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

Keni Bernardin, Tobias Gehrig, and Rainer Stiefelhagen

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. While the first system fuses audio and visual cues at the feature level, the second one incorporates them at the decision level using state-based heuristics.

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

R. Stiefelhagen et al. (Eds.): CLEAR 2007 and RT 2007, LNCS 4625, pp. 70–81, 2008.

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 probabilistic data association filter uses the audio streams from 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 using 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 maps and upper body region colors gained from all 5 camera images, resampled to 320x240 resolution, as well as upper body detection cues from the room corner cameras, person region hints from the top camera, and source localization estimates gained from the T-shaped microphone arrays.

– The foreground segmentation is made using a simple adaptive background model, which is computed on grayscale images as the running average of the last 1000 frames. The background is subtracted from the current frame and a fixed threshold is applied to reveal foreground regions.

– The color features are computed in a modified HSV space and modeled using a specially designed histogram structure, which eliminates the usual drawbacks of HSV histograms when it comes to modeling low saturation or brightness colors.

The color space is a modified version of the HSV cone. First, colors for which the brightness and saturation exceed 20% are set to maximum brightness . This reduces the effect of local illumination changes or shadows. The HSV values are subsequently discretized as follows: Let hue, sat and val be the obtained HSV values, then the corresponding histogram bin values, h, s and v, are computed as:

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

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

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

The effect is that the number of bins in the hue and saturation dimensions decreases towards the bottom of the cone. There is, e.g., only one histogram bin to model colors with zero brightness. This is in contrast to classical discretization techniques, where e.g. grayscale or nearly grayscale values, for which the hue component is either undefined or ill-conditioned, are spread over a large number of possible bins. At the large end of the cone, a maximum of 16 bins for hue, 10 bins for saturation, and 10 for brightness are used. Figure 1 shows a graphical representation of the resulting discretized HSV space.

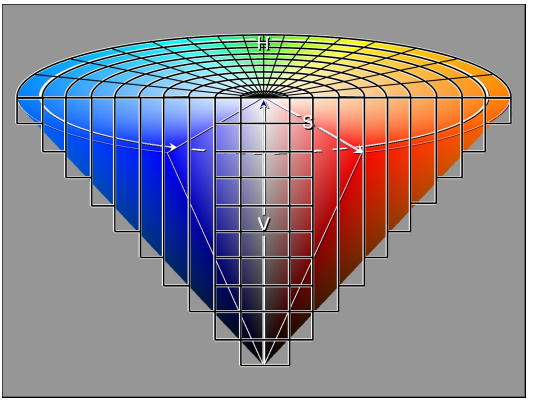

Fig. 1. The discretization of the HSV cone into histogram bins. Colors with zero saturation or brightness have a unique mapping although their hue value, e.g., is undefined.

The color features for tracking are gained from the detected upper body regions of subjects, as well as their immediatly surrounding background. One upper body and one background histogram are kept per camera and track. Upper body histograms for corner cameras are adapted with each

upper body detection from pixels inside the detection region. The upper body histograms for the top camera are obtained using colors sampled from a 60cm diameter region centered around the tracked person position. They are continuously adapted with each frame, but only after valid color histogram models for the corner cameras could be created. This is because the track is only considered reliable enough for continous color model adaptation after it has been confirmed at least once by a direct upper body detection in a corner camera. The background histograms for all views, in turn, are continuously adapted in every frame, with the adaptation 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:

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

with all operations performed bin-wise. For the *minmax* operator, the min value is set to 0 and the max value is set to the maximum bin value of the respective histogram.

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 was found to dramatically increase tracking accuracies.

