

Audio-Visual Multi-Person Tracking and Identification for Smart Environments

Keni Bernardin
Universität Karlsruhe, ITI
Am Fasanengarten 5
76131, Karlsruhe, Germany
keni@ira.uka.de

Rainer Stiefelhagen
Universität Karlsruhe, ITI
Am Fasanengarten 5
76131, Karlsruhe, Germany
stiefel@ira.uka.de

ABSTRACT

This paper presents a novel system for the automatic and unobtrusive tracking and identification of multiple persons in an indoor environment. Information from several fixed cameras is fused in a particle filter framework to simultaneously track multiple occupants. A set of steerable fuzzy-controlled pan-tilt-zoom cameras serves to smoothly track persons of interest and opportunistically capture facial close-ups for face identification. In parallel, speech segmentation, sound source localization and speaker identification are performed using several far-field microphones and arrays. The information coming asynchronously and sporadically from several sources, such as track updates and spatio-temporally localized visual and acoustic identification cues, is fused at higher level to gradually refine the global scene model and increase the system's confidence in the set of recognized identities. The system has been trained on a small set of users' faces and/or voices and showed good performance in natural meeting scenarios at quickly acquiring their identities and complementing the ID information missing in single modalities.

Categories and Subject Descriptors

I.4.8 [Image Processing and Computer Vision]: Scene Analysis

General Terms

Algorithms, Design

1. INTRODUCTION AND RELATED WORK

In recent years, there has been a growing interest in intelligent systems for indoor scene analysis. Various research projects, such as the European CHIL or AMI projects [27, 29], aim at developing smart environments, at facilitating human-machine and human-human interaction, or at analyzing meeting or conference situations. To this effect, multimodal approaches that utilize a variety of far-field sen-

sors, video cameras and microphones, to gain rich scene information and achieve robust, unobtrusive and detailed scene understanding gain more and more popularity. Related research has focused, for example, on understanding the actions of individuals or the interactions between groups of persons in the room [1, 2], estimating their head pose, their body posture, analyzing their speech, to infer higher level knowledge, produce meeting summaries [4], offer useful proactive services, etc. An essential task on the way to realizing these goals is the localization and identification of humans in the scene.

While much research has been done on indoor tracking or on multimodal person identification in the past, work has only begun on building integrated, online systems that tackle all the related subtasks without severe restrictions on the scenario or application environment. Early systems, such as the one presented by Choudhury et al. [3] for multimodal person identification were either limited to a single user, or required users to stand closely in front of the identifying sensors. Recent work focuses more and more on the use of far-field sensors and on multiple user scenarios. Yang et al. [4], show a framework including color-based person and face tracking, speech detection and localization, and audio-visual ID. The integration is, however, still made conceptually on a frame level, assuming most cues for fusion are accessible at every point in time. This restricts the application to scenarios such as e.g. a small meeting around a table, where the data association problem is not so acute.

The problem when dealing with general, unconstrained environments involving several users is that information gained from passive sensors is either too coarse or noisy to allow correct identification, or too focused and narrow to keep track of all users or capture good identification features at the right time. This is why several approaches resort to a network of sensors, using wide-view fixed cameras or microphone arrays to keep track of users in the room, and pan-tilt-zoom (PTZ) cameras to actively seek high resolution images for identification.

Tsuruoka et al. [11] present a system that tracks a lecturer in a classroom using foreground segmentation on images from a fixed camera and uses a fuzzy control scheme to steer an active camera and deliver closeup views. It is however limited to a single user standing in front of a clean background. Peixoto et al. [12] use one fixed camera and a binocular active camera system, and implement a target selection strategy based on state transitions to deal with scenarios involving several users. Trivedi et al. [16] present the concept of a smart environment in which users are audio-

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

MM'07, September 23–28, 2007, Augsburg, Bavaria, Germany.
Copyright 2007 ACM 978-1-59593-701-8/07/0009 ...\$5.00.

visually tracked by a number of sensors, their faces identified in steerable camera images, and a number of user activities are recognized. Similarly, Hampapur et al. [13] perform 2D and 3D blob tracking on fixed camera images and locate head regions by analyzing the silhouettes of tracked persons. They discuss several strategies for target selection and active camera assignment to capture good facial views. They do not, however, address the problem of identifying users in the closeup views. Just like the previous approach, they also do not tackle the problems of data association or fusion of identification cues in time or across modalities. Stillman et al. [14] show a system for tracking, face detection and recognition of multiple users using a combination of fixed and PTZ cameras. In later work [15], they extend their approach with the inclusion of microphone arrays for source localization, and show a framework, using a 3-layer hierarchical model and occupancy grids, for the fusion of multimodal tracking data coming from sensors spread throughout several rooms. While all the above mentioned approaches describe some of the needed sensors and components or discuss concepts for the fusion of their outputs, none of them present an actual integrated and real-time capable system performing all the necessary tasks of audio-visual tracking and identification, data association, active sensor control, and temporal and multimodal fusion simultaneously and robustly for several users.

In a real, dynamic environment, single modality face or speaker recognition accuracies are highly dependent on lighting conditions, head orientations, face alignment precision, room noise levels, reverberations, crosstalk, and so forth, and using the information from several types of sources over multiple time frames for identification can bring a substantial improvement. Moreover, the fact that interacting users take turns speaking, their faces are not always visible even when using several cameras for active scanning, delays between the time of image or voice capture and the availability of recognition results, etc, force us to deal with incomplete information coming sporadically from several sources with variable confidence.

Our system deals with these issues by consistently tracking all room occupants, spatio-temporally remapping identification results to person tracks, accumulating ID hypotheses and confidences for each track, and dynamically deciding on the most probable configuration of tracks and IDs for the observed scene globally. Moreover, our fusion scheme allows us to implement a simple procedure for detection of unknown persons, although the recognition modules themselves were not designed for this purpose. Our system uses five fixed cameras for tracking, several T-shaped microphone arrays for speaker localization and ID, two active cameras for face ID, and realizes an online, incremental identification of multiple persons engaged in natural interaction scenarios.

