Context Awareness triggered by Multiple Perceptual Analyzers *

Josep R. Casas, Joachim Neumann
UPC - Technical University of Catalonia

Abstract. A multitude of technologies from computer vision, acoustic signal analysis and natural language processing are used to implement multi-modal perceptual components. The output of this analysis is used to gain context awareness – a necessity when designing a computer-based service that interacts reactively and proactively with humans. This article describes the integration process and our experience in implementing one such information service, the "Memory Jog", in a particular scenario where the computer system supports a group of journalists in their daily work.

Keywords. Multi-modal analysis, Multi-sensor network, Perceptual component, Memory Jog service

Introduction

Computers are tools we use for multiple purposes as providers of information, communication and entertainment. They are great for information management, useful tools for educational tasks and valid helpers for creative design processes. As far as the functional interaction is designed around a well-defined situation, with the computer as the centre of the task and the human as the controller, computers perform tedious tasks for us and have proven to operate efficiently providing help in working scenarios.

One long-term goal in Human Computer Interfaces (HCI) has been to migrate the "natural" means that humans employ to communicate with each other into HCI [1]. Advanced HCI set the user as the centre of the interaction instead of the computer. Instead of the human controller being aware of the task performed by the computer [2], it is the computer which should "be aware" of what the humans are doing around it. This demands equipping computer systems with audio-visual sensors conveying signals from the scene into analysis modules. These analysis modules, also called "perceptual components", are computer programs that analyse the signals from the sensors in a smart environment.

The work presented in this article was developed within the CHIL project [3] and has been inspired by the underlying vision of "Computers in the Human Interaction Loop". CHIL aims to change the way we use computers by yet taking modern HCI one step further. Instead of focussing on the direct Human-Computer interaction, which might keep the Human in the Computer Interaction Loop, a CHIL-like computer service stays in the background and helps the users by understanding their needs and providing support in a natural human-to-human interaction scenario [4]. In the vision of

* This work has been partially supported by the European Union, IP 506909 (CHIL)

CHIL, the humans don't need make any unnecessary effort to "understand" the computer and its processes, because it is the computer which is forced to get into the loop of humans in order to understand their interaction.

The first pre-requisite for such an ambitious aim is to gain context awareness, i.e., detection of people, objects, events and situations in the interaction scene. The information needed to build the relevant context awareness stems from the analysis of the signals acquired in real-time from a collection of sensors. Smart scenarios for HCI base their interaction with the users on "Multimodal interface technologies" for the detailed analysis of the environment, which require the design of flexible and reconfigurable sensor networks feeding data to the perceptual analysis components, as well as actuators providing natural signals to address human users.

However, it is not sufficient to simply gather the analysis result as a growing list of detections coming from the perceptual components. The overall process of Human-Computer interaction should be managed at a higher cognition level, where ontologies of objects, situations and events must be defined. This strategy provides the necessary knowledge complementing the perceptual analysis for adequate understanding of the human-to-human interaction, in order to take the most appropriate action at any time.

In the following sections we first explain how multimodal interface technologies and a simple situation/event model complement each other to build a CHIL-like service. In section 2, we describe in detail the functionality of the Memory Jog service we have implemented. Section 3 reviews the perceptual technologies involved in the service and section 4 presents the overall software architecture allowing the integration in the service prototype.

1. Multimodal interface technologies and situation models for CHIL-like Services

Multimodal interface technologies comprise both the perceptual analysis and the display/synthesis front-end of a CHIL environment, as shown in figure 1. The perceptual analysis front-end consists of a collection of sensors and perceptual components detecting and classifying low-level features which can be later interpreted at a higher semantic level. At the synthesis side, actuators such as an Embodied Conversational Agent or ECA [5] might address the user providing (and requesting) information in a natural, human-like way.

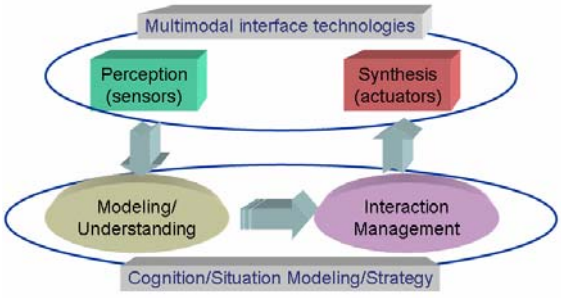


Figure 1. The chain Perception-Understanding-Interaction-Synthesis organized in the two levels of interfaces (sensors/actuators in the upper row) and the higher level of cognition (understanding/interaction management in the lower row).

