Computers in the Human Interaction Loop

A. Waibel, R. Stiefelhagen, R. Carlson, J. Casas, J. Kleindienst, L. Lamel, O. Lanz, D. Mostefa, M. Omologo, F. Pianesi, L. Polymenakos, G. Potamianos, J. Soldatos, G. Sutschet, J. Terken

1 Introduction

It is a common experience in our modern world, for us humans to be overwhelmed by the complexities of technological artifacts around us, and by the attention they demand. While technology provides wonderful support and helpful assistance, it also

A. Waibel, R. Stiefelhagen
Universität Karlsruhe (TH), Interactive Systems Labs, Karlsruhe, Germany, e-mail: ahw@cs.cmu.edu, stiefel@ira.uka.de

R. Carlson
Kungl Tekniska Högskolan, Centre for Speech Technology, Stockholm, Sweden

J. Casas
Universitat Politecnica de Catalunya, Barcelona, Spain

J. Kleindienst
IBM Research, Prague, Czech Republic

L. Lamel
LIMSI-CNRS, France

O. Lanz, M. Omologo, F. Pianesi
Foundation Bruno Kessler, irst, Trento, Italy

D. Mostefa
ELDA, Paris, France

L. Polymenakos, J. Soldatos
Athens Information Technology, Athens, Greece

G. Potamianos
Institute of Computer Science, FORTH, Crete, Greece (work performed while with the IBM T.J. Watson Research Center, Yorktown, Heights, NY, USA)

G. Sutschet
Fraunhofer Institute IITB, Karlsruhe, Germany

J. Terken
Technische Universiteit Eindhoven, Netherlands

H. Nakashima et al. (eds.), *Handbook of Ambient Intelligence and Smart Environments*,
DOI 10.1007/978-0-387-93808-0_40, © Springer Science+Business Media, LLC 2010

causes an increased preoccupation with technology itself and a related fragmentation of attention. But as humans, we would rather attend to a meaningful dialog and interaction with other humans, than to control the operations of machines that serve us. The cause for such complexity and distraction, however, is a natural consequence of the flexibility and choice of functions and features that technology has to offer. Thus flexibility of choice and the availability of desirable functions are in conflict with ease of use and our very ability to enjoy their benefits. The artifact cannot yet perform autonomously and therefore requires precise specification of every aspect of the machine's behavior under a variety of different user interface conventions. Standardization and better graphical user interfaces, multimodal human-machine dialog systems including speech, pointing, mousing have all contributed to improve this situation. Yet, they have only improved the input mode, but not provided any proactive, autonomous capabilities of the artifact, and still force the user in a human-machine interaction loop at the exclusion of other human-human interaction.

To change the limitations of present day technology, machines must be engaged implicitly and indirectly in a world of humans, or: We must put Computers in the Human Interaction Loop (CHIL), rather than the other way around. Computers should proactively and implicitly attempt to take care of human needs without the necessity of explicit human specification. If technology could be CHIL-enabled, a host of useful services could be possible. Could two people get in touch with each other at the best moment over the most convenient and best media, without phone tag, embarrassing ring tones and interruptions? Could an attendee in a meeting be reminded of participants' names and affiliations at the right moment without messing with a contact directory? Could computers offer simultaneous translation to attendees who speak different languages and do so unobtrusively just as needed? Can meetings be supported, moderated and coached without technology getting in the way? Human secretaries often work out such logistical support, reminders, and helpful assistance, and they do so tactfully, sensitively and sometimes diplomatically. Why not machines? Clearly, it is a lack of recognition and understanding of human activities, needs and desires and an absence of learning, proactive, context-aware computing services that limit machines' ability to specify their own behavior and serve people implicitly.

To change the status quo, a consortium of 15 laboratories in 9 countries has teamed up to explore what is needed to build usable CHIL computing services. Supported under the 6th Framework Program of the European Commission, the CHIL consortium has been one of the largest consortia working on this problem. It has developed a research and development infrastructure by which different CHIL services can be quickly proposed, assembled and evaluated, exploiting a User Centered Design approach to secure that the developed services answer real users' needs and demands. Examples of prototype services are 1.) the Connector (a proactive phone/communication device), 2.) the Memory Jog (for supportive information and reminders in meetings), 3.) collaborative supportive workspaces and meeting monitoring. Under the CHIL paradigm the consortium also experimented with new ideas for CHIL services based on the background of the partners (e.g. recently, a simultaneous speech translator for the lecture domain).

To realize such CHIL computing, the work of the consortium concentrated on four key areas:

- Technologies: Proactive, implicit services require a good description of human interaction and activities. This implies a robust description of the perceptual context as it applies to human interaction: Who is doing What, to Whom, Where, How, and Why. Unfortunately, such technologies – in all of their required robustness – do not yet exist, and the consortium identified a core set of needed technologies and set out to build and improve them for use in CHIL services.
- Data Collection and Evaluation: Due to the inherent challenges (open environments, free movement of people, open distant sensors, noise, ...), technological advances are made under an aggressive R&D regiment rounded off by world-wide evaluation campaigns. In support of these campaigns, data from meetings, lectures, seminars held in offices and lecture rooms has been collected at 5 different sites, and metrics for technology evaluation defined.
- Software Infrastructure: A common and defined software architecture serves to improve inter-operability among partners and offer a market driven exchange of modules for faster integration.
- Services: Based on the emerging technologies developed at different labs, using a common architecture, and within a User Centered Design framework, CHIL services are assembled and evaluated. In this fashion, first prototypes are continuously being (re-) configured and the results of user studies effectively exploited.

In the following, we review some of the highlights of each of these critical components of CHIL service design.

2 Audio-Visual Perceptual Technologies

The user-centered approach of CHIL-enabled environments requires a wide range of audio-visual (AV) perceptual technologies in order to place the Computer into the Human Interaction Loop. Such technologies are necessary for the computers to exhibit context- and content-awareness, and can be further facilitated by cognition tools, for example situation modeling, strategy and planning to decide on the most adequate interaction. The latter will be discussed in the software infrastructure section of this chapter (Sect. 4).

Multimodal interface technologies "observe" humans and their environments by recruiting signals from the multiple AV sensors of the CHIL smart rooms to detect, track, and recognize human activity. The analysis of all AV signals in the environment (speech, signs, faces, bodies, gestures, attitudes, objects, events, and situations) provides the proper answers to the basic questions of "who", "what", "where", and "when", that can drive higher-level cognition concerning the "how" and "why", thus allowing computers to engage and interact with humans in a human-like manner using the appropriate communication medium. Research work by the CHIL consortium on a number of such technologies is described next. All these technologies are rigorously evaluated on a regular basis within international evaluations, such as the Rich Transcription Spring 2006 (RT06s) and 2007 (RT07) evaluation campaigns [39, 75] and the CLEAR – Classification of Events, Activities and Relationships – evaluation [21, 85, 84] (see also Sect. 3).

2.1 Speech Recognition

Speech is the most critical human communication modality in the CHIL seminar and meeting scenarios of interest. Its automatic transcription is of paramount importance to real-time support and off-line indexing of the interaction, for example providing input to the higher-level technologies of summarization and question answering. In the spirit of ubiquitous computing, the goal of *automatic speech recognition* (ASR) in CHIL is to achieve high performance using *far-field sensors*. Therefore, despite the maturity of ASR technology, the CHIL scenarios present significant challenges for state-of-the-art systems. Such are, for example, the presence of speech from multiple speakers with varying accents and frequent periods of overlapping speech, a high level of spontaneity with many hesitations and disfluencies, and a variety of interfering acoustic events, such as knocks, door slams, steps, cough, laughter, and others.

The CHIL data are comprised of interactive seminars, held inside smart rooms equipped with numerous acoustic and visual sensors. The acoustic sensors include a number of microphone arrays located on the walls (typically, at least three T-shaped four-microphone arrays and at least one linear 64-channel array), as well as a number of table-top microphones (at least three). The latter provide the basis of the *multiple distant microphone* (MDM) primary condition of interest in the development of far-field ASR technology. In addition to these sensors, most meeting participants wear headset microphones to capture individual speech in the close-talking condition.

Three CHIL partner sites, IBM, LIMSI, and UKA-ISL, have been actively involved in the effort to develop robust ASR technology on such CHIL data. Although the sites primarily worked on their ASR systems independently of each other, all systems contain a number of standard important components. At the lower level, these include feature extraction and enhancement. ASR systems typically utilize Mel frequency cepstral coefficients (MFCCs) [24], or perceptual linear prediction (PLP) [46] features. Additional techniques include linear discriminant analysis (LDA) [44], a maximum likelihood linear transform (MLLT) [43], feature normalization steps, such as variance normalization and vocal tract length normalization (VTLN) [3], as well as a novel feature extraction technique developed by UKA-ISL, which is particularly robust to the changes in fundamental frequency [97].

At a higher level, for acoustic modeling, all systems made use of *hidden Markov models* (HMMs) estimated with the expectation maximization (EM) algorithm (maximum likelihood training) [26], followed by discriminative model training using maximum mutual information (MMI) [71], or the minimum phone error (MPE) approach [72]. A number of adaptation techniques were also employed, ranging from maximum a-posteriori estimation (MAP) [41] to maximum likelihood linear regression (MLLR) [62], feature-space MLLR (fMLLR), or speaker adaptive training (SAT) [2]. Most ASR systems used a multi-pass decoding strategy, where a word hypothesis was employed for unsupervised model adaptation prior to the next decoding pass. For language modeling, different n-gram language models (LMs), with $n = 3$ or 4, have been employed by the CHIL sites. These LMs were typically developed separately for various data sources, and were subsequently linearly interpolated to give rise to a single model. Most often, CHIL and other meeting corpora were employed for this task. An important part relevant to the language modeling work, is determining the recognition vocabulary and the pronunciations, particularly to model foreign accented speech. The use of a connectionist LM [76], shown to be performant when LM training data is limited, was also explored at LIMSI. Finally, an important aspect of all systems was the combination in the recognition process of the available information coming from the multiple available acoustic sensors. Both decision and feature fusion approaches have been employed for this purpose, for example beamforming [98] and ROVER [38].

The CHIL partner sites involved in ASR work have made steady progress in the CHIL domain, as benchmarked by yearly technology evaluations. During the first two years of CHIL, the consortium evaluated ASR internally with a first dry-run held in June 2004, followed by an "official" evaluation in January 2005. Subsequently, the three ASR CHIL sites participated in the Rich Transcription evaluations of 2006 and 2007 – RT06s [39] and RT07 [75]. These international evaluations were sponsored by NIST and attracted a number of additional external participants. The evaluations have provided a quantitative measure of progress over the years. For the far-field ASR task, for example, the best system performance improved from a word error rate of over 68% in the 2004 dry run to approximately 52% in the CHIL 2005 internal evaluation, and from 51% in RT06s down to 44% in RT07. These improvements were achieved in spite of the fact that the recognition task became increasingly more challenging: Indeed, from 2005 to 2006, multiple recording sites involving speakers with a wider range of non-native accents have been introduced into the test set. Furthermore, both in 2006 and 2007, the degree of interactivity in the data has significantly increased.

In addition to ASR, two other relevant technologies were investigated, namely *speech activity detection* (SAD) and *speaker diarization* (SPKR) – also known as the "who spoke when" problem. Both SAD and SPKR constitute crucial preprocessing components of state-of-the-art ASR systems, and are important for fusion with other CHIL perceptual technologies, for example, acoustic based speaker localization and identification. In particular, they locate the speech segments that are processed by the ASR systems, and attempt to cluster homogeneous sections, which is crucial for efficient signal normalization and speaker adaptation techniques.

SAD technology has been investigated by many CHIL partners (AIT, FBK-irst, IBM, INRIA, LIMSI, UKA-ISL, and UPC) that explored various approaches differing in a number of factors: For example, in feature selection (MFCCs, energy-based, combined acoustic and energy based features, etc.), the type of classifier used (Gaussian mixture model classifiers (GMMs), support vector machines, linear discriminants, decision trees), the classes of interest and channel combination techniques. Similarly to ASR, these components have been evaluated in separate CHIL-internal and NIST-run evaluation campaigns. In the RT06s evaluation, the best systems achieved very encouraging error rates of 4% and 8% for the conference and lecture subtasks, respectively, when calculated by the NIST diarization metric. A number of CHIL partner sites have also developed SPKR systems (AIT [73], IBM [48, 47], LIMSI [101, 102], and UPC [63]), exploring a variety of research directions, including exploiting multiple distant microphone channels, speech parametrization, as well as segmentation and clustering [63, 4, 102].

2.2 Person Tracking

The localization and tracking of multiple persons behaving without constraints, unaware of audio/video sensors, in natural, evolving and unconstrained scenarios, still poses significant challenges to the current state-of-the-art in tracking technologies.

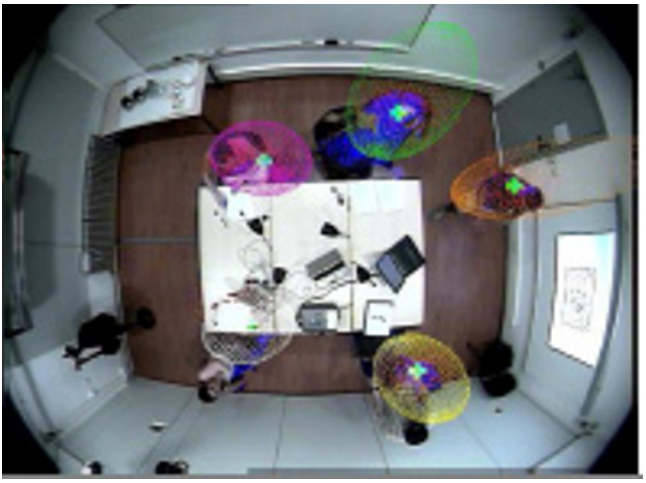


Fig. 1 Example screenshot of a person tracking system running on CHIL data (image taken from [17]).

Video-based tracking techniques have to deal with varying illumination, shadows, occlusions, changing target appearances, the lack of color constancy, and so forth. While these difficulties are inherent to some degree in all kinds of visual tracking tasks, the problem posed by occlusions becomes particularly severe here, when observing multiple persons in a small crowded space. Close proximity to the capturing cameras may cause persons to occlude others; partial occlusions by furniture, such as tables and chairs, are very common. The choice of reliable features for tracking is also not straightforward. Changes in the foreground support may be caused by person parts of varying sizes, by other objects, such as moved chairs, by projection surfaces, etc. The use of color or edge features on the other hand presumes that person-specific appearance models are available. These can, however, be difficult to initialize or maintain for every new acquired target.

