Tracking Identities and Attention in Smart Environments - Contributions and Progress in the CHIL Project

Rainer Stiefelhagen, Keni Bernardin, Hazım K. Ekenel, Michael Voit*
Interactive Systems Labs
Universität Karlsruhe (TH), Germany
{stiefel|keni|ekenel|voit}@ira.uka.de

Abstract

To provide intelligent services in a smart environments it is necessary to acquire information about the room, the people in it and their interactions. This includes, for example, the number of people, their identities, locations, postures, body and head orientations, among others. This paper gives an overview of the perceptual technology evaluations that were conducted in the CHIL project, specifically those held in the CLEAR 2006 and 2007 evaluation workshops. We then summarize the main achievements and lessons learnt in the project in the areas of person tracking, person identification and head pose estimation, all of which are critical perception components in order to build perceptive smart environments.

1. Introduction

In recent years there has been much research effort spent on building smart perceptive environments, such as smart living rooms [7], smart lecture and meeting rooms [1, 3, 27, 36, 44] or smart houses [29, 51]. Such smart spaces are usually equipped with a variety of sensors which allow for automatic acquistion of information about the users and their activities. The challenge then is to build smart spaces which support humans during their activities inside them without obliging them to concentrate on operating complicated technical devices.

In the framework of the project CHIL, Computers in the Human Interaction Loop [13, 55] , a team of fifteen international academic and industrial research labs have collaborated on developing services that aim at proactively assisting people during their daily activities and, in particular, during their interaction with others. Some prototypical services that were developed in the project include a perceptive collaborative workspace [22, 41], various services facilitating collaboration in meeting and lecture rooms [45], and a perceptive virtual office assistance system [16].

To provide such intelligent services in a smart environment it is necessary to acquire information about the room, the people in it and their interactions. This includes, for example, the number of people, their identities, locations, postures, body and head orientations, and the words they utter, among others. In the CHIL project considerable effort was thus spent in order to build novel techniques to sense who is doing what, where, with whom and how in these environments.

An important aspect for the development of such perception components is the availabilty of realistic training and evaluation data. In CHIL large audio-visual copora have been collected and annotated to facilitate the development and evaluation of the envisaged perception components. Furthermore, CHIL has organized a series of technology evaluations. These were first conducted internally in the project, before it was decided to completely open them up, and to create an international evaluation workshop called CLEAR - "Classification of Events, Activities and Relationships" [14, 14, 48, 49]. This evaluation workshop was successfully held with broad international participation in 2006 and 2007.

This paper gives an overview of the perceptual technology evaluations that were conducted in the CHIL project, specifically those held in the CLEAR 2006 and 2007 evaluation workshops. We then summarize the main achievements and lessons learnt in the project in the areas of person tracking, person identification and head pose estimation, all of which are critical perception components in order to build perceptive smart environments.

2. Perceptual Component Evaluation and Data Collection

Systematic evaluation is essential to drive 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 can be measured objectively and different approaches compared and assessed. The technology evaluations were followed by evaluation workshops, during which the systems and obtained results were discussed in detail. The first technology evaluations were held in June 2004, during the first year of the project, in order to establish baseline results of the available technologies on the real-life lecture scenario, which we wanted to address. Here, 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

*M. Voit is now with the Fraunhofer Institut for Information- and Data Processing - IITB, Karlsruhe, Germany

978-1-4244-2154-1/08/$25.00

(a) ***AIT*** **(b)** ***UKA*** **(c)** ***FBK-IRST*** **(d)** ***IBM*** **(e)** ***UPC***

Figure 1. Scenes from the five smart rooms apparent in the 2007 CHIL Interactive Seminar database

event detection. In January 2005, the first formal evaluations were then conducted, which now also included multimodal evaluation tasks (multimodal tracking, multimodal identification). Also, these evaluations were made open to external participants.

Although many researchers, research labs and in particular a number of major research projects worldwide – including the European projects CHIL, Computers in the Human Interaction Loop [13], and AMI, "Augmented Multiparty Interaction" [1], as well as the US programs VACE, "Video Analysis and Content Extraction" [53], and CALO, "Cognitive Assistant that Learns and Organizes" [10] – are working on technologies to analyze people, their activities, and their interaction, common benchmarks for such technologies are usually not available. Most researchers and research projects use their own data sets, annotations, task definitions, metrics and evaluation procedures. As a consequence, comparing the advantages of research algorithms and systems is virtually impossible. Furthermore, this 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 and to create an open international evaluation workshop called CLEAR - "CLassification of Events, Activities, and Relationships" [14, 48, 49] -, in which part of the CHIL technology evaluations took place. This was possible by joining forces with NIST (the U.S. National Institute of Standards and Technology), which organizes the technology evaluation of the US Video Analysis Content Extraction (VACE) program [53]. The goal of CLEAR is to provide an until now missing common international evaluation forum and framework for such 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 first CLEAR evaluation was conducted in spring 2006 and was concluded by a two day evaluation workshop in the UK in April 2006. Apart from CHIL, CLEAR was also supported by NIST and the VACE program. Also, CLEAR 2006 was organized in cooperation with the NIST Rich Transcription (RT) Meeting Recognition evaluation [42], which focused more on the evaluation of content-related technologies, such as speech and video text recognition. CLEAR and RT shared some of their evaluation data sets, so that the speaker-localization results generated for CLEAR could be used for the far-field speech-to-text task in RT 2006, for example. This was facilitated through the harmonization of the 2006 CLEAR and RT evaluation deadlines.