The smart room at UPC is equipped with multiple cameras and microphones [6]. Continuous room video monitoring is achieved by several calibrated cameras connected to dedicated computers, whose fields of view aim to cover completely the scene of interest, usually with a certain amount of overlap allowing for triangulation and 3D data capture for visual tracking, face localization, object detection, person identification, gesture classification and overall scene analysis. A multi-microphone system for aural room analysis deploys a flexible microphone network comprising microphone arrays, microphone clusters, table top microphones and close-talking microphones, targeting the detection of multiple acoustic events, voice activity detection, ASR and speaker location and tracking. Also for acoustic sensors a calibration step is defined, according to the purpose of having a jointly consistent description of the audio-video sensor geometry. Timestamps are added to all the acquired data for temporal synchronization.

Perceptual components are computing modules analyzing the signals provided by the network of sensors in order to detect and classify objects of interest, persons and events adding information to context awareness. In the Service implemented at the UPC smart room, context awareness consists of knowledge about the number of persons in the room, their identification, position in the room and their orientation. Objects in the room and acoustic events also add to the context awareness.

Context awareness stems from the set of detections coming from the perceptual components, which needs to be properly organized to get a correct understanding of the situation. The model of the possible situations and a well-defined strategy for the interaction should be defined to react properly (i.e. providing the needed information to the user in a timely manner) in the different situations identified by the system. A simple state-model was designed at UPC for the service at hand and the detected events were processed according to the current state of the state-model representing the knowledge introduced in the computer about the events and situations of interest. Some combinations of events were allowed to trigger a change of the state. Based on the current state, some of the detected events trigger a reaction of the Memory Jog service.

2. The Memory Jog as a CHIL-like Service

The Memory Jog Service focuses at providing information. In the specific implementation of the Memory Jog at UPC, information is provided to a group of newspaper journalists gathered together in the CHIL smart room. They have to decide within ten minutes on the front page of tomorrow's edition of their newspaper. In addition to the information provided to the journalists in the smart room, a field journalist and a late-comer propose an additional news story. In this case, the Memory Jog service makes news available that have been created elsewhere (information-shift). The Memory Jog service is also capable of providing background information about the news (information-pull) and in some occasions may decide to jump in human-to-human communication to provide a pro-active service (information-push). The design paradigm behind these three ways to provide information was to enhance human-to-human communication, i.e., the journalists are helped to freely interact with each other instead of being forced to focus on how to interact with the Memory Jog Service. The two most outstanding means of the Memory Jog to interact with the journalists are:

- A Talking Head that not only informs the journalists about available resources, points out events such as the arrival of a latecomer or news being contributed by

remote colleagues, but also facilitates information requests from the journalists in a human-like interface based on automatic speech recognition technologies.

- A remote field journalist is enabled to easily communicate with the journalists in the smart room. A Skype-based bi-directional audio communication is supported by real-time video managed by an automatic cameraman. The video is further enhanced by text annotations that reflect the context awareness of the Memory Jog.

The task of the journalists is to decide on two most important pieces of news to appear on the front page of tomorrow's newspaper. The news from which the journalists have to select from are imported by an up-to-date RSS feed. The following Figure shows an example of the resulting front page.

The UPC Newspaper

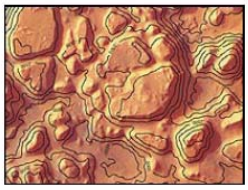
Barcelona, February 12, 2007

Red Planet 'hiking maps' produced

science

The topographic maps show an interesting region of Mars: Iani Chaos

Scientists using data from a European space probe orbiting Mars have produced new topographic maps of the Red Planet.
The "hiker's maps" provide detailed height contours and names of geological features on the Martian surface.
The European Space Agency (Esa), which compiled the maps, said it hoped the maps would become a standard reference for future research on the Red Planet.
The data, from the Mars Express spacecraft, has also been turned into 3-D models of the surface of Mars.
The topographic maps use contour lines to show the heights of the landscape.
The contour lines are superimposed upon high-resolution images of Mars, taken by the High-Resolution Stereo Camera (HRSC) aboard Mars Express.
The maps are much like those of Earth used by hikers and planning authorities.
The samples released by Esa show the Iani Chaos region of Mars because of its major topographical interest.
It is covered in individual blocks and hills that form a chaotic pattern across the landscape.
Mars Express entered orbit around the Red Planet in December 2003.
LATEST NEWS RED PLANET GUIDES VIDEO AND AUDIO



