Multi-Level Particle Filter Fusion of Features and Cues for Audio-Visual Person Tracking

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. In this paper, two multimodal systems for the tracking of multiple users in smart environments are presented. The first is a multi-view particle filter tracker using foreground, color and special upper body detection and person region features. The other is a wide angle overhead view person tracker relying on foreground segmentation and model-based blob tracking. Both systems are completed by a joint probabilistic data association filter-based source localizer using the input from several microphone arrays.
The systems are designed to estimate the 3D scene locations of room occupants and are evaluated based on their precision in estimating person locations, their accuracy in recognizing person configurations and their ability to consistently keep track identities over time.
The trackers are extensively tested and compared, for each separate modality and for the combined modalities, on the CLEAR 2007 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 [20, 21] or the VACE project in the U.S. [22], 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 obtain rich scene information gain more and more popularity. An essential building block for complex scene analysis is the detection and tracking of persons.

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, unconstrained scenarios. The most popular visual features in use are color features and foreground segmentation or movement features [2, 1, 3, 6, 7, 14], 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, the problem is to find appropriate color models for tracking. Generic color models are usually sensitive and environment-specific [4]. If no generic model is used, color models for tracked person need to be initialized automatically at some point [3, 7, 13, 14]. 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, while 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 in conjunction with a set of microphone arrays to detect speech and determine active speaker positions. For the video modality, we compare the performance of 2 approaches: A particle filter approach using several cameras and a variety of features, and a simple blob tracker relying on foreground segmentation features gained from a wide angle top view. While the former system fuses the acoustic and visual modalities at the feature level, the latter does this at the decision level using a state-based selection and combination scheme on the single modality tracker outputs. All systems are evaluated on the CLEAR'07 3D Person Tracking Database.

The next sections introduce the multimodal particle filter tracker, the single-view visual tracker, the JPDAF-based acoustic tracker, as well as the fusion approach for the single view visual and the acoustic tracking systems. Section 6 shows the evaluation results on the CLEAR'07 database and section 7 gives a brief summary and conclusion.

2 Multimodal Particle Filter-Based 3D Person Tracking

The multimodal 3D tracking component is a particle filter using features and cues from the four room corners cameras and the wide angle ceiling camera, as well as source localization hypotheses obtained from the room's microphone arrays. The tracker automatically detects and tracks multiple persons without requiring any special initialization phase or area, room background images, or a-priori knowledge about person colors or attributes, for standing, sitting or walking users alike.

2.1 Tracking Features

The features used are adaptive foreground segmentation and upper body color backprojection maps 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 substracted from the current frame and a fixed threshold is applied to detect foreground regions.
– The color features are computed in a modified HSV space and modeled in a specially designed histogram representation, which eliminates the usual drawbacks of HSV histograms when it comes to modeling low saturation or variance colors.
The color space is a modified version of the HSV cone. First, colors for which the variance and saturation values exceed 20% are set to maximum variance. This reduces the effect of illumination, shadows or pose changes. The HSV values are subsequently discretized as follows: Let hue, sat and var be the obtained HSV values, then the corresponding histogram bin values, h, s and v, are computed as:

$$v = var \tag{1}$$

$$s = sat * var \tag{2}$$

$$h = hue * sat * var \tag{3}$$

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 bin to model colors with zero variance, in contrast to classial discretuzation techniques, where there is no unique bin mapping for grayscale HSV values. A maximum of 16 bins for hue, 10 bins for saturation, and 10 for variance are used.
The color features for tracking of a person are gained for the detected upper body region of subjects, as well as their immediatly 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 valid color histograms for the corner cameras exist, the upper body histogram for the top camera is continuously adapted using colors sampled from a 60cm diameter region centered around the tracked person position. This is because the track is only considered reliable enough for continous color model adaptation after it has been confirmed at least once by an upper body detection. The background histograms for all views, in turn, 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:

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

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 immmediately 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 [8, 9]. Using calibration information, the 3D scene coordinates of the detected upper body as well as the localization uncertainty, expressed as covariance matrix, are computed from the detection window position and size. This information is later used to associate detections to person tracks and to score particles.
- Person regions are found in the top camera images through the analysis of foreground blobs, as described in [12]. 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 within a 60cm diameter region in a certain time interval. The motivation is that top view images present very little overlap between persons, making a simple spatial assignment plausible.

2.2 Intialization 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 the particle cloud, calculated as the variance in particle positions, is above a fixed activation threshold.
- 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.
- The target object is sufficiently dissimilar to its surrounding background in every view. Again, the bhattacharyya distance is used to measure similarity between the computed track histograms and the corresponding background histograms. 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 particles, calculated as the variance in their positions, exceeds a fixed limit.

