Multimodal Identity Tracking in a Smartroom

Keni Bernardin[1], Hazim Kemal Ekenel[1], and Rainer Stiefelhagen[1]

Universität Karlsruhe, ITI,
76131 Karlsruhe, Germany
{keni, ekenel, stiefel}@ira.uka.de

Abstract. The automatic detection, tracking, and identification of multiple people in intelligent environments is an important building block on which smart interaction systems can be designed. Those could be, e.g. gesture recognizers, head pose estimators or far field speech recognizers and dialog systems.
In this paper, we present a system which is capable of tracking multiple people in a smartroom environment while infering their identities in a completely automatic and unobtrusive way. It relies on a set of fixed and active cameras to track the users and get closeups of their faces for identification, and on several microphone arrays to determine active speakers and steer the attention of the system. Information coming asynchronously from several sources, such as position updates from audio or visual trackers and identification events from identification modules, is fused at higher level to gradually refine the room's situation model. The system has been trained on a small set of users and showed good performance at acquiring and keeping their identities in a smart room environment.

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 [16, 17] aim at developing smart room 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 sensors, 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 [4, 5], estimating their head pose [6], their body posture, analyzing their speech, to infer higher level knowledge, produce meeting summaries [8], offer useful proactive services, etc. An essential task on the way to realizing these goals is the location and identification of humans in the scene.

While much research has been done on indoor tracking or on 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. Choudhury et al. [7] present a person identification system based on the fusion of multimodal cues. It is however limited to a single user required to stand closely in front of the identifiying sensors. Another approach, shown by Yang et al. [8], tackles more complex scenarios including multiple users. A framework including color-based person and face tracking, speech detection, localization and segmentation, and audio-visual ID is presented. 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.

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 combination 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. [9] 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. [10] use one fixed camera and a binocular active camera system, and implement a target seletion strategy based on state transitions to deal with scenarios involving several users. Hampapur et al. [11] perform 2D and 3D blob tracking on images from two fixed cameras 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. Like the previous two

approaches, they do not, however, address the problem of recognizing users in the closeup views or of fusing identification results over time. Similarly, Stillman et al. [12] achieve tracking, face detection and recognition of multiple users using a combination of fixed and PTZ cameras, but they do not offer a framework for camera/target selection or for fusion of the identification results.

In a real, dynamic environment, face recognition accuracy on single frames is highly dependent on lighting conditions, head orientations, face alignment precision, and so forth, and using the information from several frames for identification can bring a substantial improvement. Moreover, limitations in the number of available cameras for active scanning, delays between the time of image 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 remapping ID results to person tracks, accumulating confidences, and deciding on the most probable ID for a track dynamically. 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 a fixed camera for tracking, several T-shaped microphone arrays for speech localization and two active cameras for person identification, and realizes an online, incremental identification of multiple persons in our smartroom.

In the following section, we present a detailed description of the tracking and identification system and of its various components. Section 3 then explains the developed camera selection and data fusion technique. Section 4 shows a sample recognition scenario and section 5 gives a short summary and an outlook.

2 Multimodal Identity Tracking System

This section describes our developed system for simultaneous person tracking and identification in our smartroom. It is designed to acquire the identities of several persons on the fly, i.e. without requiring them to pass through designated areas or to specifically interact with identification devices. It also keeps track of all identified persons and refines its confidence with time.

It is composed of several components, distributed over a network of computers, which seamlessly work together while dealing with latency and synchronization issues. For each of these components, which will be described in the following, fast and effective techniques had to be adopted to allow for realtime application. The first of these components is a multiperson tracker which analyzes the scene from a ceiling-mounted camera and delivers the positions of all persons present in the room. These are compared with the output of a speech detection and source localization module which uses the input from several microphone arrays to pinpoint the active speaker and provide a focus of attention for subsequent components. The person of interest is then actively focused on by a set of pan-tilt-zoom cameras which deliver high resolution images of his or

her head from several different views. On these images, a two-stage algorithm is applied to detect and align near frontal faces, which are sent to the face identification module. The recognized identities are tagged with an identification time and confidence and sent to a fusion module, where the information coming asynchronously from all other modules is merged: Visual and audio tracks are compared, identity tags from several sources are spatially and temporally matched to tracks, and confidences are accumulated in order to gain a gradually refined view of all identities in the room.