This CLEAR 2006 evaluation was a big success. Overall, around sixty people from 16 different institutions participated in the workshop, and nine major evaluation tasks, including more than 20 subtasks were evaluated [14]. Based on the success of CLEAR 2006, CLEAR 2007 took place in May 2007, in Baltimore, USA. CLEAR 2007 was again successfully held in conjunction and collocated with the NIST RT 2007 evaluations. In addition to the support from the CHIL project and NIST, CLEAR 2007 was also supported by the European project AMI [1].

Tasks	Evaluation
Person tracking (2D,3D,A,V,AV)	CHIL, CLEAR'06,'07
Person identification (A,V,AV)	CHIL, CLEAR'06,'07
Head pose estimation	CHIL, CLEAR'06,'07
Acoustic event detection	CHIL, CLEAR'06,'07
Speech reco. (CT,FF,TT mics)	RT'05,'06,'07
Speech activity detection	CHIL, RT'05,'06,'07
Speaker diarization	RT'07
Question Answering	CHIL, CLEF'07

Table 1. Overview of the tasks and evaluation workshops, which used part of the CHIL corpus

An important aspect of these technology evaluations, was to use real-life data covering the application scenarios that we wanted to address in the project. We therefore collected a number of seminars and meetings in different smart rooms that were all equipped with a range of cameras and microphones. The first collected data has been used in the CHIL consortium internal technology evaluations (June 2004 and January 2005). New corpora collected in 2005 and 2006 have been the main data set in the CLEAR evaluations during the springs of 2006 and 2007. Furthermore, the audio modality of the CHIL corpus has been part of the speech technology evaluations within the RT evaluations organized by NIST during the springs of 2005, 2006 and 2007, and was used in a pilot track in the CrossLanguage Evaluation Forum CLEF 2007 [15]. Utilization of the CHIL data set in these high-profile evaluation activities demonstrates the state-of-the-art nature of the corpus, and its contribution to advanced perception technology development, further enhanced by the numerous papers resulting from these evaluations. This data set is publicly available to the community through the language resources catalog [21] of the European Language Resources Association (ELRA).

Figure 1 shows some scenes from the 2007 CHIL Interactive Seminar database. Table 1 shows an overview of perceptual technology evaluation tasks that were conducted in the CHIL project. More details on the datasets can be found in [37].

In the next sections, contributions, progress and lessons on the person tracking, person identification and head pose estimation tasks will be discussed. While these are only a part of the perception technologies investigated in the CHIL project, and do not cover high-level behavior analysis, such as gestures, turn-taking, etc., they constitute essential building blocks and where thus extensively evaluated.

3. Person Tracking

The research on person tracking in CHIL focused mostly on the tracking of persons inside smart rooms. The goal of tracking was to determine, for all points in time, the scene coordinates of room occupants with respect to a given room coordinate frame. This is in contrast to much of the visual tracking research, where only image coordinates are estimated, and to most of the acoustic or multimodal tracking research, where only relative azimuths are determined. Whereas at the start of the project, only the tracking of a single room occupant, the main speaker in a seminar, was aimed at, from the second year already, attention was shifted to the simultaneous tracking of all room occupants on the visual side, and the consecutive tracking of alternating speakers on the acoustic side.

The smart-room sensors used in tracking include a minimum of four fixed cameras installed in the room corners, with highly overlapping fields of view, one wide angle camera fixed under the ceiling overlooking the entire room, at least 3 T-shaped 4-channel microphone arrays and one Mark III 64-channel microphone array on the room walls (see [37]). While the availability of a high number of sensors may be seen as an advantage, offering a great deal of redundancy in the captured information that could be exploited by algorithms, it must also be seen as a challenge, requiring to solve problems such as data synchronization, transfer and distributed processing, spatio-temporal fusion, modality fusion, etc. From the audio point of view, it is also worth mentioning that CHIL represents one of the first attempts to perform and systematically evaluate acoustic tracking with a distributed microphone network (DMN), for which simplifying assumptions made with linear or compact microphone arrays, such as the near field assumption, do not hold. In this sense, the CHIL-room sensor setup itself created new and interesting research problems that required the development of original tracking and data fusion techniques.