Teraflop chip hints at the future

technology

A chip with 80 processing cores and capable of more than a trillion calculations per second (teraflop) has been unveiled by Intel.
The Teraflop chip is not a commercial release but could point the way to more powerful processors, said the firm.
The chip achieves performance on a piece of silicon no bigger than a fingernail that 11 years ago required a machine with 10,000 chips inside it.
The challenge is to find a way to program the many cores simultaneously.
Current desktop machines have up to four separate cores, while the Cell processor inside the PlayStation 3 has eight (seven of them useable). Each core is effectively a programmable chip in its own right.
But to take advantage of the extra processing power, programmers need to gives instructions to each core that work in parallel with one another.
There are already specialist chips with multiple cores - such as those used in router hardware and graphics cards - but Dr Mark Bull, at the Edinburgh Parallel Computing Centre, said multi-core chips were forcing a sea-change in the programming of desktop applications.
"It's not too difficult to find two or four independent things you can do concurrently, finding 80 or more things is more difficult, especially for desktop applications.
"It is going to require quite a revolution in software

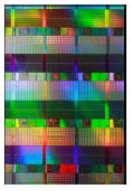


Figure 2. The final result of the work in the journalist scenario: this example shows the front page generated by a group of journalists on February 12, 2007.

3. Perceptual components and technologies

This chapter portrays hardware setup of the UPC smart room and requirements set by the integration of technologies in the Memory Jog Service. Thereafter, the perceptual components that perform the perceptual analysis (first step in Figure 1) are introduced. Subsequent sections describe additional technologies that enable the Memory Jog to actuate and thus to provide its service.

3.1. Hardware setup of the UPC smart room

The unobtrusiveness desired for a natural interaction between humans and computers sets limitations on the positioning of the sensors in the room: the acoustic technologies applied in the UPC smart room limit themselves to far-field wall-mounted microphones that allow the participants to freely move around in the room without being concerned about how and where their voices and other sounds are picked up. However, the signals that these microphones deliver show an unfavourable signal-to-noise ratio and contain a large amount of reverberation due the scarce furniture and the acoustically hard floor and walls.

The consumer-type video sensors were as well chosen and mounted to yield an unobtrusive observation of the whole room. For example, the angle of the corner-cameras is wide enough to cover each point in the room by at least two cameras. Consequently, the quality and details obtained from these lenses is limited. Even the zoom camera that points at the entrance of the room and a webcam on the console shows close-ups of faces which are not more than 60 x 80 pixels in size.

3.2. Integrating Technologies to a Service

The two multi-modal perceptual components described in the following receive the raw audio and video streams as input data. The analysis of these data is often inspired by our knowledge about human perception in vision and hearing. Most of the technologies on which both perceptual components are based have their typical applications with well established evaluation methods. However, the criteria that determine their usefulness can unexpectedly change when they are integrated in a multi-modal system that aims at acquiring context awareness for providing services in real-time. In some cases, astonishing large error rates can be tolerable if a technology is backed up by a similar one that uses a different modality. In other cases, strict criteria of synchronisation and low delay can arise as a consequence of the integration.

When humans experience the computer-driven service like the Memory Jog, another subjective bias naturally arises: unexpected actions of the service triggered by a false-positive detection of one of the technologies turn out to be far more annoying than a service not provided due to false-negative detection.

The following sections look at each of the applied technologies and describe briefly our experience with their role in the Memory Jog Service. Special focus is given to the usability of the output of their analysis rather than to implementation details or to their individual performance.