2.3 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 real-time computation speed. The system was extensively tested and achieved high accuracy rates, as shown in section 6.

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

In contrast to the above presented system, the panoramic camera tracker relies solely on the wide angle images captured from the top of the room. 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 (see Fig. 1). This system has already been tested in the CLEAR 2006 evaluation, and is described in detail in [12]. No modifications to the system or its parameters, and no tuning on the 2007 data was done, to provide for an accurate baseline for comparisons. The following gives only a brief system overview.

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.

At 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 substraction, thresholding and morphological filtering provide the foreground blobs for tracking.

The subsequent EM tracking algorithm tries to find an optimal assignment of the detected blobs to a set of active person models. Person models are composed of an image position (x, y), velocity (vx, vy), radius r and a track ID, and are instantiated or deleted based on the foreground blob support observed over a certain time window.

The approach results in a simple but fast tracking algorithm that is able to maintain several person tracks, even in the event of moderate overlap. By assuming an average height of 1m for a person's body center, and using calibration

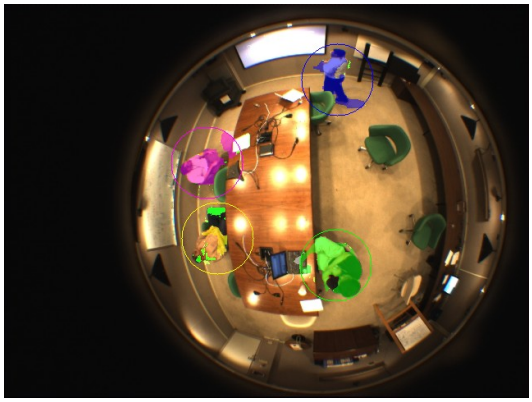

Fig. 1. The output of the top camera tracker. The colored circles represent the person models

information for the top camera, the positions in the world coordinate frame of all 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 about the recording room. It runs at a realtime factor of 0.91, at 15fps, on a Pentium 3GHz machine.

4 JPDAF-based Acoustic Source Localization

The acoustic source localization system is based on a joint probabilistic data association filter (JPDAF) [15, 19]. This is an extension to the IEKF used in previous approaches [18], that makes it possible to track multiple targets at once and updates each of the internally maintained IEKFs probabilistically. It also introduces a "clutter model" that models random events, such as door slams, footfalls, etc., that are not associated with any speaker, but can cause spurious peaks in the GCC of a microphone pair, and thus lead to poor tracking performance. Observations assigned with high probability to the clutter model do not affect the estimated positions of the active targets.

First of all the timedelays and corresponding correlation values are calculated for all possible microphone pairs within each of the T-arrays and 14 pairs of the available MarkIIIs by calculating the GCC-Phat [16, 17] of the frequencies below 8000 Hz. The timedelays are estimated 25 times per second resulting in a hamming window size of 0.08 ms with a shift size of 0.04ms. For the GCC we used a FFT size of 4096 points. The maximum search in the resulting correlation function is restricted to be in the valid range of values as conditioned on the room size and the microphone positions.

Then, for each of the seminars and each of the used microphone pairs, a correlation threshold is estimated separately by calculating the histogram of all

correlation values of that pair and seminar and using the value that is at 85% of it, i.e. the smallest value that is greater than 85% of the correlation values.

The JPDAF is then fed with one measurement vector for each time instant and microphone array that is made up of the TDOAs of those microphone pairs of that array with a correlation higher than the previously estimated one. As observation noise we used 0.02ms. The measurement vector is only used for position estimation if it has at least 2 elements.

The first step in the JPDAF algorithm is the evaluation of the conditional probabilities of the *joint association events*

$$\boldsymbol{\theta}(t) = \bigcap_{i=1}^{m_t} \theta_{ik_i}, \quad t = 0, \ldots, T \tag{5}$$

where the atomic events are defined as

$$\theta_{ik} = \{\text{observation } i \text{ originated from target } k\} \tag{6}$$

Here, k_i denotes the index of the target to which the i-th observation is associated in the event currently under consideration. In our case we chose the maximum number of targets to be $T \leq 3$ and the maximum number of measurements per step to be $m_t \leq 1$. From all the theoretically possible events only *feasible events* are further processed. A feasible event is defined as an event wherein

1. An observation has exactly one source, which can be the clutter model;
2. No more than one observation can originate from any target.

An observation is possibly originating from a target when it falls inside the target's validation region given by the innovation covariance matrix and a gating threhold of 4.0.