Another factor making the tracking tasks particularly challenging is the nature of the application scenario. Algorithms have to automatically adapt to data coming from up to five smart rooms with very different characteristics, such as room dimensions, illumination, chromatic and acoustic signature, average person-sensor distances, camera coverage, furnishing, reverberation properties, sources of noise or occlusion, etc. The scenario is that of real seminars and meetings with sometimes large numbers of occupants free to sit around tables or on rows of chairs, stand or move around, occasionally enter or leave, laugh, interrupt or occlude each other, etc, making it hard to make any assumptions but the most general ones about their behavior. This requires elaborate methods for combined (person or speech) detection and tracking, model adaptation, data association, feature and sensor selection, and so forth.

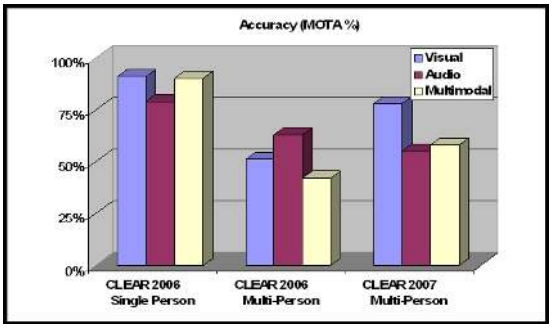

Figure 2. Best system performances for the CLEAR 2006 and 2007 3D person tracking evaluations. The MOTA score measures a tracker's ability to correctly estimate the number of objects and their rough trajectories.

The developed tracking systems have been extensively tested in increasingly challenging CLEAR evaluations, using the CHIL seminar and meeting database. 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 one site, also became increasingly difficult and varied, with 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 cross-talk. 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 CLEAR evaluations.

Concerning the visual tracking algorithms, two main approaches have been followed by the various developed 3D tracking systems: First, a model-based approach where a 3D model of the tracked object is maintained by rendering it onto the camera views and searching for supporting evidence in each view to update its parameters [2, 6, 11, 31, 33, 40]. Second, a data-driven approach where 2D trackers operate independently on the separate camera views and the 2D tracks belonging to a same target are collected into a 3D track [28, 52, 56].

In terms of performance, the model-based approach generally provides for better accuracy (MOTA) but less precision (MOTP) than the data-driven one[1]. One of its advantages is that rendering can be implemented such that it mimics the real image formation process, including perspective transformation and scaling, lens distortion, etc. In the context of multi-body tracking this is particularly advantageous, since occlusions can be handled at the rendering level by looking for supporting evidence only in the image parts where the different models are visible [31]. The difficulty here lies in the automatic initialization and update of

[1] Details on the used metrics MOTA and MOTP can be found in [48]

the person models.

The handling of occlusions and the association of tracks are the main drawbacks of the data-driven approach. For the approaches presented in CLEAR, the work-around was to detect and track faces instead of whole bodies, considerably reducing the problems of split or merged tracks, and making position estimations more accurate.

On the acoustic side, 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 [2, 9]; particle filter approaches, which approximate the belief on speaker positions by a set of samples and measure the agreement of the observed acoustic signals (their correlation value) given each sample position hypothesis [39]; approaches that feed computed time delays of arrival (TDOAs) between microphone pairs directly as observations to a Kalman or other probabilistic tracker [24, 30, 50]. The best performing system was based on a Joint Probabilistic Data Association Filter (JPDAF), which keeps track of a number of sound sources, including noise sources, resolving data associations and position updates jointly for all tracks. This allows to better handle rapid speaker switches or occasional sources of noise, as these do not disturb the track of the main target. Whereas in earlier systems developed in CHIL, Speech Activity Detection (SAD) was often performed and evaluated separately, toward the end of the project, more and more approaches featured built-in speech detection techniques [24, 43].

In the field of multimodal tracking, finally, while most initial systems combined monomodal tracker outputs in a post-processing manner, later approaches incorporated the audio and visual streams at the feature level. These were notably particle filter based trackers [6, 8, 39], as these allow for a flexible integration of features across sensors and modalities. The underlying idea is that early fusion of the data provides more accurate results, as it eliminates the effects of wrongful hard decisions made by monomodal trackers. An important point that became clear during the last 2 years of the project, however, is that multimodal fusion does not necessarily lead to higher accuracies, contrary to what may be expected in general. It is in fact highly dependent on the task and data at hand, and requires a careful balance in availability and quality of the modalities if the advantages of fusion are to be measured. For the 2007 evaluations, the tracking task required to find the active speaker using acoustic features and keep tracking him/her audio-visually while using purely visual cues during periods of silence. This prevented trackers from performing well by relying only on one modality and forced the development of truly audio-visual systems.

