

Person Tracking

Keni Bernardin¹, Rainer Stiefelhagen¹, Aristodemos Pnevmatikakis², Oswald Lanz³, Alessio Brutti³, Josep Casas⁴, Gerasimos Potamianos⁵

¹ Universität Karlsruhe (TH), ITI, Am Fasanengarten 5, 76131, Karlsruhe, Germany

² Athens Information Technology, Peania, Attiki, 19002, Greece

³ Foundation Bruno Kessler, Irst, Trento, Italy

⁴ Universitat Politecnica de Catalunya, Barcelona, Spain

⁵ IBM T.J. Watson Research Center, Yorktown Heights, New York, USA

One of the most basic building blocks for the understanding of human actions and interactions is the accurate detection and tracking of persons in the scene. In constrained scenarios involving at most one subject, or in situations where persons can be confined to a controlled monitoring space or required to wear markers, sensors or microphones, these tasks can be solved with relative ease. However, when accurate localization and tracking have to be performed in an unobtrusive or unaware fashion, using only distantly placed microphones and cameras, in a variety of natural and uncontrolled scenarios, the challenges posed are much greater. The problems faced by video analysis are those of poor or uneven illumination, low resolution, clutter or occlusion, unclean backgrounds, and that of multiple moving and uncooperative users which are not always easily distinguishable. The problems faced by acoustic techniques, which rely on arrays of distant microphones to detect and pinpoint the source of human speech, are those of ambient noise, cross talk or rapid speaker turns, reverberations, unpredictable or hardly identifiable non-human noises, silent persons, etc. The following section presents the efforts undertaken by the CHIL consortium in the visual, acoustic and audio-visual domains for the simultaneous and unobtrusive tracking of several users in close to real-life interaction scenarios. As contributions to this challenging and varied field are too numerous to be presented in detail, the discussion focuses on the main problems encountered and lessons learned, on higher level design strategies and trends that proved successful, and tries to highlight individual approaches of particular interest that show the advantages of combining multiple far-field sensory sources. A noteworthy point is that almost all investigated techniques are designed to allow for fully automatic online operation, making them usable in a wide range of applications requiring quick system reaction or feedback.

3.1 Goals, challenges and originality

During its 3,5 year period, the CHIL project has carried out a considerable amount of research on detection and tracking techniques powerful enough to accomplish its



Fig. 3.1. Example screenshot of a person tracking system running on CHIL data (image taken from [57]).

ambitious goal: The detailed, unobtrusive and unaware observation of humans engaging in natural interaction. Within the list of technological challenges addressed by the project in the “Who and Where” field, person tracking occupied a quite important position both because of the various subtasks aimed at, including acoustic, visual, and multimodal tracking, and because of the high number of partners involved in friendly competition (so-called co-opetition) to create the most reliable and performant system.

Research focused mostly on the tracking of persons inside smart rooms equipped with a number of audio-visual sensors. 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. In CHIL, the requirement, through all modalities, was to track the xy -coordinates of a person on the ground plane, with the aim of reaching precisions down to a few centimeters. The restriction to the plane is due in part to the fact that the goal is to track entire bodies, not for example heads or torsos, in part to the ease of annotation when creating the reference ground truth. Although the estimation of a person’s height was not deemed a high priority in technology evaluations, almost all developed acoustic trackers and most of the more advanced visual trackers did in fact perform this estimation in live systems. 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 (see [35] for more details on tracking tasks).

The CHIL-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 [252]). 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. One should note that the minimal required sensor setup offers only rough guidelines on the amount and placement of sensors, leaving open such parameters as the type of cameras or microphones used, camera resolution and framerate, color constancy across views, etc, while developed trackers are expected to operate on data from all CHIL-rooms alike. 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), while most of the research in this field is focused on linear or compact microphone arrays. A DMN scenario is characterized by completely different problematics than that of a compact array as, for example, the near field assumption does not hold between microphone pairs from different clusters. Moreover, in contrast to what is often implicitly done when exploiting linear arrays, in the CHIL scenario no assumption can be made on the relative position or distance of the speaker to the various sensors. 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 CHIL tracking tasks particularly challenging is the nature of the application scenario. As mentioned above, algorithms have to automatically adapt to data coming from up to five CHIL 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 (see [252] for a description of the CHIL seminar and meeting corpus). This prevents the application of conventional techniques such as visual tracking based on foreground blobs or generic color models, track initialization based on creation or deletion zones, temporal smoothing of sound source estimations, etc, and requires elaborate methods for combined (person or speech) detection and tracking, model adaptation, data association, feature and sensor selection, and so forth. The developed tracking systems have been extensively tested in a series of increasingly challenging evaluations, using the CHIL seminar and meeting database [336, 334]. These large scale, systematic evaluations, using realistic, multi-

modal data and allowing to objectively compare systems and measure progress, also constitute a noteworthy contribution to the research field.

