Hierarchical Audio-Visual Cue Integration Framework for Activity Analysis in Intelligent Meeting Rooms

Shankar T. Shivappa
University of California, San Diego
9500 Gilman Drive, La Jolla CA 92093-0434, USA
sshivappa@ucsd.edu
http://cvrr.ucsd.edu

Mohan M. Trivedi Bhaskar D. Rao

Abstract

Scene understanding in the context of a smart meeting room involves the extraction of various kinds of cues at different levels of semantic abstraction. Specifically, human activity in a scene is usually monitored using arrays of audio and visual sensors. Tasks such as person localization and tracking, speaker ID, focus of attention detection, speech recognition and affective state recognition are among them. In this paper we demonstrate a system that extracts such information by synergistically combining the information from the various tasks to support each other. We exploit the fact that the output of one kind of human activity analysis task contains valuable information for another such block and by interconnecting them, a robust system results. We demonstrate this in a smart meeting room context equipped with 3 cameras and 16 microphones. The system performs the tasks of person tracking, head pose estimation, beamforming, speaker ID and speech recognition using audio and visual cues. The novelty lies in putting together the tasks such that they can provide relevant information to one another. We evaluate the performance of our system and present results for tasks such as keyword spotting and tracking re-identification on real-world meeting scenes collected in our audio-visual testbed.

1. Introduction

Scene understanding in the context of intelligent meeting rooms involves, among others, the following important types of information to be extracted from noisy sensory input, usually audio-visual[22].

- What is in the scene? : In the specific case of human activity analysis, this points to detecting the presence or absence of humans. Detecting people in indoor scenes has been addressed using both video and audio sensors.
- Where are the objects at a particular time? : In our case, this corresponds to tracking people continuously using audio and video cues.
- Who are the people present in the scene? : This corresponds to identifying the people using speaker ID or face recognition modules and associating the corresponding tracks and speech segments with them.
- What is happening in the scene? : This corresponds to detection of specific events, keywords in the scene, enabling to draw higher level conclusions about the activity taking place in the room.

These tasks correspond to extracting semantic information at different levels of abstraction. Existing work addresses almost all these individual issues using appropriate modalities. While little is known on how humans understand and interpret the complex visual world, the consensus is that an integration of information at different levels of the semantic hierarchy has to come together for this task. In this paper we demonstrate that not only is such an hierarchical integration of audio and video cues necessary, but it is also beneficial to the performance of the individual blocks because the output of one kind of human activity analysis task contains valuable information for another such block and by interconnecting them, a robust system results. In Figure 1, we illustrate the interconnected blocks in our hierarchical fusion framework.

2. Literature Review

Algorithms for specific multimodal tasks such as person tracking [21][9][5][3], speech recognition[19][15], biometrics[1] and affective state recognition[23] have been researched. General and specific multimodal fusion schemes have also been proposed [18][6][16]. Even hierarchical fusion schemes have been investigated for specific

