Person Tracking With Audio-visual Cues Using The Iterative Decoding Framework

Shankar T. Shivappa , Mohan M. Trivedi and Bhaskar D. Rao
Department of Electrical and Computer Engineering
University of California, San Diego.
sshivappa@ucsd.edu

Abstract

Tracking humans in an indoor environment is an essential part of surveillance systems. Vision based and microphone array based trackers have been extensively researched in the past. Audio-visual tracking frameworks have also been developed. In this paper we consider human tracking to be a specific instance of a more general problem of information fusion in multimodal systems. Dynamic Bayesian networks have been the modeling technique of choice to build such information fusion schemes. The complexity and non-Gaussianity of distributions of the dynamic Bayesian networks for such multimodal systems have led to the use of particle filters as an approximate inference technique. In this paper we present an alternative approach to the information fusion problem. The iterative decoding algorithm is based on the theory of turbo codes and factor graphs used in communication systems. We modify and adapt the iterative decoding algorithm to do probabilistic inference for the problem of tracking humans in an indoor space, using multiple cameras and microphone arrays.

1. Introduction

Robust object tracking is a essential part of surveillance systems. It is the first step in facilitating detection and analysis of human activity in a monitored space. It is also an integral component of intelligent spaces, for facilitating seamless interaction between humans and computers. According to [22] an intelligent space is defined as a system where humans and computers collaborate through natural modes of interaction and person tracking is essential to facilitate such interactions. Real world surveillance systems have to work with noisy sensors in cluttered environments. Tracking humans using audio-visual cues can provide robustness to background noise and visual clutter. Tracking based on visual sensors has been widely researched[23]. Microphone array based trackers that track sound sources have also been studied by some researchers[6]. One of the advantages of a multimodal system is its robustness to environmental and sensor noises. Performance of a multimodal system should degrade gracefully when noise affects its individual modalities. This is the goal of a good information fusion scheme. Information/cues from multimodal systems can be fused at various levels. As shown in [19], even heterogenous sensors like cameras and microphone arrays can be treated in a unified manner and the cues merged at the sensor level. Feature level fusion is seen most commonly in the case of audio-visual speech recognition systems [18]. Decision level fusion offers the advantage of robustness in case of failures in single modalities. This is typically achieved by scaling the decisions of individual modalities according to their reliability (SNR) [12],[10]. In this paper we use the iterative decoding algorithm to formulate a general fusion framework for multimodal person tracking and apply it to track people in an indoor environment with multiple cameras and microphone arrays. We also present details of the experimental evaluation of the framework in our laboratory testbed. The evaluation is carefully designed to bring forth the true strengths of our framework and its weaknesses. This evaluation process is also part of our continuing research effort to improve our tracking system. In section 2, we present a survey of related research and the comparative advantage of the proposed framework. In sections 3 and 4, we present the mathematical formulation of the hidden Markov model based multimodal tracker. We describe the modeling, training and testing of such a system. In section 5, we present the laboratory testbed with multiple cameras and microphone arrays which was used for extensive experimentation and evaluation studies. The results are presented in a concise manner.

978-0-7695-3341-4/08 $25.00 © 2008 IEEE
DOI 10.1109/AVSS.2008.54

2. Person Tracking with multimodal cues

In this section we present a brief survey of related research activities in the field of multimodal person tracking. We also develop the motivation behind using iterative decoding framework to solve the tracking problem.

Fusion of information from different sensors in a multimodal system can take place at various levels. There are significant advantages to be gained by fusing the signals early on (sensor level fusion). Specifically, this approach tends to simplify the higher level processing in a multimodal system. On the other hand, delayed/decision level fusion allows us to exploit the redundancy and noise independence of the modalities to achieve robustness. For a given task, the optimal strategy is often a combination of the two. Our framework includes both early and late fusion strategies for tracking humans using multiple cameras and microphone networks. Early fusion strategy [19] is particularly interesting because it treats the sensors from different modalities in a unified manner. However the scheme lacks a strategy to exploit the noise independent nature of the multimodal sensors. The frames affected by noise in different modalities are independent, because the noise affecting heterogenous modalities like audio and video is independent. For a given task, one can exploit the nature of the problem to come up with fusion strategies that emphasize the noiseless modality.