In the following section, we present a detailed description of the system’s various tracking and identification components. Section 3 then explains the developed target selection and data fusion technique. Section 4 presents an experimental evaluation of the integrated system and section 5 gives a short summary and an outlook.

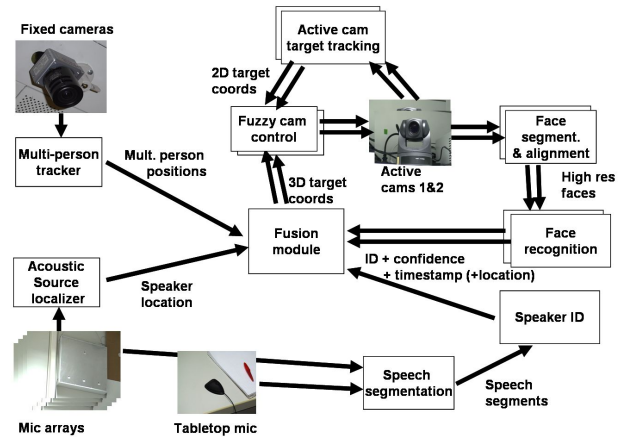


Figure 1: Overview of the multimodal identity tracking system components

2. MULTIMODAL IDENTITY TRACKING SYSTEM

This section describes the components of our multimodal tracking and identification system. The system is designed to acquire the identities of persons in an opportunistic, unaware and unobtrusive way, e.g. whenever a person speaks or faces a camera.

All components and sensors work together to achieve the goals of high room coverage, precise localization and quick and accurate identification. Each of these components is designed to be fully automatic and realtime-capable, and is seamlessly integrated in the overall system. The prime building block is a multiple camera particle filter-based person tracker delivering the positions of all persons present in the room. This information constitutes the basic scene model to which localized identification hints from other sources are mapped. Modules for speech detection and speaker identification, coupled with a source localizer using the input from several microphone arrays, deliver precisely localized ID cues whenever a speaker becomes active. A set of pan-tilt-zoom cameras focus in on the tracked persons to gain high resolution snapshots usable for face identification. The face of the focus person is automatically detected and tracked in the active camera images. This information is used in a fast fuzzy control loop, allowing for smooth tracking, high confidence identification, and pin-point 3D localization of the face. A central fusion module, described in section 3, analyzes the output of all components and implements a variety of strategies for target and camera selection. It performs spatio-temporal association of incoming identification cues to person tracks, accumulates statistics over time, and optimizes the global scene configuration according to current and past observations.

The acquisition and the processing of information are distributed over a network of computers. A total of eight Pentium IV, 3GHz machines is used: Five for the visual tracking, one for the acoustic tracking and identification, and two more for the control of active cameras and the tracking and identification of faces in closeup views.

Fig. 1 gives an overview of the system and of the interaction between its components.

2.1 Multiple Person Tracking in Fixed Camera Views

The person tracking component is a particle filter fusing the features from several fixed cameras: four SONY DFW-V500 color firewire cameras mounted in the room corners and one SCORPION SCOR-03NSC firewire color camera equipped with a 180° fisheye lens placed under the ceiling in the middle of the room, all capturing at 15fps with a resolution of 640x480 pixels. The tracker automatically detects and tracks multiple persons without requiring any special initialization phase or area, clean backgrounds or a-priori knowledge about person colors or attributes, for standing, sitting or walking users alike. The tracker has been extensively tested and shown to work robustly in a number of environments with different users and scenarios, with varying occlusion, lighting and background conditions. The following gives a short description of the system. For details, the reader is referred to [24].

2.1.1 Tracking Features

The features used are adaptive foreground segmentation and upper body color information from all 5 cameras, as well as upper body detection cues from the room corner cameras and person region hints from the top camera, computed on reduced 320x240 pixel images.

- The foreground segmentation is made using a simple adaptive background model, which is computed on grayscale images as the running average of the last 1000 frames. The background is subtracted from the current frame and a fixed threshold is applied to detect foreground regions.
- The color features are modeled in a specially designed histogram structure, which eliminates the usual drawbacks of HSV histograms when it comes to modeling low saturation or brightness colors.

The color space is a modified version of the HSV cone, where the color values are discretized as follows: Let *hue*, *sat* and *val* be the image HSV values, then the corresponding histogram bin values, *h*, *s* and *v*, are computed as:

$$v = val \quad (1)$$

$$s = sat * val \quad (2)$$

$$h = hue * sat * val \quad (3)$$

The effect is that the number of bins in the hue and saturation dimensions decreases towards the bottom of the cone and there is, e.g., only one histogram bin to model colors with zero brightness, in contrast to classical discretization techniques.

Color histograms are gained for the detected upper body region of subjects, as well as their immediate surrounding background. One upper body and one background histogram are kept per camera for every track. Upper body histograms for corner cameras are adapted with each detection hit from pixels inside the detection region. As soon as a track was confirmed by a corner camera detection, the upper body histogram for the top camera is continuously adapted with every frame using colors sampled from a 60cm diameter region centered on the track's estimated position. This

is to avoid continuously adapting spurious tracks. The background histograms for all views are continuously adapted in every frame, with the learnrate set such as to achieve a temporal smoothing window of approximately 3 seconds.

All upper body histograms are continuously filtered using their respective background histograms. Let *H* be an upper body and *H_{neg}* a background histogram. Then the filtered histogram *H_{filt}* is obtained as:

$$H_{filt} = minmax(H) * (1 - minmax(H_{neg})). \quad (4)$$

The effect of histogram filtering is to decrease the bin values for upper body colors which are equally present in the background. The motivation is that since several views are available to track a target, only the views where the upper body is clearly distinguishable from the immediately surrounding background should be used for tracking. The use of filtered histograms has shown to dramatically increase tracking accuracies.

- The upper body detection hints in fixed corner camera images are obtained by exhaustive scanning with Haar-feature classifier cascades, such as in [5, 6]. Using calibration information, the 3D scene coordinates of the detected upper body as well as the localization uncertainty, expressed as covariance matrix, is computed from the detection window position and size. This information is later used to associate detections to person tracks and in particle scoring.
- Person regions are found in the top camera images through the analysis of foreground blobs, as described in [23]. It is a simple model-based technique that dynamically maps groups of foreground blobs to possible person tracks and hypothesizes a person detection if enough foreground is found in a 60cm diameter region within a certain time interval. The motivation is that top view images present very little overlap between persons, making a simple spatial assignment plausible.