3.2 Difficulties, lessons learned, contributions

In retrospect, one of the most important problems that needed to be solved in the person tracking field was the definition of a set of metrics for the tracking of multiple objects that could be used in offline evaluations. No consensus on standardized metrics existed in the research field so far. For the first set of CHIL tracking evaluations, only simple metrics for single object tracking, based on average spatial error, were used for the visual subtask. These metrics were also different from those used for the acoustic subtask, which made performance comparisons difficult. To remedy this, a set of concise and intuitive metrics, that would be general enough to be used in all subtasks, visual, acoustic and multimodal, were defined at the end of 2005. These 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 [36] for details). They were subsequently used for all following evaluations, notably the CLEAR [336, 334] evaluations, for a broad range of CHIL or VACE [356] related tasks including acoustic, visual and multimodal 3D person tracking, 2D visual person tracking, face detection and tracking, and vehicle tracking.

A related problem was that of creating accurate ground truth labels, for person positions and speech activity. Whereas 3D position information could be easily obtained from the synchronized camera views by marking head centroids in each view and triangulating, the annotation of active speakers was somewhat more problematic. This information was not discernable in the video streams and had to be extracted from transcriptions which were made separately on the audio channels. The evaluation criteria required to identify speech segments, portions of silence, of overlapping speech, or containing active noise sources, which is why transcriptions had to be made on far-field microphone channels. In the first set of evaluations, only single person tracking was attempted and the audio and video tasks were completely separated. The advent of multimodal and multi-person tracking tasks in the 2006 evaluations made it necessary to merge audio and video annotations and tools, which until then were developed independently. The process of defining valid annotation guidelines went on until the final year of the project. On the whole it can be said that the transcription of audio-visual data for tracking no doubt constitutes one of the main lessons learned by the consortium.

Another smaller but no less important achievement was the definition of common standards for the calibration of room cameras and the distribution of camera and microphone calibration and position information.

Finally, the synchronization of all sensory streams was a big challenge, both for online operation in live demonstrations and for offline processing of recorded data.

This included the synchronization of all cameras, the synchronization of all microphone arrays, table top or close talking microphones (these were not used in tracking), of the MarkIII array with respect to the microphone clusters, and finally of all microphones with respect to the cameras. The synchronization of all T-shaped microphone clusters could easily be achieved to the sample level (at 44KHz) by capturing all channels using a single 24-channel RME Hammerfall soundcard, a solution which was adopted unanimously by all partners. The requirements for synchronization of video frames from different camera views or of cameras with respect to the MarkIII and T-shaped microphone arrays were less stringent, so that simple solutions, such as NTP synchronization of the capturing machines' clocks, was sufficient. More precise and practical solutions worth mentioning were however implemented by some project partners. One example is the use of gen-locked cameras synchronized to the frame level, coupled with a dedicated MOTU board, delivering a common synchronization signal to all cameras and microphone clusters. Such a solution was applied with success for example by UPC [274], both for online operation and for flawless recording over extended lengths of time.

3.3 Progress, approaches, results, highlights

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, 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 3.2 shows the progress made in audio, video and multimodal tracking for the last official CLEAR evaluations [336, 334].

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:

1. A model-based approach: A 3D model of the tracked object is maintained by rendering it onto the camera views, searching for supporting evidence per view, and based on that, updating its parameters. Such systems were used for example as early as 2005 by UKA-ISL for particle filter based single person tracking [257] and 2006 by FBK-irst and UPC for particle filter and voxel based multi-person tracking [214, 8]. For 2007 all approaches were further refined [37, 213, 57].

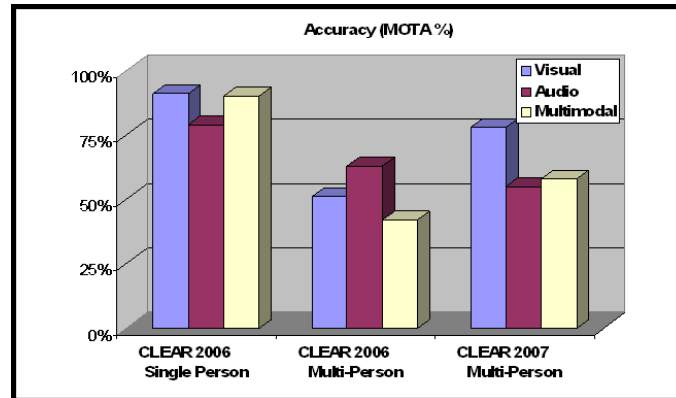


Fig. 3.2. Best system performances for the CLEAR 2006 and 2007 3D person tracking evaluations. The MOTA score measures a tracker's ability to correctly estimate the number of objects and their rough trajectories.

2. 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. Such systems were used throughout the course of the project by various partners. Notable are for example systems used in 2006 by IBM [390, 355] and in 2007 by AIT [189].