The primary constraint in audio-based localization and tracking, using distantly placed microphone arrays, is that the target person needs to be an active speaker. In most practical scenarios, this limits applications to the tracking of one person at a time, with further difficulties caused by the variety of acoustic conditions (e.g., room acoustics and reverberation time) and, in particular, the undefined number of simultaneous active noise sources and competing speakers.

In addition to single modality tracking, audio-visual tracking approaches were also researched, to measure the advantage of modality fusion. The challenges here lie in the balancing of the modalities and in the fusion algorithms themselves, as measurements from different types of sensors come with varying levels of confidence, at different frequencies, and may not always be available at all times (e.g. for visually occluded persons or for silent audience members).

One of the key problems that needed to be solved, for visual, acoustic and multimodal tracking alike, was the definition of performance evaluation procedures for multiple object trackers. The defined metrics, usable for single or multiple object tracking alike, judge a tracker's performance based on its ability to precisely estimate an object's position (Multiple Object Tracking Precision, $MOTP$) and its ability to correctly estimate the number of objects and keep correct trajectories in time (Multiple Object Tracking Accuracy, $MOTA$) (see [8] for details). They were used in the CLEAR [88, 87] evaluations, for a broad range of CHIL or VACE [93] related tasks including acoustic, visual and multimodal 3D person tracking, 2D visual person tracking, face detection and tracking, and vehicle tracking.

To face the numerous further challenges mentioned above, the CHIL consortium applied several strategies for what concerns the room design and selected sensors, as well as the algorithms employed. A distributed camera and microphone network provided a good "coverage" of each area of the room and the fusion of sensor data helped overcome the problems of occlusions or noisy observations and increase localization accuracies.

On the visual side, from the point of view of algorithm design, two distinct approaches have been followed by the various 3D tracking systems built within the consortium: First, a model-based approach: A 3D model of the tracked object is maintained by rendering it onto the camera views, searching for supporting evidence in each view, and based on that, updating its parameters [67, 59, 1, 7, 61, 17]. Second, a data-driven approach: 2D trackers operate independently on the separate camera views; then the 2D tracks belonging to a same target are collected into a 3D one [100, 92, 52].

An important advantage of the model-based approach is that model rendering can be implemented in a way that mimics the real image formation process, including effects like perspective distortion and scaling, lens distortion, etc. In the context of multi-body tracking this is particularly advantageous, since occlusions can be handled in a systematic manner. The model-based approach also makes it easier to incorporate many different types of features, such as foreground segments, color histograms, etc., increasing tracking robustness.

On the acoustic side, tracking approaches can be roughly categorized as follows: Approaches which rely on the computation of a more or less coarse global coherence field (GCF, or SRP-PHAT), on which the tracking of correlation peaks is performed [1, 14], particle filter approaches, which stipulate a number of candidate person positions at every point in time and measure the agreement of the observed acoustic signals (their correlation value) to the position hypothesis [66], and approaches that feed computed time delays of arrival (TDOAs) between microphone pairs directly as observations to a Kalman or other probabilistic tracker [54, 42]. Whereas in earlier CHIL systems, speech activity detection (SAD), necessary for correct localization, was often performed and evaluated separately, toward the end of the project, more and more approaches featured built-in speech detection techniques [42, 77].

In the field of multimodal tracking, while most initial systems performed audio and video tracking separately and combined tracker outputs in a post-processing step, a few select approaches incorporated the multimodal streams at the feature level. These were notably particle filter based trackers [66, 7, 13], as these allow for the probabilistic, flexible and easy integration of a multitude of features across sensors and modalities.

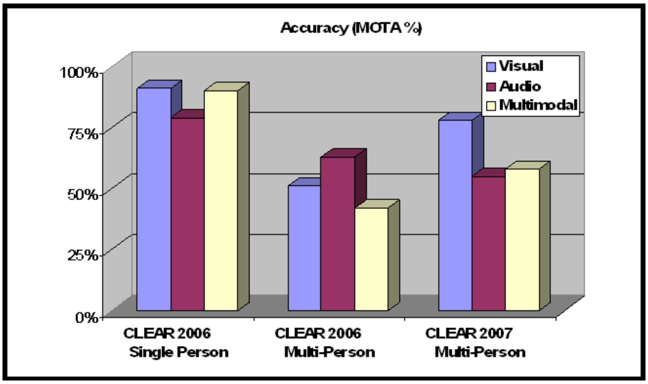


Fig. 2 Best system performances for the CLEAR 2006 and 2007 3D person tracking evaluations. The MOTA score measures a tracker's ability to correctly estimate the number of objects and their rough trajectories.

Throughout the duration of the project steady progress was made, going from single modality systems with manual or implicit initialization, using simple features, sometimes implying several manually concatenated offline processing steps and tracking at most one person, to fully automatic, self-initializing, real-time capable systems, using a combination of features, fusing several audio and visual sensor streams and capable of tracking multiple targets. Aside from the tracking tasks, which grew more and more complex, the evaluation data, which was initially recorded only at the UKA-ISL site, also became increasingly difficult and varied. This was due to the completion of four more recording smart rooms, the inclusion of more challenging interaction scenarios, the elimination of simplifying assumptions such as main speakers, areas of interest in the room, or manually segmented audio data that excludes silence, noise or crosstalk. Nevertheless, the performance of systems steadily increased over the years. Figure 2 shows the progress made in audio, video and multimodal tracking for the last official CLEAR evaluations [88, 87].

2.3 Person Identification

The challenges for AV person identification (ID) in CHIL scenarios are due to far-field, wide-angle, low-resolution sensors, acoustic noise, speech overlap and visual occlusion, unconstrained subject motion, and no position/orientation assumptions to facilitate well-posed signals (one cannot assume frontal faces, or speakers facing/talking to sensors). A sample face sequence from a smart room can be seen in Fig. 3. Clearly, employing tracking technologies (detection followed by tracking) and fusion techniques, either multi-view or multimodal (speaker ID combined with face ID for example) is a viable approach in order to improve robustness.

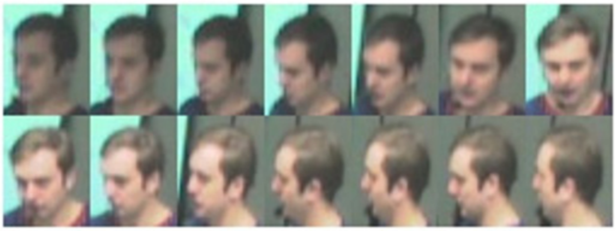


Fig. 3 A sample face sequence from a smart room.

Identification performance depends on the enabling technologies used for audio, video and their fusion, but also on the accuracy of the extraction of the useful portions from the audio and video streams. The detection process for audio involves finding the speech segments in the audio stream by employing speech activity detection (see Sect. 2.1). The corresponding process for video involves face detection, which might be a composite task, comprising body tracking, head tracking, and face detection, each subtask narrowing the search space for the face.

Developed mono- and multimodal ID systems within CHIL have been successfully evaluated in the CLEAR 2006 and 2007 campaigns [84, 85]. In the evaluations, two different training durations, 15 and 30 seconds, and four different testing durations, 1, 5, 10 and 20 seconds, were considered. The number of subjects in the

CLEAR 2006 person identification evaluation corpus was 26 and it became 28 in CLEAR 2007. A significant performance increase was observed in CLEAR 2007, especially in face recognition. It was shown that longer training and testing durations resulted in very high recognition results and that multimodality improves the robustness of person identification. This indicates that person ID in open environments is quite plausible. The correct recognition rates of the best performing systems in the CLEAR 2006 and 2007 evaluations are shown in Fig. 4.

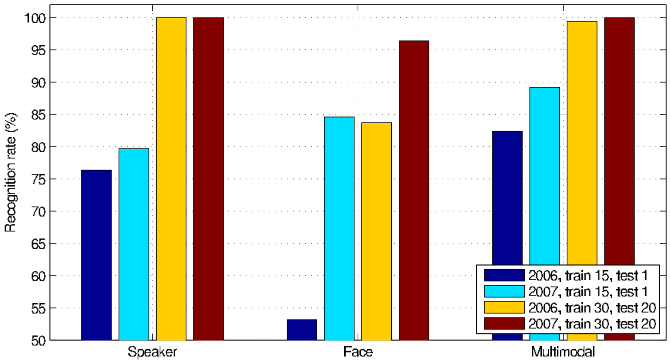


Fig. 4 Performance comparison of the person identification systems in the CLEAR 2006 and 2007 evaluations. Two out of the eight conditions are shown per evaluation; the shortest training and testing (15 and 1 seconds respectively) and the longest training and testing (30 and 20 seconds respectively) (from [94]).

2.4 Interaction Cues: Gestures, Body Pose, Head Pose, Attention

Studies in social psychology have experimentally validated the common feeling that non-verbal behavior, including but not limited to gaze and facial expression, is extremely significant in human interaction. In the analysis of group behavior a key variable is the *focus of attention* which indicates the object or person one is attending to. Another form of non-verbal communication develops along body language, including explicit gestures, such as raising a hand to catch attention, or nodding the head to manifest agreement, and body postures. In this chapter, approaches for the automated analysis of non-verbal communication in a meeting scenario will be discussed.

Although visual attention does not necessarily coincide with psychological attention, it is highly correlated with it, and it is physically observable. While the image resolution required to estimate eye gaze may not always be available in a meeting recording, a reliable indicator can still be derived from estimated head pose. CHIL

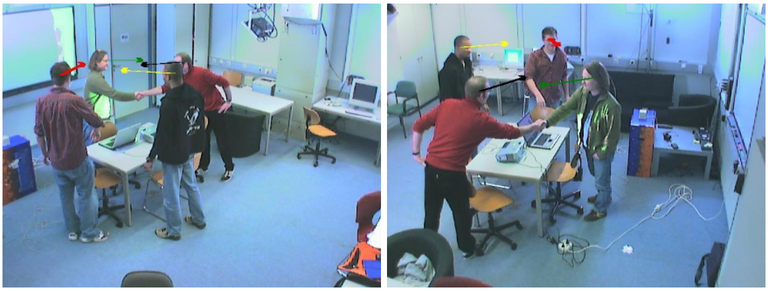


Fig. 5 Estimating head pose and focus of attention [86]. Head orientations are estimated from four camera views. These are then mapped to likely targets of attention, such as other people.

partners have investigated approaches to estimate the head orientation of people in a smart room using multiple fixed cameras (see Fig. 5). With a neural network based approach [65], it is possible to estimate relative head orientation from a single view with mean errors as low as twelve degrees. When applied to each single camera and combined by a Bayes filter to obtain an absolute value for head orientation, such an approach allowed to correctly determine who is looking at whom in a CHIL meeting recording around 70% of the time. Another option is to perform signal level based fusion using spatial and color analysis [18]. Such an approach exploits the redundancy available in multiple images, and can be extended to integrate also acoustic input.

A detailed description of a person's pose can be extracted from multiple views, e.g. using a particle filter based approach in combination with a 3D shape and appearance model for each target [60] (see Fig. 6). The state of a person includes 2D body position, horizontal and vertical head and torso orientation, and a binary value for sitting or standing pose. A coarse 3D shape model maps such a state into a set of image patches where the different body parts (head, torso, legs, arms) are expected to be visible. A likelihood score is then computed by matching those patches with a previously acquired color model of the target. This approach has the advantage of estimating a number of suitable perceptual parameters jointly (e.g. position and head orientation to determine visual attention in dynamic environments), thus reducing the impact of error propagation which can lead to failures in cascaded solutions (e.g. head alignment error impacts the accuracy of head orientation estimation).

Other visual cues that are relevant to human behavior analysis can be detected at a finer spatio-temporal scale. To pinpoint fidgeting, temporally unstable skin pixels are recorded to a Motion History Image (MHI) representing a temporal record of repetitive skin motion. Patterns of interest, such as nodding, shaking and typing, are characterized in the MHI and then used for their detection [20].

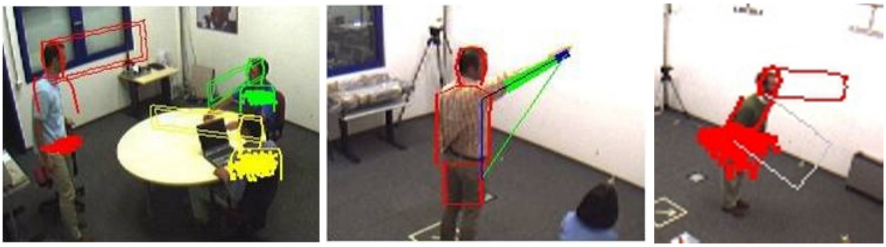


Fig. 6 The *SmarTrack* system [82] simultaneously estimates spatial location, head orientation and body pose of up to 5 persons in real time, allowing for online detection of focus of attention and macro-scale gestures such as pointing.

2.5 Activity Analysis, Situation Modeling

The recognition of activities depends on many factors such as the location and number of people, speech activity, and the location and state of certain objects. Perceptual technologies like person tracking or acoustic event detection provide important information on which the higher-level analysis of the activity can be based on. Due to the complexity of the scene there are, however, potentially relevant phenomena like for example a door being half opened, which – due to their high number and variability – cannot be addressed by manually designed detectors at large. Therefore, activity recognition may need to directly analyze the observation in order to find out what is relevant and what is not to detect a certain activity.

Using only one camera and one microphone per room, we can distinguish, for example, between typical office activities such as “paperwork”, “meeting” or “phone call”. The approach employs a multi-level HMM framework to characterize these kinds of events and situations. It is based on a low-level local feature model that integrates adaptive background modeling, optical flow and speech activity detection. The classes are trained from a relatively small number of samples. Figure 7 depicts an example of data-driven clustering of activity regions within an office. The regions labeled represent the learned areas of activity of office workers and their visitors. The evaluation of an unconstrained one week recording session showed acceptable recognition accuracy despite strongly varying illumination conditions, as shown in [96].