Object tracking has been a important component of computer vision systems. [23] is a recent survey of object tracking methodologies using vision based sensors. Audio tracking, based on speaker localization by microphone arrays has been explored too[6]. Recent research efforts are mostly directed towards integrating the audio, visual and other modalities to track objects. Several frameworks have been proposed to fuse audio-visual cues. Dynamic Bayesian networks have been a popular choice for fusion of multimodal information [3],[16],[14],[15]. Approximate inference in the dynamic Bayesian network framework, necessitated by the complexity and non-Gaussianity of the joint models, is performed by the use of particle filters [14],[9]. HMMs and graphical models have been used to model multimodal data, especially by the ambient intelligence community [8],[20],[7]. However, our approach uses the HMMs at the sensor level as opposed to the higher levels of abstraction used by the ambient intelligence community.

We present an alternative approach to fusion of multimodal cues in the form of iterative decoding, motivated by the theory of turbo codes [4]. In [21], iterative decoding was introduced as an effective means for fusion of multimodal information for human activity analysis and the intuition and advantages of the scheme have been discussed in detail. The scheme as described in [21] is not applicable to tracking as we need to solve the data association problem[2] before using iterative decoding. In this paper we present a HMM based tracking framework which specifies the tracking problem in a hierarchical manner, allowing the local sensors (camera/microphone array) to maintain track hypotheses and the global tracker to fuse the local tracks from various sensors to generate a robust estimate using iterative decoding. The same framework is also applicable to situations where multiple sensors are used to monitor disjoint spaces. In this case, one cannot expect robustness to sensor limitations as one would in the overlapping-field-of-view case.

The iterative decoding algorithm is applicable to multimodal systems that use the hidden Markov model framework [21]. The primary advantage of this scheme is the ability to use unimodal models while obtaining the performance of joint multimodal models. In the presence of background and sensor noise, the joint models are not able to separate out the noisy modalities from the clean ones. Because of this reason, the iterative decoding algorithm outperforms the joint model at low SNR. In the case of other decision level fusion algorithms like the multistream HMMs [12] and reliability weighted summation rule [13], one has to estimate the quality(SNR) of the individual modalities to obtain good performance. Iterative decoding does not need such apriori information. This is another significant advantage which we shall demonstrate in section 5 by tracking people robustly over noisy video and audio channels. In addition to these advantages of the iterative decoding scheme, already described in [21], we show that our tracking framework can use the same iterative decoding algorithm to perform data association over video occlusions and audio silence segments. We demonstrate the effectiveness of our approach through extensive experimental results on stationary and dynamic scenes with varying number of subjects and mixed composition, from our audio-visual testbed. Our use of multiple cameras and wide aperture microphone array sets us apart from the experimental setup presented in [15],[3].

The calibration of multimodal sensors is an important issue in tracking. In our proposed framework, the system only requires a rough calibration step. After this initial calibration, the system can improve its accuracy by acquiring the calibration information in an online manner from the tracks. The framework also allows for reconfiguring the sensor network without having to recalibrate the system. These aspects are however part of ongoing research and are not discussed in this paper.

In our present work, we use the iterative decoding principle on a hidden Markov model based framework for audio-visual tracking of multiple persons through multiple cameras and a network of multiple microphone arrays. The framework is modular and hence easy to expand to more number of cameras and microphone arrays or any other sensors that can localize persons. It is also applicable to sensors with overlapping and non-overlapping field of 'view'. Since

the placement of the sensors is assumed to be arbitrary but fixed, we only need a rough calibration scheme to establish the correspondence between sensors. The unimodal models considered in this paper are simple and naive. The goal of the paper is to demonstrate the fusion algorithm and its applicability to the tracking scenario.

3. Iterative decoding algorithm

Consider a hidden Markov model Λ_k for sensor k with N hidden states. For clarity, we drop the sensor index k. Λ has a parametric transition density. The hidden state q_t corresponds to the true location of the object at time t in the same space as the observations of sensor k. Thus the hidden states are, in a Bayesian sense, the filtered observations. The conditional distribution of the observation o_t when the hidden state is q_t is assumed to be Gaussian. Now, the decoding problem is to estimate the optimal state sequence $Q_1^T = \{q_1, q_2 \dots q_T\}$ of the HMM based on the sequence of observations $O_1^T = \{o_1, o_2 \dots o_T\}$.

The Maximum aposteriori probability state sequence is provided by the BCJR (Bahl Cocke Jelinek and Raviv) algorithm[1]. The MAP estimate for the hidden state at time t is given by $\hat{q}_t = \arg\max P(q_t, O_1^T)$. The BCJR algorithm computes this using the forward and backward recursions.