Both the acquisition and the processing of information are distributed over a network of computers. A total of four Pentium IV, 3GHz machines is used: One for the visual and acoustic tracking each and two more for the control of active cameras and the analysis of closeup views.

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

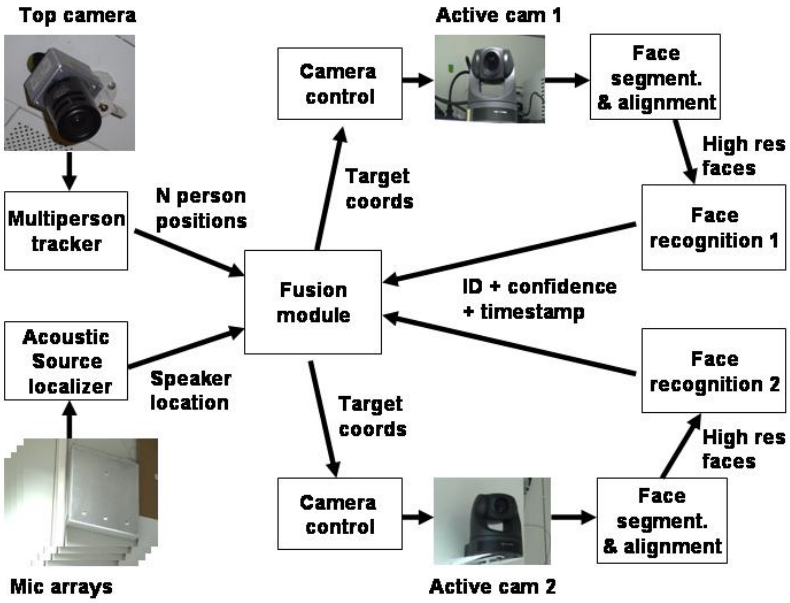

Fig. 1. Overview of the multimodal identity tracking system components

2.1 Multiple Person Tracking in Wide Angle Views

The multiple person tracking module was designed to work on wide angle images captured from the top of the room. The advantage of such views is that they reduce the chance of occlusion by objects or overlap between persons. The drawback is that detailed analysis of the tracked persons is difficult as a top

view offers only few good features for identification. This is why, in our system, a combined approach is followed.

The person tracker module processes images captured by a SCORPION SCOR-03NSC firewire color camera equipped with a 180° fisheye lens at 15fps with a resolution of 640x480 pixels.

The tracking algorithm is essentially composed of a simple but fast algorithm for foreground blob segmentation followed by a more complex EM algorithm based on person models:

First, foreground patches are extracted from the images by using a dynamic background model. The background model is created on a few images of the (preferably empty) room and is constantly adapted with each new image with an adaptation factor α. Background subtraction and thresholding yield an initial foreground map, which is morphologically filtered. A connected component analysis provides the foreground blobs for tracking. Blobs below a certain size are rejected as segmentation errors.

The subsequent EM tracking algorithm tries to find an optimal assignment of the detected blobs to a set of active person models, instantiating new models or deleting unnecessary ones if need be. A person model, in our case is composed of a position (x, y), a velocity (vx, vy), a radius r and a track ID. In our implementation, the radius was set to a fixed value, to avoid wrong estimates resulting from merged tracks, shadows, etc. The procedureis as follows:

- Expectation: For each person model M_i, update $(x, y)_{M_i}$ according to $(vx, vy)_{M_i}$. If the overlap between two models exceeds a maximum value, fuse them.
- Maximization steps: For each pixel p in each foreground blob B_j, find the person model M_k which is closest to p. If the distance is smaller than r_{M_k}, assign p to M_k.
- Iteratively assign whole blobs to person models: For every foreground blob B_j whose pixels were assigned to at most one model M_k, assign B_j to M_k and use all pixels from B_j to compute a position update for M_k. Subsequently, consider all assignments of pixels in other blobs to M_k as invalid. Repeat this step until all unambiguous mappings have been made. Position updates are made by calculating the mean of assigned pixels $(x, y)_m$ and setting $(x, y)_{M_{k,new}} = \alpha_M (x, y)_m + (1 - \alpha_M) (x, y)_{M_k}$, with α_M the learnrate for model adaptation.
- For every blob whose pixels are still assigned to several models, accumulate the pixel positions assigned to each of these models. Then make the position updates based on the respectively assigned pixels only. This is to handle the case that two person tracks coincide: The foreground blobs are merged but both person models still subsist as long as they do not overlap too greatly, and can keep track of their respective persons when they part again.
- For each remaining unassigned foreground blob, initialize a new person model, setting its (x, y) position to the blob center. On the other hand, if a model stays unassigned for a certain period of latency, delete it.