3.3. Multi-Modal Perceptual Component: Person and Object Tracking

Person Tracking is based on an Acoustic Localizer and a multi-camera 3D Person and Object tracker. The latter detects regions of interest (e.g., persons, chairs or laptops) via foreground segmentation in each of the five cameras. A three-dimensional representation of these regions of interest is obtained by a ShapeFromSilhouette algorithm [7] that receives the binary foreground masks from all five cameras. These 3D regions of interest are consequently labelled and tracked over time. To resolve ambiguities (people crossing, someone sitting on a chair, etc.), a colour histogram is acquired from each person and object. The output of the multi-camera 3D Person and Object Tracker is enriched by (a) an algorithm that is able to distinguish between an object and a person -assuming an average range of physical properties of adult humans- and (b) an algorithm that analyses human body posture [8] (standing, sitting, etc.) with a standard model of the human body that is aligned to the 3D regions of interest earlier classified as 'person'.

The real-time multi-microphone acoustic localization and tracking system [9] is based on the cross-power spectrum phase from three T-shaped microphone arrays. It is robust to the speaker head orientation and provides one or more 3D localizations with detected acoustic activity. The output of the multi-microphone acoustic localization is enriched by an Acoustic Event Classifier [10] that is based on a combination of ASR features and acoustic features. Typical events such as door opening/closing, phone ringing, chair moving, speech, cough, laugh, etc. can be detected.

The combination of video-based and audio-based tracking systems allows the system to gain a basic understanding of what happens in the smart room[1]:

a) A person of interest (e.g. the latecomer) can be tracked in the room. This location is used to direct the talking head and an automatic cameraman (cf. section 3.5) to his current position.
b) The position of all participants can be used to guesstimate changes of the state of the session, e.g. between the states "people enter", "meeting starts" or "coffee break".
c) The position of sudden acoustic event can be determined. The automatic cameraman has been configured to capture these events by choosing the camera that is positioned furthest from the location of the acoustic event.
d) In the current implementation of the context awareness, the detection of a latecomer is based on a multitude of criteria amongst which the first two depend on the Person and Object Tracking: increase of the number of person, appearing of a new object close to the door, detection of the acoustic signal of a door-knock, a door-slam or steps close to the door.

The delay requirements of the Person Tracker are very relaxed in the application described here. Even a delay of 500 ms in the reaction of the system might be unnoticed. Since the precision of the multi-camera tracker is in the order of a few centimetres, a less precise acoustic localisation of the present speaker is tolerable since the origin of the acoustic energy can be re-mapped to the position of the nearest person detected by the visual tracker.

[1] This requires proper handling of timestamps. We chose to synchronize the clocks of the involved computers in a network time protocol cluster (NTP peers) and to always forward the timestamp corresponding to the acquisition of the signal (audio or video frame) to the next processing stage.

3.4. Multi-Modal Perceptual Component: Person Identification

Person Identification is based on Face ID [11] and Speaker ID technologies [12]. The Face ID algorithm is applied to faces that are captured either close to the door or images that are captured by the webcam that is mounted on top of the monitor at the console.

In a pre-processing step, a Face Detector is applied on those parts of the image that have previously been classified as 2-dimensional regions of interest in a binary foreground mask in the pre-processing stage of the Person and Object Tracker. Multiple face-like regions collected at different time instants are then analysed to select a frontal view of the face for further processing by the Face Identification technology. Face ID matches these frontal views against faces stored in a database. If no frontal view is available, the algorithm is capable to base the Face ID on side and profile views, although the identification is less reliable.

The second technology, Speaker ID, provides real-time information about the identity of an active speaker. The SpeakerID algorithm is based on the comparison of Gaussian mixture models. Apart from the poor signal-to-noise ratio and the reverberation in the signals obtained from the far-field microphones, this technology is challenged by the necessity to separate the microphone signal into segments uttered by a single speaker. In an unscripted scenario, this would require the detection of speech activity, the detection of a shift of speakers and the detection of the number of persons speaking simultaneously. In order to circumvent this segmentation problem, the ID of the active speaker is only determined during the usage of the dialogue system (cf. section 3.4), because in this situation the signal naturally stems from a single speaker talking in a quiet background.

Instead of simply listing all recognized detections of the multi-modal Person Identification system, the ID output is assigned to the corresponding person in the Person Tracker. This allows for accumulating the IDs obtained for a person from both Speaker ID and Face ID in the course of the session. The Person Identification technology also provides a valuable feedback to the 3D Person Tracker whether it is still tracking the same person.

Since Person Identification was also used to allow the Talking Head to address the session participants with their name, reducing false positive IDs was emphasised during debugging and error minimization. For the same purpose, both the audio and the video based ID technologies have incorporated a model-class for unknown IDs.