In a smart room it may also be necessary to determine the type of interactions among the people that are present in the room in order to perform context-dependent actions. To this aim, a system based on [50], which analyzes activities using stochastic parsing, has been developed by adapting it to the specific requirements of smart room environments. The system can be divided into three modules: the tracking system, the event generator and the parser. The tracking system [58] takes as input the multi-camera video sequence, reconstructs the 3D objects in the room using a foreground detection for each camera, and performs the tracking of the various objects. This tracker output is used, together with some configuration information of

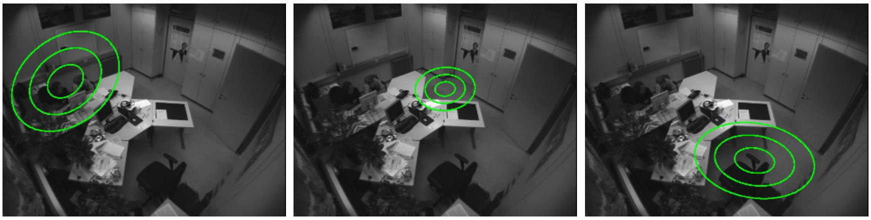


Fig. 7 The first three Gaussian mixture components obtained from data driven clustering represent typical activity regions in the office room.

the room, to detect events using simple grammars. The actual activity recognition is performed by the parser: given a chain of events and a stochastic context-free grammar, it finds the chain derivation with the maximum probability. The activities recognized by the system are: “meeting”, “presentation”, “conversation between two people”, “leave an object” and “take an object” (Fig. 8).

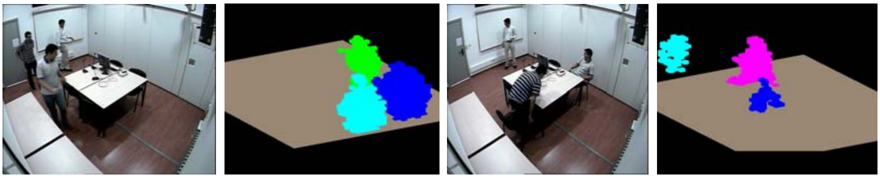


Fig. 8 Two snapshots from a presentation, shown as camera image and 3D blob representation. The left shows an example of the detected event “person enters”, the right shows an example for the events “person at whiteboard” and “person sits”.

Activities that involve the motion of body limbs cannot be recognized using only person/object tracking. Therefore, a view-independent approach has been developed to recognize human gestures from multiple cameras [16]. In contrast to other systems that perform a classification based on the fusion of features coming from different views, our system performs data fusion to obtain a 3D representation of the scene, and then feature extraction and classification. Motion descriptors introduced by [11] are extended to 3D and a set of features based on 3D invariant statistical moments are computed. A simple body model is fit to the 3D data to capture in which body part the gesture occurs. Classification is thus performed by jointly analyzing the motion features and the pose information from the body model. Finally, a Bayesian classifier is employed to perform recognition over a small set of actions. The actions relevant to the smart room scenario are “raising hand”, “sitting down” and “standing up”. However, we have tested the system including other actions such as “waving hands”, “crouching down”, “punching”, “kicking” and “jumping”.

Our experience from CHIL shows that it is hard to find a common methodology for activity recognition that fits all application domains. Our current systems are dedicated to certain domains like office situations and meetings. They define a small set of activities that are relevant within their domain, and they proved to be able to recognize these on in-domain test data. Future work could on one hand try to extend the application domains, and on the other hand it could aim for a more detailed analysis of the activities within one domain.

2.6 Audio-Visual Output Technologies

Conveying information from the computer to the user in a natural way is another key ability on the way to realizing unobtrusive CHIL services. Multimodal output in innovative combinations, such as far-field acoustic and visual modalities, makes it possible to convey messages without disturbing other meeting participants. Experiments in CHIL show that this can be achieved by employing highly directive audio beams produced by properly designed steerable arrays. When the signal is accompanied by the video of lip-synchronized talking heads [9, 10] to improve the naturalness of interaction, the combined modalities increase intelligibility in noisy environments, a frequent occurrence in CHIL scenarios. Similarly, adding a lip-synchronized talking head makes it possible to deliver voice messages more quietly without decrease in intelligibility from the perspective of the addressee of a message, resulting in less noise pollution for non-addressees [80, 90].

2.7 Interaction

Designing spoken dialog systems in coherence with a human metaphor, so that the computer is perceived to have human-like conversational abilities, allows us to profit from a number of characteristics of spoken interaction that are unexploited or underexploited in current systems. This, however, increases the demands on a number of conversational abilities [32].

One of the most obvious differences between human-human conversations and the vast majority of spoken human-machine interactions is the interaction flow. Spoken human-human interaction is characterized by a large number of non-lexical sounds (hums and grunts, filled pauses, false starts, breath, smacks, etc.). These are an essential part of human interaction, and are used for a great number of reasons, such as adding focus to part of an utterance or controlling who should speak next. When these are removed, as they habitually are on the computer side of human-computer communication, the human metaphor suffers and the interaction becomes more similar to GUI manipulation than to human speaking. As this is contrary to the spirit of CHIL, where humans should not adapt to computers but rather the other way around, efforts were undertaken to model more human behavior – both for production and for perception. Real-time prosodic analysis is used for improved interaction control and precise timing of feedback, which in turn allows the dialog system to barge in when it is appropriate, and users to do the same without causing the interaction to become strange and interrupted [45, 34, 33]. In CHIL dialogs, elliptical clarifications are utilized to focus on problematic fragments in order to make the dialog more natural and efficient, which makes prosody and context then more important [81, 95]. The human metaphor can be strengthened further by adding a visual embodiment to the system. The concept of embodied conversational agents (ECAs) – animated agents that are able to interact with a user in a natural way using speech, gesture and facial expression – holds the potential of a new level of naturalness in human-computer interaction, where the machine is able to convey and interpret verbal as well as non-verbal communicative acts, ultimately leading to more robust, efficient and intuitive interaction. A series of studies performed within CHIL were aimed at exploiting embodiment for interactional purposes [31, 49].

2.8 Natural Language

Intelligent solutions for human information needs in a smart room scenario also require natural language processing technologies such as summarization and question answering (QA). Due to the fact that speech is the most natural way of human communication, spontaneous speech single document summarization (SSSDS) can help in digesting large amounts of information produced by speakers. The major aim of speech summarization is to distill relevant information from speech data, which can be communicated efficiently to the user at a later time. Summarization can be considered as a variant of speech understanding. An important issue when summarizing the output of an automatic speech recognition system is that the document to be summarized has no punctuation or capitalization.

Searching textual content for short text snippets that are relevant to natural language questions is also a key objective in CHIL. QA can be regarded as a direct extension to search engines (e.g. Google) to handle cases when the type of the expected answer is known. For example, answers to questions of the form: “Who discovered/wrote/developed NAME?” are not expected to be a list of URLs but rather just the name of the author of the required work. QA can enable automated customer service, where many of the questions are of the form: “How do I perform ACTION?” Standard information retrieval techniques are not sufficient to answer such questions, because the answer is usually hidden deep inside technical manuals of hundreds of pages. Last but not least, in the context of the CHIL project, QA is one of the engines behind the Memory Jog service (see Sect. 5). Within this service, a QA system is employed to search both the transcriptions of a meeting and background knowledge offered by documents relevant to this meeting on the web.

3 Technology Evaluations and Data Collection

Systematic evaluation is essential to drive the rapid progress of a broad range of audio-visual perceptual technologies. Within the CHIL project, such evaluations were undertaken on an annual basis, so that improvements could be measured objectively and different approaches compared and assessed.

CHIL has organized a series of technology evaluations with subsequent evaluation workshops, during which the systems and obtained results were discussed in detail. A first project-internal technology evaluation was held in June 2004. Here baseline results of available technologies used in the real-life lecture scenarios CHIL was initially focusing on, were established. Already twelve evaluation tasks were conducted, including face and head tracking, 3D person tracking, face recognition, head pose estimation, hand tracking and pointing gesture recognition, speech recognition (close-talking and far-field), acoustic speaker tracking, speaker identification, acoustic scene analysis and acoustic event detection.

Then, in January 2005, a more formal evaluation was organized, which now also included multimodal evaluation tasks (multimodal tracking, multimodal identification). External research groups were also invited to participate. Also in 2005, CHIL research teams took part in NIST's (National Institute of Standards and Technology, USA) Rich Transcription (RT) Meeting Evaluation. This annual evaluation series focuses on the evaluation of content-related technologies such as speech and video recognition. CHIL contributed test and training data sets in the lecture room domain, and participated in the Speaker Location Detection task.

Many researchers, research labs and, in particular, a number of current major research projects worldwide are working on technologies for analyzing people's activities and interactions. However, common evaluation standards, common metrics and benchmarks for such technologies, as for example for person tracking, are missing. It is therefore hardly possible to compare the developed algorithms and systems. Hence most researchers rely on using their own data sets, annotations, task definitions and evaluation procedures. This again leads to a costly multiplication of data production and evaluation efforts for the research community as a whole. In order to overcome this situation, we decided in 2005 to completely open up the project's evaluations by creating an open international evaluation workshop called CLEAR, Classification of Events, Activities, and Relationships [88, 21], and transferred part of the CHIL technology evaluations to CLEAR. CLEAR's goal is to provide a common international evaluation platform and framework for a wide range of multimodal technologies, and to serve as a forum for the discussion and definition of related common benchmarks, including the definition of common metrics, tasks and evaluation procedures. The creation of CLEAR was possible by joining forces with NIST, who, amongst many other evaluations, also organizes the technology evaluation of the US Video Analysis Content Extraction (VACE) program [93]. As a result, CLEAR could be jointly supported by CHIL, NIST and the VACE program. The first CLEAR evaluation was conducted in spring 2006 and was concluded by a two day evaluation workshop in the UK in April 2006. CLEAR 2006 was organized in cooperation with RT06s [74]. This additionally allowed for sharing data between both evaluations by also harmonizing the 2006 CLEAR and RT evaluation deadlines. For example, the speaker-localization results generated for CLEAR were also used for the far-field ASR task in RT06s.

The CLEAR 2006 evaluation turned out to be a big success. Overall, around sixty people of 16 different institutions participated in the workshop, and nine major evaluation tasks, including more than 20 subtasks were evaluated [21]. Following the success of CLEAR 2006, CLEAR 2007 took place in May 2007, in Baltimore, USA, again organized in conjunction and collocated with the NIST RT 2007 evaluation.

An important aspect of the technology evaluations was using real-life data representing situations for which the applications and technologies in CHIL were developed, annotated with the necessary information for various modalities. Therefore, a number of data sets was created which enabled the development and evaluation of audio-visual perception technologies. The data was collected inside smart rooms at five different locations. All of these rooms were equipped with a broad range of cameras and microphones.

After an initial data set collected in 2004 for the first CHIL internal technology evaluation, new data sets in slightly modified scenarios were collected in 2005 and 2006. These sets constituted the main test and training data sets used in the CLEAR 2006 and 2007 evaluations. The audio modalities (close-talk and far-field) of the 2005, 2006 and 2007 data sets were also part of the RT05s, RT06s and RT07 evaluations. Additionally, the audio portion was employed in a pilot track in the Cross Language Evaluation Forum, CLEF 2007 [22].

The utilization of the CHIL data sets in such high-profile evaluation activities demonstrates the state-of-the-art nature of the corpus, and its contribution to the development of advanced perception technology. Moreover, the numerous scientific publications produced as results of the evaluations are further enhancing the importance of the corpus. The CHIL data sets are publicly available to the community through the language resources catalog [36] of the European Language Resources Association (ELRA).

3.1 Data Collection

The CHIL corpus consists of multi-sensory audio-visual recordings inside smart rooms. The corpus has been collected over the past few years in five different recording sites and it contains data of two types of human interaction scenarios: lectures and small meetings.

3.1.1 Sensor Setup

Five smart rooms have been set up as part of the CHIL project, and used in the data collection efforts. They are located in five countries: Germany (UKA-ISL), Greece (AIT), Italy (ITC), Spain (UPC), and USA (IBM). These five smart rooms are medium-size meeting or conference rooms with a number of audio and video sensors installed, and with supporting computing infrastructure. The multitude of recording sites provides desirable variability in the CHIL corpus, since the smart rooms obviously differ from each other in their size, layout, acoustic and visual environment (noise, lighting characteristics), as well as sensor properties (location, type). Nevertheless, it is crucial to produce a homogeneous database across sites to facilitate technology development and evaluations. Therefore, a minimum common hardware and software setup has been specified concerning the recording sensors and resulting data formats. All five sites comply with these minimum requirements, but often contain additional sensors. Such a setup consists of:

- A set of common audio sensors, namely:
  - A 64-channel linear microphone array;
  - Three 4-channel T-shaped microphone clusters;
  - Three table-top microphones;
  - Close-talking microphones worn by the lecturer and the meeting participants.
- A set of common video sensors, that include:
  - Four fixed cameras located at the room corners;
  - One fixed, wide-angle panoramic camera;

This set is accompanied by a network of computers to capture the sensory data, mostly through dedicated data links, with data synchronization information provided by the Network Time Protocol (NTP). A schematic diagram of such a room including its sensors is depicted in Fig. 9.

3.1.2 Scenarios

Two types of interaction scenarios constitute the focus of the CHIL corpus: lectures and meetings. In both cases, a presenter gives a seminar in front of an audience, but the two scenarios differ significantly in the degree of interactivity between the audience and the presenter, as well as the number of participants. The seminar topics are quite technical in nature, spanning the areas of audio and visual perception technologies, but also biology, finance, etc. The language used is English; however most subjects exhibit strong non-native accents, such as Italian, German, Greek, Spanish, Indian, Chinese, etc.

