (LDA). Then, a decision tree classifier is used to obtain the final speech/non-speech decision. Both the LDA transformation matrix and the decision tree parameters are estimated during the training phase. In the main far-field microphone task of the CHIL evaluations our SAD system achieved 11.78% detection error rate 10]. Due to low computational requirements of the system, nearly a hundred of such systems can be running simultaneously real-time in our smart-room. SAD system is part of the yearly CHIL service demonstrations together with other UPC technologies.
Speaker identification	Speaker identification provides information about the identity of the active speaker. Our SID system is based on Gaussian mixture modeling. As acoustic features we use Mel-frequency cepstral coefficients and frequency filtering features. During the training phase, a Gaussian mixture model (GMM) for the training feature vectors corresponding to the given speaker is estimated; one GMM is estimated for each speaker, and a silence model is also built. During testing, the likelihood that a sequence of feature vectors was produced by the given GMM is calculated for each speaker model and the speaker ID is chosen as the model with the largest likelihood. In the regular evaluations performed within the CHIL project, we achieved a performance of 98.3% correct identifications in the task where only 5 seconds of signal from a far-field microphone is available and the identities of 11 lecturers are needed to be differentiated.
AcousticLocalzator	The speaker localization and tracking technology offers at each timestamp a 3-dimmensional position of the active acoustic source or several sources; it can be a speaking person, but also a ringing phone or moving chair. The UPC

Technology	Description
	SLT system is based on the cross-power spectrum phase approach [11], which we showed is quite robust to the speaker head orientation 12] if using an appropriate distribution of microphone arrays. We use three T-shaped microphone arrays and one linear array (Mark III), so that there is a microphone array at each wall in the room. When evaluating out SLT technology within the CHIL evaluations 10, we achieved 80% of correct localizations, where the localization is considered as a correct if the distance between the system's output and the reference is lower than 50cm. Our SLT system is running real-time in UPC's smart-room and it is being regularly shown in the CHIL service demonstrations complementing the other UPC technologies.
Acoustic event classification	The objective of acoustic event detection is to detect and classify various acoustic events that may occur in a smart-room, such as door opening/closing, phone ringing, chair moving, and also vocal tract produced non-speech sounds such as cough, laugh, etc. AED is a relatively new area and at UPC we currently focus on the investigation of appropriate features and classification/detection methods. In our publications 13-15], we compare and combine ASR features and acoustic features. Also, we showed that the support vector machine approach provides a good classifier alternative to the more common approaches such as Gaussian mixture models.
Wavplayer	The wavplayer is the voice of the Memory Jog service. This perceptual component can synthesize speech and play pre-recorded messages in a polite way.

The following figure shows the cross-connection between a subset of the abovementioned analysis technologies in a smartflow map (see the following section on Software Architecture for an explanation of the distributed data flows)

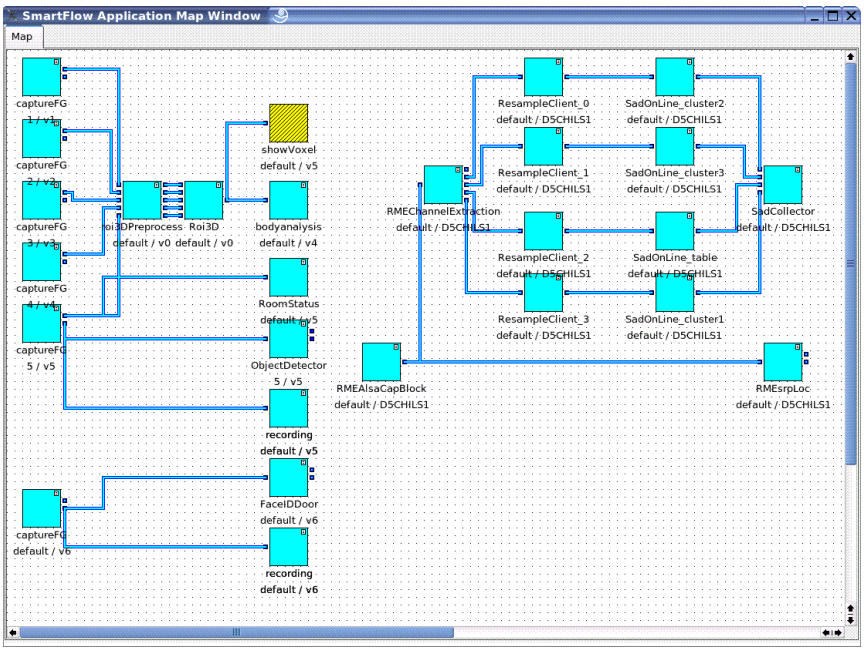


Fig. 2. Smartflow map of the client programs that were running simultaneously on 10 computers