An important advantage of the model-based approach is that rendering can be implemented in a way that it 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 at the rendering level. This way, the update is done by looking for supporting evidence only in the image parts where the different models are visible, thus occlusions are handled in a systematic manner. Systems that do so are [177, 273], and the FBK-irst tracker [214]. The real novelty of the FBK-irst approach is that it can do the update in a probabilistic framework (particle filtering) more efficiently than previous approaches: the computational complexity grows at most quadratically with the number of targets, rather than exponentially. This makes the system run robustly in real time with 7 occluding targets. The disadvantage, of course, is that the person models have to be initialized and occasionally updated, which in some situations may be tricky (e.g. 5 people entering simultaneously into the monitored room).

The handling of occlusions and the association of (possibly split or merged) tracks are the main drawbacks of the data-driven approach. There is not enough information in the independent camera views to efficiently address them. The work-around in this case is to work with faces and or heads instead of bodies. In this case the initialization problems of the direct approach are diminished as initialization is handled by face detectors. Also, 2D face tracking can be complemented by detection to increase precision.

In terms of performance, the model-based approach generally provides for better accuracy (MOTA) but less precision (MOTP) than the data-driven one. In terms of speed, the data-driven approach is relatively slow, owing to the face detectors which constitute a bottleneck in the system.

On the acoustic side, most approaches built on the computation of the generalized cross correlation (GCC-PHAT) between microphone pairs of an array. Other methods relied on calculating the mutual information between the microphones [345] in order to integrate information about the reverberant characteristics of the environment. Tracking approaches can be roughly categorized as follows:

1. 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. Such an approach was used for example by UPC [8], in combination with Kalman filtering, or by FBK-irst [50], using simple temporal filtering of the peaks.
2. 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. Such a system was used for example by UKA-ISL [256] in the 2006 evaluations for visual, acoustic and multimodal tracking alike.
3. Approaches that feed computed time delays of arrival (TDOAs) between microphone pairs directly as observations to a Kalman or other probabilistic tracker. This approach was repeatedly followed for example by UKA-ISL, using an iterated extended Kalman filter [194] or a joint probabilistic data association filter (JPDAF) [147]. The advantage of the JPDAF approach is that it is designed to keep track of a number of sound sources, including noise sources, by maintaining a Kalman filter for each of them. The association of observations, in the form of peaks in correlation between microphone pairs, to tracks and the update of track positions and uncertainties is made jointly for all tracks using a probabilistic framework. This allows to better handle rapid speaker switches or occasional sources of noise, as these do not disturb the track of the main target. Another point of merit of the UKA-ISL tracker was the incorporation of automatic speech activity detection (SAD) into the tracking framework itself. In this case, the activity of a target is detected by analyzing its state error covariance matrix. Whereas in earlier CHIL systems, SAD was often performed and evaluated separately, towards the end of the project, more and more approaches featured built-in speech detection techniques [147, 313].

It should be mentioned at this point that a direct correlation between speech source localization performance and automatic far-field speech recognition performance could be shown [385]. In these experiments, which based on CHIL seminar data, the output of an audio-visual, particle filter based speaker tracking system was passed to a beamforming module. The beamformer filtered the data captured by the MarkIII linear array before passing it on for speech recognition. In the experiments, increasing the localization accuracy led to a measurable decrease in the word error rate (WER).

In the field of multimodal tracking, finally, efforts started in the second half of 2005. 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 [256, 37, 48], as these allow for the probabilistic, flexible and easy integration of a multitude 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 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. For the 2006 evaluations, the multimodal tracking task was split into two conditions. The first required to audio-visually track all participants of a meeting. The second required to audio-visually track only the active speakers in predetermined segments of clean speech. The result was that accuracies for the first condition were no better than using visual tracking alone, as the inclusion of acoustic features could only help improve accuracies for one participant out of many for select points in time. The achievable slight improvements did not fall into balance when computing global averages. Accuracies for the second condition, in turn, were no better than using acoustic tracking alone, as the limitation to time frames of clean speech produced an imbalance in favor of the acoustic modality which nearly eliminated the advantages of using visual features. Owing to this, the multimodal task was redesigned for the 2007 evaluations, with the inclusion of silence and noise segments in the tracking data, and the requirement to continuously track the last known speaker. Systems were thus obliged to select the right target person using acoustic features, track audio-visually, and bridge periods of silence using visual cues. With the so-defined task and data, the advantages of fusion become more clearly visible, as could be shown for example in [37].

On a final note, it is important to mention that many of the CHIL tracking systems were designed with realistic online operation in mind. In addition to being continuously tested in offline evaluations, the developed techniques were also integrated with great success in a number of realtime demonstrations, not only limited to the inside of CHIL smart rooms. These include the simultaneous determination of a speaker's position and orientation, using only microphone clusters [49], the robust visual tracking of multiple persons through occlusions (the "SmarTrack" system [214]), which was demonstrated outside of CHIL-rooms, notably at the IST conference in Helsinki in 2006, the simultaneous audio-visual tracking and identification of multiple persons [38] which required the online integration of several of tracking and ID components, the various Memory Jog Service services and their demonstrations (see Chapter 17, where accurate situation models of the CHIL-room were constructed based on tracker output).