Multi- and Single View Multiperson Tracking for Smart Room Environments

Keni Bernardin[1], Tobias Gehrig[1], and Rainer Stiefelhagen[1]

Interactive Systems Lab
Institut für Theoretische Informatik
Universität Karlsruhe, 76131 Karlsruhe, Germany
{keni, tgehrig, stiefel}@ira.uka.de

Abstract. Simultaneous tracking of multiple persons in real world environments is an active research field and several approaches have been proposed, based on a variety of features and algorithms. In this work, we present 2 multimodal systems for tracking multiple users in a smart room environment. One is a multi-view tracker based on color histogram tracking and special person region detectors. The other is a wide angle overhead view person tracker relying on foreground segmentation and model-based tracking. Both systems are completed by a joint probabilistic data association filter-based source localization framework using input from several microphone arrays.
We also very briefly present two intuitive metrics to allow for objective comparison of tracker characteristics, focusing on their precision in estimating object locations, their accuracy in recognizing object configurations and their ability to consistently label objects over time.
The trackers are extensively tested and compared, for each modality separately, and for the combined modalities, on the CLEAR 2006 Evaluation Database.

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 [17, 18] or the VACE project in the U.S. [19], 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 gain more and more popularity. An essential building block for complex scene analysis is the detection and tracking of persons in the scene.

One of the major problems faced by indoor tracking systems is the lack of reliable features that allow to keep track of persons in natural, evolving and unconstrained scenarios. The most popular visual features in use are color features and foreground segmentation or movement features [2, 1, 3, 6, 7, 16], each with their advantages and drawbacks. Doing e.g. blob tracking on background subtraction maps is error-prone, as it requires a clean background and assumes only persons are moving. In real environments, the foreground blobs are often fragmented or merged with others, they depict only parts of occluded persons or are produced by shadows or displaced objects. When using color information to track people, the problem is to create appropriate color histograms or models. Generic color models are usually sensitive and environment-specific [4]. If no generic model is used, one must at some point decide which pixels in the image belong to a person to initialize a dedicated color histogram [3, 7, 15, 16]. In many cases, this still requires the cooperation of the users and/or a clean and relatively static background. On the acoustic side, although actual techniques already allow for a high accuracy in localization, they can still only be used effectively for the tracking of one person, and only when this person is speaking. This naturally leads to the development of more and more multimodal techniques.

Here, we present two multimodal systems for the tracking of multiple persons in a smart room scenario. A joint probability data association filter is used to in conjunction with a set of microphone arrays to determine active speaker positions. For the video modality, we investigate the advantages and drawbacks of 2 approaches, one relying on color histogram tracking in several corner camera images and subsequent triangulation, and one relying on foreground blob tracking in wide angle top view images. For both systems, the acoustic and visual modalities are fused with a state-based selection and combination scheme on the single modality tracker outputs. The systems are evaluated on the CLEAR'06 3D Multiperson Tracking Database, and compared using the MOTP and MOTA metrics, which will also be briefly decribed.

The next sections introduce the multi-view and single-view visual trackers, and the jpdaf-based acoustic tracker. Section 6 gives a brief explanation of the used metrics. Section 7 shows the evaluation results on the CLEAR database, while section 8 gives a brief summary and concludes.

2 Multi-View Person Tracking using Color Histograms and Haar-Classifier Cascades

The developed system is a 3D tracker that uses several fixed cameras installed at the room corners [11]. It is designed to function with a variable number of cameras, with precision increasing as the number of cameras grows. It performs tracking first separately on each camera image, using color histogram models. Color tracks are initialized automatically using a combination of foreground maps and special object detectors. The information from several cameras is then fused to produce 3D hypotheses of the persons' positions. A more detailed explanation of the system's different components is given in the following.

2.1 Classifier Cascades and Foreground Segmentation

A set of special object detectors is used to detect persons in the camera images. They are classifier cascades that build on haar-like features, as decribed in [9, 8]. For our implementation, the cascades were taken from the OpenCV [20] library. Two types of cascades are used: One trained to recognize frontal views of faces(*face*), and one to recognize the upper body region of standing or sitting persons (*upper body*). The image is scanned at different scales and bounding rectangles are obtained for regions likely to contain a person. By using these detectors, we avoid the drawbacks of creation/deletion zones and are able to initialize or recover a track at any place in the room.