4. Person Identification

Person identification is one of the most important perceptual technologies for smart environments. Using the identity information, the space can be personalized according to the person's preferences. In order to perform identification naturally and implicitly, the sensors distributed in the environment should continuously monitor the space, and capture audiovisual data of the persons unobstrusively when they appear. That is, the person identification system is required to operate in the background without getting attention and cooperation from the persons. This brings many challenges both for audio-based and image sequence-based identification: Depending on the location of the person and his/her distance to the sensors, the received signals vary. Reverberations, large attenuations and background noise degrade audio signal quality. Large variations in illumination, face resolution and head pose are difficulties inherent in the visual modality. Moreover, persons can have different facial expression, hair styles, make-ups, etc. Nevertheless, identification can be still done robustly by utilizing multiple sensors available in the environment.

Concerning the task of (acoustic) *speaker identification*, features used in the various systems include Mel Frequency Cepstral Coefficients (MFCC) [5, 18, 23, 47], perceptually-weighted Linear Prediction coefficients (PLP) [47] or frequency filtering (FF) [34, 35]. Some systems also perform post-processing of the features, such as cepstral mean normalization and feature warping [5, 18]. All systems construct speaker models using Gaussian Mixture Models (GMM). Model training is done either using the available training speech for each speaker independently, or by maximum a posteriori (MAP) adaptation of a universal background model (UBM) [5]. The classification is done using a maximum-likelihood classifier.

The systems utilize the multiple audio streams, either by improving the quality of the signal prior to constructing or testing models, performing some sort of beamforming [5, 34], or doing post-decision fusion of the matching scores derived from each stand-alone microphone [34]. Some of the systems use speech activity detector to extract speech segments from the audio signal.

Concerning the task of *face identification*, systems utilized image sequences provided by the multiple cameras in the room. In the initial project-internal evaluations, faces were automatically detected and identified [19]. In the following CLEAR evaluations, face bounding boxes and eye center positions at every 200 ms were provided. Some systems use only the annotated faces [19], while others do interpolation between the labels and use all the faces available in the video data [47]. The faces are aligned, either using the eye centers [19,46] or the face bounding box [18,47]. To obtain robustness against registration errors, some systems generate additional aligned images, either by modifying the eye position labels [19], or by modifying the face bounding box labels [18].

One approach performs local appearance-based face recognition using discrete cosine transform (DCT) [18, 20]. PCA-based approaches were also tested [19, 46, 47]. A modified version of the weighted Euclidean is employed as the distance metric [47]. Linear discriminant analysis (LDA) has been also tested in the past [19, 46], without much success, as the face images were found to be linearly non-separable. For this reason sub-class LDA has been utilized [47]. Gaussian modeling of intrapersonal variations is also used to evaluate the probability that the difference of a gallery face from a probe face is indeed intrapersonal [47]. All the systems use a nearest neighbor classifier, although the distance from class centers has been used by some systems in the past [19, 46].

Decisions obtained from all camera views are combined using the weighted sum rule [18, 19, 46, 47]. The weights

Figure 3. Example screenshot of a face identification system running on Interactive Seminar data (Image taken from [18]). In the bottom right corner the extracted face is shown.

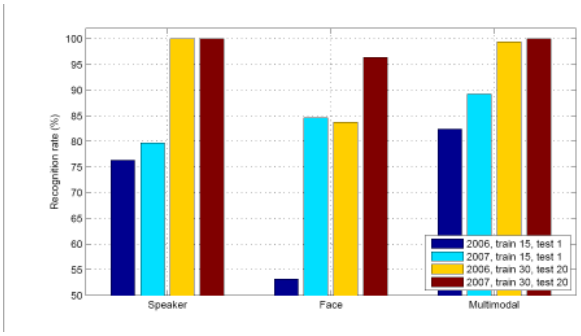

Figure 4. Performance evolution of the person identification systems in the CLEAR 2006 and 2007 evaluations (from [55]). Two out of the eight conditions are shown per evaluation; the shortest training and testing (15 and 1 seconds respectively) and the longest training and testing (30 and 20 seconds respectively)

are determined according to the separation of the best two matches. Some systems also combine different feature extraction methods [46, 47].

Concerning the task of *multimodal person identification*, decision fusion was performed to combine individual modalities. The weighted sum rule was used, with weights determined either proportional to the individual modalities' performance or to the separation of the best two matches at each modality. To perform the latter approach, three different methods were utilized. These are, the ratio between the closest and second closest matches [46,47], a histogram equalization of the monomodal confidence scores [17] and a non-parametric model of the distribution of the correct matches with respect to the confidence differences between the best two matches [18, 20].

Given the development and evaluation of the various speaker recognition systems in the project, the following could be concluded: the use of beamforming techniques to produce a single audio-signal from multiple microphone channels performed worse than post decision fusion approaches. Concerning the features, it turned out that stand-alone PLP features, or those obtained by the combination of PLP and MFCC into a single feature vector using PCA are better than stand-alone MFCC features. Finally, using a UBM with MAP adaption performed better than direct estimation of speaker models.