2.1.2 Initialization and Termination Criteria

To detect persons and initialize tracking, a fixed number of “scout” particle filter trackers are maintained. These are randomly initialized in the room and their particles are scored using the foreground, color, and detection features described above. A track is initialized when the following conditions are met:

- The average weight of a scout's particles exceeds a threshold *T*, set such that initialization is not possible based on the foreground feature alone, but requires the contribution of at least an upper body detection or person region hint.
- The spread of its particle cloud, calculated as the variance in particle positions, is below a fixed limit.
- The target object's color is balanced throughout all camera images. For this, color histograms are computed in each view by sampling the pixel values at the scout's particles' projected 2D coordinates, and histogram similarity is measured using the bhattacharyya distance [26].

- The target object is sufficiently dissimilar to its surrounding background in every view. Again, the bhat-tacharyya distance is used to measure similarity between the computed track histogram and the corresponding background histogram. For the latter, colors are sampled in each view from a circle of 60cm diameter, centered around the scout track’s position and projected to the image planes. This measure helps to avoid initializing faulty tracks on plane surfaces, triggered e.g. by false alarm detections or shadows.

Tracks are deleted when their average weight, considering only color, detection and person region contributions, falls below a certain threshold, or the spread of their particle cloud exceeds a fixed limit.

2.1.3 Particle Filtering

The tracking scheme presented here uses as separate particle filter tracker for each person. Each particle represents a hypothesized (x, y, z) person position in the scene. After scoring, normalization, and resampling, the mean of the particle’s positions is taken as the track center. Propagation is then done by adding gaussian noise to the resampled particle’s positions in the following way: The particles are first split into 2 sets. The first set comprises the highest scoring particles, the “winners” of the resampling step, and contains at most half of the particle mass. The rest of the particles comprises the second set. The speed of propagation is then adjusted differently for each set, such that the high scoring particles stay relatively stable and keep good track of still targets, while the low scoring ones are heavily spread out to scan the surrounding area and keep track of moving targets. A total of only 75 particles is used per track. The tracking system implementation is distributed over a network of 5 machines to achieve realtime computation speed. The system was extensively tested and achieved high accuracy rates, as shown in section 4.

2.2 Speech Detection, Localization and Speaker Identification

In parallel to the visual tracking of all room occupants, speech localization and speaker identification is performed on a separate machine. For localization, four T-shaped microphone arrays installed on the room walls are used. For speaker identification, only one microphone from one of the arrays or a table-top microphone is used. The three subtasks are accomplished as follows:

- Speech detection and segmentation. This is currently done by thresholding in the power spectrum. Speech segments of more than 1 second length are extracted and fed to the identification module.
- Speaker localization and tracking. This is done by estimating time delays of arrival between microphone pairs using the Phase Transform variant of the Generalized Cross Correlation function (GCC-PHAT):

$$R_{12}(\tau) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{X_1(e^{j\omega\tau})X_2^*(e^{j\omega\tau})}{|X_1(e^{j\omega\tau})X_2^*(e^{j\omega\tau})|} e^{j\omega\tau} d\omega$$

where $X_1(\omega)$ and $X_2(\omega)$ are the Fourier transforms of the signals of a microphone pair in a microphone array.

The localization approach used here feeds the computed time delays between the various microphone pairs directly to a Kalman Filter, which performs the tracking in a unified way. The details of the source localizer can be found in [17]. Its output is the 3D position of the active speaker in the scene.

- Speaker Identification. The component for speaker ID is based on the approach presented in [21]. Speakers are modeled using a 32-component Gaussian Mixture Model (GMM). The inputs to the GMMs are the MFCC coefficients computed on the segmented speech from one audio channel. For each speaker, one set of GMMs is trained offline on a 30 second speech segment. The recognition itself is made on segments of 1 to 5 seconds, with longer segments being broken down into smaller ones, to allow for intermediate identification results. Cepstral mean subtraction and feature warping are performed on the audio signal to reduce channel, noise and reverberation effects. The output of the speaker ID module is the identity of the speaker as well as the corresponding GMM’s a-posteriori probability for the analyzed segment, which is used as confidence measure.

A history of speech source location estimates is kept for the duration of a speech segment. Similarly, for the same time window, a record is kept in the fusion module of the positions of all visually tracked persons. The visual and acoustic tracks are then compared to associate the recognized speaker ID to the best matching person track.

2.3 Acquisition of High Quality Face Images for Identification

Based on the current room occupant configuration, and the confidence in each person’s identity, the fusion module decides on the persons of interest and on the active cameras to be used to acquire frontal views of their faces for identification. The focus persons’ 3D scene positions are sent to the active camera tracking and control modules where several subtasks are accomplished:

- The fuzzy control of the camera’s pan, tilt, and zoom factor to focus in on the desired region.
- The detection and tracking of the target person’s head and/or body in the camera image, and the feedback of this information to the fuzzy controller to allow for high quality closeups and smooth tracking.
- The detection and alignment of frontal faces whenever available from the person track and the reprojection of detection hits to the 3D scene.
- The frame-based identification of the person of interest. The determined ID, together with a frame-level confidence are combined with the 3D location information and sent back to the fusion module.

The face acquisition task is accomplished by two SONY EVI-D70P cameras mounted on the room walls. They are placed such as to offer good views of a person giving a presentation in front of the projection board or coming in the door, but also offer good coverage of the audience and the rest of the room. Each camera is connected to a separate machine running dedicated components for fuzzy control, detection, tracking, alignment and identification of faces.

2.3.1 Active Camera Person Tracking

Although the location of the focus person is given by the fusion module, the active camera acquisition system was designed to autonomously detect the person of interest's face in the moving camera image itself, to estimate its position, speed, size and size variation, and to track the person's face and upper body using only this information until instructed to switch focus to another person or until a tracking error is detected. This allows for much stabler and higher quality views to be captured than would be possible based on external track information alone. A detailed description of the autonomous active camera tracker module can be found in [20]. The following gives only a brief overview.