978-1-4244-3993-5/09/$25.00 ©2009 IEEE

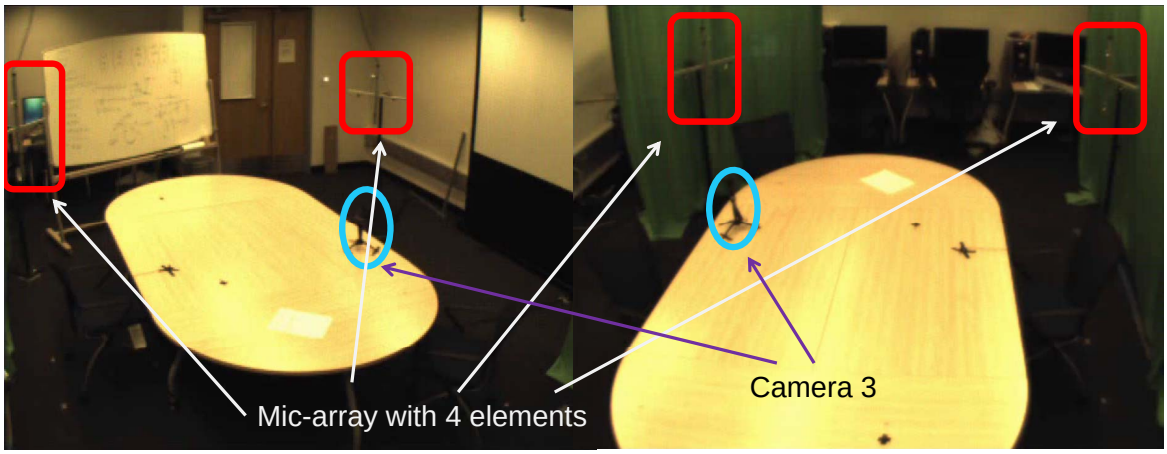

Figure 2. The audio-visual testbed - sensor configurations.

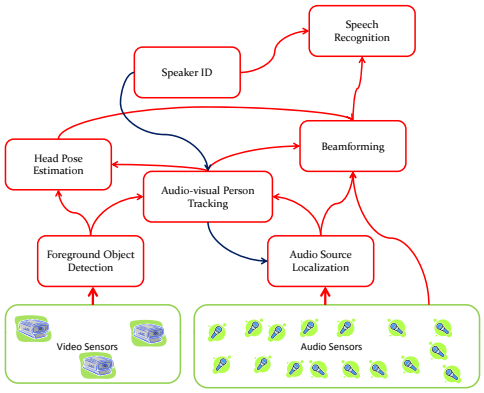

Figure 1. The hierarchical fusion of multimodal information.

tasks [24][17][7][4]. In [4], the authors develop a probabilistic integration framework for fusion of audio visual cues at the track and identity levels. This is an example of fusion at multiple levels of abstraction. Similarly, in [20], the utility of head pose estimation and tracking for speech recognition from distant microphones is explored. In [12], the authors use video localization to enhance the performance of the beamformer for better speech reconstruction from far field microphones. The utility of hierarchical fusion to develop robust human activity analysis algorithms is quite evident from these existing examples. In this paper we develop a hierarchical fusion framework and explore the relationship between tasks such as person tracking, speech recognition, beamforming, speaker identification, head pose estimation and key word spotting. We demonstrate that these tasks can be synergetically performed and the whole is greater than the sum of the parts. The rest of the paper is organized as follows. We describe the individual tasks, the challenges, algorithms and the additional cues that can enhance performance of these algorithms. We then describe the fusion process in the particular context of an intelligent meeting room. We then describe the experimental testbed and provide performance evaluation results from specific tasks.

3. Audio-visual scene analysis tasks

3.1. Person tracking

Tracking persons is necessitated by the non-obtrusive nature of the observing sensors. An intelligent space should be capable of functioning without imposing restrictions on the natural behavior of its inhabitants. This makes the tracking of humans necessary. In this paper we restrict ourselves to the problem of tracking multiple persons in an indoor space equipped with multiple cameras and microphones. Also, we assume that in the general case, the output of the tracker will be the 3D co-ordinates of the multiple subjects. This necessitates the calibration of the audio and video sensors with respect to one another and the world co-ordinates. We discuss this briefly in Section 5.

The tracker is expected to initiate and maintain tracks based on the sensory observations. A simple background subtraction and blob tracking scheme is implemented in each video field of view. Using triangulation and the calibration parameters of the cameras, the 3-D world co-ordinates of the subjects are obtained. The video based tracker could confuse tracks due to occlusions. Also, we would like to relax the conditions of uniform lighting because in many meeting scenarios, especially during presentations, we have drastic variations in ambient lighting situations.

The SRP-PHAT algorithm [8] is used to localize the active speaker. We detect only one source at a time by searching over a uniform grid of points in the world co-ordinates. Unlike the video tracks, audio localization cues on the other

hand are not always available, being robustly estimated when a single dominant speaker is active. This leads us to the data association problem - which track should be associated with the particular active speaker location. We associate the audio clues with the video tracks that are close in the spatiotemporal domain. This association is then used to disambiguate the merging and diverging tracks in the video domain using the iterative decoding algorithm.

3.2. Iterative Decoding Algorithm

In [21], the iterative decoding framework is used to track multiple persons in the sensor co-ordinates, demonstrating improvements in the multimodal tracker's performance when using audio and video sensors. In this paper we use the algorithm to track multiple people in the 3D world co-ordinates. This is possible because our sensors are calibrated with respect to the 3D world co-ordinates and this makes the algorithms simpler and more intuitive compared to [21].