The following observations have been derived for face recognition: Selecting just the frontal faces (using the provided labels), instead of using all the available samples, is detrimental to performance. The best system employs local appearance using DCT. Experiments using the same face extraction and normalization methods lead to the following ranking of feature extraction methods: Intrapersonal modeling (Bayesian) $>$ Subclass LDA $>$ PCA $>$ LDA; LDA/Kernel PCA (KPCA) combination $>$ KPCA $>$ LDA $>$ PCA.

Multimodal systems have provided improved correct classification rates over the single modality-based systems in all training-testing conditions, which indicates that face images and speech signal provide complementary cues for person identification.

Through the project the systems' performance improved significantly. In CLEAR 2006 evaluations, there were 26 subjects in the database. The segments that contain speech signal have been selected without paying attention to the facial image quality in the video. On the other hand in CLEAR 2007, the database size has increased to 28 subjects and during the segment selection, facial image quality has been also taken into account. The correct recognition rates of the best performing systems in CLEAR 2006 and 2007 evaluations are shown in Figure 4.

5. Estimation of Head Pose and Focus of Attention

Observing and understanding human interaction was an important goal in the CHIL project. An important cue for the analysis of human interaction are the dynamics of peoples' head orientations. Head orientation can for example be used to determine people's direction of attention, whether they have looked onto a specific object or other persons in the surrounding, and is thus useful to better understand human interaction. It can also be used to build better (attentive) user interfaces.

In the CHIL project, we have been working on head pose estimation in single-camera setups - which can for example be useful if someone sits in front of a display or specific devices -, as well as multi-camera scenarios, such as they were available in our smart rooms. An advantage of the multi-camera scenario is that it allows for a much greater working area, since people can be observed in almost the whole room, and that the information gathered from the various cameras can be fused in order to obtain more robust head orientation estimates. Estimating head poses in smart rooms is, however, still a difficult task: first, the heads of multiple people need to be correctly tracked in order to find correct correspondences in the different camera views. Second, depending on the user's position in the room, his or her head size may greatly vary in the different camera views. Finally, a challenge is to best fuse the obtained head pose estimates.

Concerning the work on monocular datasets for head pose estimation, two different approaches using neural networks to estimate head pose were investigated in the project [26, 54]. One system outputs a direct estimation of the observed head orientation in either horizontal or vertical direction [54], the other uses further refinement steps by means of using the coarse estimate from the network to search among refining pose graphs constructed around gaussian receptive fields [26]. Both systems were evaluated in CLEAR

2006 on the monocular *Pointing04* dataset [25]. Here, both systems performed with state-of-the art results, and the best system achieved an error of 10.1° for pan and 12.6° for tilt. In CLEAR 2007, the system presented in [54] was also evaluated on the AMI meeting corpus [4] and lead to good result (9.5° for pan and 13.8° for tilt).

Given the multi-camera sensor setup of our smart-rooms, a logical step is to build systems that make us of all the camera views to estimated people's head orientations. This has the advantage that a much bigger working area can be covered, and also multiple head pose estimates can be fused to get more robust results. In CHIL, thus a number of multi-camera head pose estimation systems were developed [12, 32,54]. Both signal level [12] and decision level [54] fusion schemes were investigated.

In one system, the output of neural-network based head pose estimations in each camera view, were fused in a Bayesian filter, in order to compute a combined head pose estimate [54]. The neural network was trained to output the likelihoods over a range of relative angles, related to the corresponding camera's line of sight. The Bayes filter, in which each state describes one possible head rotation angle in the room coordinate system, the final a-posteriori density was computed by taking into account the respective angle's previous posterior likelihood, its probability to change into a given new state (temporal smoothing) as well as the mean of all cameras' estimates for the individual angles in room-coordinates. This fusion scheme of course requires processing power to necessarily run multiple classifiers (one neural network for each camera) instead of only a single one, that uses a joint feature vector, gathered from all views. However, it has the advantage, that more data can be used to train the neural networks (namely the images from all cameras), and that the system can be easily extended to incorporate more camera views, since no additional training is required.

In addition, signal-level fusion techniques for multi-camera head pose estimation were investigated [12]. Here, synthetic reconstructions of different head poses are parsed through by comparing these templates with the new, currently achieved query vector. Corresponding skin patches from all camera views are mapped onto a 3D ellipsoid, approximating the observed head's shape. The intuition behind this approach is to combine all views into a reconstructed depiction of skin colour, as distributed on the ellipsoid as the currently observed head orientation allows to capture over all distinct camera views. A planar representation of this head approximation (both in shape and colour) then results in a final feature vector that can be interpreted as a saliency map, used for assigning it to the best matching pose template in a stored database.