– Using the updated model positions, calculate new velocity estimates (vx, vy).
– Repeat the procedure from step 1.

The two stage approach results in a fast tracking algorithm that is able to initialize and maintain several person tracks, even in the event of moderate overlap. Relying solely on foreground maps as features, however, makes the system relatively sensitive to situations with heavy overlap. This could be improved by including color information, or with e.g. temporal templates, as proposed in [1].

By assuming an average height of 1m for a person's body center, and using calibration information for the top camera, the positions in the world coordinate frame of all N tracked persons are calculated and output.

2.2 Speech Detection and Localization

In parallel to the visual tracking of all room occupants, acoustic source localization is performed on a separate machine to estimate the position of the active speaker. For this, the system relies on the input from four T-shaped microphone clusters installed on the room walls. They allow a precise localization in the horizontal plane, as well as a rough height estimation. Two subtasks are accomplished:

– Speech detection and segmentation. This is currently done by thresholding in the power spectrum, but techniques more robust to non-speech noise and cross-talk are already being experimented with.
– Speaker localization and tracking. This is done by estimating time delays of arrival between microphone pairs using the Generalized Cross Correlation function (GCC):

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

where $X_1(\omega)$ and $X_2(\omega)$ are the Fourier transforms of the signals of a microphone pair in a microphone array.
As opposed to other approaches, where speaker positions are first calculated for microphone pairs, and the resulting positions later combined, this approach uses a Kalman filter that directly receives as input the correlation results from the various microphone pairs, and performs the tracking in a unified probabilistic way, thereby achieving more robust and accurate results. The details of the source localizer can be found in [13].

The output of the speaker localization module is the tracked position of the active speaker in the world coordinate frame. This position is compared in the fusion module to those of all visually tracked persons in the room and the person closest to the speech source is chosen as focus of interest.

2.3 Acquisition of High Quality Face Images for Identification

Once a person of interest has been determined, several parallel processes are launched to acquire frontal views of his or her face for identification. For this, two subtasks are accomplished: The automatic control of active cameras for the acquisition of closeup views and the detection and alignment of frontal faces in the captured images.

The first part is accomplished by two SONY EVI-D70P cameras mounted on the room walls. They are placed such as to offer best views of a presenter as he is talking to the audience or facing the projection board, but also offer good coverage of the rest of the room. Each camera is connected to a separate machine running dedicated components for automatic camera control and for detection, alignment and identification of faces in its images.

The second part is done by a two stage algorithm using appearance based object detectors. Because of the dynamic nature of the images, algorithms that require a static background, such as foreground segmenters, or complex initialization and slow object movement, such as contour trackers, are not applicable. The detectors used here are cascades of classifiers built on haar-like features, as described in [2, 3]. They offer fairly good detection rates and are fast enough for realtime use.

In a first pass, a face detector is used to find occurences of nearly frontal faces in the image. The image is scanned at several scales and bounding rectangles for face candidates are returned. In our implementation, the generally available frontal face classifier cascade included in the OpenCV [18] library was used. Its advantage is that it was not tuned to a specific environment and is fairly robust to lighting and background changes. It does however deliver a moderate amount of false detections or tilted faces, which are not suitable for identification.

This is why the inside of the detected rectangle is again scanned in a second pass with specially trained classifiers to recognize eye regions. These dedicated "eye cascades" have been trained on face images recorded in our smartroom. Only if both tests are passed, and two eyes can be detected, reasonably situated inside the face rectangle, the thereby aligned face is passed on for recognition. This two stage approach guarantees an extremely high precision rate with practically 0% false detections. Fig. 2 shows the face detection and alignment process.