For face detection, boosted cascades of classifiers based on Haar-like features are used, as described in [5, 6]. This technique was chosen as it is relatively insensitive to the heavy camera motion, and the resulting color and lighting changes in the scene. Once a face was detected, a track is initialized and subsequent searching is limited to the area surrounding the track.

From the detection window, a color histogram is built in HSV space for the person's face region, H , the upper body region, Hb , and the surrounding background, H_{neg} and Hb_{neg} . The face and upper body histograms are filtered as described in [7]:

Given that after normalization, $H(x)$ can be seen as modeling $P(x|Face)$, and $H_{neg}(x)$ as modeling $P(x|\neg Face)$, for a given pixel x , we can, by applying Bayes' rule, obtain the likelihood ratio of a pixel belonging to the face as

$$\frac{P(Face|x)}{P(\neg Face|x)} \sim \frac{P(x|Face)}{P(x|\neg Face)} = \frac{H(x)}{H_{neg}(x)}. \quad (5)$$

For ease of representation, we directly compute the histogram $H_{filt} = H/H_{neg}$ which will, after normalization, be used to calculate backprojection maps on the input image. Similarly, a filtered histogram Hb_{filt} is built for the upper body. Two separate meanshift [8] trackers are run in parallel, on the face and upper body backprojection maps. Since the face track is usually very reliable and allows estimation of the face size and therefore its distance to the camera, priority is given to the face tracker. Upon failure of the face track, e.g. when the person turns away, the upper body track is used to keep the target in view, although no reliable size information is available. Color histograms are updated whenever a new face detection shows sufficient overlap with the tracked face window. Furthermore, the backprojection maps are used upon detection and adaptation to assess the quality of the built color model. Only if the average backprojection value inside the detection window exceeds a certain threshold is the detection considered valid and the learnrate for adaptation is calculated proportional to the detection quality. Figure 2 shows the backprojection maps for the face and upper body models.

2.3.2 Fuzzy Control

The advantages of fuzzy control over other techniques, such as PID controllers, etc, is that expert knowledge can be used, encoded in the fuzzy rules, to simulate the natural behavior of a human operator [10]. It allows for much smoother camera handling in stable situations, while maintaining the ability to react quickly and keep the target in the image in emergency situations.

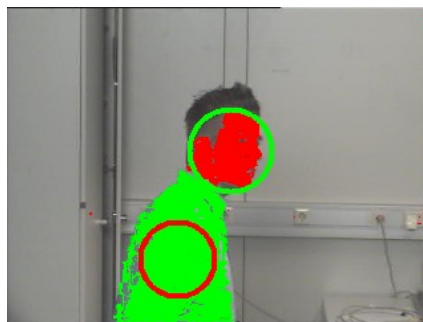


Figure 2: Backprojection maps for face (red) and upper body (green)

Here, the fuzzy controller must accomplish two main tasks: First, it must quickly position the camera to observe the new target region, whenever a switch of focus person is made. Second, it must smoothly keep track of the focus person after it was detected in the image and focus in until the desired face size is reached.

For the first part, the 3D scene location obtained from the fusion module is used to compute the expected position, speed, and size of the target's face (or outside of) the image, which serve as inputs to the fuzzy controller. In the second case, the inputs are the tracked image coordinates of the face, its size, as well as the gradients of these values.

The output of the fuzzy controller are the required pan, tilt and zoom speeds for the camera. Experiments have shown that using gradients and angle speeds allows for much more dynamic and smooth control as absolute positioning would, as the camera can adapt its rotation and zoom to match the relative speed of the target.

The behavior of the system is determined by a set of fuzzy rules, which for conciseness will not be discussed in detail here. The set includes basic rules for adjusting pan and tilt, based on the image position and speed of the target, as well as more complex rules encoding specific behavior. For example, when the track is close to the left image border and is moving left, the camera should not only quickly move to the left, but also zoom out slowly, as a wider angle of view will automatically help to keep the track in the image. On the other hand, if the target is close to the right border and is moving left, nothing should be done, as the target's motion will automatically bring it closer to the image center.

The fuzzy rules have been designed manually, and the fuzzy sets empirically adjusted to yield satisfactory results on a range of test scenarios.

2.3.3 Face Alignment and 3D Scene Reprojection

Once the active camera has zoomed in on the focus person, high resolution snapshots of frontal views of the face are taken, aligned for face identification, and their 3D scene position is estimated. The detection of frontal faces is made as described in section 2.3.1. This first step does however deliver a moderate amount of tilted faces, whereas the identification is made only on upright frontal faces. This is why the inside of the detected face rectangle is again scanned in a second pass with Haar-feature classifier cascades specially trained to recognize eye regions. Only if two eyes can be detected, reasonably situated inside the face rectangle, is the aligned face passed on for recognition. Although this may



Figure 3: Face detection and alignment in active camera images. In a first pass, a frontal face is detected by scanning the region surrounding the person track. In a second pass, the inside of the face rectangle is scanned with a specialized eye detector. If two eyes can be found, the face is aligned and passed on for recognition. The procedure guarantees extremely low false alarm rates

cause some faces to be discarded because both eyes could not be detected, the two stage approach does guarantee extremely high precision rates with practically no false alarms. Fig. 3 shows the face detection and alignment process.

To estimate the 3D scene location of the aligned face from its image coordinates and size, it is necessary to recompute the intrinsic and extrinsic parameters of the moving camera at every point in time. This is also useful to compare tracked 2D face positions from the active camera image with specified 3D focus person positions from the fusion module and detect tracking errors. Here, the camera parameters are continuously updated using rotation and zoom values obtained from the camera. These are read directly through the camera’s RS-232 serial interface. An initial calibration of the camera is performed in its rest position ($pan = tilt = 0^\circ$) using standard calibration techniques [9, 30], yielding initial values for the camera position in the scene T_{init} and its base rotation R_{init} , as well as focal length estimates at 9 discrete zoom steps $f_{x,0} \dots f_{x,8}$. The camera rotation matrix is then continuously updated from the latest pan and tilt information, by multiplying the initial rotation matrix with a “correction matrix” R_{corr} (Eq. 6),

$$R_{act} = R_{init} \cdot \begin{pmatrix} \cos(\beta) & \sin(\alpha) \sin(\beta) & -\cos(\alpha) \sin(\beta) \\ 0 & \cos(\alpha) & \sin(\alpha) \\ \sin(\beta) & -\sin(\alpha) \cos(\beta) & \cos(\alpha) \cos(\beta) \end{pmatrix} \quad (6)$$

with α the camera pan angle and β the tilt angle.