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 world co-ordinates. We consider discrete hidden states, by selecting a grid of points, to make the problem tractable. Thus the hidden states are, in a Bayesian sense, the quantized and 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[2]. 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, $m = 1, 2 \dots N, m' = 1, 2 \dots N$

Observe that these variables allow us to estimate the MAP hidden state,

$$\lambda_t(m) = P(q_t = m, O_1^T) \tag{5}$$
$$= \alpha_t(m) \cdot \beta_t(m) \tag{6}$$
$$\hat{q}_t = \arg\max P(q_t, O_1^T) = \arg\max \lambda_t(m) \tag{7}$$

We can then establish the recursions,

$$\alpha_t(m) = \sum_{m'} \alpha_{t-1}(m') \cdot \gamma_t(m', m) \tag{8}$$
$$\beta_t(m) = \sum_{m'} \beta_{t+1}(m') \cdot \gamma_{t+1}(m, m') \tag{9}$$
$$\lambda_t(m) = \alpha_t(m) \cdot \beta_t(m) \tag{10}$$

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.

$$\lambda_t^{(2)}(m) = P(q_t = m, O_1^T, {Z^{(1)}}_1^T) \tag{11}$$
$$\alpha_t^{(2)}(m) = P(q_t = m, O_1^t, {Z^{(1)}}_1^t) \tag{12}$$
$$\beta_t^{(2)}(m) = P(O_{t+1}^T, {Z^{(1)}}_{t+1}^T | q_t = m) \tag{13}$$
$$\gamma_t^{(2)}(m', m) = P(q_t = m, o_t, Z_t^{(1)} | q_{t-1} = m') \tag{14}$$

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 $P(q_{1,t} =$

$n|q_{2,t} = m)$ which is assumed to be approximated by a rescaled Gaussian kernel with zero mean and whose covariance corresponds to the association of close tracks in the spatial domain. This is possible because states m and n correspond to 3D co-ordinates and we choose a Gaussian-looking discrete distribution to model the conditional density.

Thus, the iterative decoding algorithm provides us with a probabilistic information for the fusion of audio and video tracks.

The tracking ambiguities can be greatly resolved using speaker ID information as elaborated in [4]. However, in that work, the authors consider the output from the far field microphones directly, to evaluate the speaker ID. In our system, we augment the tracker information with speaker ID obtained by beamforming towards the dominant source. The beamformer uses the position information from the tracks and head pose information to reconstruct good quality speech from far field microphones. The overall flow diagram is shown in Figure 1. Thus we see that the tracker performance can be improved by augmenting it with information obtained by the fusion of several other cues. This is a theme that will repeat at various stages of our paper.

3.3. Head pose estimation

The video based head pose estimator adds valuable information to human activity analysis[14]. Head pose is useful in attentional studies, face recognition as well as for effective beamforming[20][13]. Headpose estimation can also be helpful in face recognition based systems [11][4]. In this paper we use the head pose estimates for effective beamforming, thus fusing video cues for better audio processing.

In our experiments we use a simple ellipse fitting method to estimate the head pose. Figure 6 illustrates the process. A skin tone detector is used to extract the face pixels. An ellipse is fit to the skin pixels and the orientation of the ellipse gives us an estimate of the head pose of the subject relative to the camera. We only consider the horizontal direction of the person head and ignore the vertical tilt. A linear regression model is used to map the orientation of the best-fit ellipse with the actual head direction. The skin tone detector is prone to false positives. This is improved by using a mean shift tracker for reducing the search space in the camera view. Also, the ellipse method gives only a coarse estimate of the head pose. The numerical results and analysis are provided in Section (5).

3.4. Beamforming

Beamforming allows us to reconstruct good quality speech from far field microphones. Speech reconstruction from distant microphones in reverberant microphone is a challenging task[10]. But it is essential for better performance of the speech recognizer and speaker ID systems, while using far field microphones. Since we use microphone arrays with arbitrary geometry and wide aperture, we train a delay, filter and sum beamformer using the a stochastic gradient descent algorithm. The structure of the beamformer is shown in Figure 3.