The forward recursion variable $\alpha_t(m)$, the backward recursion variable $\beta_t(m)$, the joint likelihood of the hidden state and the observation sequence $\lambda_t(m)$ and the recursion variable $\gamma_t(m', m)$ are defined as follows,

$$\lambda_t(m) = P(q_t = m, O_1^T) \tag{1}$$
$$\alpha_t(m) = P(q_t = m, O_1^t) \tag{2}$$
$$\beta_t(m) = P(O_{t+1}^T | q_t = m) \tag{3}$$
$$\gamma_t(m', m) = P(q_t = m, o_t | q_{t-1} = m') \tag{4}$$

where, $m = 1, 2 \dots N, m' = 1, 2 \dots N$

Then establish the recursions,

$$\alpha_t(m) = \sum_{m'} \alpha_{t-1}(m') \cdot \gamma_t(m', m) \tag{5}$$
$$\beta_t(m) = \sum_{m'} \beta_{t+1}(m') \cdot \gamma_{t+1}(m, m') \tag{6}$$
$$\lambda_t(m) = \alpha_t(m) \cdot \beta_t(m) \tag{7}$$

At the first sensor, we decode the hidden states using the observations from the first modality. We obtain the aposteriori probabilities, $\lambda_t^{(1)}(m) = P(q_t = m, O_1^T)$.

In the second sensor, these aposteriori probabilities, $\lambda_t^{(1)}(m)$ are utilized as extrinsic information in decoding the hidden states from the observations of the second modality. Thus the aposteriori probabilities in the second stage of decoding are given by $\lambda_t^{(2)}(m) = P(q_t = m, O_1^T, {Z^{(1)}}_1^T)$ where $Z_t^{(1)} = \lambda_t^{(1)}$ is the extrinsic information from the first sensor.

$$\lambda_t^{(2)}(m) = P(q_t = m, O_1^T, {Z^{(1)}}_1^T) \tag{8}$$
$$\alpha_t^{(2)}(m) = P(q_t = m, O_1^t, {Z^{(1)}}_1^t) \tag{9}$$
$$\beta_t^{(2)}(m) = P(O_{t+1}^T, {Z^{(1)}}_{t+1}^T | q_t = m) \tag{10}$$
$$\gamma_t^{(2)}(m', m) = P(q_t = m, o_t, Z_t^{(1)} | q_{t-1} = m') \tag{11}$$

Then the recursions do not change, except for the computation of $\gamma_t^{(2)}(m', m)$.Since the extrinsic information is independent of the observations from the second modality,

$$\gamma_t^{(2)}(m', m) = P(q_{2,t} = m, o_{2,t}, Z_t^{(1)} | q_{2,t-1} = m')$$
$$\gamma_t^{(2)}(m', m) = P(q_{2,t} = m | q_{2,t-1} = m') \cdot P(o_{2,t} | q_{2,t} = m) \cdot P(Z_t^{(1)} | q_{2,t} = m)$$

$$\gamma_t^{(2)}(m', m) = P(q_{2,t} = m | q_{2,t-1} = m') \cdot P(o_{2,t} | q_{2,t} = m) \cdot \sum_n \{P(Z_t^{(1)} | q_{1,t} = n) P(q_{1,t} = n | q_{2,t} = m)\}$$

where $q_{2,t}$ and $o_{2,t}$ correspond to the hidden state and observation at time t for modality 2.

Assuming that $P(Z_t^{(1)} | q_{1,t}) = 1$ if $q_{1,t} = \arg\max_n Z_{t,n}^{(1)}$ and 0 otherwise, where $Z_{t,n}^{(1)}$ is the nth component of the vector $Z_t^{(1)}$, which corresponds to a hard decision rule, we are now left with the evaluation of $P(q_{1,t} = n | q_{2,t} = m)$. In section 5, we describe the process of sensor calibration by which we obtain this distribution.

We proceed likewise till we decode the hidden states of the last sensor from the extrinsic information of the previous sensor. In the next iteration, we use the extrinsic information of the last sensor to decode the hidden states of the first sensor again, using the extrinsic information of the last. Then the second iteration proceeds as the first, with updated state sequences. Finally we threshold the overall log-likelihood of the track combinations to select the surviving tracks.