3.5. Dialogue System

The Dialogue System allows a human-like verbal interaction with the Memory Jog System. It is based on two components: A commercially available 2D animation of a talking head (PeoplePutty® by Haptec [13]) and an ASR based dialogue system that utilizes the Cambridge University Engineering Department's HTK [14] recognizer.

Figure 3. Face of the PeoplePutty® Talking Head: The parameters of the talking head software allow adapting its voice, emotions and look to be appropriate for the Dialogue system according to the context.

The speech synthesis part of the talking head was enhanced by a politeness delay unit that acquired a speech activity flag and obliged the speech synthesis engine not to interrupt a human-to-human conversation[2].

The visual representation of the talking head was projected on one of the walls of the smart room - next to the graphical user interfaces (GUI) used by the journalists to publish the front page (cf. Figure 2 and 7).

The talking head also serves as the voice of the Memory Jog, e.g. giving indications about how to use the GUIs, commenting on acoustic events, welcoming the participants or a latecomer, pointing out when the participants run out of time, congratulating upon a successful contribution, saying good bye, etc.

The KTH based speech recognition was trained for two different grammars: the trigger sentences and the commands. The possible phrases of the two grammars are listed in Figure 4.

[2] However, we implemented a message-specific timeout after which the talking head was allowed to interrupt human-to-human communication. The adequate duration of this timeout depends on the amount and duration of pauses at speaker change and thus might depend on the cultural background of the meeting participants.

Trigger sentences

[PLEASE]	CHIL ROOM	[PLEASE]
[PLEASE]	CHIL SERVICE	[PLEASE]
[PLEASE]	CHIL SYSTEM	[PLEASE]
[PLEASE]	CHIL ASSISTANT	[PLEASE]

Commands

	NOTHING	
	NOTHING NOTHING	
[PLEASE]	SHOW ME THE FRONT PAGE OF OUR COMPETITOR	[PLEASE]
[PLEASE]	SHOW ME THE FRONT PAGE OF OUR COMPETITORS	[PLEASE]
[PLEASE]	SHOW ME YESTERDAYS FRONT PAGE	[PLEASE]
	WHO IS OUR FIELD JOURNALIST	[PLEASE]
[PLEASE]	SHOW ME YESTERDAYS BUSINESS NEWS	[PLEASE]
[PLEASE]	SHOW ME YESTERDAYS ENTERTAINMENT NEWS	[PLEASE]
[PLEASE]	SHOW ME YESTERDAYS POLITICS NEWS	[PLEASE]
[PLEASE]	SHOW ME YESTERDAYS SCIENCE NEWS	[PLEASE]
[PLEASE]	SHOW ME YESTERDAYS TECHNOLOGY NEWS	[PLEASE]

Figure 4. The trigger sentences and the commands understood by the dialogue system. Most of the sentences can be accompanied by an optional "please".

When not active, the Dialogue System is listening in the human-to-human conversation to detect one of the trigger sentences. If a trigger sentence is detected with a high enough confidence level, the Talking Head utters "Yes, please?" as positive feedback to inform the person in the room who raised the question. When the following command is understood, the corresponding action is initiated. If the Dialogue System wrongly detects a trigger sentence, the user can re-set the Dialogue System by uttering a simple "nothing".

For the Field Journalist who connects from a remote location, an annotated videoconferencing system was developed. The screen of the field journalist's laptop is shown in the following Figure.