Finally, an approach for joint tracking and pose estimation was investigated [32]. Here, both 2D body position and velocity, as well as horizontal and vertical head orientation are jointly estimated in a Bayesian approach. In every frame step, a hypothesized body position is updated along its corresponding velocity component according to the time elapsed between these two frames. For accounting uncertainty and ambiguity, a particle filter allows to propagate numerous hypotheses (particles), and low-dimensional shape and appearance models (color histgrams) for different body parts are used to compute each hypothesis' likelihood. Such an integrated approach might provide for faster sytems, as well as better generalisation in the face of misaligned head regions.

Figure 5. Example camera view with automatically estimated head orientations (from [38]).

In the CLEAR 2007 multi-view head pose estimation task, the best performance was achieved with the system using neural networks and a late integration approach using a Bayes filter. It provided mean errors as low as 8.5° for pan and 12.5° for tilt

The developed systems for estimating head orientations can be used to determine the user's focus of attention, i.e. to determine at whom or what someone was looking. In addition to the head pose estimation systems, which were evaluated in the CLEAR workshops, we have also worked on approaches to estimate the persons' focus of attention. In the work presented in [38], a Bayesian framework was used to map observed head orientations on the most likely targets (the other participants). To model the individual head orientation styles of people and to account for different seating of people, a clustering technique was used. Overall, the correct focus target could be estimated in approximately 70% of all video frames in a meeting recorded in a smart room (see Figure 5).

6. Conclusion

From 2004 to 2007 a team of fifteen research labs has collaborated in the CHIL - Computers in the Human Interaction Loop - project in order to build perceptive proactive services that can support humans during their activities and interaction with others. To this end a number of perceptual components have been developed and thoroughly evaluated. The project's perceptual technology evaluations have led to the creation of the international CLEAR evaluation workshop, which was successfully held in 2006 and 2007, and has brought together a broad number of international research groups working on systems for audio-visual perception of people. An important contribution of the CLEAR evaluations, is the fact that they provided an international forum for the discussion and harmonization of related evaluation tasks, including the definition of procedures, metrics and guidelines for the collection and annotation of necessary multimodal datasets. Also, significant multimedia datasets and evaluation benchmarks have been made available to the reseach community. This paper gave an overview of the perceptual technology evaluations conducted in CHIL and the evaluations conducted within CLEAR. We also summarized the project's work

and progress on some of the most crucial perception components, namely person tracking, person identification and estimation of head pose. Further details on investigated higher-level perception and analysis of human interaction conducted in the project can be found in [55] or in the publications available on [13].

Acknowledgements

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] AMI: Augmented Multi-party Interaction. http://www.amiproject.org.

[2] A. Abad, C. Canton-Ferrer, C. Segura, J. L. Landabaso, D. Macho, J. R. Casas, J. Hernando, M. Pardas, and C. Nadeu. Upc audio, video and multimodal person tracking systems in the clear evaluation campaign. In *Multimodal Technologies for Perception of Humans, Proceedings of the First International CLEAR Evaluation Workshop*, Southampton, UK, 2007. Springer LNCS 4122.

[3] G. D. Abowd, C. Atkeson, A. Feinstein, C. Hmelo, R. Kooper, S. Long, N. Sawhney, and M. Tani. Teaching and learning as multimedia authoring: The classroom 2000 project. In *Proceedings of the ACM Multimedia'96 Conference*, pages 187–198, November 1996.

[4] S. O. Ba and J. M. Odobez. Evaluation of head pose tracking algorithm in indoor environments. In *International Conference on Multimedia & Expo, ICME 2005*, Amsterdam, Netherlands, 2005.

[5] C. Barras, X. Zhu, C.-C. Leung, J.-L. Gauvain, and L. Lamel. Acoustic Speaker Identification: The LIMSI CLEAR'07 System. In *Proceedings of the Second International CLEAR Evaluation Workshop - CLEAR'07*, Baltimore, May 2007. Springer LNCS 4625.

[6] K. Bernardin, T. Gehrig, and R. Stiefelhagen. Multi-Level Particle Filter Fusion of Features and Cues for Audio-Visual Person Tracking. In *Proceedings of the Second International CLEAR Evaluation Workshop*, Baltimore, MD, USA, 2007. Springer LNCS 4625.

[7] B. Brumitt, B. Meyers, J. Krumm, A. Kern, and S. Shafer. Easyliving: Technologies for intelligent environments. In *Handheld and Ubiquitous Computing*, September 2000.

[8] R. Brunelli, A. Brutti, P. Chippendale, O. Lanz, M. Omologo, P. Svaizer, and F. Tobia. A generative approach to audio-visual person tracking. In *Multimodal Technologies for Perception of Humans, Proceedings of the First International CLEAR Evaluation Workshop*, pages 55–68, Southampton, UK, 2007. Springer LNCS 4122.

[9] A. Brutti. A person tracking system for chil meetings. In *Multimodal Technologies for Perception of Humans, Proceedings of the Second International CLEAR Evaluation Workshop*, Baltimore, MD, USA, 2007. Springer LNCS 4625.