In the *Lecture* scenario, the presenter talks in front of an audience of typically ten to twenty people having little interaction in the form of a few question-answering turns, mostly toward the end of the presentation. As a result, the audience region is quite cluttered and of little activity and interest. The focus in lecture analysis lies therefore on the presenter. As a consequence, only the presenter has been annotated in the lecture part of the CHIL corpus and is the subject of interest in the associated

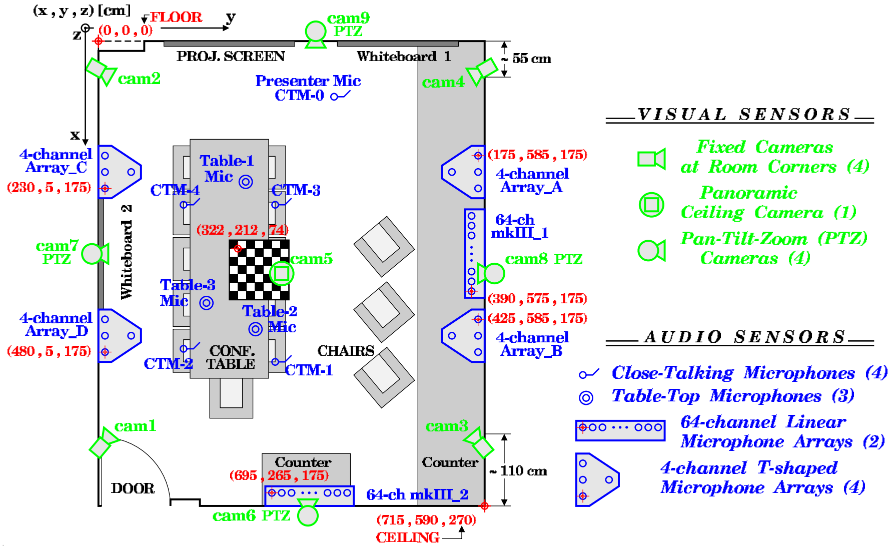


Fig. 9 Schematic diagram of the IBM smart room, one of the five installations used for recording the CHIL corpus. The room is approximately $7 \times 6 \times 3m^3$ in size and contains 9 cameras and 152 microphones for data collection.

evaluation tasks. A total of 46 lectures are part of the CHIL corpus (see also Table 1), most of which are between 40 and 60 minutes long.

Site	# Seminars	Epoch	Type	Evaluation Campaigns
UKA	12	2003/2004	Lectures	CHIL Internal Evals., CLEF07
UKA	29	2004/2005	Lectures	CLEAR06, RT05s, RT06s, CLEF07
ITC	5	2005	Lectures	CLEAR06, RT06s
UKA	5	2006	Meetings	CLEAR07, RT07
ITC	5	2006	Meetings	CLEAR07, RT07
AIT	5	2005	Meetings	CLEAR06, RT06s
AIT	5	2006	Meetings	CLEAR07, RT07
IBM	5	2005	Meetings	CLEAR06, RT06s
IBM	5	2006	Meetings	CLEAR07, RT07
UPC	5	2005	Meetings	CLEAR06, RT06s
UPC	5	2006	Meetings	CLEAR07, RT07

Table 1 Details of the 86 collected lectures/non interactive seminars (upper table part) and meetings/interactive seminars (lower part) that comprise the CHIL corpus. The table depicts the recording site, year, type, and number of collections, as well as the evaluation campaigns where the data were used.

In the *Meeting* scenario, the audience is small, between three and five people, and the attendees mostly sit around a table, all wearing close-talking microphones. There exists significant interaction between the presenter and the audience, with numerous questions and often a brief discussion among meeting participants. A few of these meetings are scripted to provide a more interesting, challenging, and dynamic interaction, with participants entering or leaving the room or congregating during a short coffee break. In addition, a significant number of acoustic events is generated to allow more meaningful evaluation of the corresponding technology. Clearly, in such a scenario all participants are of interest to meeting analysis. Therefore, the CHIL corpus provides annotations for all. Such data have been recorded by AIT, IBM and UPC in 2005, as well as all five sites during 2006. A total of 40 meetings are part of the CHIL corpus (see also Table 1), most of which are approximately 30 minutes in duration.

3.2 Corpus Annotation

For the collected data to be useful in technology evaluation and development, it is crucial that the corpus is accompanied by appropriate annotations. The CHIL consortium has devoted significant effort in identifying useful and efficient annotation schemes for the CHIL corpus and providing appropriate labels in support of these activities. As a result, the data set contains a rich set of annotations in multiple channels of both audio and visual modalities.

Data recording in the CHIL smart room results in multiple audio files containing signals recorded by close-talking microphones (near-field condition), table-top microphones, T-shaped clusters, and Mark III microphone arrays [68] (far-field condition), in parallel. The recorded speech as well as environmental acoustic events were carefully segmented and annotated by human transcribers.

Video annotations were manually generated using an ad-hoc tool. The tool allows displaying one picture every second, in sequence, for four cameras. To generate labels, the annotator performs a number of clicks on the head region of the persons of interest, i.e. the lecturer only in the non-interactive seminar (lecture) scenario, but all participants in the interactive seminar (meeting) scenario. In particular, the annotator first clicks on the head centroid (e.g., the estimated center of the person's head), followed by the left eye, the right eye, and the nose bridge (if visible). In addition, the annotator delimits the person's face with a bounding box. The 2D coordinates of the marked points within the camera plane are saved to the corresponding label file. This allows the computation of the 3D head location of the persons of interest inside the room, based on camera calibration information.

In addition to 2D face and 3D head location information, part of the lecture recordings were also labeled with gross information about the lecturer's head pose. In particular, only eight head orientation classes were annotated, as this was deemed to be a feasible task for human annotators, given the low-resolution captured views of the lecturer's head. The head orientation label corresponded to one of eight discrete orientation classes, ranging from a 0^o to 315^o angle, with an increment of 45^o. Overall, nineteen lecture videos were annotated with such information.

4 Software Infrastructure

CHIL aims to develop next-generation services that assist humans in a natural and unobtrusive way by utilizing the state-of-the-art in machine perception and context modeling. The CHIL software infrastructure makes this vision conceivable by implementing three key architectural pieces: the CHIL Reference Architecture [25], a set of CHIL Tools for service authoring, and the catalog of CHIL-compliant components.

The CHIL Reference Architecture (also called the Ice Cube) provides a collection of structuring principles, specifications and Application Programming Interfaces (APIs) that govern the assemblage of CHIL components into highly distributed and heterogeneous interoperable systems. The CHIL Ice Cube is designed as a layered architecture model (see Fig. 10). Each level represents an abstraction of the information processed and has specific data flow characteristics such as latency and bandwidth, based on the functional requirements from the CHIL system design phase. Building software infrastructure for smart context-aware services is an iterative process, where most of the software architecture craft resides in the discovery of suitable conceptual abstractions and the refinement of their functional relationships. We now describe the basic layers of the CHIL Ice Cube.

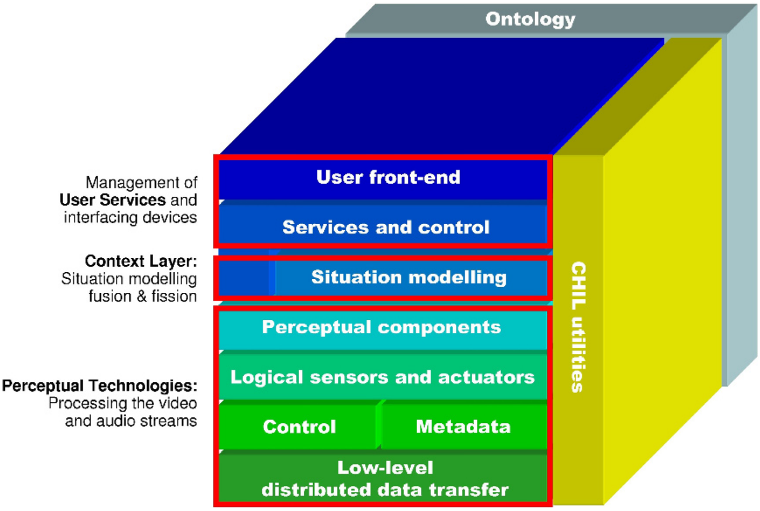


Fig. 10 The CHIL Reference Architecture (CHIL Ice Cube).

The upper layers – User Services – implemented as software agents, manage interactions with humans by means of various interaction devices. The components at this level are responsible for communication with the user and with presenting the appropriate information at appropriate time-spatial interaction zones. The User Services utilize the contextual information available from the situation modeling layer. In projects like CHIL, where a number of service developers concentrate on radically different services, it is of high value that a framework ensures reusability in the scope of a range of services. To this end, we have devised a multi-agent framework that:

- Facilitates integration of diverse context-aware services developed by different service providers.
- Facilitates the development of services by leveraging basic functionalities (e.g. sensor and actuator control) available within the smart rooms.
- Allows augmentation and evolution of the underlying infrastructure independent of the services installed in the room.
- Controls user access to services and supports service personalization through maintaining appropriate profiles.
- Enables discovery, involvement and collaboration of services.

The middle layer – Situation Modeling – is the place where the situation context received from audio and video sensors is processed and modeled. The context information acquired by the components at this layer helps CHIL services to respond better to varying user activities and environment changes. For example, the situating modeling answers questions such as: “Is there a meeting going on in the smart room?”, “Who is the person speaking at the whiteboard?”, “Has this person been in the room before?”. In the CHIL Reference Architecture, the Situation Modeling layer is implemented by the SitCom framework. Specifically, this layer is also a collection of abstractions representing the environment context in which the user interacts with the application. It thus maintains an up-to-date state of objects (people, artifacts, situations) and their relationships. The situation model acts as a inference engine, which watches for the occurrence of certain situations in the environment and triggers respective events and actions.

The lower layers – Perceptual Technologies – host perceptual components that extract meaningful events from continuous streams of video and audio signals. All kinds and variations of perceptual components live in this layer. In CHIL we actually counted a total of 64 such components provided by 8 different project partners. These components process the sensor signals from one modality (body trackers, face recognizers) or modality combinations (audio-visual speech recognition) and deliver events to upper layers via a dedicated communication mechanism, the CHiLiX middleware. The CHIL Reference Architecture defines the API contract (Access, Subscriber, Control, Introspection, and Admin APIs) as well as the life cycle (Unregistered, Registered, Launched, Running) to which the perceptual components must adhere to be considered CHIL-compliant. The compliance specifies how these perceptual components operate, “advertise” themselves, subscribe to receiving a specific sensor data stream and how they forward their extracted context to the higher layers of the CHIL Ice Cube.

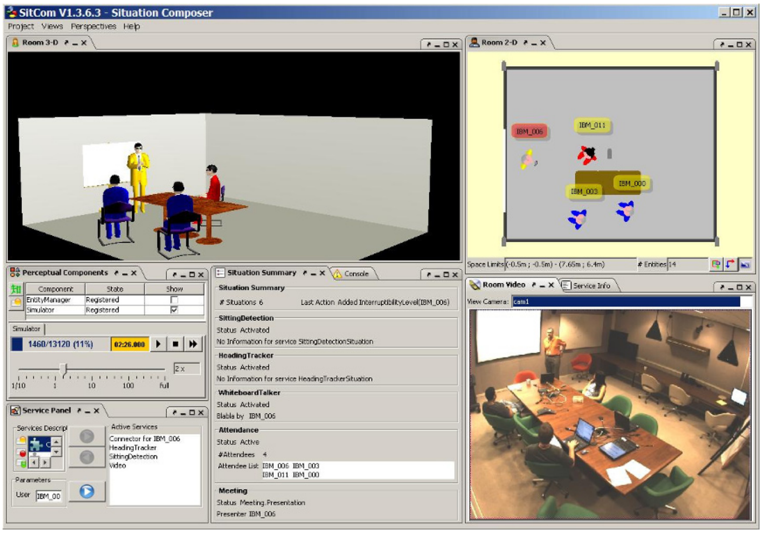
The vertical columns CHIL utilities and Ontology provide support across layers. The CHIL utilities provide global timing and other basic services that are relevant to all layers. The Ontology provides a definition of CHIL concepts and a directory service that provides access to information at any layer. The CHIL Ontology is implemented in a Knowledge Base Server. It is worth emphasizing that the layers are not strictly isolated. For example, the multimodal service residing at the user front-end can render a video stream originating from the Logical sensors and actuators at the bottom of the Ice Cube.

Separating the architecture into different abstract functional layers enabled the implementation of CHIL components that are self describing, self regulating, self repairing and self configuring. Via feedback and evaluation from developers, we are enhancing the CHIL architecture with features such as dynamic service look-up and invocation, system reconfiguration and adaptation. We strongly believe that these non-proprietary APIs, tools and methods, available as part of the CHIL technology catalog [19], can be embraced by a larger community of pervasive service researchers and developers.

4.1 SitCom - The Situation Composer

For designing, debugging and running the situation model, we have implemented a framework called SitCom [40]. SitCom (Situation Composer) is a simulator tool and runtime for the development of context-aware applications and services. SitCom gets information from perceptual and other context acquisition components and allows the composition of situation models into hierarchies to provide event filtering, aggregation, and combination. Through its IDE controls, SitCom also facilitates the capture and creation of situation machines and subsequent realistic rendering of the scenarios in a 3D visualization module.

The conceptual goal of SitCom is to extract semantically higher information from the stream of environment observations. Technically, this is achieved through a set of situation machines (SM), where each SM is being implemented as a finite state graph. Situation machines can be organized into hierarchical structures. A state of the SM is determined by a set of entities in the entity repository that mirror real objects in the environment and by the current states of other SMs. The actual implementation of a particular situation machine can be part of the Java runtime or it can be a standalone module using a remote API. In Fig. 11, we can see the middle lower panel showing the currently active set of situation machines, along with their state information, both a 2D and a 3D visualization, and the corresponding video frame from the actual meeting recording.



Space Limits (-0.5m ; -0.5m) - (7.65m ; 6.4m) # Entities 14

Component	State	Show
EntityManager	Registered	□
Simulator	Registered	☑

1460/13120 (11%) 02:26.000 2 x

Active Services
Connector for IBM_006
HeadingTracker
SittingDetection
Video

Parameters
User IBM_00

Situation Summary
Situations 6 Last Action Added InterruptibilityLevel(IBM_006)

SittingDetection
Status Activated
No Information for service SittingDetectionSituation

HeadingTracker
Status Activated
No Information for service HeadingTrackerSituation

WhiteboardTalker
Status Activated
Blabla by IBM_006

Attendance
Status Active
#Attendees 4
Attendee List IBM_006 IBM_003 IBM_011 IBM_000

Meeting
Status Meeting.Presentation
Presenter IBM_006

View Camera: cam1

Fig. 11 SitCom running the situation model on meeting data recorded in the IBM smart room.

4.2 Agent Infrastructure

The CHIL software agent infrastructure covers the upper two layers of the CHIL architecture. Agents and components close to the user including the User Profile are situated in the User Front-end layer. Agents in the Services and Control layer are further subdivided in basic agents including the communication ontology and specific service agents. Services include reusable basic services as well as complex higher-level services composed of suitable elementary services.