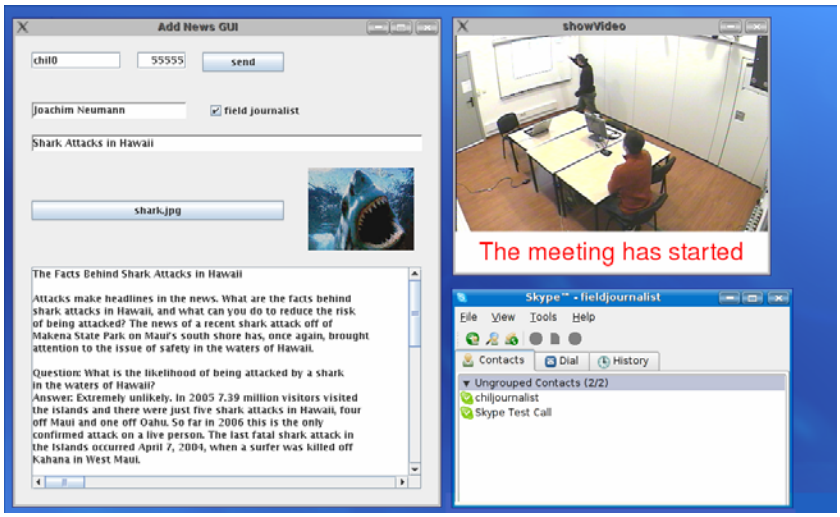


Figure 5. Screenshot of the Field Journalist's laptop: in the lower right, the Skype-based bi-directional audio communication allows talking to the journalists in the room. The upper right shows the real-time video stream from one of the cameras of the meeting room. An automatic cameraman is choosing the optimal camera from five possible angles. This decision is based on the location of the last acoustic event and smoothed by a hysteresis to avoid rapid camera-changes. The real-time video streaming also displays annotations in the form of subtitles that explain the situation, e.g., "people enter", "interaction with ASR", "sound of keys", "front page published", etc or as shown here: "The meeting has started". On the left side of the screen, a graphical user interface allows the field journalist of add a piece of news (a test and an image) to the decision GUI of the journalists in the room.

3.6. Personalized Question-Answering System

The Journalists in the smart room have an advanced Question-Answering System [15] at hand that allows then to ask questions related to the news of previous weeks. In comparison to a Google based search engine, this system is capable of giving the exact answer for factual questions instead of merely listing promising text-snippets.

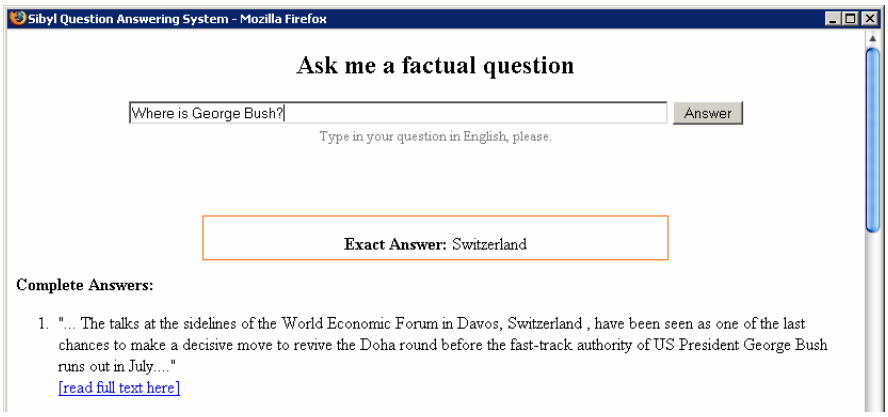


Figure 6. The front end of the personalized Question-Answering system runs in any browser: questions like "Where is George Bush" are answered with a specific location if such a location has been mentioned in the news of the last weeks.

To personalize the output of the Question-Answering System, the result of the Face ID obtained from the webcam at the console is utilised. For this purpose, the field of expertise of each Journalist has been pre-configured. For example, a Journalist working on business news would receive a different answer to the question "Who is meeting in New York?" than a journalist responsible for entertainment news. In order to allow a journalist to access the news from all fields, the automatic selection of specific areas from which the Question-Answering System gains its knowledge can be manually overridden in a small GUI.

3.7. Decision GUI

With this graphical user interface, the journalists decide on the front page. It allows editing the automatically downloaded news text, pre-viewing the resulting front page and also initiates the final publishing stage (cf. Figure 2).