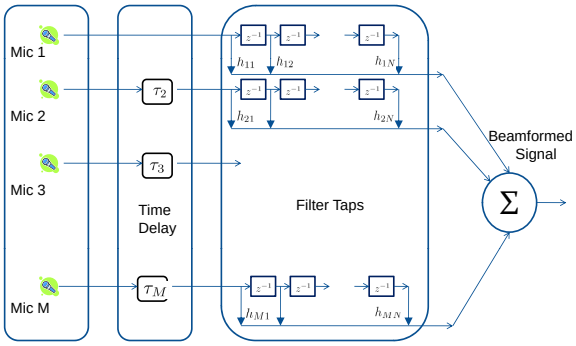

Figure 3. The delay, filter and sum beamformer. The filter taps are trained for specific location and head orientation using a stochastic gradient descent algorithm to match the beamformer output to a close talking microphone. After the taps are determined, the close talking microphone is not needed for beamforming.

The beamformer requires the appropriate delays to align the microphone outputs. In [12], the authors explain the benefits of using video based tracking to enhance speaker localization for effective beamforming. In our setup, we use the beamformer with the smoothed delay estimates obtained from the tracker's output, combined with the active speaker localization. Note that speaker localization by itself is quite noisy and using the tracking information improves the accuracy of the beamformer. Also, in [20], the authors discuss the significance of speaker's head orientation in the beamformer accuracy. In our set up, we use both the location and head orientation information for beamforming.

3.5. Speaker ID

Speaker ID provides valuable information for the tracker in allowing the tracker to recover from failures that cannot be resolved otherwise. For example, it allows us to proceed through drastic lighting changes. In [4], this is called ' identity tracking'. We use a Gaussian mixture model based maximum likelihood classifier to build our speaker recognition system. The models are trained using close talking microphones. Beamformer output is used for recognition. Also, all the speech from a whole speech segment corresponding to a sentence or phrase spoken by a speaker is used for the speaker ID task, boosting the recognition accuracy compared to frame-wise recognition. The details are explained in Section 4. Our basic speaker ID module is based on Gaussian mixture models (GMM) and maximum likelihood classification. The feature vectors are 39 dimensional MFCC co-efficients as used in standard speech recognition

tasks. Each GMM has 4 mixture components with diagonal co-variance matrices.

3.6. Speech Recognition and keyword spotting

Speech recognition in unconstrained meetings is a challenging task. We use a commercially available speech recognition software. In order to boost the performance, we have also trained the models for keyword spotting. This works much better than unconstrained speech recognition which has very poor performance in a multiparty interaction without close talking microphones. The keyword spotting accuracy is one of our main evaluation metrics because it involves the accuracy of the whole system.

4. Hierarchical fusion framework

In Figure 1, we present a flow diagram of the fusion of multimodal cues. The audio and video signals provide the person location information and this is fused in the audio-visual tracking step to come up with robust estimates of the 3D co-ordinates of the subjects. The tracking information is augmented with the speaker ID when available and this betters the re-identification of the tracks in ambiguous cases. The location and head pose estimates are fused for effective beamforming. The reconstructed clean speech from the beamformer is used by the speaker ID module which identifies the active speaker. The speech recognizer uses both the speaker ID and the reconstructed speech to recognize full speech or spot keywords in the utterance.

Thus, when the various blocks for audio-visual human activity analysis are put together, there is a whole range of fusion possibilities to make the system more robust and effective. This fusion hierarchy is the main contribution of this paper. In our experience, there is no other set up which combines audio and visual cues at such varying levels of abstractions to achieve a set of tasks.

The fundamental information extracted from the audio and video sensors are the person locations. For every video frame and the corresponding audio frame, we detect and track foreground objects and sound sources. The tracking module is described in detail in Section 3.1. The output of this module is a set of tracks that are identified with a track ID $\Gamma_i, 1 \leq i \leq N$. In the second pass, we use an adaptive threshold based voice activity detector on audio frame energy for all the microphone outputs to separate speech frames from non speech frames. This is a standard preprocessing step and we do not describe it further here. For the speech frames, the output of the SRP source localizer (Section 3.1) is used to associate the speech frame with one of the active tracks using the nearest neighbor approach.