4 Software Architecture

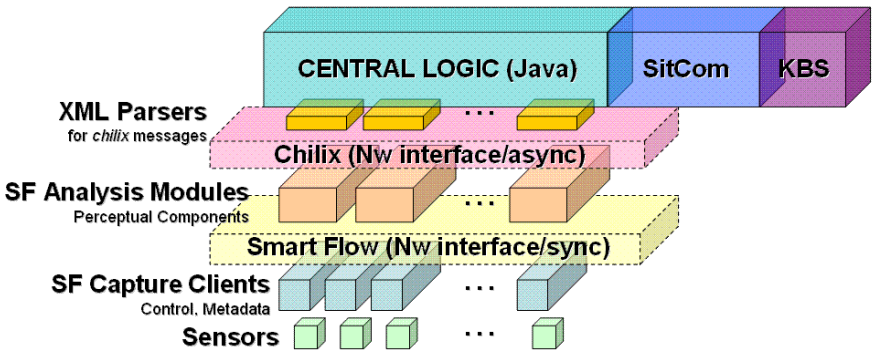


Fig. 3. Block diagram of the software architecture

The distributed computing environment smartflow allows to create and to run different algorithms transparently into a network of different computers. For the UPC Smart Room the NIST Smart Space [8] system has been adopted as the infrastructure for distributed computing. Smart Space provides a client-server mechanism to configure a network of computers. This means that audiovisual modules can be run from any computer (allowing the distribution the complexity of algorithms) and that data communication between modules is handled by the Smart Space server.

The output of each of these smartflow (SF) analysis stages is fed asynchronously into the common central logic framework. This framework guarantees that multimodal audiovisual algorithms can seamlessly access the data captured by audio and video sensors. All analysis data is sent as XML message to a central logic.

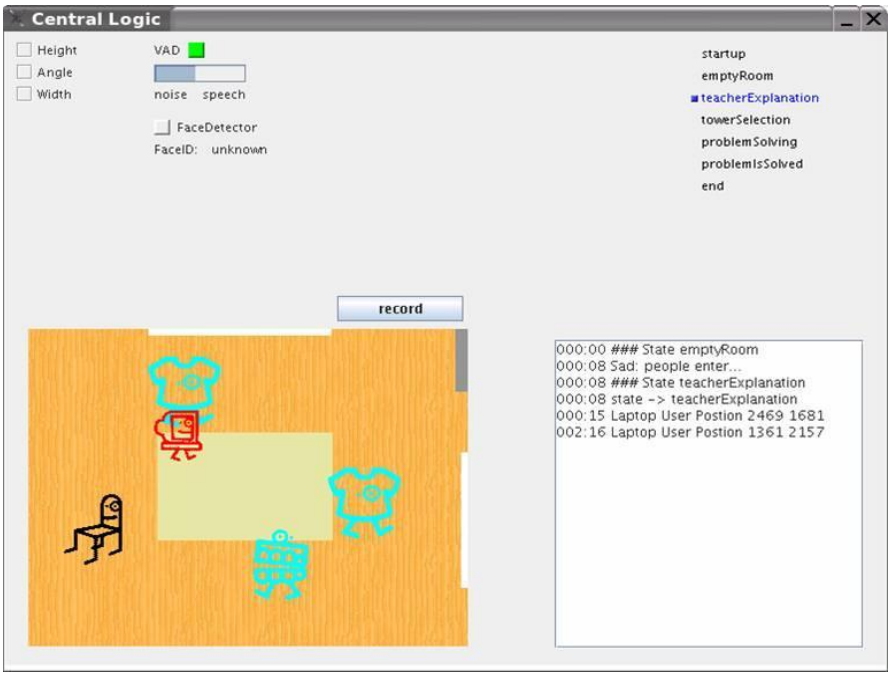


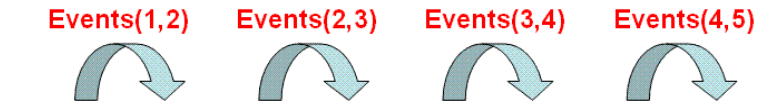
Fig. 4. : GUI of the central logic software that receives the information of all perceptual components.

Although the software architecture has been designed with real-time applications in mind, it provides a mechanism to use audiovisual technology modules in both real time and in non-real time situations. Non-real time scenarios allow capturing data from the room and processing it at a later time. This allows developing audiovisual algorithms or to test non-real time or slow algorithms. The software architecture also provides modules to read and write all captured or processed data into disk and to

reproduce the data at a later time at, for instance, different rates. The audiovisual modules do not need to be modified to adapt to these situations.