The focal length itself is not directly readable and is interpolated for the current camera zoom step from the discrete values $f_{x,0}$ to $f_{x,8}$, using a 4th order polynomial function. Even though interpolation introduces some imprecision, the maximum observed deviation error comprised only a few pixels, which is completely sufficient for our purpose.

Using the up to date intrinsic and extrinsic camera parameters, and assuming a standard eye distance of 7cm, the 3D location of each aligned and identified frontal face is computed. This additional information is useful to avoid misalignments and false associations when mapping active camera face ID cues to person tracks in the global scene model.

2.3.4 Face Recognition

For the identification of faces inside the active camera views, the approach presented in [18, 19] is used. It uses a face representation based on local appearance and is less sensitive to lighting changes and occlusion than methods based on global appearance. An aligned and normalized face image is divided into blocks of 8x8 pixels. Each block is then represented by its DCT coefficients. The top-left DCT coefficient is discarded since it only represents the average intensity value of the block. From the remaining DCT coefficients, the ones containing the highest information are extracted via zig-zag scan. The so obtained DCT coefficients for each block are concatenated to construct the feature vector for identification. This is fed to a nearest neighbor classifier using the normalized vector correlation as distance metric. To derive a confidence measure for the computed ID, the distances of the test feature vector to the 10 nearest training sample vectors are used. These distances are first sorted, min-max normalized, their complement is computed, and after an additional normalization, they represent the confidences for the 10 most probable IDs. Currently, only the best frame-based ID and its confidence are passed on to the fusion module.

The feature vectors used in training of the classifier were obtained by automatically capturing sample images for each subject at different points in the room using the active cameras, and applying the same alignment and decomposition techniques described above. Training was done offline using roughly 70-180 images per person (captured at about 10fps).

3. DATA FUSION AND OPTIMAL ID ASSIGNMENT

The main idea behind the design of the ID Tracking system is to actively seek, opportunistically capture, and fuse reliable cues for recognition whenever they become available, and to keep tracking identified persons while attention is focused to unknown ones. Audio-visual ID cues of varying accuracy coming from the different system modules are expected to arrive in an irregular way. In a natural, unconstrained scenario, good facial shots for identification can seldom be acquired. Similarly, speaker localization and identification is not possible for persons that remain silent, and is very inaccurate in the presence of noise or cross-talk. This raises the need for a fusion technique that handles incomplete and very sporadic information.

In the current implementation, the fusion module uses the person track information from the multiple camera tracker as the basis upon which the scene model is updated and association of ID cues is performed. The scene model is composed of a number of active person models, and some optional information such as the position of the entrance door and of the whiteboard. A person model comprises the person’s 3D location, and a histogram of identification cues that were assigned to it over time. This “ID histogram” has as many bins as audio-visually trained in subjects, and the values accumulated in the respective bins are the confidences given by the face or speaker ID modules.

The mapping of ID cues to person models is made in the following way: As visual identification is made at frame-level, the face ID cues, which are tagged with their 3D location estimates, are directly mapped to the currently tracked persons based on spatial overlap. For speaker identification,

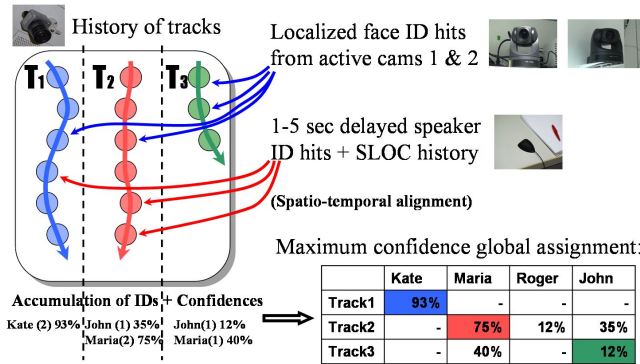


Figure 4: The process of mapping localized ID cues to person tracks. A spatio-temporal mapping is made for face and speaker identification cues. Confidences are accumulated in time and a global assignment of IDs to tracks is made that optimizes the overall confidence level

on the other hand, recognition is delayed until whole speech segments of 1 to 5 seconds have been processed, which causes ID cues for a track to come at a sensitively late time. Therefore, a temporal matching also has to be made. This is done by keeping a 5 second history of all person tracks, comparing with the source localization history of the identified speech segment, and selecting the track with the highest spatio-temporal overlap.

Once the association of an identification cue to a track has been made, the identification confidence is accumulated in the track’s ID histogram, by adding its value to that of the corresponding ID’s bin. To limit the temporal scope for subsequent optimization, a ring buffer-like mechanism ensures that only the last 10 ID hits are stored and used in the accumulation. The selection of the definitive ID is not made for each track independently, e.g. based on the highest accumulated score. Rather, a global optimization procedure is used to determine the ID mapping jointly for all tracks. If the person tracks and the set of accumulated IDs are considered the rows and columns of a matrix, with the respectively accumulated confidences as values, it quickly becomes evident that this is a fundamental combinatorial optimization problem, specifically a maximum weight assignment problem, for which standard solutions exist. Here, Munkre’s algorithm [25], which has polynomial runtime complexity, was chosen. The optimal assignment is recomputed every time a new identification hit is received. The advantage of this technique is that assignment of the same ID to several persons is excluded, as the system will change the hypothesized ID for one track based on new information for another track. Furthermore, temporal fusion can be performed even for sporadically received ID cues. Figure 4 shows the process of spatio-temporal mapping and ID assignment.