Let t be the frame index of a speech frame, that is, the audio energy of one of the microphones exceeds the threshold (adaptively set) at frame t. Let the output the of the SRP source localizer be the source location $\overline{x}_s = (x_s, y_s, z_s)$ and the indices of active tracks at time t be T_t. Then, the frame is assigned to track $i = \arg\min_{j \epsilon T_t} d(\overline{x}_s, \overline{x}_{(j,t)})$ where $\overline{x}_{(j,t)}$ denotes the jth track's location at frame t and d is the Euclidean distance metric. We then group together adjacent speech frames assigned to same tracks. This corresponds to forming sentences or phrases that are spoken by the same speaker. This grouping is necessary to increase the performance of the speaker ID module as described in Section 3.5. Note that we are restricting the case to deal with one dominant speaker at any time. For each speech segment we calculate the TDOA and beamform (Section 3.4) to obtain the reconstructed speech. Using this reconstructed speech segment, we ID the speaker as described in Section 3.5. In our present framework we follow a conservative approach and use the speaker ID to index the tracks. Significant improvements are obtained in the re-identification of the persons through merging, occluding and re-entering tracks.

Next, we describe the fusion of information for effective beamforming. In order to reconstruct the speech from far field microphones, the beamformer described in Section 3.4 needs the correct TDOA values to align the signals of the microphones to the reference microphone. For a particular speaker location, there is a fixed TDOA vector. However, small changes in the speaker's mouth position lead to corresponding changes in the TDOA vector. Hence the location information from the tracker is too coarse to calculate the exact TDOA vector. For a set of M microphones, assuming that the first microphone is our reference, the TDOA vector is $M-1$ dimensional vector $\overline{\tau} = (\tau_2, \tau_3, \ldots \tau_M)$, where τ_j is the relative delay between the signals received at the reference microphone and the jth microphone. This is estimated using the GCC-PHAT algorithm [21]. However, an interesting observation is that the TDOA vector evaluated per frame, over a speech segment, is noisy. Using longer frames improves the robustness but makes the process computationally expensive. In our framework the location estimate from the tracker is used to discard the noisy TDOA vectors. Only those TDOA vectors that are within a certain neighborhood of the computed TDOA vector are retained for further processing. At this stage, we make a simplifying assumption that during the course of a speech segment, the source is stationary and we can use the average TDOA vector, computed over all clean frames in the speech segment to align the signals in our beamformer, Also, the beamformer filters are trained for specific locations and hence the speaker location from the tracker is necessary for using the appropriate filter co-efficients. Also, the filters can be trained for specific locations and particular head orientations. In our present scenario, if the head pose information is unavailable, we use head pose agnostic beamformer taps for the particular location. We have described the fusion of the location information from the tracker and the head pose

information for effective beamforming.

The reconstructed speech signal from the beamformer is also used in our keyword spotting experiment. Keyword spotting is one way to address the issue of "what is happening in the scene?". The performance of the keyword spotter is an indicator of the effect of beamforming on reducing the reverberations in the far-field microphone signals. Also, the keyword spotter is a first step in using the spoken words to draw inferences about the scene. However we do not explore such possibilities in the current paper. Moreover, other tasks such as face recognition could be added to the current framework, but such extensions are not described here.

The flow of information described above is summarized in Figure 4. In Section 5, we present a realworld testbed where the performance of such a framework is demonstrated through experimental evaluations.

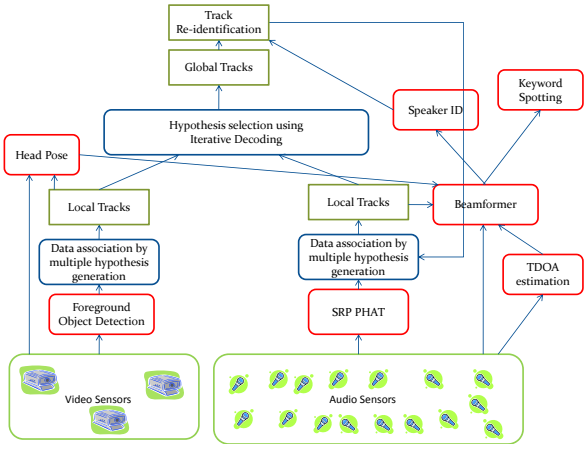

Figure 4. Flowchart summarizing the exchange of audio and visual cues at multiple levels of semantic abstraction.