Based on the output of the sensors, mono- and multi-modal analysis technologies carry out scene-analysis tasks. This knowledge is fed into a situation model in the central logic in order to understand what is going on in the room. Knowledge about the status of the situation model and the collected data allow services to be provided to the participants of a meeting taking place in the Smart Room. The situation model contains information about:

- The state of the meeting: empty room, Teacher explanation, Tower selection, Problem solving, Solution found.
- The state of each participant: ID information, speaking / non-speaking, position changes, gestures.
- The state of objects: location and classification of objects on the table
- Acoustic events.



State 1	State 2	State 3	State 4	State 5
Empty Room	**Teacher Explanation**	**Tower Selection**	**Problem Solving**	**Solution Found**
Actions(1)	Actions(2)	Actions(3)	Actions(4)	Actions(5)
	startStateChangeTimer	playMessage(m1)	startStateChangeTimer playMessage(m2-m7) startSysHintTimer callTeacher showHint startTeacherButtonTimer	playMessage(m8-m11)
Events(1)	Events(2)	Events(3)	Events(4)	Events(5)
doorOpened faceDetecte	stateChangeTimerEn doorOpen+3sec multimodal	towerSelected	teacherPressButton handRaised stateChangeTimerEnd sysHintTimerEnd teacherButtonTimerEnd	doorOpened

Fig. 5. : State model which is implemented in the central logic.

5 The Memory Jog Service

Service provided by the system is called “Memory Jog”, because it helps participants in the Smart Room by providing background information and memory assistance. The basic concept of the Memory Jog is based on information shift in time and space. For example, the Memory Jog seeks, finds and retrieves information on demand using the Question & Answering technology. The Memory Jog can also translate requests/notice/advice from one user to another user (information shift). This information is not only provided unobtrusively, it can be provided reactively

(on request) or proactively (automatically). To achieve this, the service needs to be context- and content-aware. The implementation of the service further strives at providing the correct information in a polite manner. The following table illustrates the various moments in which the Memory Jog service can be active in a typical situation:

Table 2. List of events that are noticed by the system (in parenthesis its explained how the system learns about the event). The right column shows the action taken by the system.

Event noticed… (how detected?)	**…and react**
Start of lab session (multimodal detection)	starts perception & analysis
Teacher selects and explains the task (interaction with GUI)	Stores information in database, initialized internal timers (for Q&A)
Teacher leaves (multimodal detection)	starts interacting w/students through its voice (a pre-recorded message is played)
A questions is asked (interaction with GUI)	The system answers and notes down if a relevant information has been requested
An assumptions about the task is made(interaction with GUI)	The system notes down the assumption (highlight for Teacher GUI, cf. Table 1)
Someone raises his hand (video technologies)	The system calls teacher through the Teacher GUI (cf. Table 1)
The teacher responds to the student's request or he does not respond to the students request (timeouts in the Teacher GUI)	The system informs students about the teachers arrival or it hives a pre-recorded hint.
The progress of the students is slow	The system sends a pre-recorded hint (Highlight, cf. Table 1). This service is proactive.
The students have reached a solution (interaction with GUI)	The system notes down (highlight for Teacher GUI, cf. Table 1)
The students are leaving the room (multimodal detection)	The system plays a goodbye message and gives further instructions