2.4 Face Recognition

The recognition is made using a local appearance based face representation approach. A detailed description of the technique can be found in [14, 15]. A detected 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 removed from the representation 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 DCT coefficients

Fig. 2. Face detection and alignment in active camera images. In a first pass, the face area is found by a frontal face detector. In a second pass, the region inside the detected 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

obtained from each block are concatenated to construct the feature vector which is used by a nearest neighbor classifier with the normalized correlation d as distance metric:

$$d = \frac{f_{training} \cdot f_{test}}{\|f_{training}\| * \|f_{test}\|}$$

The distance of the feature vector to its assigned class is subsequently also used to derive a confidence measure for the identification.

The class representative vectors for the nearest neighbor algorithm were obtained by applying the same decomposition technique described above on a set of training images. The images were captured automatically by the active cameras at different points in the room, using the detection and alignment technique described in section 2.3, and training was done offline. Currently, nine users, mostly students and members of our lab were trained in, using approximately 70 training images per person. The biggest challenge to the recognition algorithm is the completely unconstrained nature of our scenario, as variations in lighting, in head orientation, in face sizes, and in the daily appearance of users have to be coped with. The recognizer is also not able to distinguish between unknown users and known but poorly recognizable ones. While this can lead to faulty recognition results for single images, these errors are corrected later on by the fusion module which accumulates the ID tags and confidences for a person to produce a combined hypothesis.

3 Asynchronous Fusion and Identity Tracking

The main goal of our system is to realize an unobtrusive identification and tracking of all room occupants by actively seeking and fusing the best cues for

recognition whenever they become available. As good facial shots are not easy to acquire, this raises the need for a fusion technique that deals with incomplete information coming in an irregular way.

The fusion module analyzes tracks from the multiperson tracker and position estimates from the source localizer and, using a selection strategy, decides which camera to point at which user to achieve quick identification. It also sporadically receives ID tags from several modules which it must map back to person tracks.

The currently implemented selection strategy is quite simple: Whenever available, the acoustic localization estimates are compared to the visual tracks, and if necessary an attention switch is made and both cameras are focused on the active speaker. This strategy assumes speakers are the most important actors and tries to achieve high recognition rates for them first. However, other strategies are also thinkable:

- Achieve accurate recognition for all room occupants as fast as possible. This would prioritize participants that have not been identified yet.
- Try to refresh all identities of all participants as regularly as possible. This is a good strategy if e.g. our confidence in the tracker's accuracy is low.
- Focusing two cameras on one person increases the chances to get a frontal face. Alternatively, split cameras among users, possibly choosing the best camera for a user e.g. using head orientation estimates.
- In situations where one person often speaks, keep one camera on the speaker and use the other to examine the audience.
- Define regions of high priority in the room, e.g. the door, to quickly identify new persons entering the room, etc.

Once the cameras were steered and face images captured, the fusion module waits to receive ID tags coming from the different identification modules and matches them to their respective tracks. As delays resulting from processing steps and network latency can cause ID tags for a track to come at a sensitively late time, a temporal matching has to be made also. This is currently done by keeping a history of the selected foci of attention and resynchronizing with the received ID tags based on image time stamps.

By doing this, the ID tags and their confidences can be accumulated for every track to improve recognition accuracy. Currently, the last 10 ID tags for a track are considered. The confidence scores for every hypothesized identity are accumulated, normalized and the identity with the highest score is chosen if this score surpasses 50%.

This procedure also serves to recognize unknown persons. When an untrained face is repeatedly presented to the ID module, it typically outputs a hypothesis with low confidence or a series of different hypotheses. Therefore, the accumulated confidence never reaches 50% for any identity and the track ID is marked as unknown. Note that in this way, the ID tracker system only outputs track identities in which it is confident, and can in time recover from initial wrongful decisions. Fig. 3 shows an example output of the ID tracking system.

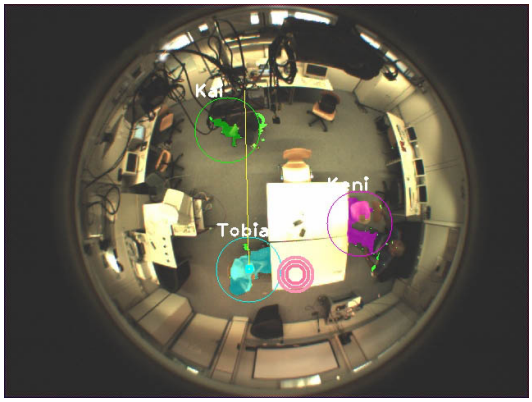

Fig. 3. The output of the identity tracking system. The colored circles represent the person models. The identities for recognized persons are printed on top of the respective tracks. The yellow line points at the actual focus of attention of the active cameras