- The upper body detections in the fixed corner camera images are obtained by exhaustive scanning with Haar-feature classifier cascades, such as in [8,9]. Using camera 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, update color models, 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 tracking algorithm 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 for a given time interval. The motivation for detecting person regions in this way is that top view images present very little overlap between persons, making a simple spatial assignment of blobs to tracks plausible.
- The acoustic features fed to the particle filter are the 3D source localization estimates from the JPDAF tracker described in section 4. Currently, only the active source with the smallest localization uncertainty at a given time is considered. As the source is assumed to be located at the mouth region of a speaking subject, its 3D coordinates are used to associate it to tracks, score particles, and to deduce appropriate upper body regions in each view for updating color models. Acoustic features can therefore lead to the initialization of person tracks with increasingly accurate color models even in the absence of visual upper body detections.

2.2 Intialization and Termination Criteria

For automatic detection of persons and initialization of tracks, 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 person track is initialized when the following conditions are met:

- The average weight of a scout's particle cloud exceeds a fixed activation threshold T. This threshold is set such that initialization is not possible based on the foreground feature alone, but requires the contribution of at least an upper body detection, person region hint or source localization estimate.
- The spread of the particle cloud, calculated as the variance in particle positions, is below a threshold T_2.
- The tracked 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 target object 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 last condition helps to avoid initializing faulty tracks on plane surfaces, triggered e.g. by false alarm detections or shadows, or when the target's upper body color is not distinct enough from the surrounding background to allow stable tracking.

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

2.3 Particle Filtering

The tracking scheme used here maintains a separate particle filter tracker for each person. In contrast to conventional implementations, in this framework a particle represents the hypothesized (x, y, z) scene coordinates of one single point on the target object, not necessarily its center. Consequently, the foreground and color feature scores of a particle in each camera image are not computed using a projected kernel or person window, but rather by using only one pixel value: The particle's 3D position, shifted by -20cm in the z-axis, is projected to the image. The corresponding pixel coordinates and color value are then used, together with the foreground segmentation map and the track's color histogram, to derive the foreground and color scores respectively. The particle's 3D position, on the other hand, is used together with available upper body detection, person region hint or acoustic source positions and uncertainties to derive a detection score. A weighted combination of these scores using predefined fixed weights then yields the final particle score. In this way, the computational effort for an individual particle's score is kept at a minimum, notably requiring no time consuming histogram comparisons or backprojections.

After scoring, normalization and resampling, the mean of the particles' positions is taken as the track center. Propagation is then done by adding gaussian noise to the resampled 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.

Since the system maintains a separate particle filter for every track, some mechanism is required to avoid initializations on already supported tracks or accumulations of filters on the same track. This is accomplished as follows: A "repellent" region of 60cm diameter is defined around the center of each track. The weight of all other tracks' particles which fall into the repellent region are then set to 0 if their current weight does not exceed the repellent track's average weight. Additionally, absolute priority is given to valid tracks over scout tracks. In this way, particles from distinct tracks which share the same space are penalized and tracks with higher confidence repel less confident ones.

The system implementation is distributed over a network of 5 machines to achieve real-time computation speed. It was extensively tested and achieved high accuracy rates, as shown in section 6. Figure 2 shows a graphical output of the particle filter tracking system.