The CHIL Agent, the CHIL Agent Manager, the communication ontology and the Service Agent form the fundamental part of the agent framework. The CHIL Agent is the basic abstract class for all agents in the CHIL environment. It provides methods for agent administrative functionality, for registration and deregistration of services provided by agent instantiations, and additional supporting functions like creating and sending ontology-based messages, extracting message contents and logging. The CHIL Agent Manager is a central instance encapsulating and adding functionality to the JADE [51] Directory Facilitator (DF). Other CHIL agents register their services with the Agent Manager including required resources for carrying out these services. The Service Agent is the abstract base class for all specific service agents.

Service Agents provide access to the service functionality. Based on a common ontology they map the syntactical level of the services to the semantic level of the agent community enabling the use of other Service Agents to supply their own service. The Travel Agent can arrange and rearrange itineraries according to the user's profile and current situation; it uses a simulated semantic web interface to gather the

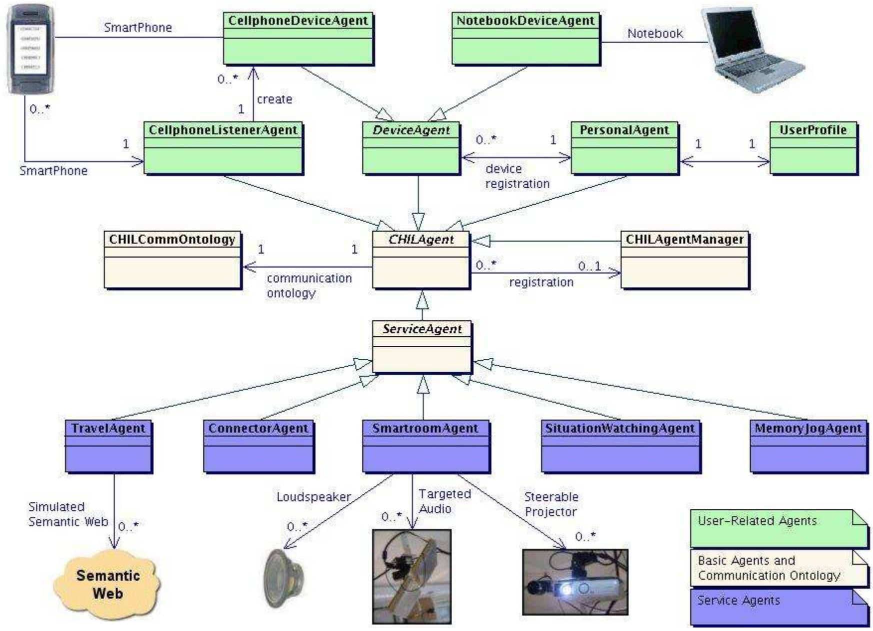


Fig. 12 CHIL agent infrastructure.

required data from several online resources on the internet. The Connector Agent comprises the core Connector Service functionality: managing intelligent communication links, both bilateral and multilateral, it stores pending connections and finds a suitable point in time for these connections to take place. The Smart Room Agent controls the service-to-user notification inside the smart room using appropriate devices such as loudspeakers, steerable video projectors and targeted audio. The Situation Watching Agent provides access to the situation model. Finally, the Memory Jog Agent is the central agent for the Memory Jog Service, providing pertinent information to the participants.

As proposed in the FIPA Abstract Architecture Specification [37], the information exchange between agents is completely based upon a specific communication ontology in order to ensure that the semantic content of tokens is preserved across agents. Distributed development of multiple services and the necessity to understand each other increase the importance of such a common semantic concept. Together with a FIPA-compliant implementation of the agent communication and an agent coordination mechanism, the communication ontology forms a sophisticated technique for agent collaboration.

In order to allow the distributed development of services and to facilitate their integration and configuration, a plugin mechanism has been designed that enables the implementation of service specific code in pluggable handlers, keeping the agent itself service independent. Three types of pluggable handlers have been considered to be necessary: “Setup handlers” are responsible for service specific initialization in the setup phase of an agent, “Event handlers” are registered for certain events from outside the agent world (e.g. the user’s GUI, a perceptual component, the situation model or a web service), and “Pluggable responders” are triggered by incoming messages from other agents.

5 CHIL Services

Several prototypical services were developed that instantiate the CHIL vision of perceptive and proactive services supporting human-human interaction. These services make use of the perceptual technologies and are integrated using the software architecture described in the previous section. Chosen domains for applications were lectures and small office meetings (up to 4-5 participants). The development process took a User-Centered Design (UCD) approach, which is characterized by taking into consideration the user perspective in all stages of the design process. CHIL services address intrinsically social scenarios, where two or more people interact to reach both individual and shared (group-level) goals. They aim to unobtrusively and seamlessly offer better chances to people for managing meeting-related activities. On the basis of an analysis of primary processes, services were proposed for supporting task-related activities such as discussing, producing and sharing material, accomplishing common goals (the Cooperative Workspace) and providing background information and memory assistance (the Memory Jog), for managing the social interaction and fostering a better quality of the interaction (the Relational Report and the Relational Cockpit), and for managing contacts with the outside world (the Connector). These goals are pursued by means of services that perceive the ongoing communication, reason about what they perceive and act to meet people’s needs and objectives, either autonomously or on demand.

User-Centered Design (UCD) was adopted as a general methodology to secure that the services developed provide value to people, e.g., in terms of performance, perceived helpfulness, etc. The consistency and appropriateness of the initial service concepts, as emerged from brainstorming within the consortium, were validated by means of interviews and focus groups with potential end users. This stage produced a set of clearer requirements, modifications and refinements of the original concepts, which were used for the design of the various services, followed by the implementation of the first prototypes. Formative evaluations of the designs were performed by means of user studies conducted with mock-ups, simulations, and prototypes. Despite the fact that the usefulness of simulations and mock-ups can be limited by the status of the perceptual components involved – continuous monitoring of people locations, focus of attention, acoustic events, etc., is not an easy job for humans to perform, or simulate in mock-ups – at least in some cases it was possible to both validate the services’ conception and collect information for the later stages of the UCD cycle. For instance, it turned out that people find a computer’s feedback about their social behavior as acceptable as that from a human expert, and that such a feedback can actually influence behavior, making it more appropriate to the goals of the meeting.

Insights from the initial phases were fed into the following implementation phase, leading to the deployment of the working prototypes for the services described above. The latter were tested in summative evaluations, with two purposes. The first purpose was to establish whether and to what extent our services enhanced productivity, both in terms of effectiveness (quality of solution) and efficiency (speed to arrive at the solution). The second purpose was to establish whether and to what extent our services enhanced the satisfaction of meeting participants, in terms of dimensions such as ease of use, usefulness, involvement, group engagement, etc. These data were used to model people' intention to use the various services in terms of networks where the psychological and motivational dimensions are posited to causally affect intention to use. The model was realized by means of a questionnaire measuring the relevant dimensions by means of appropriate psychometric scales, and was tested by applying Partial Least Square Structural Equation Modeling.

5.1 The Memory Jog

CHIL produced a number of prototype services in order to validate the technologies, but also to manifest the benefits of context-aware implicit services. One of these services of the CHIL project was Memory Jog (MJ), a complex integrated context-aware service that was developed in-line with the CHIL Reference Architecture. Apart from the CHIL Reference Architecture, the Memory Jog service made extensive use of perceptual components of the CHIL technology catalog [19]. MJ was developed and demonstrated within the CHIL smart rooms [83].

From a functional perspective the MJ service is a pervasive non-obtrusive human-centric assistant for lectures, meetings and presentations, which supports:

- Continuous room monitoring enabling identification and tracking of persons in the room. To this end the service leverages face detector, body tracker and face identification components.
- Context-Aware provision of information concerning past events and access to associated information. Information is provided upon user request, but also in a proactive and non-obtrusive manner (i.e. based on both "push" and "pull" modes). Moreover, information can be provided to participants physically present in the meeting/lecture, as well as remote participants who attend the meeting/lecture via a networked device.
- Context-aware access to actuating services (such as display services and targeted audio). These services are appropriately invoked based on the status of the user and his/her surrounding environment. The MJ supports a range of automatic, ambient and context-aware invocations of services (e.g., context-aware opening and display of slides).
- Ambient video recording of the meeting from a multi-camera system. A multi-camera selection algorithm optimally selects the best camera stream for the cur-

rent speaker, while recordings are tagged according to contextual states. In this way, the MJ system integrates the functionality of an automated "intelligent" cameraman, without however any human intervention.

- A "Show me the video" functionality enabling searching and accessing past video recordings, according to the collection of past contextual states used to tag the recordings. Hence, participants can recall past video segments pertaining to their context.
- A "What happened while I was away?" functionality providing a summary of events that occurred during a user's short periods of absence.
- Context-aware integration with additional third-party utilities such as Skype and Google Calendar. This functionality enables the services to initiate Skype sessions and Google searches, according to the end-user's needs.

From an integration perspective the MJ service leverages structuring principles and middleware libraries specified in the CHIL Reference architecture. Specifically, the CHiLiX middleware bridge [27] is used to integrate perceptual components and service elements, which are hosted on different platforms. At the heart of the MJ implementation is also the situation modeling approach defined in the CHIL Reference Architecture. In particular, the MJ is supported by a situation model which defines a graph of situations comprising the context states of interest for the service, along with the possible transitions between these states. Situation states are triggered by combinations of perceptual components including body trackers, face and speech identification components, speech recognizers, acoustic localization components, face detectors, as well as speech activity detectors. These components provide elementary context cues (e.g., person's location), which are then combined toward identifying the more sophisticated contextual states (e.g., a person returning in the room, a person conducting a presentation, a person addressing a question). Perceptual component combinations for triggering the states of the situation model, as well as the situation model are depicted in Fig. 13 which is derived from [27]. According to the CHIL situation modeling approach, a situation is triggered when all perceptual component outputs are close to the target values.

Implementing the MJ required thorough design and testing of the situation model. This was challenging given that most of the underlying perceptual components were in their infancy when the development started. The SitCom tool [27] facilitated the parallel evolution of the development of perceptual components and situation models. Service developers were able to systematically simulate contextual states and models, until they provided a model that meets end-user requirements, while also being in-line with the functionalities of perceptual components. Note however that the rule based approach adopted for situation modeling in CHIL can be error prone, as a result of failing or inaccurate perceptual components. To this end, the CHIL project has also investigated and implemented statistical modules for situation machines (see [83] and references therein). Such statistical modules can handle and resolve conflicting input in cases of uncertainty.

In addition to the CHIL situation modeling approach, MJ also makes use of the CHIL Knowledge Base Server (KBS) [83]. The KBS provides directory services for all the components of the MJ service. Sensors, devices, perceptual components,

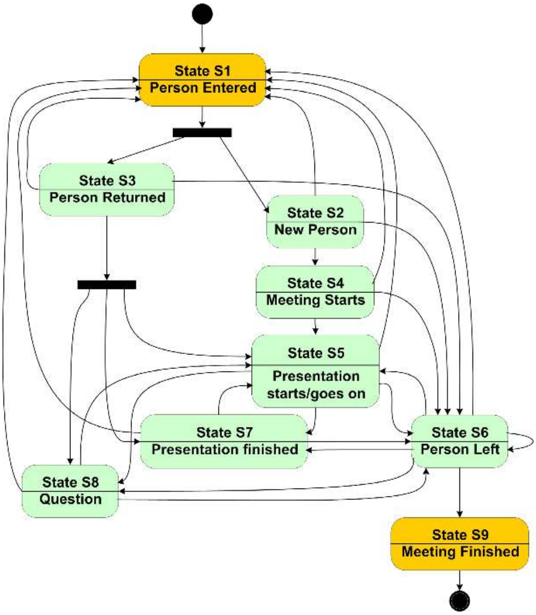


Fig. 13 The situation model for monitoring meetings as adopted by the AIT Memory Jog service.

actuating services, as well as the smart room configuration are registered within the semantic KBS. Through querying the KBS perceptual components can automatically get a reference to their required sensors, while they can also access information about the room's physical objects. Similarly, situation models can automatically discover their required perceptual components. The KBS facilitates discovery and use of the various MJ components (e.g., sensors, perceptual components, actuating services) in a nearly "plug n' play" fashion.

Overall, the MJ is a manifestation of a non-trivial pervasive service that consists of several hardware, software and middleware components in a highly distributed and heterogeneous environment. This service has validated a great number of CHIL perceptual processing components, while also providing insights into the benefits on implicit human centric services. At the same time, the MJ developments are a first class manifestation of CHIL contributions in the area of integrated development of pervasive systems for smart spaces [27] .

5.2 The Collaborative Workspace

The Collaborative Workspace (CW) is an infrastructure for fostering cooperation among participants. The system provides a multimodal interface for entering and manipulating contributions from different participants, e.g., by supporting joint discussion of minutes or joint accomplishment of a common task, with people proposing their ideas and making them available on the shared workspace, where they are discussed by the whole group.

Technologies to support human-human collaboration have always been a hot topic for computer science. Meetings in particular represent a stimulating topic since they are a common, yet at the same time problematic, human activity. One of the seminal studies to inform design of technology to support meetings is [91]. The aim of that work was to inform the design of an application to support remote meetings, yet the author described observations to real face-to-face meetings. The list of work on remote meetings published since then is long enough to discourage any attempt at synthesis. In the last years, the emergence of hardware able to support, at least partially, multi-users raised the interest in technologies to support face-to-face collaboration [64]. Shen and colleagues in [79] proposed the use of DiamondTouch, a real multi-user touch device to support collocated interaction. They proposed a horizontal circular interface to solve the problems of the different users' point of view around the table while users manipulated the objects on the projected display by touching the devices with the fingers. Kray and colleagues [55] discuss several issues that arise in the design of interfaces for multiple users interacting with multiple devices with focus on user and device management, technical concerns and social concerns. The Collaborative Workspace is realized as a horizontal device and implemented as a top-projected interface that turns a standard wooden table into an active surface (see Fig. 14, left). It also integrates person tracking facilities to detect the (possibly changing) location of the users around the table, moving toolbars accordingly. The main objects are virtual sheets of paper that can be created, edited, rotated and shared by the participants and used as generic textual or graphical documents. They are acted upon by means of an electronic pen (see Fig. 14, right). In order to have a more natural way of rotating documents, the Rotate and Translate algorithm proposed by Kruger and colleagues [56] was implemented. This technique allows the rotation and the movement of the window with a single gesture by dragging the window from a border. The system is also augmented with an agenda that allows the participants to keep track of the time spent on each item. Furthermore, by dragging documents on an item it will store them and use them for a semi-automatic generation of the minutes. In its simplest form, the minutes contain just the indication of the time spent, and, if any, the documents generated from a template. A more complex use of the agenda tool, that leads to a more elaborated report, is by using a special type of document, called the "note". Notes look pretty similar to textual documents with the difference that they also have space to store other documents as attachments. When dragged into an agenda item, however, the text of the note becomes part of the report. A careful use of the notes, therefore, allows the users to generate rich reports during the meetings.