4 Example Scenario

Figures 4 and 5 shows an example run of our identity tracker for a short scenario involving 3 users. The users enter the room and engage in conversation before they sit down together at the table. Upon entrance, they are automatically tracked by the person tracker module. User A is targeted by the active cameras and immediately recognized as "Ken". Then, user B starts to talk and the focus of attention is switched to him. The system can not, however capture a clean face shot for recognition, the identifiaction confidence is low and he is marked as "unknown". After a while, user C goes to the presentation area and starts to talk. The active cameras focus on his face, but the first identification attempts wrongfully classify him as "Kai". As frontal face shots keep getting captured, though, the confidence for the correct ID, namely "Toby", rises. The system output gradually passes from "Kai" to "unknown", to "Toby". Finally, user B speaks again. The system focuses on his face and acquires a few good facial views, definitely classifying him as "Kai". Once all users are identified, the system keeps tracking them and updates their identities everytime they take turns speaking.

5 Conclusion and Outlook

In this paper, a system for the simultaneous tracking and incremental identification of multiple users in a smartroom scenario is presented. The system relies on a variety of sensors and processing units distributed over a network of computers. Visual tracks gained from a wide angle top view camera and speaker localization cues delivered by a combination of microphone arrays are

Fig. 4. Example scenario involving 3 users in our smartroom

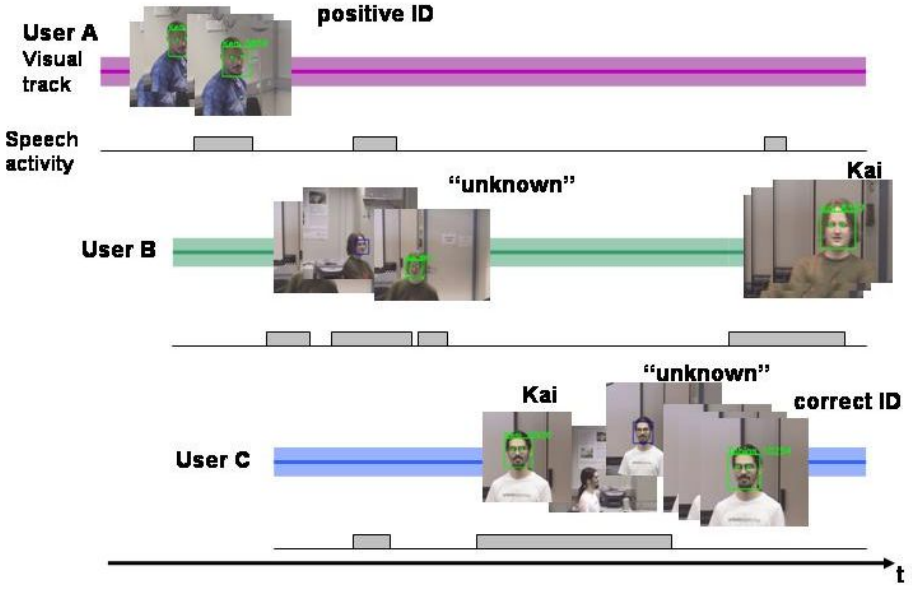

Fig. 5. Example identification scenario. The colored bars represent the visual person tracks in time. Below them, the speech activity for the respective users is depicted. The system immediately recognizes user A from a series of high quality face shots. As user B speaks, attention switches, but the system cannot capture good enough images for definite identification. Confidence is low and the user is marked as unknown until, later on, better views can be obtained. User C is first misrecognized as "Kai". As correct ID results eventually come in, though, the fusion module revises its confidence for user C, labeling him as unknown, until good enough confidence can be reached to make a definite and correct ID

used to sequentially select persons of interest. Active cameras mounted on the room walls are used to focus in on the speakers and gain high quality closeup views. A two stage algorithm detects frontal faces in the captured images and aligns them with respect to the eye positions. Parallel processes recognize the cropped faces through a local appearance based method using DCT coefficients. The resulting IDs are sent back to a fusion module, which realizes spatial and temporal alignment, accumulates confidences for person tracks, recognizes unknown persons and gradually updates all known person identities. The system actively seeks good cues for identification in an unobtrusive way, requires no specific interaction with the users and functions in realtime at a framerate of 15fps.