Applying Bayes' rule, the conditional probability of $\boldsymbol{\theta}(t)$ can be expressed as

$$P\{\boldsymbol{\theta}(t)|\mathcal{Y}_t\} = \frac{P\{\mathbf{Y}(t)|\boldsymbol{\theta}(t), \mathcal{Y}_{t-1}\}P(\boldsymbol{\theta}(t)}{P\{\mathbf{Y}(t)|\mathcal{Y}_{t-1}\}} \tag{7}$$

where the marginal probability $P\{\mathbf{Y}(t)|\mathcal{Y}_{t-1}\}$ is computed by summing the joint probability in the numerator of (7) over all possible $\boldsymbol{\theta}(t)$. The conditional probability of $\mathbf{Y}(t)$ required in (7) can be calculated from

$$P\{\mathbf{Y}(t)|\boldsymbol{\theta}(t), \mathcal{Y}_{t-1}\} = \prod_{i=1}^{m_t} p(\mathbf{y}_i(t)|\theta_{ik_i}(t), \mathcal{Y}_{t-1}) \tag{8}$$

The individual probabilities on the right side of (8) can be easily evaluated given the fundamental assumption of the JPDAF, namely,

$$\mathbf{y}_i(t) \sim \mathcal{N}(\hat{\mathbf{y}}_{k_i}(t|\mathcal{Y}_{t-1}), \mathbf{R}_{k_i}(t)) \tag{9}$$

where $\hat{\mathbf{y}}_{k_i}$ and $\mathbf{R}_{k_i}(t)$ are, respectively, the predicted observation and innovation covariance matrix for target k_i. The prior probability $P\{\boldsymbol{\theta}(t)\}$ in (7) can be

readily evaluated through combinatorial arguments [15, §9.3] using a detection probability of 85%. Once the posterior probabilities of the joint events $\{\boldsymbol{\theta}(t)\}$ have been evaluated for all targets together, the state update for each target can be made separately according to the update rule of the PDAF [15, §6.4].

As the JPDAF can track multiple targets, it was necessary to formulate rules for deciding when a new track should be created, when two targets should be merged and when a target should be deleted. The JPDAF is initially started with no target at all. A new target is created every time a measurement can not be assigned to any previously existing target. A new target is always initialized with a start position in the middle of the room and a height of 163.9cm and a diagonal state error covariance matrix with a standard deviation that is essentially the size of the room for x and y and 1m for z. This initialization is allowed to take only 0.1s otherwise the target is immediately deleted. The initialization is said to be finished when the target is detected as active. This is when the volume of the error ellipsoid given by the state error covariance matrix is smaller than a given threshold. If a target didn't receive any new estimates for 5s, it is labeled as inactive and deleted. If two targets are less than 25cm apart from each other for at least 0.5s the target with the larger error volume is deleted.

To allow speaker movement, the process noise covariance matrix is dynamically set to a multiple of the squared time since the last update. For stability reasons, the process noise as well as the error state covariance matrix are upper bounded.

Since the filters used for each of the targets is built on top of the IEKF, there are at most 5 local iterations for each update.

The selection of the active speaker out of the maintained targets is done by choosing the target with the smallest error volume that has a height between 1m and 1.8m. Additionally, an estimate is only output when it is a valid estimate inside the physical borders of the room. If no estimate is output for 1.8 seconds, the timestamp of the last output plus 0.9s is output. If an estimate is available after 0.9s since the last output, the current estimate is duplicated at the timestamp of the last output plus 0.9s.

The JPDAF algorithm used here is a fully automatic two-pass batch algorithm, since the correlation thresholds are first estimated on the whole data and then the position is estimated using the precalculated time delays. If those correlation thresholds would be used from previous experiments, it would be a fully automatic one-pass online algorithm. The algorithm runs at realtime factor 1.98 on a Pentium 4, 2.66GHz machine.

5 State-Based Decision-Level Fusion

For the panoramic camera system, the fusion of the audio and video modalities is done at the decision level. Track estimates coming from the top camera visual tracker and the JPDAF-based acoustic tracker are combined using a finite state machine approach, which considers their relative strengths and weaknesses. The visual trackers can be very accurate at determining a person's position. In

scenarios involving several persons and requiring automatic intialization, they can, however, fail to detect persons completely for lack of observable features, poor discernability from the background, or overlap with other persons, when using color, shape or motion features. The acoustic tracker, on the other hand, can precisely determine a speaker's position only in the presence of speech, and does not produce accurate estimates for several simulaneous speakers or during silence intervals.

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

- 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 assume the visual tracker has recovered the previously undetected speaker and keep ouputting the position of the last overlapping visual track.

The results of the so developed multimodal tracker are presented in section 6.