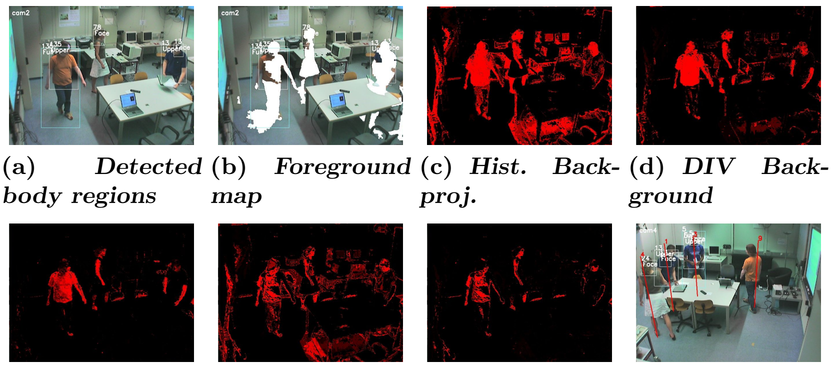
Further, to reduce the amount of false detector hits, a preprocessing step is made on the image. It is first segmented into foreground regions by performing background subtraction using an adaptive background model. The foreground regions are then scanned using the classifier cascades. This combined approach offers two advantages: The cascades, on the one hand, increase robustness to segmentation errors, as foreground regions not belonging to persons, such as moved chairs, doors, shadows, etc, are ignored. The foreground segmentation, on the other hand, helps to decide which of the pixels inside a detection rectangle belong to a person, and which to the background. Knowing exactly which pixels belong to the detected person is useful to create accurate color histograms and improve color tracking performance.

2.2 Color Histogram Tracking and 2D Hypotheses

Whenever an object detector has found an upper or a full body in the image, a color histogram of the respective person region is constructed from the foreground pixels belonging to that region, and a track is initialized. The actual tracking is done based only on color features by using the meanshift algorithm [5] on histogram backprojection images. Care must be taken when creating the color histograms to reduce the negative effect of background colors that may have been mistakenly included in the person silhouette during the detection and segmentation phase. This is done by histogram division, as proposed in [12]. Several types of division are possible (division by a general background histogram, by the histogram of the background region immediately surrounding the person, etc, see Fig. 1). The choice of the best technique depends on the conditions at hand and is made automatically at each track initialization step, by making a quick prediction of the effect of each technique on the tracking behavior in the next frame.

To ensure continued tracking stability, the histogram model for a track is also adapted every time a classifier cascade produces a detection hit on that track. Tracks that are not confirmed by a detection hit for some time are deleted, as they are most likely erroneous.

The color based tracker, as described above, is used to produce a 2D hypothesis for the position of a person in the image. Based on the type of cascade that triggered initialization of the tracker, and the original size of the detected region, the body center of the person in the image and the person's distance from the camera are estimated and output as hypothesis. When several types of trackers (face and upper body) are available for the same person, a combined output is produced.



(a) ***Detected body regions*** (b) ***Foreground map*** (c) ***Hist. Back-proj.*** (d) ***DIV Background***

(e) ***DIV Background***^{2} (f) ***DIV Border*Backg*** (g) ***DIV Border*Backg***^{2} (h) ***Tracker output***

Fig. 1. Color histogram creation, filtering and tracking. *a)* Face, upper and full body detections (*rectangles*) in one camera view. *b)* Foreground segmentation (in *white*). Only foreground pixels inside the rectangles are used. *c)* Histogram backprojection for the upper body track of the leftmost person. *d), e), f) and g)* Effects of different types of histogram division. *Background*: Overall background histogram. *Border*: Histogram of the background region immediately surrounding the detected rectangle. *h)* Tracker output as seen from another view