4. Multimodal tracking framework

We are interested in tracking multiple targets(people) in a space instrumented with multiple sensors. Each sensor detects targets in its field of view and maintains an exhaustive list of possible hypotheses. Human motion in indoor environment is highly non-linear and hence at the sensor level there is not enough information to reject the false hypotheses. Once the information from other sensors is also available, a composite tracker can evaluate the likelihood of

each hypothesis and the hypotheses with high likelihood are selected and tracked more accurately. This process is graphically depicted in figure 1. The process described in figure 1 is intuitive and the iterative decoding algorithm gives us a statistical framework to implement it. The tracks are then passed back to the sensors as initial hypotheses to continue tracking. In figure 2, we present a high level view of the tracking system.

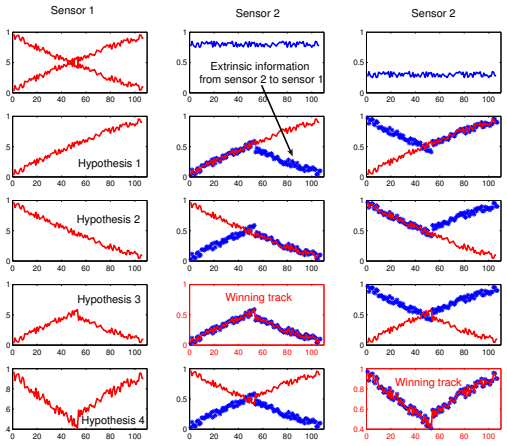

Figure 1. The disambiguation of confusable hypothesis using the iterative decoding scheme is illustrated here. The first graph shows the tracks as seen in one of the sensors. The next four images in the first column present the possible hypotheses that are plausible according to the first sensor alone. The second and third columns have two tracks in the field of view of sensor 2. The extrinsic information that these tracks provide sensor 1 are shown, superimposed with the hypotheses and the winning tracks are indicated.

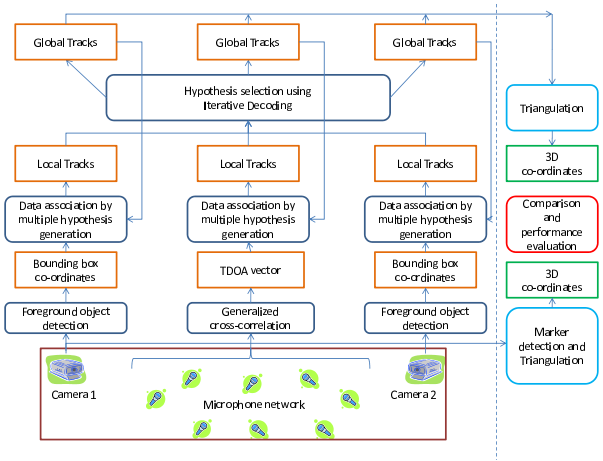

Figure 2. Proposed framework

4.1. Object detection

4.1.1 Foreground object detection for the cameras

In our experiment we choose a simple foreground object detection scheme. The foreground pixels in a frame are detected by background subtraction. They are then fused into reliable blobs by morphological operations. We fit a bounding rectangle to each distinct blob. The pixel co-ordinates of the center of the ith rectangle, (x_{it}, y_{it}) and the area of the rectangle $[a_{it}]$ are the components of the observation vector $o_{it}^T = [x_{it} y_{it} a_{it}]$. For every frame at time t, for the jth camera, we maintain a list of the M_j detected foreground objects $o_{it}^j, 1 \leq i \leq M_j$.

4.1.2 Sound source detection for the microphone network

We use the time delay of arrival(TDOA) estimates between pairs of microphones to estimate the location of the sound source. We use the generalized cross correlation based phase transform (GCC-PHAT) framework [17][5] to locate sound sources if present. This technique has been the preferred method of TDOA estimation is established literature[14][19] as it has shown to be robust to reverberations. For simplicity, the TDOA estimates are computed on time windows of audio samples corresponding to the interval between the camera frames. We have a vector of TDOA values between each microphone i and the reference microphone r, given by $\vec{\tau} = (\tau_{1r}, \tau_{2r} \ldots \tau_{mr})$. The TDOA components for the first sound source are the components of the observation vector $o_{1,t}$ corresponding to the microphone network. Thus we reduce a microphone network to a 3-d localizer similar to a camera. Note that the use of the SRP-PHAT technique [11] would allow us to detect multiple sound sources simultaneously. We could then have M_t detected sources at each time instance and a list of observations, $o_{i,t}, 1 \leq i \leq M_t$. In the current paper, we limit ourselves to finding only one sound source at a time. After the observations are extracted for each frame of audio, the cameras and the microphone network are treated equivalently as in [19]. Our audio setup however differs from [19] in the arrangement of microphones. Traditional microphone arrays (linear/planar/spherical) have only angular resolution. The spacing out of the microphone arrays in 3-d space to form an ad-hoc microphone array network provides us with 3-d resolution in the audio space.