Future efforts will go into improving the performance of the various system components: Adding more features to the person tracking module to increase robustness; improving the smoothness of camera control and the hit rate of the face detectors; extending the face ID modules to recognize profile views; etc.

Another planned improvement is the inclusion of acoustic speaker identification cues. This would allow to recognize speakers even if no good facial view can be acquired for long periods of time.

It should also be worthwhile to experiment with other priority management and camera selection strategies, e.g. decoupling cameras or providing feedback from the active camera images to steer the selection and control process.

Finally, complementing the system to allow it to automatically recognize, cluster, and train in unknown faces in an unsupervised way will be the next big step towards making it generally usable for a wide range of real life applications.

6 Acknowledgement

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).

References

1. Rania Y. Khalaf and Stephen S. Intille, *"Improving Multiple People Tracking using Temporal Consistency"*, Massachusetts Institute of Technology, Cambridge, MA, MIT Dept. of Architecture House_n Project Technical Report, 2001.
2. Rainer Lienhart and Jochen Maydt, *"An Extended Set of Haar-like Features for Rapid Object Detection"*. IEEE ICIP 2002, Vol. 1, pp. 900–903, Sep. 2002.
3. Paul Viola and Michael Jones, *"Rapid Object Detection using a Boosted Cascade of Simple Features"*. Accepted Conference On Computer Vision And Pattern Recognition, 2001.
4. Iain McCowan, Daniel Gatica-Perez, Samy Bengio, Guillaume Lathoud, Mark Barnard, Dong 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.
5. R. Stiefelhagen, *"Tracking Focus of Attention in Meetings"*. IEEE International Conference on Multimodal Interfaces - ICMI 2002, pp. 273-280, Pittsburgh, 2002.
6. M. Voit, K. Nickel, R. Stiefelhagen, *"Multi-view Head Pose Estimation using Neural Networks"*. Second Workshop on Face Processing in Video (FPiV'05), in association with IEEE Second Canadian Conference on Computer and Robot Vision (CRV 2005), 9-11 May 2005, Victoria, BC, Canada.
7. Tanzeem Choudhury, Brian Clarkson, Tony Jebara and Alex Pentland, *"Multimodal Person Recognition using Unconstrained Audio and Video"*. Second Conference on Audio- and Video-based Biometric Person Authentication '99 (AVBPA '99), pages 176-181, Washington DC
8. Jie Yang, Xiaojin Zhu, Ralph Gross, John Kominek, Yue Pan, Alex Waibel, *"Multimodal people ID for a multimedia meeting browser"*. Proceedings of the 7th ACM International Conference on Multimedia '99, Orlando, FL
9. Shinji Tsuruoka, Toru Yamaguchi, Kenji Kato, Tomohiro Yoshikawa, Tsuyoshi Shinogi, *"A Camera Control Based Fuzzy Behaviour Recognition of Lecturer for Distance Lecture"*. Proceedings of the 10th IEEE International Conference on Fuzzy Systems, December 2001, Melbourne, Australia.
10. P. Peixoto, J. Batista, H. Araujo, *"A surveillance system combining peripheral and foveated motion tracking"*. Proceedings of the Fourteenth International Conference on Pattern Recognition. Volume 1, 16-20 Aug. 1998 Page(s):574 - 577 vol.1
11. Arun Hampapur, Sharath Pankanti, Andrew W. Senior, Ying-li Tian, Lisa Brown, Ruud 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.
12. 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 Institute of Technology, Graphics, Visualization, and Usability Center, 1998.
13. 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.
14. H.K.Ekenel, R. Stiefelhagen, *"Local Appearance based Face Recognition Using Discrete Cosine Transform"*. 13th European Signal Processing Conference (EUSIPCO), Antalya Turkey, September 2005.
15. H.K.Ekenel, R. Stiefelhagen, *"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.
16. CHIL - Computers In the Human Interaction Loop, http://chil.server.de
17. AMI - Augmented Multiparty Interaction, http://www.amiproject.org
18. OpenCV - Open Computer Vision Library, http://sourceforge.net/projects/opencvlibrary