6 Evaluation on the CLEAR'07 3D Person Tracking Database

The above presented systems for visual and multimodal tracking were evaluated on the CLEAR'07 3D Person Tacking Database. This database comprises recordings from 5 different CHIL smartrooms, involving 3 to 7 persons in a small meeting scenario, for a total of 200 min.

Table 1 shows the results for the Single- and Multi-view visual systems (Particle, Top), for the acoustic tracker (JPDAF), as well as for the corresponding mutimodal systems (PFusion, TFusion). For details on the Multiple Object Tracking Precision (MOTP) and Multiple Object Tracking Accuracy (MOTA) metrics, the reader is referred to [11].

As Table 1 shows, the particle filter tracker clearly outperforms the baseline top view system, while still remaining competitive in terms of computation speed.

Factors that still affect tracking accuracies can be summed up in 2 categories:

Table 1. Evalution results for the 3D person tracking systems

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%

- Detection errors: In some cases, participants showed no significant motion during the length of the sequence, were only hardly distinguishable from the background using color information, or could not be detected by the upper body detectors, due to low resolution, difficult viewing angles or body poses. This accounts for the relatively high amount of missed persons. The adaptation of the used detectors on CLEAR recording conditions or the inclusion of more varied features for detection could help alleviate this problem.
- False tracks: The scarce availability of detection hits for some targets leads to a system design that aggressively initializes person tracks whenever a detection becomes available and keeps them for extended periods of time even in the absence of such. This unfortunately can lead to a fair amount of false tracks which can not be distinguished from valid tracks and effectively eliminated based on color or foreground features alone. Again, the design of more reliable person detectors should help reduce the number of false tracks.

The MOTP numbers range from 222mm for the top camera visual system to 140mm for the acoustic tracker. The acoustic tracker reached an accuracy of 55%, with the main source of errors being localization uncertainty. On the visual and the multimodal side, the particle filter fusion approach (70% / 55%) outperformed the baseline approach (58%, 51%). One can also see that for the particle filter based feature-level fusion approach, the addition of the visual modality could help improve tracking accuracy, compared to acoustic tracking alone, although in the multimodal case, track of the speaker had to be kept also during silence segments.

7 Summary

In this work, two systems for multimodal tracking of multiple users are presented. A joint probabilistic data association filter for source localization is used in conjunction with two distinct systems for visual tracking: One particle filter using multiple camera images, based on foreground and color features and upper body detection and person region cues. The other using only a wide angle overhead view, and performing model based tracking on foreground segmentation features. Two fusion scheme are presented, one at feature level, inherent in the particle filter approach, and one at decision level, 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 2007 3D Person Tracking Database. High accuracies of up to 70% could be reached, with position errors below 15cm.

8 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 CVPR, 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, *"Tracking of Multiple Speakers with Probabilistic Data Association Filters"*. CLEAR Workshop, Southampton, UK, April 2006.
11. Keni Bernardin, Alexander Elbs and Rainer Stiefelhagen, *"Multiple Object Tracking Performance Metrics and Evaluation in a Smart Room Environment"*, Sixth IEEE International Workshop on Visual Surveillance, in conjunction with ECCV2006, May 13th 2006, Graz, Austria
12. Keni Bernardin, Tobias Gehrig, Rainer Stiefelhagen *"Multi- and Single View Multiperson Tracking for Smart Room Environments"*. CLEAR Evaluation Workshop 2006, Southampton, UK, April 2006
13. 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.

14. 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.
15. Y. Bar-Shalom, *"Tracking and data association"*, Academic Press Professional, Inc., San Diego, CA,USA, 1987.
16. C. H. Knapp and G. C. Carter, *"The Generalized Correlation Method for Estimation of Time Delay"*, IEEE Trans. Acoust. Speech Signal Proc., vol. 24, nr. 4, pp. 320–327, August 1976.
17. M. Omologo and P. Svaizer, *"Acoustic Event Localization Using a Crosspower-spectrum Phase Based Technique"*, Proc. ICASSP, vol. 2, pp. 273–276, 1994.
18. U. Klee, T. Gehrig and J. McDonough, *"Kalman Filters for Time Delay of Arrival-Based Source Localization"*, EURASIP Journal on Applied Signal Processing, 2006.
19. Tobias Gehrig and John McDonough, *"Tracking Multiple Simultaneous Speakers with Probabilistic Data Association Filters"*, Proc. Workshop on Machine Learning and Multimodal Interaction, 2006
20. CHIL - Computers In the Human Interaction Loop, http://chil.server.de
21. AMI - Augmented Multiparty Interaction, http://www.amiproject.org
22. VACE - Video Analysis and Content Extraction, http://www.ic-arda.org
23. OpenCV - Open Computer Vision Library, http://sourceforge.net/projects/opencvlibrary