4.2. Multiple hypotheses generation - local tracking

The object detection module associated with each camera(or a network of microphone arrays) detects the foreground objects (or sound sources) in each frame. In the presence of multiple objects of interest, all distinguishable objects are detected by each sensor. False positive errors

could occur in the presence of background noise or clutter. False negative errors could occur due to occlusions. The tracking framework will address both these issues.

Consider frames from time $t = 1 \ldots T$. We start with a list of features of detected objects $o_{i,t}, 1 \leq i \leq M_t$ at time t, where M_t is the number of detected objects at time t. We start with a set of initial track values for track j $l_{j,0}$. At each timestep, the tracks are updated according to the rule $l_{j,t} = \{o_{i,t} | d(l_{j,t-1}, o_{i,t}) \leq r\}$, where $d(x, y)$ is the Euclidean distance between x and y. If more than one observation lies within Euclidean distance r from $l_{j,t-1}$, the old track is split to account for each such observation. If no observation lies within radius r, we assign the past value $l_{j,t-1}$ to the track. This corresponds to occlusions or the object leaving the field of 'view' of the sensor. We can see that this is a very simple data association framework and would result in a lot of false positives, as it maintains tracks corresponding to all the possibilities in case of any occlusions or merging and diverging of tracks. Only those possibilities are discarded where the data association can be completed without ambiguity based on nearest neighbors. In the next step, using the information from other tracks, we reject the hypotheses that are unlikely under a probabilistic joint model.

4.3. Multiple hypotheses selection and filtering-global tracking

Consider the set of all hypotheses $h^k = l_j | 1 \leq j \leq N_k$ from sensor k which has N_k hypotheses. In the global tracking step, we consider all possible combinations of these hypotheses, one from each sensor. There are $\prod_k N_k$ such combinations. We evaluate the likelihood of each combination $C = (l^1_{j_1}, l^2_{j_2} \ldots l^N_{j_N})$ under the iterative decoding framework with HMM λ_k for sensor k. Spurious tracks have a low likelihood and are discarded. The remaining tracks are then passed down to the local trackers to use as initial tracks for the next time window.

5. Experimental validation and evaluation

In this section we present the details of our laboratory testbed with multiple cameras and microphone arrays. We also describe the results from the validation and evaluation of the iterative decoding based person tracker.

The experimental set up consists of 2 rectilinear cameras and an network of 8 microphones. The cameras have significantly overlapping field of view and different perspectives. The cameras have a resolution of 640x480 pixels and capture frames, synchronously, with each other and the microphones, at 30 fps. The audio signal is sampled at 44.1kHz. The microphone arrays are suspended from the ceiling at fixed but arbitrary locations. The cameras have a overlapping field of view with different perspectives.

The camera and microphone locations are assumed to be arbitrary but fixed. Hence we need a rough calibration step to establish a relationship between the state space of different sensors. In the iterative decoding algorithm presented in section 3, we are left with problem of estimating $P(q_{1,t} = n | q_{2,t} = m)$ for sensor pair $(1, 2)$. There are efficient ways of learning and storing this distribution by using decision trees, piecewise linear approximations and kernel based density estimation techniques. In our experiments we use a simple kernel density estimation scheme to estimate the conditional distribution $P(q_{1,t} = n | q_{2,t} = m)$, by first estimating the joint distribution $P(q_{1,t} = n, q_{2,t} = m)$ from a set of training points collected during the calibration step. In order to collect training points, we have an initial calibration step where a single person carrying a sound source walks around the space monitored by the sensors. Tracking is now trivial as there is only one object. The observations from several frames are used to estimate the joint distribution $P(q_{1,t} = n, q_{2,t} = m)$ using a Gaussian kernel of appropriate bandwidth for smoothing.