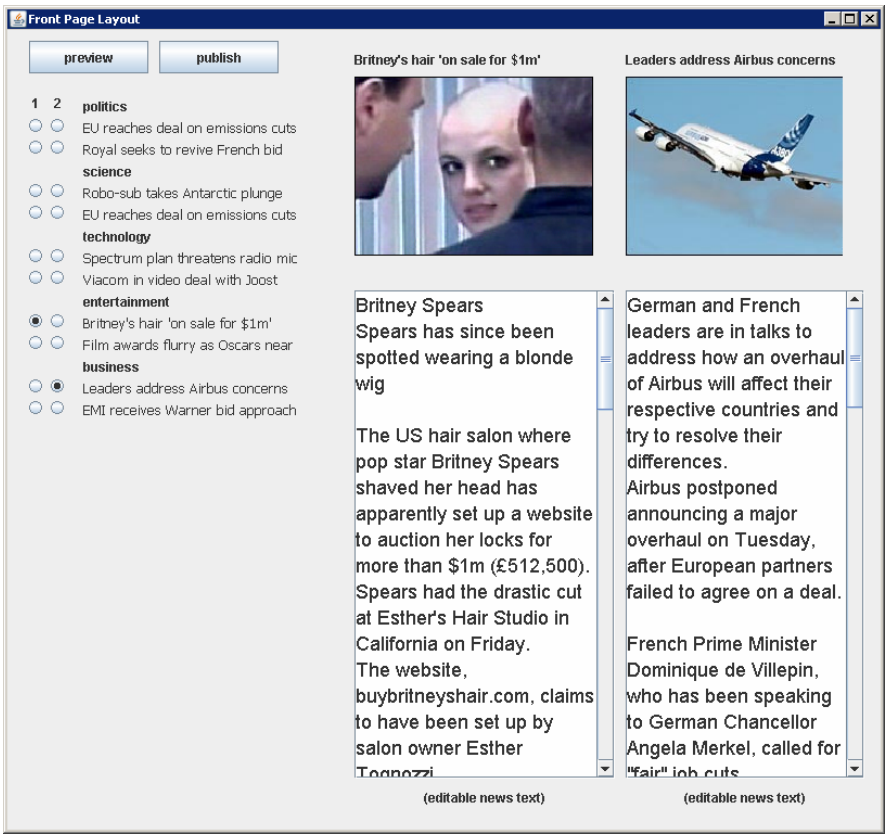


Figure 7. Screenshot of the Decision graphical user interface: this GUI is the main tool of the field journalist. If shows and allows editing of two selected top news. Clicking on the related image directs the journalists to the internet page with the source of the information. The GUI is also projected on one of the walls of the smart room to allow the journalists to move freely in the room while discussing the news.

3.8. RSS feed of BBC news

A Bash script is executed every five minutes on one of the Linux servers to download the latest RSS feeds[3] from the BBC world news. The downloaded web pages are processed with the text-based Lynx internet browser (lynx.browser.org). The text of the news and a corresponding image are extracted with awk and made available via Samba to the computers that display the graphical user interfaces for the journalists.

4. Software Architecture

Most of the technologies described in the previous section require a high-bandwidth data stream of several hundreds of Megabytes per second and the results of some

[3] For example: newsrss.bbc.co.uk/rss/newsonline_world_edition/europe/rss.xml

dozens of analysis algorithms need to be collected in a thread-safe manner be a single application which we call the Central Logic. The software architecture chosen in the UPC smart room is based on NIST smartflow system [16] and KSC Socket messaging system. This is illustrated in the following Figure 8.

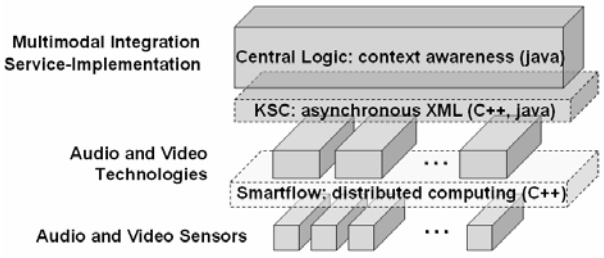


Figure 8. Software architecture in the UPC smart room.

The lower level of the software architecture consists of the video and audio sensors. The signal capture software is implemented as smartflow clients in the computers with the corresponding acquisition hardware. The resulting data streams are transferred as smartflow flows into other computers that can either pre-process the data streams (as in the case of the foreground segmentation which is part of the multi-camera Person Tracker) or directly analyze the raw data streams (as in the case of the Speech Activity Detection audio technology). [17]

Smartflow also provides a mechanism to dynamically decide on which computer in the local area network a specific technology should run – while assuring the correct handling of the data streams between the involved technologies.