In order to evaluate the prototype, several project managers working at FBK-irst in various research projects in the Information Technology field were requested to hold their meetings in a special room equipped with the tabletop device and video recording infrastructure. Overall, 16 meetings by 5 different groups were recorded during the period from December 2004 to June 2005. The main purpose of the meet-

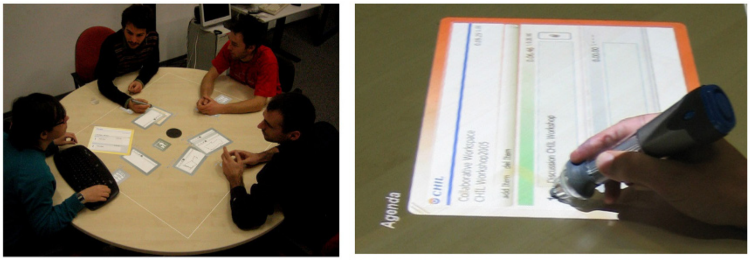


Fig. 14 Left, the Collaborative Workspace. Right, a snapshot of the interface.

ings was project coordination, decision taking, sharing of information, planning of future individual work and so on. It is worth noting that the small groups observed can actually be defined as teams, i.e. groups of people who are task interdependent and share a goal or purpose [53].

Two broad dimensions emerged from the observations: the patterns of technology use and the styles of interaction that the groups employed. While the patterns of use make evident the lack or the redundancy of some of the functionalities thus providing direct insights for a future re-design, the style of interaction revealed how the collocated interface may help in structuring the group style of interaction and how the interface promotes certain kinds of collaboration. More details on the results of the evaluation and the design process can be found in [94].

The most important pattern of use relates to the use of the documents. For two of the five groups, the documents served mainly as a whiteboard. The documents were enlarged to reach about the surface of the interface and used to support the communication by drawing diagrams and schemes. In this way the documents foster communication and the common understanding of a problem by providing a shared attentional focus for participants. In other meetings, documents were used to simulate the functionalities not provided by the system. For example, a calendar and a timeline have been sketched in order to plan future activities. Both these aspects relate to the concept of appropriation [29] which is defined as the process by which a technology is adopted and adapted into a working practice. Appropriation is claimed to be endemic to collaborative work and it is a twofold process: from one side the users modify their behavior to fit the system constraints while from the other side they re-invent new uses to accomplish tasks not foreseen by the designers.

In some cases the tabletop device has been circumvented by traditional technologies and artifacts, such as whiteboard, pens and, above all, paper-based resources. In any ecological situation, the technology has to “negotiate” its space with other artifacts. Harper and Sellen [78] demonstrated, for example, the importance of the paper in supporting collaborative work. In our observations too, paper and electronic tools co-exist and contribute to the meeting activities. Paper is mainly used at the personal level during the meeting, usually to take personal notes. Yet paper resources also served to introduce in the meeting preparation the work done before (as handouts or as personal notes). The necessity of linking the meeting to the activity of the team before and after the meeting itself emerged in several interviews.

Three main working styles have been identified observing the allocation of inputs between participants and the functionalities employed: the division of labor, the specialization in roles and the centralized management. These patterns contribute to understanding the modalities through which the small groups implicitly or explicitly coordinate their actions in order to carry out a collaborative activity with the collocated interface. Also, this work is significant as a first step toward the identification of learning and repairing processes that groups display to overcome system limitations or unexpected behavior of the system.

The CW proved to be flexible enough to make it possible for each group to accommodate it into their work practices while not preventing the teams to keep using their normal tools (from sheets of papers to personal laptops). Yet, the introduction of the system required the groups to engage in a sense-making process that was sometimes disrupting for the cohesion and the efficiency of the meeting. In designing this kind of system it is crucial to consider both their usability and their impact on established working practices.

5.3 Managing the Social Interaction: The Relational Cockpit and the Relational Report

The availability of rich multimodal information can be exploited to provide group-oriented services. Our approach to service development was inspired by the observation that often professional coaches are invoked to facilitate group sessions and provide feedback to team members about their behavior, in order to help the team improve the social interaction and the effectiveness of their meetings. Two services were developed and evaluated. The Relational Cockpit (RC) monitors and provides feedback during a meeting about participants' speaking time and eye gaze behavior. The Relational Report (RR) consists of a report about the social behavior of individual participants that is generated from multimodal information, and privately delivered to them after the meeting [70]. In both cases, the underlying idea is that the individuals, the group(s) they are part of, and the whole organization might benefit from an increased awareness of participants about their own behavior during meetings.

5.3.1 The Relational Cockpit

The Relational Cockpit aimed to provide non-obtrusive feedback about participants' social behavior during a meeting. This aim inspired us to adopt the constraint that the service should not provide in-depth feedback (this was to be reserved for the Relational Report), but instead feedback that could be easily grasped from a peripheral display. The focus of our research was on meetings where all participants are supposed to contribute and reach consensus about a course of action. Based on literature reports, analyses of meeting behavior and interviews, we identified aspects of non-verbal behavior that are indicative of the social inclusion of meeting participants as most suited for feedback. In the first place, the speaking time of individual participants provides information about dominance patterns, leading to the monopolization of the conversation by certain participants and the exclusion of other participants from the conversation. In the second place, the eye gaze patterns of meeting participants provide information about the turn-giving and turn-taking behavior during a meeting. Since the current speaker has the privilege to select the next speaker and eye gaze plays an important role in identifying the addressee of an utterance, meeting participants who are not gazed at by the current speaker will feel neglected and excluded.

Based on this reasoning, we devised a service that provided feedback about speaking time and eye gaze patterns to meeting participants. Speaking time was obtained from a speech diarization component (indicating who is speaking when). Eye gaze patterns were estimated from a perceptual component that measured head orientation (the physical arrangement of the meeting participants around the table was such that shifts in eye gaze across participants required a change in head orientation). The extracted speaking time and eye gaze patterns were visualized in the form of colored circles that were projected on the table before each participant and that would grow as the meeting evolved. The circle for speaking time reflected the amount of time that a participant had been speaking since the beginning of the meeting. A relatively large circle, compared to other participants, would identify that participant as an "over-participator", i.e. someone who had been speaking proportionally much. The circle for eye gaze reflected the amount of time that a participant had been gazed at by the speaker since the beginning of the meeting, summing across the other participants. A relatively large circle, compared to other participants, would mean that the participant had been gazed at proportionally much by the other participants when they had been speaking. There was a third circle for each participant, reflecting how much visual attention he or she received while speaking from the other participants, but we will leave this out of consideration in the further treatment.

Reliability analyses were conducted by comparing post-hoc manual annotations with the automatically extracted speaking time and eye gaze patterns. It was found that, while the accuracy of the automatically extracted patterns was only modest at a micro-level, there was satisfactory agreement between the manually and automatically extracted patterns for the cumulative patterns. This means that the duration of individual speaking turns and shifts in eye gaze were not extracted very accurately, but summing over larger stretches of the conversation the total speaking time and gaze time for individual participants were estimated at a satisfactory level.

Two user tests were conducted to test the impact of the service on the social interaction during meetings [57, 89]. Summed across the two tests, thirty-one groups of four participants tried to reach an agreement about pre-specified topics requiring negotiation in two sessions. One session was a normal meeting, in the other session the group received feedback about its social interaction. Each session lasted about twenty minutes. The order of sessions with and without feedback was balanced across groups. It was found that participants who contributed relatively much ("over-participators") in the session without feedback used less speaking time in the session with feedback and that participants who contributed relatively little ("under-participators") in the session without feedback used more speaking time in the session with feedback. Further inspection of the data suggested that the effects were mainly due to the feedback on speaking time. There was no evidence that feedback on eye gaze patterns led speakers to change their eye gaze patterns such that they would result in better inclusion of "under-participators".

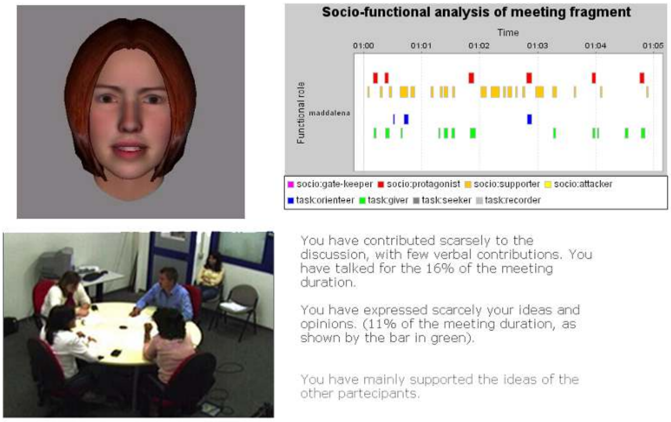
5.3.2 The Relational Report

The success of a meeting is often hindered by the participants' behavior: professionals agree that as much as 50% of meeting time is unproductive and that up to 25% of meeting time is spent discussing irrelevant issues [30]. In order to improve the productivity of meetings, external interventions such as facilitators and training experiences are commonly employed. Facilitators participate in the meetings as external elements of the group and their role is to help participants maintain a fair and focused behavior as well as directing and setting the pace of the discussion. Training experiences aim at increasing the relational skills of individual participants by providing an offline (with respect to meetings) guidance – or coaching – so that the team eventually will be able to overcome or to cope with its disfunctionalities.

The Relational Report (RR) aims at providing coaching support in an automatic way, by monitoring groups, and generating individual reports about the participants' relational behavior. The system does not keep verbatim trace of what people said or did during the meeting, and does not produce minutes; rather it makes available more qualitative, meta-level interpretations of people's social behavior. The reports are delivered privately to each participant after the meeting, with the purpose of informing them about their behavior, this way stimulating reflective processes and, ultimately behavioral change [12]. Support for the idea that automatic coaching might be as effective as human coaching comes from studies that have shown how people do not rate differently reports about their own relational behavior according to whether they were generated by a human expert or by a system on dimensions such as perceived usefulness of the report, reliability, intrusiveness, and acceptability [70].

The RR composes its reports from information about the functional relational roles the given person has played during the meeting. Building on Benne and Sheats [6] and Bales [5] works, we identified two sets of role labels: the Task Area, roles related to facilitation and coordination tasks as well as to technical expertise of members; the Socio Emotional Area, which is concerned with the relationships between group members and the "functioning of the group as a group". The Task Area functional roles include: the Orienteer, who orients and keeps the group focused; the Giver, who provides factual information and answers to questions; the Seeker, who requests information and clarifications; the Follower, who just listens, without actively participating. The Socio-Emotional functional roles include: the Attacker; who deflates the status of others, expressing disapproval; the Protagonist; who takes the floor and drives the conversation; the Supporter, who shows a cooperative attitude demonstrating understanding and attention; the Neutral, played by those who passively accept the ideas of the others, serving as an audience in group discussion; see [69].

Information concerning the functional relational roles played by participants is assembled by the system from low level audio-visual inputs, in particular, speech activity and fidgeting (that is, the overall level of hand and body activity). Those features were exploited in experiments aiming to investigate and compare the performances of a number of classifiers for role detection: Support Vector Machines, Hidden Markov Models and the Influence Model [99, 28] ; the best results were yielded by the Influence Modeling, with an accuracy of 75% for both “task” and “socio” roles. The RR uses this information to compose multimedial reports (see Fig. 15).



You have contributed scarsely to the discussion, with few verbal contributions. You have talked for the 16% of the meeting duration.

You have expressed scarcely your ideas and opinions. (11% of the meeting duration, as shown by the bar in green).

You have mainly supported the ideas of the other partecipants.

Fig. 15 An example of a multimedia relational report for one meeting participant.

The textual part describes the behavior in an informative rather than normative way, helping the user accomplish the first step of the reflective process – namely, the return to experience [12]. It is based on a general-purpose schema-based text planner [15] which accesses a repository of declarative discourse schemata derived from the analysis of the actual reports written by the psychologists. To improve effectiveness and emotional involvement, a virtual character reads the report with emotional facial expressions appropriate to the content (e.g., a sad expression is used when something unpleasant, e.g., a serious contrast with a colleague, is being recalled). When appropriate, the presentation is enriched with short audio-video clips from the actual meeting, which exemplify the information presented. Finally, a graphical representation of the participant’s behavior is also provided, yielding a more explicit feedback about the system’s internal interpretation of what was monitored during the meeting. The graphical part was inspired from visualizations discussed by Bales [5], and expresses at a glance the interaction behavior of a participant by profiling the amount of time spent on each task- and socio-emotional roles. More information about the presentation engine can be found in [70].

5.4 The Connector: A Virtual Secretary in Smart Offices

Much of the communication at the workplace – via the phone as well as face-to-face – occurs in inappropriate contexts, disturbing meetings and conversations, invading personal and corporate privacy, and more broadly breaking social norms. This is because both, callers and visitors in front of closed office doors, face the same problem: they can only guess the other person's current availability for a conversation.

The Connector Service was designed to facilitate more socially appropriate communication at the workplace by acting similarly to a context-aware Virtual Secretary. Concordant with the CHIL paradigm, it aims toward understanding a person's situation and activity in smart offices, and passes on important contextual information to callers and visitors in order to facilitate more informed human decisions about how and when to initiate contact [23].

Deploying such a context-aware Virtual Secretary requires an instrumented environment, able to sense situations, recognize people and track activities. At UKA-ISL, cameras were installed in "smart" offices (Fig. 16), and a set of perception technologies was employed as described previously in order to detect basic social situations: whether the office was empty, if the occupant was alone, or if he was in a face-to-face meeting with others.

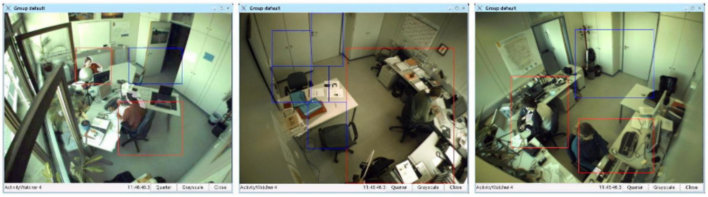


Fig. 16 Camera snapshots of different smart offices with occupants present.