[10] CALO - Cognitive Agent that Learns and Organizes, http://caloproject.sri.com/.

[11] C. Canton-Ferrer, J. Salvador, J. Casas, and M.Pardas. Multi-person tracking strategies based on voxel analysis. In *Multimodal Technologies for Perception of Humans, Proceedings of the Second International CLEAR Evaluation Workshop*, Baltimore, MD, USA, 2007. Springer LNCS 4625.

[12] C. Canton-Ferrer, C. Segura, J. R. Casas, M. Pards, and J. Hernando. Audiovisual Head Orientation Estimation with Particle Filters in Multisensor Scenarios. *EURASIP Journal on Advances in Signal Processing*, 2007.

[13] CHIL - Computers In the Human Interaction Loop, http://chil.server.de.

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

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

[16] M. Danninger and R. Stiefelhagen. A perceptive office assistant system. In *Proceedings of the ACM International Conference on Multimedia*, Vancouver, Canada, 2008.

[17] P. Ejarque, A. Garde, J. Anguita, and J. Hernando. On the use of genuine-impostor statistical information for score fusion in multimodal biometrics. *Annals of Telecommunications, Special Issue on Multimodal Biometrics*, 62(1-2):109–129, Apr. 2007.

[18] H. K. Ekenel, Q. Jin, and M. Fischer. Isl person identification systems in clear 2007. In *Multimodal Technologies for Perception of Humans, Proceedings of the Second International CLEAR Evaluation Workshop*, Baltimore, May 2007. Springer LNCS 4625.

[19] H. K. Ekenel and A. Pnevmatikakis. Video-based face recognition evaluation in the chil project - run 1. In *Proceedings of the International Conference on Face and Gesture Recognition*, pages 85–90, 2006.

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

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

[22] V. Falcon, C. Leonardi, E. Not, F. Pianesi, and M. Zancanaro. Observing multimodal behaviour to support group dynamics. In *In Workshop on User-centred Design and Evaluation of Services for Human-Human Communication and Collaboration. In conjunction with ICMI'05*, Trento, Italy, October 2005.

[23] J.-L. Gauvain and C. Lee. Maximum a Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains. *IEEE Trans. on Speech and Audio Processing*, 2(2):291–298, Apr. 1994.

[24] T. Gehrig and J. McDonough. Tracking multiple speakers with probabilistic data association filters. In *Multimodal Technologies for Perception of Humans, Proceedings of the First International CLEAR Evaluation Workshop*, Southampton, UK, 2007. Springer LNCS 4122.

[25] N. Gourier, D. Hall, and J. L. Crowley. Estimating face orientation from robust detection of salient facial features. In *Proceedings of Pointing 2004, ICPR, International Workshop on Visual Observation of Deictic Gestures*, Cambridge, UK, 2004.

[26] N. Gourier, J. Maisonasse, D. Hall, and J. L. Crowley. Head Pose Estimation on Low Resolution Images. In *Multimodal Technologies for Perception of Humans, Proceedings of the First International CLEAR Evaluation Workshop*. Springer LNCS 4122, April 2006.

[27] A. Janin, J. Ang, S. Bhagat, R. Dhillon, J. Edwards, J. Macias-Guarasa, N. Morgan, B. Peskin, E. Shriberg, A. Stolcke, C. Wooters, and B. Wrede. The icsi meeting project: Resources and research. In *NIST 2004 Meeting Recognition Workshop*, Montreal, May 2004. AMI-06.

[28] N. Katsarakis, F. Talantzis, A. Pnevmatikakis, and L. Polymenakos. The ait 3d audio / visual person tracker for clear 2007. In *Multimodal Technologies for Perception of Humans, Proceedings of the Second International CLEAR*

Evaluation Workshop, Baltimore, MD, USA, 2007. Springer LNCS 4625.

[29] C. D. Kidd, R. J. Orr, G. D. Abowd, C. G. Atkeson, I. A. Essa, B. MacIntyre, E. Mynatt, T. E. Starner, and W. Newstetter. The aware home: A living laboratory for ubiquitous computing research. In *International Workshop on Cooperative Buildings - CoBuild'99.*, 1999.

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

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

[32] O. Lanz and R. Brunelli. Dynamic head location and pose from video. In *IEEE Conf. Multisensor Fusion and Integration*, 2006.

[33] O. Lanz, P. Chippendale, and R. Brunelli. An appearance-based particle filter for visual tracking in smart rooms. In *Multimodal Technologies for Perception of Humans, Proceedings of the Second International CLEAR Evaluation Workshop*, Baltimore, MD, USA, 2007. Springer LNCS 4625.

[34] J. Luque and J. Hernando. Robust speaker identification for meetings: Upc clear07 meeting room evaluation system. In *CLEAR'07 Evaluation Campaign and Workshop - Classification of Events, Activities and Relationships*, Baltimore, May 2007.