2.3 Fusion and Generation of 3D Hypotheses

The 2D hypotheses produced for every camera view are triangulated to produce 3D position estimates. For this, the cameras must be calibrated and their position relative to a general room coordinate system known. The lines of view (LOV) coming from the optical centers of the cameras and passing through the 2D hypothesis points in their respective image planes are intersected. When no exact intersection point exists, a residual distance between LOVs, the triangulation error, can be calculated. This error value is used by an intelligent 3D tracking algorithm to establish likely correspondences between 2D tracks (as in [13]). When the triangulation error between a set of 2D hypotheses is small enough, they are associated to form a 3D track. Likewise, when it exceeds a certain threshold, the 2D hypothesis which contributes most to the error is dissociated again and the 3D track is maintained using the remaining hypotheses. The tracker requires a minimum of 2 cameras to produce 3D hypotheses, and becomes more robust as the number of cameras increases.

Once a 3D estimate for a person's position has been computed, it is further used to validate 2D tracks, to initiate color histogram tracking in camera views where the person has not yet been detected, to predict occlusions in a camera view and deactivate the involved 2D trackers, and to reinitialize tracking even in the absence of detector hits.

The developed multiperson tracker draws its strength from the intelligent fusion of several camera views. It initializes its tracks automatically, constantly adapts its color models and verifies the validity of its tracks through the use of special object detectors. It is capable of tracking several people, regardless if they are sitting, moving or standing still, in a cluttered environment with uneven lighting conditions.

3 Single-View Model-Based Person Tracking on Panoramic Images

In contrast to the above presented multi-view system, a single-view tracker working on wide angle images captured from the top of the room was also designed. The advantage of such images 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 person-specific features are hard to observe.

The tracking algorithm is essentially composed of a simple but fast 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 initial images of the 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 model radius was estimated automatically using the calibration information for the wide angle camera and rough knowledge about the room height. The procedureis as follows:

- For all person models M_i, verify their updated positions $(x, y)_{M_i}$. If the overlap between two models exceeds a maximum value, fuse them.
- 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 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 assigned 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. Make the model active, only if it subsist for a minimum period of time. On the other hand, if a model stays unassigned for a certain period of latency, delete it.
- 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.

The system makes no assumptions about the environment, e.g. no special creation or deletion zones, about the consistency of a person's appearance or the surrounding room. It runs in realtime, at 15fps, on a Pentium 3GHz machine.

4 A JPDAF Source Localizer for Speaker Tracking

In parallel to the visual tracking of all room occupants, acoustic source localization was performed 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 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, Kalman or particle-filter based, this approach uses a Joint Probabilistic Data Association Filter that directly receives as input the correlation results from the various microphone pairs, and performs the tracking in a unified probabilistic way for multiple possible target hypotheses, thereby achieving more robust and accurate results. The details of the source localizer can be found in [10].

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 a combined hypothesis is produced.

5 State-Based Fusion

The fusion of the audio and video modalities is done at the decision level. Track estimates coming from the visual and acoustic tracking systems are combined using a finite state machine approach, which considers the relative strengths and weaknesses of each modality. The visual trackers are generally very accurate at determining a person's position. In multiperson scenarios they can, however, miss persons completely because their faces are too small or invisible, or because they are not well discernable from the background by color, shape or motion. The acoustic tracker on the other hand can precisely determine a person's position when this person speaks. In the current implementation, it can, however, only track one active speaker at a time can produce no estimates when several or no persons are speaking.

Based on this, the fusion of the acoustic and visual tracks is made using a finite state machine weighing the availability or reliability of the single modality tracks.

For multimodal tracking, two main conditions are to be evaluated: For condition A, only the position of the active speaker in a multi-participant scenario is to be estimated. Fo condition B, on the other hand, all participants have to be tracked. Consequently, the states for the fusion of modalities differ slightly depending on the task condition. For condition A, they are as follows:

- State 1: An acoustic estimate is available, for which no overlapping visual estimate exists. Here, estimates are considered overlapping if their distance is smaller than 500mm. In this case, assume the visual tracker has missed the speaking person and ouptut the acoustic hypothesis. Store the last received acoustic estimate and keep outputting it until an overlapping visual estimate is found.
- State 2: An acoustic estimate is available, and a corresponding visual estimate exists. In this case, output the average of the acoustic and visual positions.
- State 3: After an overlapping visual estimate had been found, an acoustic estimate is no longer available. In this case, we consider the visual tracker has recovered the previously undetected speaker and keep ouputting the position of the last overlapping visual track.