After accumulation and fusion, the recognition accuracies for known persons are generally very high. This makes it possible to detect unknown persons, even though the separate recognition modules were not designed to do so, by thresholding the final accumulated confidence (in our case to 30%). Low single modality confidences or inconsistent identification across time or modalities, typical for unknown

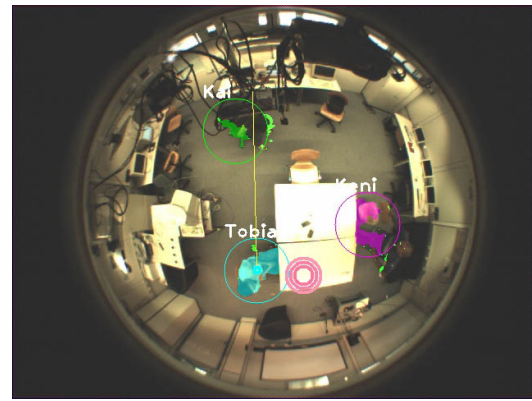


Figure 5: The output of the identity tracking system. The colored circles represent the person models. The identities for recognized persons are printed on top of the respective tracks. The yellow line points at the actual focus of attention of the active cameras

users, will yield a low overall confidence and the track ID is marked as unknown. Note that in this way, the ID tracking system only outputs track identities in which it is confident, and can in time recover from initial wrongful decisions. Fig. 5 shows an example output of the ID tracking system.

Based on the actual configuration of persons in the scene and on the recognition confidence for each, persons of interest are determined and the best active cameras for observation are selected. The currently implemented target and camera selection strategy does the following:

- It consecutively scans the locations of all participants using all active cameras simultaneously to increase the chances of capturing a frontal face. A more efficient technique could be devised by the inclusion of some form of head pose estimation, which would guide the choice of the best camera for each participant, but this was not done at this point. Targets of attention are switched regularly at 10 second intervals.
- Whenever a person track is near the entrance door, the active camera offering best views of the door area is immediately steered to capture the faces of eventual newcomers.
- Whenever a person is found to stay near the whiteboard for a certain period of time, another active camera, offering best views of the eventual presenter, is dedicated to following that track.

Other strategies, such as focusing on the active speaker, etc, have been experimented with, but will not be further considered in this paper.

4. EXPERIMENTAL EVALUATION

In the first part of this section, experimental results for the separate system components are first shown, where available. The second part then presents an evaluation of the integrated ID tracking system on a small test scenario and discusses the advantages of multimodal fusion.

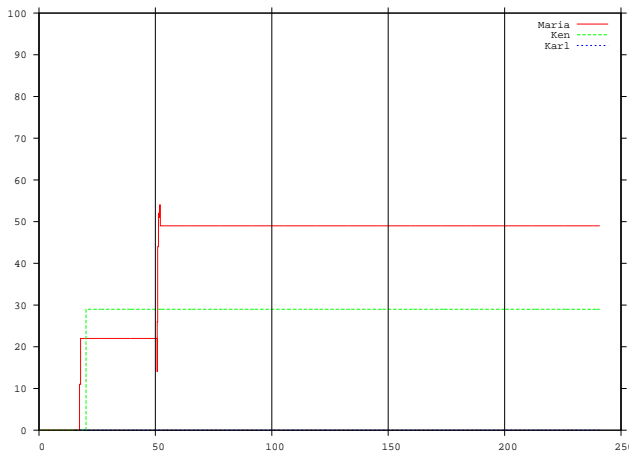


Figure 6: Temporal evolution of identification confidences, using only visual face recognition cues. The horizontal axis represents the elapsed time in seconds since the launching of the system. The vertical axis represents identification confidences in percent. Frontal faces could only rarely be acquired causing identification cues to be sparse. The 3rd user was not identified at all, as his face could never be detected and aligned in the camera image

The multiple camera person tracking system presented in section 2.1 has been evaluated on the CLEAR2007 [28] development set, comprising over 100 minutes of recordings made at 5 different sites. It reached a tracking precision (MOTP) of 15cm and a high tracking accuracy (MOTA) of 73% (A detailed explanation of the MOT metrics can be found in [22, 23]). The same system as for the offline evaluation is used with very slight modifications in our online application.

The face recognition and speaker identification components were evaluated on the CLEAR2006 [28] database, figuring faces and voices for 26 individuals, recorded at multiple sites. Speaker identification was tested on pre-segmented speech and achieved very low error rates of 3.08% for 5 second test segments. The face recognition approach was evaluated on very small face images (down to 15x15 pixels) captured with fixed far-field cameras and reached error rates from 40% for 1 second test segments to 16% when fusing hypotheses over 20 second segments. The high recognition accuracy of the acoustic system make it a very useful tool in scenarios such as ours. The drawback is that users can only be recognized when they take turns to speak, which can cause significant delays before all participants are identified. The drawback of the visual system is that users will not be recognized until they face one of the cameras.

The integrated system was evaluated on a sample interaction scenario involving 3 users entering the room and engaging in conversation in a meeting-like setting. The users were sitting around a meeting table, occasionally standing up to give explanations in front of the whiteboard. Although the recording for this experiment was not a direct excerpt of the CLEAR database, it was made in the same room under the very same recording conditions. The reason the large offline database could not be used is that the active camera steering and face capturing can not be performed on pre-

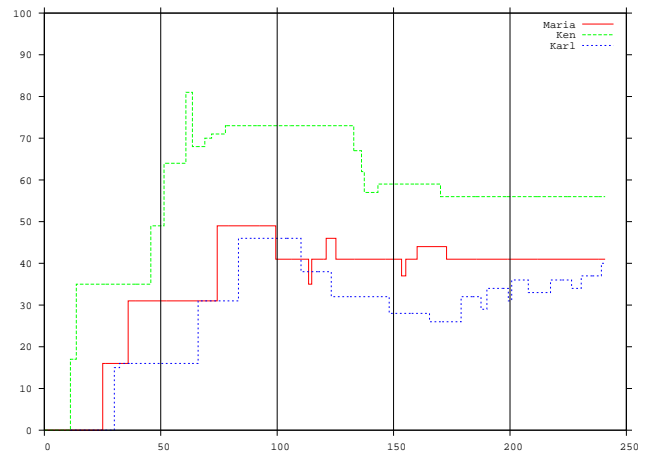


Figure 7: Temporal evolution of identification confidences, using only acoustic speaker ID cues. Results are much more consistent than in the visual case. All 3 users were correctly identified (confidence level above 30%) within the first 80 seconds of the start of recording

recorded data. The database for identification comprised 7 users. Training of their faces was done using 180 close-up images per user, captured automatically in an unaware fashion over the course of several months, whereas training of their voices was done on data from a controlled one minute recording session using only one of the microphone array microphones. The ID tracking system was started as the meeting had already begun, to prevent it from capturing easy face snapshots as users enter the room. To demonstrate the effectiveness of the individual modalities and of their combination, the output user identities and the evolution of the system's confidence in the identity for each track, using the visual modality, the acoustic modality, or both, were recorded for a 4 minute segment of the recording.