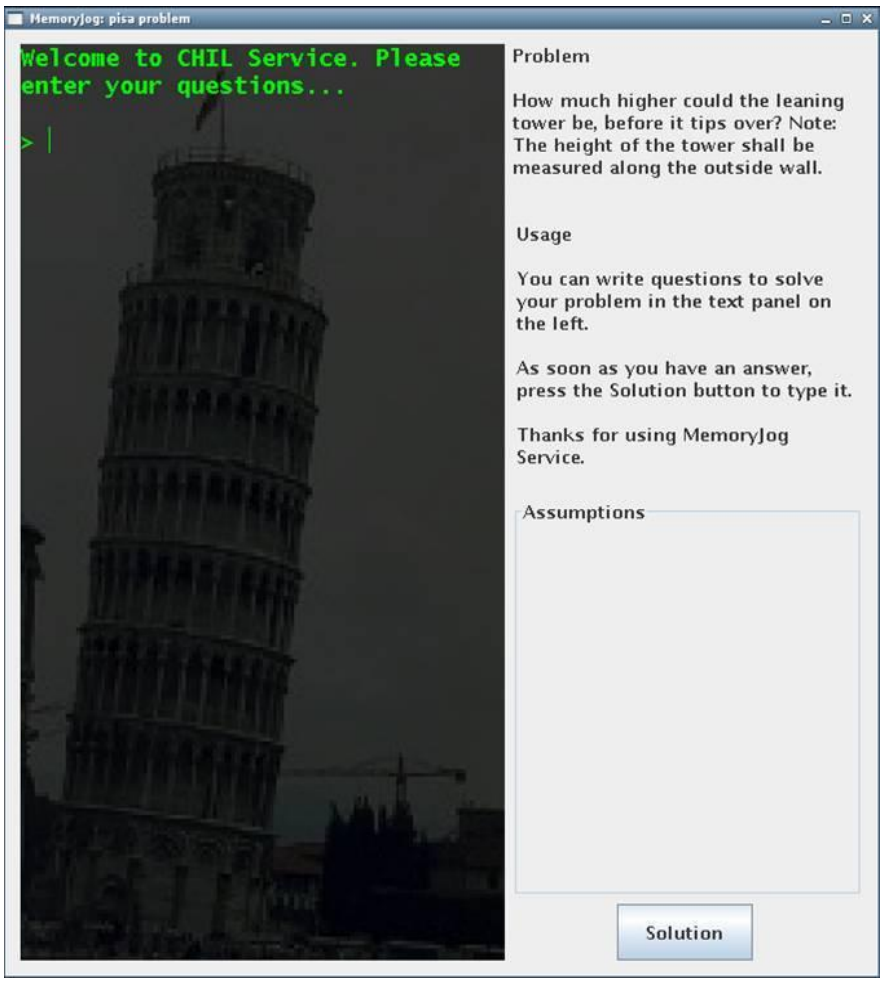


Fig. 6. : Graphical user interface of the Memory Jog service.

6 Conclusion

The application of the multiple sensors and analysis modules in a combined set-up demonstrates that multi-modal integration can be utilized beneficial in a service given to humans (in our case students) in a unobtrusive service that is only possible in a multimodal integration of sensor networks.

References

1. Josep R. Casas, R. Stiefelhagen, et al, “Multi-camera/multi-microphone system design for continuous room monitoring,” CHIL-WP4-D4.1-V2.1-2004-07-08-CO, CHIL Consortium Deliverable D4.1, July 2004.
2. J-L. Landabaso, L-O. Xu, M. Pardas, Robust Tracking and Object Classification Towards Automated Video Surveillance, Proc. of International Conference on Image Analysis and Recognition ICIAR 2004, Porto, Portugal, September 29 - October 1, 2004, Proceedings, Part II, p. 463– 470
3. J. L. Landabaso, M. Pardàs, L.-Q. Xu, Hierarchical Representation of Scenes using Activity Information, Proc of ICASSP 2005, March 18-23, Philadelphia, USA.
4. Josep R. Casas, O. Garcia, et al, “Initial multi-sensor selection strategy to get the best camera/microphone at any time,” CHIL-WP4-D4.2-V2.0-2004-10-18-CO, CHIL Deliverable D4.2, October 2004.
5. O. Garcia, J.R. Casas, “Functionalities for mapping 2D images and 3D world objects in a Multicamera Environment,” 6th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS), Montreux, Switzerland, April, 2005.
6. A. Laurentini, “The visual hull concept for silhouette-based image understanding,” IEEE Trans, Pattern Anal. Mach. Intell., 16(2):150–162, 1994.
7. J.L. Landabaso, M. Pardas, “Foreground regions extraction and characterization towards real-time object tracking,” In Proceedings of Joint Workshop on Multimodal Interaction and Related Machine Learning Algorithms (MLMI ’05), 2005. 3
8. NIST smart space system, http://www.nist.gov/smartspace
9. Padrell J., Macho D., Nadeu C., “Robust Speech Activity Detection Using LDA Applied to FF Parameters”, Proc. ICASSP’05, Philadelphia, PA, USA, March 2005.
10. Macho D., Padrell J., Abad A., Nadeu C., Hernando J., McDonough J., Wölfel M., Klee U., Omologo M., Brutti A., Svaizer P., Potamianos G., Chu S.M., “Automatic Speech Activity Detection, Source Localization, and Speech Recognition on the CHIL Seminar Corpus”, Proc. ICME 2005, Amsterdam, The Netherlands, July 2005.
11. Omologo M., Svaizer P., “Acoustic event localization using a crosspower-spectrum phase based technique,” in Proc. ICASSP’94, Adelaide, 1994.
12. Abad A., Macho D., Segura C., Hernando J., Nadeu C., “Effect of Head Orientation on the Speaker Localization Performance in Smart-room Environment”, Proc. INTERSPEECH – EUROSPEECH 2005, Lisbon, Portugal, September 2005.
13. Temko A., Macho D., Nadeu C., “Selection of features and combination of classifiers using a fuzzy approach for acoustic event classification”, Proc. of 9th European Conference on Speech Communication and Technology, Interspeech 2005, Lisbon, Portugal, September 2005.
14. Temko A., Macho D., Nadeu C., “Improving the performance of acoustic event classification by selecting and combining information sources using the fuzzy integral”, Lecture Notes in Computer Science (LNCS), vol. 3869, February 2006
15. Temko A., Nadeu C., "Classification of Acoustic Events using SVM-based Clustering Schemes", Pattern Recognition, in press, Elsevier, 2006