5. Experimental testbed, datasets and results

The experimental testbed is shown in Figure 2. It is equipped with multiple cameras and microphones of which 3 rectilinear cameras and 16 microphones are used in the experiments described in this paper. The audio and video feeds are completely synchronized. Two of the cameras and all the microphones are calibrated with respect to the world co-ordinates. The third camera provides us with the close up view of the subjects in its field of view, for accurate head pose estimation. The microphones are arranged in 4 sets of 4, located at the 4 corners of the meeting room. The datasets used to evaluate the system consist of 5 to 10 minute long clips of meetings involving 3 to 4 people. The total duration of the test set is 62 minutes. The video frame rate is 15 fps and we use audio frames that are 66ms long, to match the frame rates.

From practical considerations, we have made some simplifying assumptions in out meeting setup. We assume that majority of the meeting interactions occur when the participants are seated or presenting next to the screen. Based on this observation, we have four areas in the meeting room that are considered for training the beamformer filter taps. We use a close talking microphone when training the beamformer for these particular locations. Also, when seated, the participants in the meeting are most likely to face the other participants or the presenter, when speaking. Thus, for each location, there are three head poses that we train our beamformer for. To demonstrate the utility of the head pose estimates, we restrict our focus to one of the locations and use the third camera to obtain a close-up view for head pose estimation. Thus, the tracker results are for the entire space. The beamformer results are averaged over the four locations and the head pose based beamformer results are for the one location monitored by the third camera. The framework itself is general and these restrictions are only for the results presented in this paper.

In Figure 5, we present some snapshots from the audio-visual tracker. In Table 1, we present the results of the various tracker configurations. The results correspond to the percentage of times a track was successfully maintained through an occlusion. Also, the number of true tracks in any meeting scene was 3 or 4. Before using speaker ID, the average number of tracks that would be detected in one meeting scene was 11.3. This was due to track merging, re-entry and occlusions. However, using speaker ID, we were able to resolve the issue of track re-identification in 88% of cases. The errors correspond to the speaker ID errors that led to tracks being tagged with the wrong speaker ID.

The speaker ID module was evaluated separately. The number of people in our dataset is 20. The per-frame performance is presented in Table 2. While using this in our algorithm, we apply the speaker ID module to speech segments rather than frames of speech. Each speech segment is an average of 15 frames long. On such long segments of speech, the speaker ID module performs with 98% accuracy on the close talking microphone and 95% accuracy on the beamformed signal. However, there are many short speech segments that contribute to tracking errors.

The beamformer is illustrated in Figure 7. During our experiments, the average SNR of a farfield mic was 11dB. The SNR at the output of the beamformed mic was 16dB. The beamformer can also undo the channel effects to a great extent. This is important in reverberant environments and we illustrate this in a keyword spotting experiment.

The headpose estimation using ellipse fitting is coarse. In our experiments, the root mean squared error of the head pose estimate was 21^o. However, for our beamformer, we are interested in three orientations of the head that correspond to -60^o, -10^o and 30^o. Thus, the high MSE of the head pose module is not an issue in our current setup. However, in order to be generalizable, a more accurate head pose

Table 1. Audio-visual person tracker's performance before and after incorporating the speaker ID information.

1 camera	1 camera and 16 microphones	1 camera and 16 mics. and Speaker ID	2 cameras and 16 microphones and	2 cameras and 16 mics. and Speaker ID
49%	56%	80%	76%	85%

Table 2. The speaker ID module's performance on our 20 person dataset

Nearfield mic	Farfield mic	Beamformed signal
89%	56%	78%

estimator is necessary.

In Table 3, we show the results of a keyword spotting experiment conducted on the output of the final beamformed signal. A set of 20 commonly occurring words were chosen for the experiment. The keyword spotting was implemented in the commercially available Dragon Naturally speaking software. Because we do not have access to the parameters used in the software, it is not feasible to present ROC curves for the keyword spotting task.