Figure 6 shows the results for visual identification. Within the first 25 seconds, the two first users are captured and identified. Recognition accuracies are below the 30% margin, though, causing them to stay classified as “unknown”. After second 50, a quick series of ID hits cause confidences for user 2 to rise to 50%, and she is correctly identified. Although several incorrect frame-level ID hits accompanied the correct ones, their effect was attenuated by the temporal fusion, causing confidence levels to drop only slightly. The remainder of the sequence showed the problem inherent in the visual-only identification approach: As the users made no effort to face the active cameras, even with high magnification, their faces could only be rarely captured. For user 3, a face alignment was never possible, and he was missed for the entire sequence. This problem should be overcome in the future by a better steerable camera coverage, or by the inclusion of techniques for the alignment and identification of non-frontal views of faces.

Figure 7 shows, in contrast, the more accurate and stable results obtained with the acoustic modality. Confidence levels for user 1 start rising from second 15, and after 20 seconds, he is correctly identified. Users 2 and 3 are also correctly recognized after 35 and 60 seconds. Confidence levels keep rising as the users take turns speaking, but they also

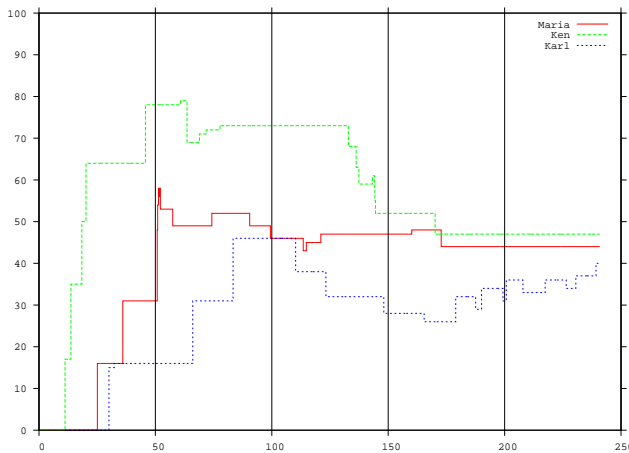


Figure 8: Temporal evolution of identification confidences, using both modalities. Recognition speed could be increased considerably for the first 2 users, compared to the acoustic only case. The maximum confidence levels over the length of the recording sequence were also sensibly higher, showing the advantages of multimodal fusion

occasionally drop, as cross-talk, laughter, spurious speech segmentation, etc, cause incorrect identification hits to be produced. The results are on the whole much smoother, as the individual identification cues and confidences themselves are computed on whole segments of several seconds.

Even though visual identification cues were hard to obtain in our test sequence, the combination of audio and visual cues for identification showed some advantages over acoustic analysis alone. The results are depicted in figure 8. Compared to the acoustic identification case, confidences rise much faster, and attain higher maximum values. The confidence level for user 1 reaches 64% after already 20 seconds, and 78% at second 45. Confidences for user 2 reach a maximum of 58% at second 51 before falling back slightly to 49%. The recognition for user 3 is made solely based on speaker ID cues, and results are therefore identical to the acoustic case. On the whole, the system rapidly acquired the users' identities as soon as identification cues became available, made very few data association mistakes, and kept correct identities and relatively stable confidence levels for the remainder of the sequence.

5. CONCLUSION AND OUTLOOK

In this paper, a system for the simultaneous tracking and incremental multimodal identification of multiple users in a smart environment was presented. The system fuses person track information, localized speaker ID and high definition visual ID cues opportunistically to achieve a high level of versatility and robustness. Visual tracks are obtained from a fully automatic particle filter based multiple camera tracker using foreground, color, and upper body detection cues as features. Speaker localization and identification cues are delivered by a combination of Kalman filter tracking and GMM-based ID using microphone arrays and tabletop microphones. Fuzzy controlled pan-tilt-zoom cameras mounted on the room walls autonomously detect and track the faces of persons of interest, and focus in to

gain high resolution facial views. A three stage approach aligns and identifies frontal faces in the captured images, and computes their 3D scene coordinates. The localized ID cues gained from the audio and visual recognition modules are sent back to a fusion module, which spatio-temporally maps them to the visual person tracks, performs a global assignment of IDs to tracks that maximizes overall confidence levels, recognizes unknown persons, and gradually refines all track identities. The implemented target and camera selection strategy focuses the system's attention to the different room occupants, to newcomers, and to speakers in the present area. The system actively seeks reliable cues for identification in an unobtrusive way, requires no initialization procedure or interaction with the users, and functions in realtime.

Future efforts will go into improving the face or body detection rate in all cameras by the addition of more features, extending the face ID recognizer to handle non-frontal views, automatically clustering and learning new faces or voices based on complementary information from the different modalities, and including head pose estimation to enhance the effectiveness of camera selection. As additional extension, a more unified fusion scheme, which integrates all observations, visual, acoustic tracks and identification cues in a single particle-filter framework, is also planned.

6. ACKNOWLEDGMENTS

The work presented here was partly funded by the *European Union* (EU) under the integrated project CHIL, *Computers in the Human Interaction Loop* (Grant number IST-506909). The authors also wish to thank Florian van de Camp, Hazim Ekenel, Tobias Gehrig and Qin Jin for their invaluable contributions.