The Connector as Virtual Secretary mediated all phone calls as well as face-to-face interactions in the smart offices. Knowing about the current situation, it could warn senders at the time of the call when the current situation was likely to be inappropriate for talking on the phone.

Virtual Secretary [synthesized voice]: *"Hello Jane. Thank you for calling Bob Smith. This is his Digital Secretary. Bob is currently in a meeting with Eric. Your call is important. Please hang on, or press 1, to leave a message. Instead, you may* *now press 2 to be connected to the office phone. To schedule a call at the next available time, please press 4.”*

Similarly, once a visitor was detected in front of the office, the Virtual Secretary provided him or her with important contextual information about the situation inside the office (Fig. 17). The goal was to prevent untimely disruptions from taking place, such as during meetings or phone conversations. In order to detect and identify people, a small webcam was placed on the screen that greeted visitors in front of the office door. A few seconds were enough for the face recognition system [35] to accurately detect and identify members of our research lab (approximately 15 persons) and distinguish them from unknown persons.

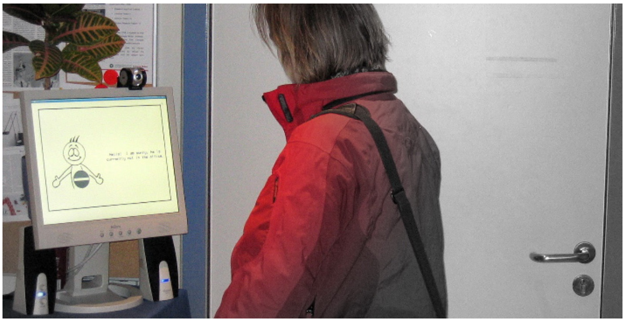


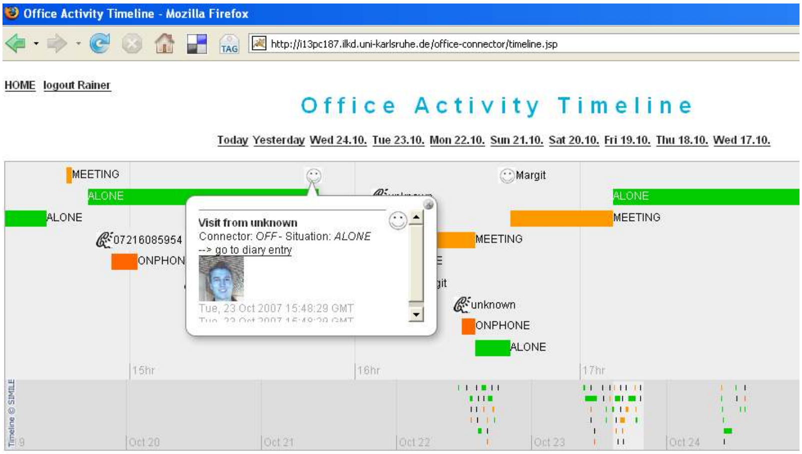
Fig. 17 A visitor consults the Virtual Secretary at the office door. A small webcam was placed on top of the screen from which the Virtual Secretary greeted visitors.

The Virtual Secretary was installed in front of the office of a senior researcher, who was very frequently occupied, and mediated all his phone calls and meetings over a period of three weeks. The system provided a web-based diary of all office activity, as shown in Fig. 18. At the end of each day, the participant was asked to browse this diary and annotate each interruption according to how available he had been and how appropriate the interruption was.

Observations and interviews showed that the system was running robustly over this extended period of time and was easy to use, even without any prior experience with it. A total number of 150 contact attempts (76 were phone calls) were protocoled by the system. Analyzing ratings of appropriateness of all interruptions empirically proved that there were significantly fewer inappropriate workplace interruptions when the Virtual Secretary mediated phone calls and direct interactions [23].

6 Conclusion

Implicit, proactive computing services as proposed in CHIL, are a way to free humans from unnecessary and unwanted preoccupation with technological artifacts.


Office Activity Timeline - Mozilla Firefox
http://i13pc187.ilkd.uni-karlsruhe.de/office-connector/timeline.jsp
HOME logout Rainer
Office Activity Timeline
Today Yesterday Wed 24.10. Tue 23.10. Mon 22.10. Sun 21.10. Sat 20.10. Fri 19.10. Thu 18.10. Wed 17.10.
MEETING
ALONE
ALONE
07216085954
ONPHON
Margit
ALONE
MEETING
MEETING
unknown
ONPHONE
ALONE
15hr
16hr
17hr
Timeline © SIMILE
Oct 20
Oct 21
Oct 22
Oct 23
Oct 24

Visit from unknown
Connector: *OFF* - Situation: *ALONE*
--> go to diary entry
Tue, 23 Oct 2007 15:48:29 GMT

Fig. 18 The office activity diary was shown as interactive time line widget.

We can think of CHIL computing as a whole new human-centered way of opening up access and deployment of computing services. They allow the attention of the human beneficiary of the services to shift away from human-machine interaction to (machine-supported) human-human interaction, in the interest of improved productivity and user satisfaction. Such proactive services, however, require considerably broader, better and more robust perceptual capabilities to suitably judge user activities, intent and needs. Under CHIL, we have proposed, implemented, improved and evaluated such perceptual processing over publicly available benchmarks, and we were able to integrate these technologies into multiple service prototypes efficiently under a common software architecture. It is our hope that this will seed and promote a drive for better and more appealing clutter-free computing.

Acknowledgement

The work presented here was funded by the *European Union* (EU) under the integrated project CHIL, *Computers in the Human Interaction Loop* (Grant number IST-506909).

References

[1] Abad, A., Canton-Ferrer, C., Segura, C., Landabaso, J.L., Macho, D., Casas, J.R., Hernando, J., Pardas, M., Nadeu, C.: UPC Audio, Video and Multimodal Person Tracking Systems in the CLEAR Evaluation Campaign. In: Multi-

modal Technologies for Perception of Humans, Proceedings of the First International CLEAR Evaluation Workshop. Springer LNCS 4122, Southampton, UK (2006)

[2] Anastasakos, T., McDonough, J., Schwartz, R., Makhoul, J.: A compact model for speaker adaptation training. In: Proc. Int. Conf. Spoken Language Process. (ICSLP), pp. 1137–1140. Philadelphia, PA (1996)

[3] Andreou, A., Kamm, T., Cohen, J.: Experiments in vocal tract normalisation. In: Proc. CAIP Works.: Frontiers in Speech Recognition II (1994)

[4] Anguera, X., Wooters, C., Hernando, J.: Acoustic beamforming for speaker diarization of meetings. IEEE Trans. Audio Speech Language Process. **15**(7), 2011–2022 (2007)

[5] Bales, R.F.: Interaction process analysis: a method for the study of small groups. University of Chicago press (1976)

[6] Benne, K.D., Sheats, P.: Functional roles of group members. Journal of Social Issues 4 pp. 41–49 (1948)

[7] Bernardin, K., Gehrig, T., Stiefelhagen, R.: Multi-Level Particle Filter Fusion of Features and Cues for Audio-Visual Person Tracking. In: Multimodal Technologies for Perception of Humans, Proceedings of the International Evaluation Workshops CLEAR 2007 and RT 2007, *LNCS*, vol. 4625, pp. 70–81. Springer, Baltimore, MD, USA (2007)

[8] Bernardin, K., Stiefelhagen, R.: Evaluating multiple object tracking performance: The CLEAR MOT metrics. EURASIP Journal on Image and Video Processing, Special Issue on Video Tracking in Complex Scenes for Surveillance Applications (2008)

[9] Beskow, J., Karlsson, I., Kewley, J., Salvi, G.: SYNFACE - A talking head telephone for the hearing-impaired, pp. 1178–1186. Springer-Verlag (2004)

[10] Beskow, J., Nordenberg, M.: Data-driven synthesis of expressive visual speech using an mpeg-4 talking head. In: Proceedings of Interspeech 2005. Lisbon (2005)

[11] Bobick, A.F., Davis, J.W.: The recognition of human movement using temporal templates. IEEE Trans. on Pattern Analysis and Machine Intelligence **23**(3), 257–267 (2001)

[12] Boud, D., Keogh, R., (Eds.), D.W.: Reflection: Turning experience into learning. Kogan Page, London (1988)

[13] Brunelli, R., Brutti, A., Chippendale, P., Lanz, O., Omologo, M., Svaizer, P., Tobia, F.: A generative approach to audio-visual person tracking. In: Multimodal Technologies for Perception of Humans, Proceedings of the First International CLEAR Evaluation Workshop, pp. 55–68. Springer LNCS 4122, Southampton, UK (2006)

[14] Brutti, A.: A person tracking system for CHIL meetings. In: Multimodal Technologies for Perception of Humans, Proceedings of the International Evaluation Workshops CLEAR 2007 and RT 2007, *LNCS*, vol. 4625. Springer, Baltimore, MD, USA (2007)

[15] Callaway, C., Not, E., Stock, O.: Report generation for post-visit summaries in museum environments. In: O. Stock, M. Zancanaro (eds.). PEACH: Intelligent Interfaces for Museum Visits. Springer (2007)

[16] Canton-Ferrer, C., Casas, J.R., Pardàs, M.: Human model and motion based 3D action recognition in multiple view scenarios (invited paper). In: 14th European Signal Processing Conference, EUSIPCO. EURASIP, University of Pisa, Florence, Italy (2006). ISBN: 0-387-34223-0

[17] Canton-Ferrer, C., Salvador, J., Casas, J., M.Pardas: Multi-person tracking strategies based on voxel analysis. In: Multimodal Technologies for Perception of Humans, Proceedings of the International Evaluation Workshops CLEAR 2007 and RT 2007, *LNCS*, vol. 4625, pp. 91–103. Springer, Baltimore, MD, USA (2007)

[18] Canton-Ferrer, C., Segura, C., Casas, J.R., Pardàs, M., Hernando, J.: Audiovisual head orientation estimation with particle filters in multisensor scenarios. EURASIP Journal on Advances in Signal Processing (2007)

[19] The CHIL technology catalogue. `http://chil.server.de/servlet/is/5777/`

[20] Chippendale, P., Lanz, O.: Optimised meeting recording and annotation using real-time video analysis. In: Proc. 5th Joint Workshop on Machine Learning and Multimodal Interaction, MLMI08. Utrecht, The Netherlands (2008)

[21] CLEAR – Classification of Events, Activities, and Relationships Evaluation and Workshop: `http://www.clear-evaluation.org`

[22] The CLEF Website: `http://www.clef-campaign.org/`

[23] Danninger, M., Stiefelhagen, R.: A context-aware virtual secretary in a smart office environment. In: Proceedings of the ACM Multimedia 2008. Vancouver, Canada (2008)

[24] Davis, S., Mermelstein, P.: Comparison of parametric representations of monosyllabic word recognition in continuously spoken sentences. IEEE Trans. on Acoustics, Speech, and Signal Process. **28**(4), 357–366 (1980)

[25] D2.2 functional requirements and chil cooperative information system software design, part 2, cooperative information system software design. Available on http://chil.server.de

[26] Dempster, A.P., Laird, M.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm. J. of the Royal Statistical Society Series B (methodological) **39**, 1–38 (1977)

[27] Dimakis, N., Soldatos, J., Polymenakos, L., Curin, J., Fleury, P., Kleindienst, J.: Integrated development of context-aware applications in smart spaces. IEEE Pervasive Computing **7**(4), 71–79 (2008)

[28] Dong, W., Lepri, B., Cappelletti, A., Pentland, A., Pianesi, F., Zancanaro, M.: Using the influence model to recognize functional roles in meetings. In: Proceedings of the International Conference on Multimodal Interaction ICMI2007. Nagoya, Japan (2007)

[29] Dourish, P.: The appropriation of interactive technologies: Some lessons from placeless documents. Computer Supported Cooperative Work (2003)

[30] Doyle, M., Straus, D.: How To Make Meetings Work. The Berkley Publishing Group, New York, NY (1993)

[31] Edlund, J., Beskow, J.: Pushy versus meek - using avatars to influence turn-taking behaviour. In: Proceedings of Interspeech 2007 ICSLP, pp. 682–685. Antwerp, Belgium (2007)

[32] Edlund, J., Gustafson, J., Heldner, M., Hjalmarsson, A.: Towards human-like spoken dialogue systems. Speech Communication **50**(8-9), 630–645 (2008). URL http://www.speech.kth.se/prod/publications/files/3145.pdf

[33] Edlund, J., Heldner, M.: Exploring prosody in interaction control. Phonetica **62**(2-4), 215–226 (2005)

[34] Edlund, J., Heldner, M.: Underpinning /nailon/: automatic estimation of pitch range and speaker relative pitch. In: C. Müller (ed.) Speaker Classification. Springer/LNAI (2007)

[35] Ekenel, H.K., Stiefelhagen, R.: Analysis of local appearance-based face recognition: Effects of feature selection and feature normalization. In: CVPR Biometrics Workshop. New York, USA (2006)

[36] ELRA Catalogue of Language Resources: http://catalog.elra.info

[37] FIPA: The foundation for intelligent physical agents. http://www.fipa.org

[38] Fiscus, J.G.: A post-processing system to yield reduced word error rates: Recogniser output voting error reduction (ROVER). In: Proc. Automatic Speech Recognition and Understanding Works. (ASRU), pp. 347–352. Santa Barbara, CA (1997)

[39] Fiscus, J.G., Ajot, J., Michel, M., Garofolo, J.S.: The Rich Transcription 2006 Spring meeting recognition evaluation. In: S. Renals, S. Bengio, J.G. Fiscus (eds.) Machine Learning for Multimodal Interaction, vol. 4299, pp. 309–322. LNCS (2006)

[40] Fleury, P., Cuřín, J., Kleindienst, J.: SitCom - development platform for multimodal perceptual services. In: Proceedings of the 3nd International Conference on Industrial Applications of Holonic and Multi-Agent Systems, pp. 106–113. Regensburg, Germany (2007). V. Marik, V. Vyatkin, A.W. Colombo (Eds.): HoloMAS 2007, LNAI 4659