[35] J. Luque, R. Morros, A. Garde, J. Anguita, M. Farrús, D. Macho, F. Marqués, C. Martinez, V. Vilaplana, and J. Hernando. Audio, video and multimodal person identification in a smart room. In *CLEAR*, pages 258–269, 2006.

[36] I. McCowan, S. Bengio, D. Gatica-Perez, G. Lathoud, F. Monay, D. Moore, P. Wellner, and H. Bourlard. Modeling human interaction in meetings. In *ICASSP 2003*.

[37] D. Mostefa, N. Moreau, K. Choukri, G. Potamianos, S. M. Chu, A. Tyagi, J. R. Casas, J. Turmo, L. Christoforetti, F. Tobia, A. Pnevmatikakis, V. Mylonakis, F. Talantzis, S. Burger, R. Stiefelhagen, K. Bernardin, and C. Rochet. The chil audiovisual corpus for lecture and meeting analysis inside smart rooms. In *Language Resources and Evaluation*, number 41 in Springer, 2007.

[38] M.Voit and R.Stiefelhagen. Tracking Head Pose and Focus of Attention with Multiple Far-field Cameras . In *International Conference On Multimodal Interfaces - ICMI 2006*, Banff, Canada, November 2006.

[39] K. Nickel, T. Gehrig, H. K. Ekenel, J. McDonough, and R. Stiefelhagen. 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*, Southampton, UK, 2007. Springer LNCS 4122.

[40] K. Nickel, T. Gehrig, R. Stiefelhagen, and J. McDonough. A Joint Particle Filter for Audio-visual Speaker Tracking. In *Proceedings of the Seventh International Conference On Multimodal Interfaces - ICMI 2005*, pages 61–68. ACM Press, October 2005.

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

[42] The Rich Transcription 2006 Spring Meeting Recognition Evaluation Website: http://www.nist.gov/speech/tests/rt/rt2006/spring.

[43] C. Segura, A. Abad, C. Nadeu, and J. Hernando. Multispeaker localization and tracking in intelligent environments. In *Multimodal Technologies for Perception of Humans, Proceedings of the Second International CLEAR Evaluation Workshop*, Baltimore, MD, USA, 2007. Springer LNCS 4625.

[44] The NIST smart space project. http://www.nist.gov/smartspace/.

[45] J. Soldatos, N. Dimakis, K. Stamatis, and L. Polymenakos. 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, March 2007.

[46] A. Stergiou, A. Pnevmatikakis, and L. Polymenakos. A decision fusion system across time and classifiers for audio-visual person identification. In *CLEAR*, pages 223–232, 2006.

[47] A. Stergiou, A. Pnevmatikakis, and L. Polymenakos. The ait multimodal person identification system for clear 2007. In *CLEAR'07 Evaluation Campaign and Workshop - Classification of Events, Activities and Relationships*, Baltimore, May 2007.

[48] R. Stiefelhagen, K. Bernardin, R. Bowers, J. Garofolo, D. Mostefa, and P. Soundararajan. The CLEAR 2006 Evaluation. In R. Stiefelhagen and J. Garofolo, editors, *Multimodal Technologies for Perception of Humans, Proceedings of the first International CLEAR evaluation workshop, CLEAR 2006*, number 4122 in Springer Lecture Notes in Computer Science, pages 1–45, 2007.

[49] R. Stiefelhagen, K. Bernardin, R. Bowers, R. Rose, M. Michel, and J. Garofolo. The CLEAR 2007 Evaluation. In R. Stiefelhagen, R. Bowers, and J. Fiscus, editors, *Multimodal Technologies for Perception of Humans, Joint Proceedings of the CLEAR 2007 and RT 2007 Evaluation Workshops*, number 4625 in Springer Lecture Notes in Computer Science, pages 1–34, 2008.

[50] F. Talantzis, A. Constantinides, and L. Polymenakos. Estimation of direction of arrival using information theory. *IEEE Signal Processing Letters*, 12(8), 2006.

[51] E. M. Tapia, S. S. Intille, and K. Larson. Activity recognition in the home setting using simple and ubiquitous sensors. In *Proceedings of PERVASIVE 2004*, volume LNCS 3001, pages 158–175. Springer, 2004.

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

[53] VACE - Video Analysis and Content Extraction, https://control.nist.gov/dto/twiki/bin/view/Main/WebHome.

[54] M. Voit, K. Nickel, and R. Stiefelhagen. Head Pose Estimation in Single- and Multi-view Environments - Results on the CLEAR'07 Benchmarks. In *Multimodal Technologies for Perception of Humans, Proceedings of the Second International CLEAR Evaluation Workshop*. Springer LNCS 4625, 2007.

[55] A. Waibel and R. Stiefelhagen, editors. *Computers in the Human Interaction Loop*. Human-Computer Interaction Series. Springer, 2008. to appear.

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