7. REFERENCES

- [1] I. McCowan, D. Gatica-Perez, S. Bengio, G. Lathoud, M. Barnard, D. Zhang, "Automatic Analysis of Multimodal Group Actions in Meetings". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 27, no. 3, pp. 305-317, March, 2005.
- [2] R. Stiefelhagen, "Tracking Focus of Attention in Meetings". *IEEE International Conference on Multimodal Interfaces (ICMI)*, Pittsburgh, 2002.
- [3] T. Choudhury, B. Clarkson, T. Jebara and A. Pentland, "Multimodal Person Recognition using Unconstrained Audio and Video". *Second Conference on Audio- and Video-based Biometric Person Authentication '99 (AVBPA '99)*, pp 176-181, Washington DC
- [4] J. Yang, X. Zhu, R. Gross, J. Kominek, Y. Pan, A. Waibel, "Multimodal people ID for a multimedia meeting browser". *Proceedings of the 7th ACM International Conference on Multimedia '99*, Orlando, FL
- [5] P. Viola and M. Jones, "Rapid Object Detection using a Boosted Cascade of Simple Features". *International Conference on Computer Vision And Pattern Recognition*, 2001.
- [6] R. Lienhart and J. Maydt, "An Extended Set of Haar-like Features for Rapid Object Detection". *IEEE ICIP 2002*, Vol. 1, pp. 900-903, Sep. 2002.

- [7] K. Nickel and R. Stiefelbogen, "Pointing Gesture Recognition based on 3D tracking of Face, Hands and Head Orientation". 5th International Conference on Multimodal Interfaces, Vancouver, Canada, Nov. 2003.
- [8] D. Comaniciu and P. Meer, "Mean Shift: A Robust Approach Toward Feature Space Analysis". IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 24, No. 5, May 2002.
- [9] R. Y. Tsai, "A versatile camera calibration technique for high-accuracy 3d machine vision metrology using off-the-shelf tv cameras and lenses". IEEE Journal of Robotics and Automation, RA-3(4), pp. 323-344, August 1987.
- [10] Earl Cox, "Fuzzy fundamentals". IEEE Spectrum, 1992, pp. 58-61
- [11] S. Tsuruoka, T. Yamaguchi, K. Kato, T. Yoshikawa, T. Shinogi, "A Camera Control Based Fuzzy Behaviour Recognition of Lecturer for Distance Lecture". 10th IEEE International Conference on Fuzzy Systems, December 2001, Melbourne, Australia.
- [12] P. Peixoto, J. Batista, H. Araujo, "A surveillance system combining peripheral and foveated motion tracking". 14th International Conference on Pattern Recognition, Vol. 1, pp. 574-577, Aug. 1998
- [13] A. Hampapur, S. Pankanti, A. W. Senior, Y.-L. Tian, L. Brown, R. M. Bolle, "Face Cataloger: Multi-Scale Imaging for Relating Identity to Location". IEEE Conference on Advanced Video and Signal Based Surveillance (AVSS 2003), July 2003, Miami, FL.
- [14] S. Stillman, R. Tanawongsuwan, and I. Essa, "A system for tracking and recognizing multiple people with multiple cameras". Technical Report GIT-GVU-98-25, Georgia Inst. of Tech., Graphics, Visualization, and Usability Center, 1998.
- [15] S. Stillman and I. Essa, "Towards reliable multimodal sensing in aware environments" Perceptual User Interfaces (PUI) Workshop, 2001.
- [16] M. Trivedi, I. Mikic and S. Bhonsle, "Active Camera Networks and Semantic Event Databases for Intelligent Environments". IEEE Workshop on Human Modeling, Analysis and Synthesis, June 2000.
- [17] T. Gehrig, K. Nickel, H. K. Ekenel, U. Klee, and J. McDonough, "Kalman Filters for Audio-Video Source Localization". IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, Oct. 2005.
- [18] H. K. Ekenel, R. Stiefelbogen, "Local Appearance based Face Recognition Using Discrete Cosine Transform". 13th European Signal Processing Conference (EUSIPCO), Antalya Turkey, September 2005.
- [19] H. K. Ekenel, R. Stiefelbogen, "A Generic Face Representation Approach for Local Appearance based Face Verification". CVPR IEEE Workshop on Face Recognition Grand Challenge Experiments, San Diego, CA, USA, June 2005.
- [20] K. Bernardin, F. v.d. Camp, R. Stiefelbogen, "Automatic Person Detection and Tracking using Fuzzy Controlled Active Cameras". 7th International Workshop on Visual Surveillance, Minneapolis, USA, June 22nd, 2007.
- [21] H. K. Ekenel, Q. Jin, "ISL Person Identification Systems in the CLEAR Evaluations". Proc. of the First International CLEAR Evaluation Workshop, Southampton, UK, April 2006.
- [22] K. Bernardin, A. Elbs, R. Stiefelbogen "Multiple Object Tracking Performance Metrics and Evaluation in a Smart Room Environment". 6th IEEE International Workshop on Visual Surveillance, VS 2006, Graz, Austria, May 2006
- [23] K. Bernardin, T. Gehrig, R. Stiefelbogen "Multi- and Single View Multiperson Tracking for Smart Room Environments". CLEAR Evaluation Workshop, Southampton, UK, April 2006
- [24] K. Bernardin, T. Gehrig, R. Stiefelbogen "A Particle Filter Fusion Framework for Audio-Visual Multiperson Tracking". To appear in Proc. of CLEAR Evaluation Workshop 2007, Baltimore, MD, May 2007
- [25] J. Munkres, "Algorithms for the Assignment and Transportation Problems". Journal of the Society of Industrial and Applied Mathematics, Vol. 5(1), pp. 32-38, March 1957.
- [26] T. Kailath, "The Divergence and Bhattacharyya Distance Measures in Signal Selection". IEEE Transactions on Communication Technology, vol. 15, pp. 52-60, Feb. 1967
- [27] CHIL - Computers In the Human Interaction Loop, <http://chil.server.de>
- [28] CLEAR - Classification of Events, Activities and Relationships, <http://www.clear-evaluation.org/>
- [29] AMI - Augmented Multiparty Interaction, <http://www.amiproject.org>
- [30] Jean-Yves Bouguet, "Camera Calibration Toolbox for Matlab", http://www.vision.caltech.edu/bouguetj/calib_doc