During the initial calibration phase, a person carrying a sound source walks around the room. A snapshot from the calibration process is presented in figure 5. From the audio signals, the TDOA vector corresponding to the sound source is computed and from the video frames, the (x, y) pixel co-ordinate of the foreground object is obtained. Note that our calibration step establishes correspondences between sensors in the sensor co-ordinate system. We do not consider the calibration of the cameras and microphone arrays to the world co-ordinate system. However the iterative decoding scheme and the HMM based tracking framework presented are also applicable to trackers that are calibrated to the world co-ordinate system.

5.1. Ground truth estimation

In order to obtain the ground truth, we use standard chessboard pattern based camera calibration techniques to calibrate the cameras with respect to the world co-ordinates. A sound source with a bright source of light is moved around the monitored space. By triangulation, the position of the light is accurately determined at each frame. The positions of the microphones are then optimized to match the TDOA values obtained at each frame with those computed from the sound source co-ordinates. This calibration allows us to obtain visual-marker based ground-truth estimates for comparison of our results. The location estimate from the triangulation procedure was compared with the actual location measurement. The standard deviation of the error was 2.4 cm on a test set that involved 100 different spots distributed in the room.

5.2. Evaluation

The dataset for the evaluation of our proposed system consists of scenes with 1 to 4 subjects. Both male and female subjects are included in the dataset. The individual segments range from 1 minute to 5 minutes in duration. There are clips where the subjects are either involved in a normal conversation or are moving around the monitored space. The results are presented for tracking using various subsets of the available sensors. This indicates how the system will perform in the case of sparse sensor configurations, discussed earlier. In table 1, we show the percentages of frames when the global tracker successfully resolves the ambiguity during occlusions and noisy detections based on the information from the other sensors. In figure 3, we show one of the tracks from a clip and the associated groundtruth. The root mean squared error between the track and the ground truth is 11cm. In figure 4, we see that TDOA estimates obtained from the marker based triangulation matches with the measured TDOA for most of the frames.

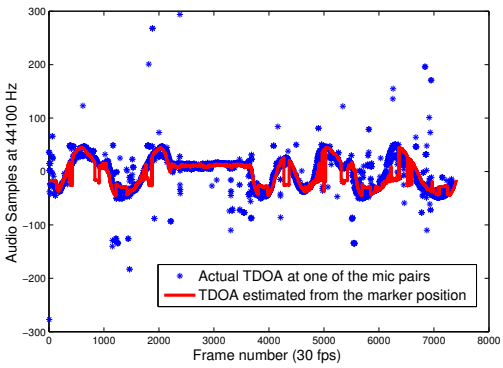

Figure 4. Estimated TDOA values from marker based estimation and the actual measured values for one of the test segments.

6. Concluding remarks

We have developed a multimodal tracking framework which combines information from simple local trackers in an intuitive manner to arrive at the correct global track. We have also presented a probabilistic framework based on iterative decoding to achieve this information fusion. This scheme has been applied to a simple tracking scenario where background subtraction based local trackers are used to track humans in an indoor space. The experimental results ar encouraging. The next logical step is to augment the capability of the local trackers with more cues like color and texture information. In fact, even more elaborate features like person ID and speaker ID can be associated with tracks in the same framework. Also, the audio sensors in the present implementation can only track a single source. We plan to extend this to tracking multiple audio sources in the future. Reconfigurability of sensors and improving sensor correspondence accuracy through online learning are part of our ongoing research efforts, in addition to extensive evaluation of the tracking framework in the audio-visual testbed described above.

References

[1] L. Bahl, J. Cocke, F. Jelinek, and J. Raviv. Optimal decoding of linear codes for minimizing symbol error rate. *IEEE Transactions on Information Theory*, IT-20(2):284–287, Mar. 1974.

[2] Y. Bar-Shalom and T. E. Fortmann. *Tracking and Data Association*. Academic Press, 1988.

[3] M. Beal, N. Jojic, and H. Attias. A graphical model for audiovisual object tracking. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 25(7):828–836, 2003.

[4] C. Berrou, A. Glavieux, and P. Thitimajshima. Near shannon limit error-correcting coding and decoding:turbo-codes. In *Proceedings of the IEEE International Conference on Communications*, Geneva, Switzerland, May 1993.

[5] M. Brandstein and H. Silverman. A robust method for speech signal time-delay estimation in reverberant rooms. 1997.

[6] M. Brandstein and D. Ward. *Microphone Arrays*. Springer, 2001.