For condition B, where all participants must be tracked, the acoustic estimate serves to increase the precision of the closest visual track, whenever available. The states are:

– State 1: An acoustic estimate is available, for which no overlapping visual estimate exists. In this case, assume the visual tracker has missed the speaking person and ouptut the acoustic hypothesis additionally to the visual ones. Store the last received acoustic estimate and keep outputting it until an overlapping visual estimate is found.
– State 2 and State 3 are similar to condition A, with the exception that here, all other visual estimates are output as well.

Using this fusion scheme, two multimodal tracking systems were designed: System1, fusing the jpdaf acoustic tracker with the single-view visual tracker, and System2, fusing it with the multi-view tracker. Both systems were evaluated on conditions A and B, and the results compared in section 7. To allow better insight in the evaluation scores, the following section first gives a brief overview of the used metrics.

6 Multiple Object Tracking Metrics

Defining good measures to express the characteristics of a system for continuous tracking of multiple objects is not a straightforward task. Various measures exist and there is no consensus in the literature on the best set to use. Here, we propose a small expressive set of metrics and show a systematic procedure for their calculation. A more detailed discussion of these metrics can be found in [14].

Assuming that for every time frame t a multiple object tracker outputs a set of hypotheses $\{h_1 \dots h_m\}$ for a set of visible objects $\{o_1 \dots o_n\}$, we define the procedure to evaluate its performance as follows:

Let the correspondence between an object o_i and a hypothesis h_j be valid only if their distance $dist_{i,j}$ does not exceed a certain threshold T, and let $M_t = \{(o_i, h_j)\}$ be a dynamic mapping of object-hypothesis pairs.

Let $M_0 = \{\}$. For every time frame t,

1. For every mapping (o_i, h_j) in M_{t-1}, verify if it is still valid. If object o_i is still visible and tracker hypothesis h_j still exists at time t, and if their distance does not exceed the threshold T, make the correspondence between o_i and h_j for frame t.
2. For all objects for which no correspondence was made yet, try to find a matching hypothesis. Allow only one to one matches. To find optimal correspondences that minimize the overall distance error, Munkre's algorithm is used. Only pairs for which the distance does not exceed the threshold T are valid. If a correspondence (o_i, h_k) is made that contradicts a mapping (o_i, h_j) in M_{t-1}, replace (o_i, h_j) with (o_i, h_k) in M_t. Count this as a mismatch error and let mme_t be the number of mismatch errors for frame t.
3. After the first two steps, a set of matching pairs for the current time frame is known. Let c_t be the number of matches found for time t. For each of theses matches, calculate the distance d_t^i between the object o_i and its corresponding hypothesis.

4. All remaining hypotheses are considered false positives. Similarly, all remaining objects are considered misses. Let fp_t and m_t be the number of false positives and misses respectively for frame t. Let also g_t be the number of objects present at time t.
5. Repeat the procedure from step 1 for the next time frame. Note that since for the initial frame, the set of mappings M_0 is empty, all correspondences made are initial and no mismatch errors occur.

Based on the matching strategy described above, two very intuitive metrics can be defined: The Multiple Object Tracking Precision ($MOTP$), which shows the tracker's ability to estimate precise object positions, and the Multiple Object Tracking Accuracy ($MOTA$), which expresses its performance at estimating the number of objects, and at keeping consistent trajectories:

$$MOTP = \frac{\sum_{i,t} d_{i,t}}{\sum_t c_t} \tag{1}$$

$$MOTA = 1 - \frac{\sum_t (m_t + fp_t + mme_t)}{\sum_t g_t} \tag{2}$$

The $MOTA$ can be seen as composed by 3 error ratios:

$$\overline{m} = \frac{\sum_t m_t}{\sum_t g_t}, \quad \overline{fp} = \frac{\sum_t fp_t}{\sum_t g_t}, \quad \overline{mme} = \frac{\sum_t mme_t}{\sum_t g_t},$$

the ratio of misses, false positives and mismatches in the sequence, computed over the total number of objects present in all frames.