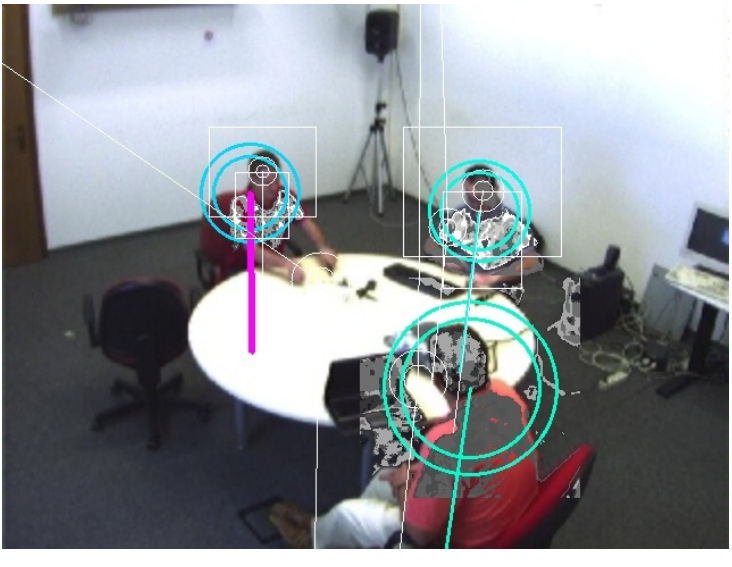

Fig. 2. A graphical output of the particle filter based tracking system. The white rectangles show upper body detection hits and the areas used for sampling foreground and background colors. The histogram backprojection values are shown for each track, superimposed on the image as grayscale values, with brighter colors indicating higher probabilities of belonging to the person track.

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. 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 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 for association of blobs to 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 using an fixed adaptation factor α. Background subtraction, thresholding and morphological filtering provide the foreground blobs for tracking.

The subsequent EM 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 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, a FFT size of 4096 points is used. 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 filter 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.

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, which considers their relative strengths and weaknesses. Although the visual tracker is able to keep several simultaneous tracks, in scenarios requiring automatic initialization it can fail to detect persons completely for lack of observable features, poor discernability from the background, or overlap with other persons. 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 below 500mm. In this case, it is assumed the visual tracker has missed the speaking person and the acoustic hypothesis is output. The last received acoustic estimate is stored and continuously output until an overlapping visual estimate is found.
- State 2: An acoustic estimate is available, and a corresponding visual estimate exists. In this case, the average of the acoustic and visual estimates is output.
- State 3: After an overlapping visual estimate had been found, an acoustic estimate is no longer available. In this case, it is assumed the visual tracker has recovered the previously undetected speaker and the position of the last overlapping visual track is continuously output.

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

The above presented systems for visual, acoustic 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, with a total length of 200 minutes.

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

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

System	$MOTP$	$\overline{miss}$	$\overline{falsePos}$	$mism.$	$(A-)MOTA$
Particle Filter	155mm	15.09%	14.50%	378	69.58%
Top Tracker	222mm	23.74%	20.24%	490	54.94%
JPDAF	140mm	20.60%	24.78%	-	54.63%
Particle Filter Fusion	151mm	20.17%	20.02%	121	58.49%
Decision Level Fusion	159mm	21.80%	23.38%	370	50.78%

As can be seen in Table 1, the particle filter tracker clearly outperforms the baseline top view system, while still remaining competitive in terms of computational speed.

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

- Detection errors: In some cases, participants showed no significant motion during the length of the sequence, rarely spoke, 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 rather high percentage of misses.

The adaptation of the used detectors on CLEAR recording conditions or the inclusion of more varied features for detection should help alleviate this problem.
- False tracks: The scarce availability of detection hits for some targets lead to a system design that aggressively initializes person tracks whenever a detection becomes available and keeps tracks alive 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 based feature level fusion approach (70%, 58%) clearly outperformed the baseline approach (55%, 51%). In the particle filter approach, the fusion of both modalities could improve tracking accuracy, compared to acoustic only tracking results, even though in the multimodal subtask the speaker additionally had to be tracked through periods of silence.

7 Summary

In this work, two systems for the multimodal tracking of multiple users were presented. A joint probabilistic data association filter for source localization is used in conjunction with two distinct systems for visual tracking: The first is a particle filter using foreground, color, upper body detection and person region cues from multiple camera images. The other is a blob tracker using only a wide angle overhead view, and performing model based tracking on foreground segmentation features. Two fusion scheme were 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 tracking accuracies of up to 70% and position errors below 15cm could be reached.