[7] O. Brdiczka, J. Maisonnasse, and P. Reignier. Automatic detection of interaction groups. In *ICMI '05: Proceedings of the 7th international conference on Multimodal interfaces*, pages 32–36, New York, NY, USA, 2005. ACM.

[8] O. Brdiczka, J. Maisonnasse, P. Reignier, and J. L. Crowley. Detecting small group activities from multimodal observations. *International Journal of Applied Intelligence*, 2007.

[9] N. Checka, K. W. Wilson, M. R. Siracusa, and T. Darrell. Multiple person and speaker activity tracking with a particle filter. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 5. IEEE, 2004.

[10] Y. Chen and Y. Rui. Real-time speaker tracking using particle filter sensor fusion. *Proceedings of the IEEE*, 92, 2004.

[11] J. H. DiBiase, H. F. Silverman, and M. S. Branstein. Robust localization in reverberant rooms. *Microphone Arrays: Signal Processing Techniques and Applications*, 2001.

[12] S. Dupont and J. Luettin. Audio-visual speech modeling for continuous speech recognition. *IEEE Transactions on Multimedia*, 2(3), Sept. 2000.

[13] E. Erzin, Y. Yemez, A. M. Tekalp, A. Ercil, H. Erdogan, and H. Abut. Multimodal person recognition for human-vehicle interaction. *IEEE Multimedia Magazine*, 13, 2006.

[14] D. Gatica-Perez, G. Lathoud, J. Odobez, and I. McCowan. Audiovisual probabilistic tracking of multiple speakers in meetings. *IEEE Transactions on Audio, Speech and Language Processing*, 15, 2007.

[15] T. Hospedales and S. Vijayakumar. Bayesian structure inference for multisensory scene understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2008.

Table 1. Results

	1 camera	Microphones	1 camera and microphones	2 cameras	2 cameras and microphones
1 subject	95%	76%	95%	98%	98%
2 subjects	53%	42%	68%	85%	87%
3 subjects	38%	40%	65%	80%	83%
4 subjects	33%	34%	55%	69%	73%

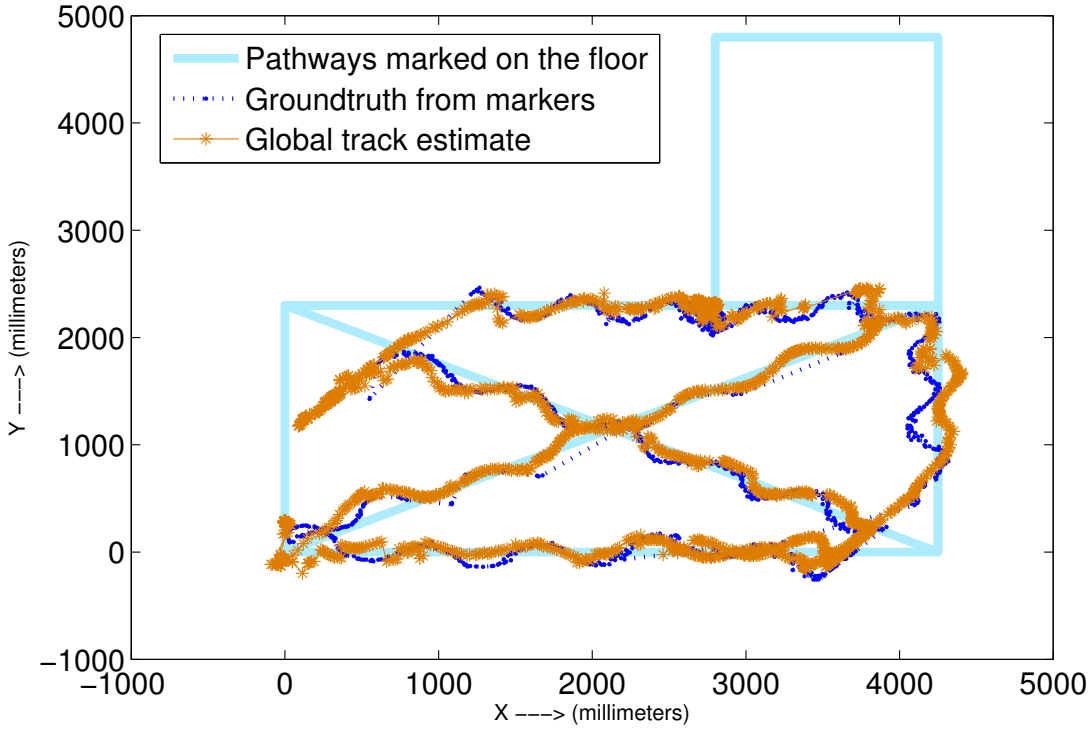

Figure 3. A track and its associated ground truth in the world co-ordinates.