Alternatively, to compare systems for which measurement of identity mismatches is not meaningful, an addtitional measure, the $A - MOTA$ can be computed, by ignoring mismatch errors in the global error computation:

$$A - MOTA = 1 - \frac{\sum_t (m_t + fp_t)}{\sum_t g_t} \tag{3}$$

7 Evaluation on the CLEAR'06 3D Multiperson Tracking Database

The above presented systems for visual and multimodal tracking were evaluated on the CLEAR'06 3D Multiperson Tacking Database. This database comprises recordings from 3 different CHIL smartrooms, involving up to 6 persons in a seminar scenario, for a total of approx. 60 min.

Tables 1 and 2 show the results for the Single- and Multi-view based systems, System1 and System2, for the visual and the mutimodal conditions A and B:

As Table 1 shows, the single view tracker clearly outperforms the multi-view approach. As the scenario involved mostly people sitting around a table and occasionally walking, they were very clearly distinguishable from a top view, even when using simple features such as foreground blobs for tracking. The

Table 1. Evalution results for the visual and multimodal B conditions

System	$MOTP$	$\overline{m}$	$\overline{fp}$	$\overline{mme}$	$MOTA$
1:Visual	217mm	27.6%	20.3%	1.0%	51.1%
1:AV CondB	226mm	26.1%	20.8%	1.1%	52.0%
2:Visual	203mm	46.0%	24.9%	2.8%	26.3%
2:AV CondB	223mm	44.4%	25.8%	3.3%	26.4%

Table 2. Evalution results for the multimodal A condition

System	$MOTP$	$\overline{m}$	$\overline{fp}$	$\overline{mme}$	$MOTA$
1:AV CondA	223mm	51.4%	51.4%	2.1%	-5.0%
2:AV CondA	179mm	51.4%	51.4%	5.3%	-8.2%

multi-view approach, on the other hand, had more moderate results, stemming from the considerably more difficult video data. The problems can be summed up in 2 categories:

– 2D tracking errors: In several seminars, participants were only hardly distinguishable from the background using color information, or detectable by the face and body detectors, due to low resolution. This accounts for the relatively high amount of missed persons.
– Triangulation errors: The low angle of view of the corner cameras and the small size of most recording rooms caused a considerable amount of occlusion in most seminars, which could not be completely resolved by the triangulation scheme. A more precise distance estimation, based on the size of detection hits could help avoid many of the occured triangulation errors, and reduce the false positive count.

In all cases, the average MOTP error was about 20cm, making the MOTA the more interesting metric for comparison. As can also be seen, although the addition of the acoustic modality could bring a slight improvement in tracking accuracy, the gain is minimal, as it could only help improve tracking performance for the speaking person at each respective point in time.

Compared to these results, the scores for condition A are relatively low. Both systems produced a high amount of miss errors (around 50%), as the correct speaker could not be selected from the multiple available tracks. It is noticeable that in case the correct speaker was tracked, though, the multi-view System2 achieved a higher precision, reaching 18cm, as compared to 20cm for System1. This suggest that for the tracking of clearly identifiable persons (such as the presenter in the seminars), the multi-view, face and body-detector based approach does have its advantages.

8 Summary

In this work, 2 systems for multimodal tracking of multiple users is presented. A joint probabilistic data association filter for source localization is used in conjunction with 2 distinct systems for visual tracking: One using multiple camera images, based on color histogram tracking and haar-feature classifier cascades for upper bodies and faces. The other using only a wide angle overhead view, and model based tracking on foreground segmentation features. A fusion scheme is presented, using a 3-state finite-state machine to combine the output of the audio and visual trackers. The systems were extensively tested on the CLEAR 2006 3D Multiperson Tracking Database, for the visual and the audio-visual conditions A and B.