The KSC message server and the KSC client library allow sending results of data analysis asynchronously to other technologies or to the Central Logic in a convenient fashion. In order to gain context awareness in the Central Logic framework, some hundreds of KSC messages have to be processed asynchronously in the multi-treaded Central Logic. A multitude of simple voting processes and if-then rules determine shifts of the underlying stage model that comprises the states “people enter”, “instruction”, “meeting”, “dialogue system”, “decision” and “end”. Depending on the present system stage, actions of the Memory Jog Service are triggered by the incoming events: the corresponding visualisations are updated, subtitles are added to the annotated videoconferencing system and the talking head informs the journalist about an event or turns its head towards the location of interest.

5. Conclusion

Our experience with the integration of multiple technologies into two multi-modal perceptual components and an integrated Memory Jog Service that interacts with humans in our smart room was very positive. The technology developers were challenged by the computational demands of a real-time implementation of their technology, signals from unobtrusive sensors and “noise” from a real-world scenario. These constrains tested the technologies at their limits of performance, processing delay and robustness. Still, the system has been successfully put to test with real-users

visiting the Smart Room at UPC. A demonstration of the Memory Jog Service is available at the UPC smartroom.

Additionally, the implementation of the multi-modal perceptual components and the integrated service sparked numerous ideas of how to combine technologies in a new way to introduce a higher level of robustness in real-world applications. Subjects that have experienced the Memory Jog Service agree that the service was helpful and some added ideas to our wish list of future services.

References

[1] R. Sharma, V.I. Pavlovic, T.S. Huang, Toward multimodal human-computer interface, *Proceedings of the IEEE* **66** (1998), 853-869

[2] K. Vredenburg, S. Isensee, C. Righi, *User-centered design: An integrated approach*, Prentice Hall, Englewood Cliffs, NJ, 2001

[3] CHIL Project website, http://chil.server.de

[4] R. Stiefelhagen, H. Steusloff, A. Waibel, CHIL - Computers in the Human Interaction Loop, *Proceedings of 5th International Workshop on Image Analysis for Multimedia Interactive Services*, Lisbon, Portugal, April 2004

[5] *Embodied Conversational Agents*. Justine Cassell, Joseph Sullivan, Scott Prevost, and Elizabeth Churchill Editors, MIT Press, Cambridge/London, 2000

[6] J.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

[7] J. L. Landabaso, M. Pardas, Foreground Regions Extraction and Characterization Towards Real-Time Object Tracking, *Machine Learning for Multimodal Interaction*, Springer LNCS **3869** (2006), 241-249

[8] C. Canton-Ferrer, J.R. Casas, M. Pardàs, M. Human Model and Motion Based 3D Action Recognition in Multiple View Scenarios, European Signal Processing Conference (EUSIPCO), Florence, Italy, September 2006

[9] A. Abad et al, UPC Audio, Video and Multimodal Person Tracking Systems in the CLEAR Evaluation Campaign, *Multimodal Technologies for Perception of Humans* (CLEAR), Springer LNCS **4122** (2007), 93-104

[10] A. Temko et al, Evaluation of Acoustic Event Detection and Classification Systems, *Multimodal Technologies for Perception of Humans* (CLEAR), Springer LNCS **4122** (2007), 311-322

[11] V. Vilaplana, C. Martinez, J. Cruz, F. Marques, Face Recognition Using Groups of Images in Smart Room Scenarios, IEEE International Conference on Image Processing (ICIP), Atlanta (GA), USA, October 2006

[12] J. Luque,et al, Audio, video and multimodal person identification in a smart room, *Multimodal Technologies for Perception of Humans* (CLEAR), Springer LNCS **4122** (2007), 258-269

[13] Haptek's PeoplePutty website, http://www.haptek.com

[14] S.J. Young, The HTK Hidden Markov Model Toolkit: Design and Philosophy, TR 152, Cambridge University Engineering Dept, Speech Group, 1993

[15] M. Surdeanu, D. Dominguez-Sal, P. Comas, Performance Analysis of a Factoid Question Answering System for Spontaneous Speech Transcriptions, Proceedings of the 10th International Conference on Speech Communication and Technology (Interspeech), Pittsburgh (PA), USA September 2006

Formatted: English U.K.

Formatted: English U.S.

[16] NIST Smart Flow System: http://www.nist.gov/smartspace/nsfs.html
[17] J. Neumann, Multimodal Integration of Sensor Network, 3rd IFIP Conference on Artificial Intelligence Applications & Innovations (AIAI), Athens, Greece, June 2006