[16] A. Jaimes and N. Sebe. Multimodal human computer interaction: A survey. In *Proceedings of the IEEE International Workshop on Human Computer Interaction in conjunction with ICCV*, Beijing, China, Oct. 2005.

[17] C. H. Knapp and G. C. Carter. The generalized correlation method for estimation of time delay. *IEEE Transactions on Acoustic, Speech and Signal Processing*, 24, 1976.

[18] C. Neti, G. Potamianos, J. Luettin, I. Matthews, H. Glotin, and D. Vergyri. Large-vocabulary audio-visual speech recognition: A summary of the Johns Hopkins summer 2000 workshop. In *Proceedings of IEEE Workshop Multimedia Signal Processing*, Cannes, 2001.

[19] A. ODonovan and R. Duraiswami. Microphone arrays as generalized cameras for integrated audio visual processing. In *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, Minneapolis, USA, 2007.

[20] N. M. Oliver, B. Rosario, and A. Pentland. A bayesian computer vision system for modeling human interactions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):831–843, 2000.

[21] S. Shivappa, B. Rao, and M. Trivedi. An iterative decoding algorithm for fusion of multi-modal information. *Eurasip Journal on Advances in Signal Processing, Special Issue on Human-Activity Analysis in Multimedia Data*, 2007.

[22] M. M. Trivedi, K. S. Huang, and I. Mikic. Dynamic context capture and distributed video arrays for intelligent spaces. *IEEE Transactions on Systems, Man and Cybernetics*, 35, Jan. 2005.

[23] A. Yilmaz, O. Javed, and M. Shah. Object tracking: A survey. *ACM Comput. Surv.*, 38(4), 2006.

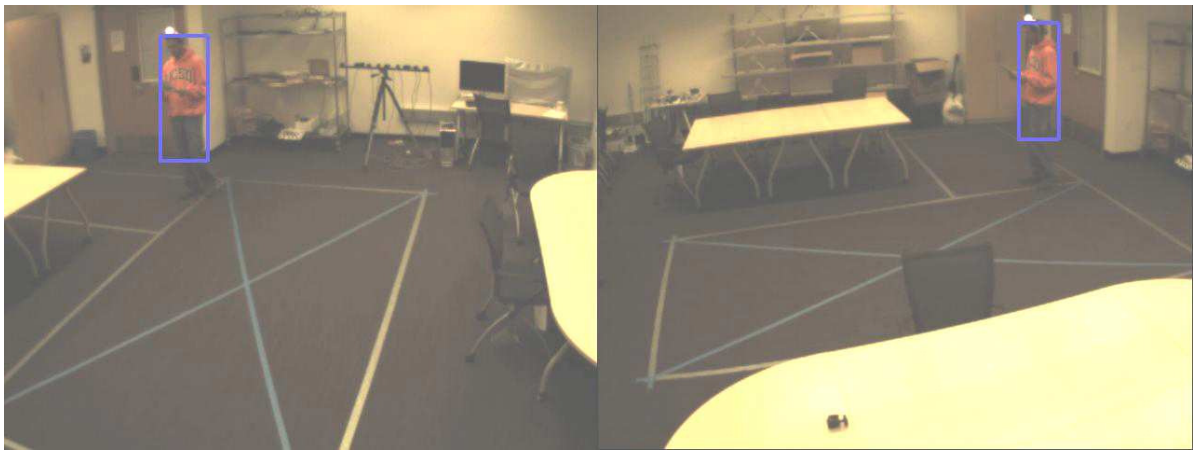

Figure 5. A snapshot from the calibration process is shown here (both the camera views are presented).

Figure 6. A successful tracking scenario where the track is maintained through merging, occlusion and divergence of foreground objects is shown here. By using the information from the other sensors through the iterative decoding scheme, the global tracker is able to maintain the track through an occlusion. Note that we are not using velocity constraints here and hence the global tracker has to rely on information from other sensors to maintain the tracks over such occlusions.