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. Khalaf, R.Y., Intille, S.S.: Improving Multiple People Tracking using Temporal Consistency. MIT Dept. of Architecture House_n Project Technical Report (2001)
2. Niu, W., Jiao, L., Han, D., Wang, Y.-F.: Real-Time Multi-Person Tracking in Video Surveillance. In: Pacific Rim Multimedia Conference, Singapore (2003)

3. Mittal, A., Davis, L.S.: M2Tracker: A Multi-View Approach to Segmenting and Tracking People in a Cluttered Scene Using Region-Based Stereo. In: Heyden, A., Sparr, G., Nielsen, M., Johansen, P. (eds.) ECCV 2002. LNCS, vol. 2350, pp. 18–33. Springer, Heidelberg (2002)
4. Checka, N., Wilson, K., Rangarajan, V., Darrell, T.: A Probabilistic Framework for Multi-modal Multi-Person Tracking. In: Workshop on Multi-Object Tracking (CVPR) (2003)
5. Comaniciu, D., Meer, P.: Mean Shift: A Robust Approach Toward Feature Space Analysis. IEEE PAMI 24(5) (May 2002)
6. Haritaoglu, I., Harwood, D., Davis, L.S.: W4: Who? When? Where? What? A Real Time System for Detecting and Tracking People. In: Third Face and Gesture Recognition Conference, pp. 222–227 (1998)
7. Raja, Y., McKenna, S.J., Gong, S.: Tracking and Segmenting People in Varying Lighting Conditions using Colour. In: 3rd. Int. Conference on Face & Gesture Recognition, p. 228 (1998)
8. Viola, P., Jones, M.: Rapid Object Detection using a Boosted Cascade of Simple Features. In: IEEE CVPR (2001)
9. Lienhart, R., Maydt, J.: An Extended Set of Haar-like Features for Rapid Object Detection. In: IEEE ICIP 2002, September 2002, vol. 1, pp. 900–903 (2002)
10. Gehrig, T., McDonough, J.: Tracking of Multiple Speakers with Probabilistic Data Association Filters. In: CLEAR Workshop, Southampton, UK (April 2006)
11. Bernardin, K., Elbs, A., Stiefelhagen, R.: Multiple Object Tracking Performance Metrics and Evaluation in a Smart Room Environment. In: Sixth IEEE International Workshop on Visual Surveillance, in conjunction with ECCV 2006, Graz, Austria, May 13th (2006)
12. Bernardin, K., Gehrig, T., Stiefelhagen, R.: Multi- and Single View Multiperson Tracking for Smart Room Environments. In: CLEAR Evaluation Workshop 2006, Southampton, UK, April 2006. LNCS, vol. 4122, pp. 81–92 (2006)
13. Tao, H., Sawhney, H., Kumar, R.: A Sampling Algorithm for Tracking Multiple Objects. In: International Workshop on Vision Algorithms: Theory and Practice, pp. 53–68 (1999)
14. Wren, C., Azarbayejani, A., Darrell, T., Pentland, A.: Pfinder: Real-Time Tracking of the Human Body. IEEE Transactions on Pattern Analysis and Machine Intelligence 19(7), 780–785 (1997)
15. Bar-Shalom, Y.: Tracking and data association. Academic Press Professional, Inc., San Diego (1987)
16. Knapp, C.H., Carter, G.C.: The Generalized Correlation Method for Estimation of Time Delay. IEEE Trans. Acoust. Speech Signal Proc. 24(4), 320–327 (1976)
17. Omologo, M., Svaizer, P.: Acoustic Event Localization Using a Crosspower-spectrum Phase Based Technique. In: Proc. ICASSP, vol. 2, pp. 273–276 (1994)
18. Klee, U., Gehrig, T., McDonough, J.: Kalman Filters for Time Delay of Arrival-Based Source Localization. EURASIP Journal on Applied Signal Processing (2006)
19. Gehrig, T., McDonough, J.: Tracking Multiple Simultaneous Speakers with Probabilistic Data Association Filters. LNCS, vol. 4122, pp. 137–150 (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`