[41] Gauvain, J.L., Lee, C.: Maximum a Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains. IEEE Trans. on Speech and Audio Processing **2**(2), 291–298 (1994). URL ftp://tlp.limsi.fr/public/map93.ps.Z

[42] Gehrig, T., McDonough, J.: Tracking multiple speakers with probabilistic data association filters. In: Multimodal Technologies for Perception of Humans, Proceedings of the First International CLEAR Evaluation Workshop. Springer LNCS 4122, Southampton, UK (2006)

[43] Gopinath, R.: Maximum likelihood modeling with Gaussian distributions for classification. In: Proc. Int. Conf. Acoustics Speech Signal Process. (ICASSP), pp. 661–664. Seattle, WA (1998)

[44] Haeb-Umbach, R., Ney, H.: Linear discriminant analysis for improved large vocabulary continuous speech recognition. In: Proc. Int. Conf. Acoustics Speech Signal Process. (ICASSP), vol. 1, pp. 13–16 (1992)

[45] Heldner, M., Edlund, J., Carlson, R.: Interruption impossible. In: M. Horne, G. Bruce (eds.) Nordic Prosody: Proceedings of the IXth Conference, Lund 2004, pp. 97–105. Peter Lang, Frankfurt am Main (2006)

[46] Hermansky, H.: Perceptual linear predictive (PLP) analysis of speech. J. Acoustic Society America **87**(4), 1738–1752 (1990)

[47] Huang, J., Marcheret, E., Visweswariah, K.: Improving speaker diarization for CHIL lecture meetings. In: Proc. Interspeech, pp. 1865–1868. Antwerp, Belgium (2007)

[48] Huang, J., Marcheret, E., Visweswariah, K., Potamianos, G.: The IBM RT07 evaluation systems for speaker diarization on lecture meetings. In: Multimodal Technologies for Perception of Humans, Proceedings of the International Evaluation Workshops CLEAR 2007 and RT 2007, *LNCS*, vol. 4625, pp. 497–508. Springer, Baltimore, MD, USA (2007)

[49] Hugot, V.: Eye gaze analysis in human-human communication. Master thesis, KTH Speech, Music and Hearing (2007)

[50] Ivanov, Y.A., Bobick., A.F.: Recognition of visual activities and interactions by stochastic parsing. IEEE Transactions on Pattern Analysis and Machine Intelligence **22**, 852–872 (2000)

[51] JADE: Java Agent DEvelopent Framework. `http://jade.tilab.com`

[52] Katsarakis, N., Talantzis, F., Pnevmatikakis, A., Polymenakos, L.: The AIT 3D audio / visual person tracker for CLEAR 2007. In: Multimodal Technologies for Perception of Humans, Proceedings of the International Evaluation Workshops CLEAR 2007 and RT 2007, *LNCS*, vol. 4625, pp. 35–46. Springer, Baltimore, MD, USA (2007)

[53] Katznbach, J., Smith, D.: The Wisdom of Teams. Creating the High Performance Organisations. Harvard Business School Press, Cambridge, MA (1993)

[54] Klee, U., Gehrig, T., McDonough, J.: Kalman filters for time delay of arrival-based source localization. Journal of Advanced Signal Processing, Special Issue on Multi-Channel Speech Processing (2006)

[55] Kray, C., Wasinger, R., Kortuem, G.: Concepts and issues in interfaces for multiple users and multiple devices. In: Proceedings of the Workshop on Multi-User and Ubiquitous User Interfaces (MU3I), IUI/CADUI (2004)

[56] Kruger, R., Carpendale, M., Scott, S., Tang, A.: Fluid integration of rotation and translation. In: Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI 2005). Portland, Oregon (2005)

[57] Kulyk, O., Wang, C., Terken, J.: Real-time feedback based on nonverbal behaviour to enhance social dynamics in small group meetings. In: MLMI'05: Proceedings of the Joint Workshop on Multimodal Interaction and Related Machine Learning Algorithms, *LNCS*, vol. 3869, pp. 150–161 (2006)

[58] Landabaso, J.L., M. Pardas, M.: Foreground regions extraction and characterization towards real-time object tracking. In: Machine Learning for Multimodal Interaction (MLMI), vol. 3869, pp. 241–249. Springer LNCS (2006)

[59] Lanz, O.: Approximate Bayesian Multibody Tracking. IEEE Transactions on Pattern Analysis and Machine Intelligence **28**(9), 1436–1449 (2006)

[60] Lanz, O., Brunelli, R.: Dynamic head location and pose from video. In: IEEE Conf. Multisensor Fusion and Integration (2006)

[61] Lanz, O., Chippendale, P., Brunelli, R.: An appearance-based particle filter for visual tracking in smart rooms. In: Multimodal Technologies for Perception of Humans, Proceedings of the International Evaluation Workshops CLEAR 2007 and RT 2007, *LNCS*, vol. 4625, pp. 57–69. Springer, Baltimore, MD, USA (2007)

[62] Leggetter, C.J., Woodland, P.C.: Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. Computer Speech and Language **9**(2), 171–185 (1995)

[63] Luque, J., Anguera, X., Temko, A., Hernando, J.: Speaker diarization for conference room: The UPC RT07s evaluation system. In: Multimodal Technologies for Perception of Humans, Proceedings of the International Evaluation Workshops CLEAR 2007 and RT 2007, *LNCS*, vol. 4625, pp. 543–554. Springer, Baltimore, MD, USA (2007)

[64] Morris, M., Piper, A., Cassanego, A., Huang, A., Paepcke, A., Winograd, T.: Mediating group dynamics through tabletop interface design. IEEE Computer Graphics and Applications pp. 65–73 (2006)

[65] M.Voit, R.Stiefelhagen: Tracking head pose and focus of attention with multiple far-field cameras. In: International Conference On Multimodal Interfaces - ICMI 2006. Banff, Canada (2006)

[66] Nickel, K., Gehrig, T., Ekenel, H.K., McDonough, J., Stiefelhagen, R.: An audio-visual particle filter for speaker tracking on the CLEAR'06 evaluation dataset. In: Multimodal Technologies for Perception of Humans, Proceedings of the First International CLEAR Evaluation Workshop. Springer LNCS 4122, Southampton, UK (2006)

[67] Nickel, K., Gehrig, T., Stiefelhagen, R., McDonough, J.: A Joint Particle Filter for Audio-visual Speaker Tracking. In: Proceedings of the Seventh International Conference On Multimodal Interfaces - ICMI 2005, pp. 61–68. ACM Press (2005)

[68] The NIST MarkIII Microphone Array: `http://www.nist.gov/smartspace/mk3_presentation.html`

[69] Pianesi, F., Zancanaro, M., Lepri, B., Cappelletti, A.: A multimodal annotated corpus of consensus decision making meetings. The Journal of Language Resources and Evaluation **41**(3–4) (2007)

[70] Pianesi, F., Zancanaro, M., Not, E., Leonardi, C., Falcon, V., Lepri, B.: Multimodal support to group dynamics. Personal and Ubiquitous Computing **12**(2) (2008)

[71] Povey, D., Woodland, P.: Improved discriminative training techniques for large vocabulary continuous speech recognition. In: Proc. Int. Conf. Acoustics Speech Signal Process. (ICASSP). Salt Lake City, UT (2001)

[72] Povey, D., Woodland, P.C.: Minimum phone error and I-smoothing for improved discriminative training. In: Proc. Int. Conf. Acoustics Speech Signal Process. (ICASSP), pp. 105–108. Orlando, FL (2002)

[73] Rentzeperis, E., Stergiou, A., Boukis, C., Pnevmatikakis, A., Polymenakos, L.C.: The 2006 Athens Information Technology speech activity detection and speaker diarization systems. In: Machine Learning for Multimodal Interaction, vol. 4299, pp. 385–395. LNCS (2006)

[74] The Rich Transcription 2006 Spring Meeting Recognition Evaluation Website: `http://www.nist.gov/speech/tests/rt/2006-spring`

[75] Rich Transcription 2007 Meeting Recognition Evaluation. `http://www.nist.gov/speech/tests/rt/2007`

[76] Schwenk, H.: Efficient training of large neural networks for language modeling. In: IJCNN, pp. 3059–3062 (2004)

[77] Segura, C., Abad, A., Nadeu, C., Hernando, J.: Multispeaker localization and tracking in intelligent environments. In: Multimodal Technologies for Perception of Humans, Proceedings of the International Evaluation Workshops CLEAR 2007 and RT 2007, *LNCS*, vol. 4625, pp. 82–90. Springer, Baltimore, MD, USA (2007)

[78] Sellen, A., Harper, R.: The Myth of the Paperless Office. MIT Press (2001)

[79] Shen, C., Vernier, F., Forlines, C., Ringel, M.: Diamondspin: An extensible toolkit for around-the-table interaction. In: ACM Conference on Human Factors in Computing Systems (CHI) (2004)

[80] Siciliano, C., Williams, G., Beskow, J., Faulkner, A.: Evaluation of a multilingual synthetic talking face as a communication aid for the hearing impaired. In: Proc of ICPhS, XV Intl Conference of Phonetic Sciences, pp. 131–134. Barcelona, Spain (2003)

[81] Skantze, G., House, D., Edlund, J.: User responses to prosodic variation on fragmentary grounding utterances in dialogue. In: Proceedings Interspeech 2006, pp. 2002–2005. Pittsburgh, PA (2006)

[82] *SmarTrack* - a *SmarT* people *Track*er. Patent pending. Online at `http://tev.fbk.eu/smartrack/`

[83] Soldatos, J., Dimakis, N., Stamatis, K., Polymenakos, L.: A Breadboard Architecture for Pervasive Context-Aware Services in Smart Spaces: Middleware Components and Prototype Applications. Personal and Ubiquitous Computing Journal **11**(2), 193–212 (2007). URL `http://www.springerlink.com/content/j14821834364128w/`

[84] Stiefelhagen, R., Bernardin, K., Bowers, R., Garofolo, J., Mostefa, D., Soundararajan, P.: The CLEAR 2006 Evaluation. In: Multimodal Technologies for Perception of Humans, Proceedings of the First International CLEAR Evaluation Workshop, CLEAR 2006, no. 4122 in Springer LNCS, pp. 1–45. Southampton, UK (2006)

[85] Stiefelhagen, R., Bernardin, K., Bowers, R., Rose, R.T., Michel, M., Garofolo, J.: The CLEAR 2007 Evaluation. In: Multimodal Technologies for Perception of Humans, Proceedings of the International Evaluation Workshops CLEAR 2007 and RT 2007, *LNCS*, vol. 4625, pp. 3–34. Springer, Baltimore, MD, USA (2007)

[86] Stiefelhagen, R., Bernardin, K., Ekenel, H., McDonough, J., Nickel, K., Voit, M., Woelfel, M.: Audio-visual perception of a lecturer in a smart seminar room. Signal Processing - Special Issue on Multimodal Interfaces **86**(12) (2006)

[87] Stiefelhagen, R., Bowers, R., Fiscus, J. (eds.): Multimodal Technologies for Perception of Humans, Proceedings of the International Evaluation Workshops CLEAR 2007 and RT 2007, *LNCS*, vol. 4625. Springer, Baltimore, MD, USA (2007)

[88] Stiefelhagen, R., Garofolo, J. (eds.): Multimodal Technologies for Perception of Humans, First International Evaluation Workshop on Classification of Events, Activities and Relationships, CLEAR'06. No. 4122 in LNCS. Springer, Southampton, UK (2006)

[89] Sturm, J., van Herwijnen, O.H., Eyck, A., Terken, J.: Influencing social dynamics in meetings through a peripheral display. In: ICMI '07: Proceedings of the 9th international conference on Multimodal interfaces, pp. 263–270. ACM, New York, NY, USA (2007)

[90] Svanfeldt, G., Olszewski, D.: Perception experiment combining a parametric loudspeaker and a synthetic talking head. In: Proceedings of Interspeech, pp. 1721–1724 (2005)

[91] Tang, J.C.: Finding from observational studies of collaborative work. International Journal of Man-Machine Studies **34**(2), 143–160 (1991)

[92] Tyagi, A., Potamianos, G., Davis, J.W., Chu, S.M.: Fusion of multiple camera views for kernel-based 3D tracking. In: Proc. IEEE Works. Motion and Video Computing (WMVC). Austin, Texas (2007)

[93] VACE - Video Analysis and Content Extraction, `http://iris.usc.edu/Outlines/vace/vace.html`

[94] Waibel, A., Stiefelhagen, R. (eds.): Computers in the Human Interaction Loop. Human-Computer Interaction. Springer (2009)

[95] Wallers, Å., Edlund, J., Skantze, G.: The effects of prosodic features on the interpretation of synthesised backchannels. In: E. André, L. Dybkjaer, W. Minker, H. Neumann, M. Weber (eds.) Proceedings of Perception and Interactive Technologies, pp. 183–187. Springer, Kloster Irsee, Germany (2006)

[96] Wojek, C., Nickel, K., Stiefelhagen, R.: Activity recognition and room level tracking in an office environment. In: IEEE Int. Conference on Multisensor Fusion and Integration for Intelligent Systems. Heidelberg, Germany (2006)

[97] Wölfel, M.: Warped-twice minimum variance distortionless response spectral estimation. In: Proc. EUSIPCO (2006)

[98] Wölfel, M., McDonough, J.: Combining multi-source far distance speech recognition strategies: Beamforming, blind channel and confusion network combination. In: Proc. Interspeech (2005)

[99] Zancanaro, M., Lepri, B., Pianesi, F.: Automatic detection of group functional roles in face to face interactions. In: Proceedings of the International Conference of Multimodal Interfaces ICMI-06 (2006)

[100] Zhang, Z., Potamianos, G., Senior, A.W., Huang, T.S.: Joint face and head tracking inside multi-camera smart rooms. Signal, Image and Video Processing pp. 163–178 (2007)

[101] Zhu, X., Barras, C., Lamel, L., Gauvain, J.L.: Speaker diarization: from Broadcast News to lectures. In: Machine Learning for Multimodal Interaction, vol. 4299, pp. 396–406. LNCS (2006)

[102] Zhu, X., Barras, C., Lamel, L., Gauvain, J.L.: Multi-stage speaker diarization for conference and lecture meetings. In: Multimodal Technologies for Perception of Humans, Proceedings of the International Evaluation Workshops CLEAR 2007 and RT 2007, *LNCS*, vol. 4625, pp. 533–542. Springer, Baltimore, MD, USA (2007)