The results show that under fairly controlled conditions, as can be expected of meeting situations with relatively few participants, an overhead wide angle view analysis can yield considerable advantages over more elaborate multicamera systems, even if only simple features, such as foreground blobs are used. Overall, an accuracy of 52% could be reached for the audio-visual task, with position errors below 23cm.

9 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).

References

1. Rania Y. Khalaf and Stephen S. Intille, *"Improving Multiple People Tracking using Temporal Consistency"*, MIT Dept. of Architecture House_n Project Technical Report, 2001.
2. Wei Niu, Long Jiao, Dan Han, and Yuan-Fang Wang, *"Real-Time Multi-Person Tracking in Video Surveillance"*, Pacific Rim Multimedia Conference, Singapore, 2003.
3. A. Mittal and L. S. Davis, *"M2Tracker: A Multi-View Approach to Segmenting and Tracking People in a Cluttered Scene Using Region-Based Stereo"*, European Conf. on Computer Vision, LNCS 2350, pp. 18-33, 2002.
4. Neal Checka, Kevin Wilson, Vibhav Rangarajan, Trevor Darrell, *"A Probabilistic Framework for Multi-modal Multi-Person Tracking"*, Workshop on Multi-Object Tracking (CVPR), 2003.
5. Dorin Comaniciu and Peter Meer, *"Mean Shift: A Robust Approach Toward Feature Space Analysis"*. IEEE PAMI, Vol. 24, No. 5, May 2002.
6. Ismail Haritaoglu, David Harwood and Larry S. Davis, *"W4: Who? When? Where? What? A Real Time System for Detecting and Tracking People"*. Third Face and Gesture Recognition Conference, pp. 222–227, 1998.
7. Yogesh Raja, Stephen J. McKenna, Shaogang Gong, *"Tracking and Segmenting People in Varying Lighting Conditions using Colour"*. 3rd. Int. Conference on Face & Gesture Recognition, pp. 228, 1998.

8. Paul Viola and Michael Jones, *"Rapid Object Detection using a Boosted Cascade of Simple Features"*. IEEE Int. Conference On Computer Vision And Pattern Recognition, 2001.
9. 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.
10. T. Gehrig, J. McDonough, *"A Joint-Probabilistic Data Association Filter Based Source Localization Technique"*. CLEAR 2006.
11. Alexander Elbs, *"Mehrpersonentracking mittels Farbe und Detektorkaskaden"*. Diplomarbeit. Institut für Theoretische Informatik, Universität Karlsruhe, August 2005.
12. Kai Nickel and Rainer Stiefelhagen, *"Pointing Gesture Recognition based on 3Dtracking of Face, Hands and Head Orientation"*, 5th International Conference on Multimodal Interfaces, Vancouver, Canada, Nov. 2003.
13. Dirk Focken, Rainer Stiefelhagen, *"Towards Vision-Based 3-D People Tracking in a Smart Room"*, IEEE International Conference on Multimodal Interfaces, Pittsburgh, PA, USA, October 14-16, 2002, pp. 400-405.
14. Keni Bernardin, Alexander Elbs and Rainer Stiefelhagen, *"Multiple Object Tracking Performance Metrics and Evaluation in a Smart Room Environment"*, accepted Sixth IEEE International Workshop on Visual Surveillance, in conjunction with ECCV2006, May 13th 2006, Graz, Austria
15. Hai Tao, Harpreet Sawhney and Rakesh Kumar, *"A Sampling Algorithm for Tracking Multiple Objects"*. International Workshop on Vision Algorithms: Theory and Practice, pp. 53–68, 1999.
16. Christopher Wren, Ali Azarbayejani, Trevor Darrell, Alex Pentland, *"Pfinder: Real-Time Tracking of the Human Body"*. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol 19, no 7, pp. 780–785, July 1997.
17. CHIL - Computers In the Human Interaction Loop, http://chil.server.de
18. AMI - Augmented Multiparty Interaction, http://www.amiproject.org
19. VACE - Video Analysis and Content Extraction, http://www.ic-arda.org
20. OpenCV - Open Computer Vision Library, http://sourceforge.net/projects/opencvlibrary/