6. Concluding Remarks

We have presented a novel hierarchical fusion strategy for the fusion of audio and visual cues at different levels of abstraction. This not only facilitates a holistic approach to scene understanding but also provides performance improvements by fusion relevant cues from different blocks. The results are promising. There are many open issues to be addressed in the near future including the extension of the system to include more blocks like face recognition, gesture recognition etc. Similar fusion approaches can be investigated in other scenes such as intelligent vehicles. Such an extension will also facilitate the comparison of fusion strategies in different scenes and might eventually lead to the development of a general scheme for hierarchical fusion of multimodal cues.

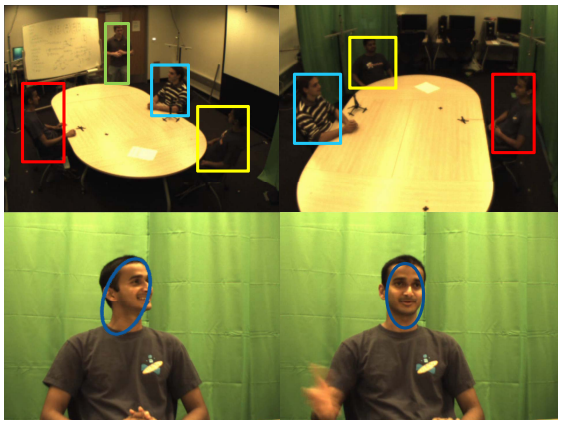

Figure 5. Snapshots from a typical meeting scene in the dataset. The three views from cameras are shown and the head pose estimation algorithm is illustrated. The coloring of the boxes correspond to the identities of the subjects.

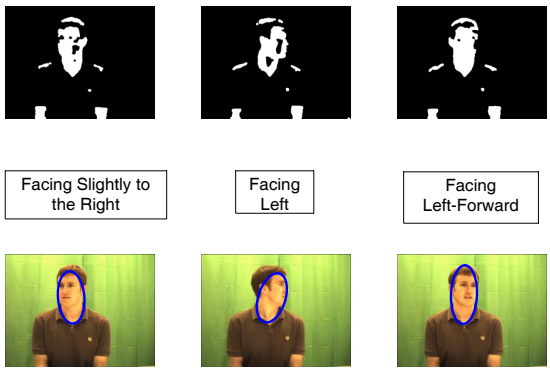

Figure 6. Head pose estimation using ellipse fitting on skin tone pixels. The first row of images shows the extracted skin pixels.

References

[1] P. S. Aleksic and A. K. Katsaggelos. Audio-visual biometrics. *Proceedings of the IEEE*, 94(11):2025–2044, Nov. 2006. 1

[2] 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. 3

[3] K. Bernardin, T. Gehrig, and R. Stiefelhagen. Multi-Level Particle Filter Fusion of Features and Cues for Audio-Visual Person Tracking. In *CLEAR Evaluation Workshop 2007*, 2007. 1

[4] K. Bernardin, R. Stiefelhagen, and A. Waibel. Probabilistic integration of sparse audio-visual cues for identity tracking. In *MM '08: Proceeding of the 16th ACM international conference on Multimedia*, pages 151–158, New York, NY, USA, 2008. ACM. 2, 4

[5] 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*, 2004. 1

[6] T. Choudhury, J. M. Rehg, V. Pavlovic, and A. Pentland. Boosting and structure learning in dynamic bayesian networks for audio-visual speaker detection. In *ICPR '02: Proceedings of the 16 th International Conference on Pattern Recognition (ICPR'02) Volume 3*, page 30789, Washington, DC, USA, 2002. IEEE Computer Society. 1

[7] P. Dai and G. Xu. Context-aware computing for assistive meeting system. In *PETRA '08: Proceedings of the 1st international conference on PErvasive Technologies Related to Assistive Environments*, pages 1–7, New York, NY, USA, 2008. ACM. 2

Table 3. Keyword spotting results using clean speech from the beamformer. The insertion, substitution and deleting error rates are presented, lower rate are better. Note that the last column has significant improvement over the second and third.

	Near field microphone	Far field microphone	Beamformed waveform (head pose agnostic)	Beamformed waveform (head pose specific)
Insertion	20%	40%	30%	25%
Deletion	6%	26%	25%	8%
Substitution	2%	4%	4%	4%

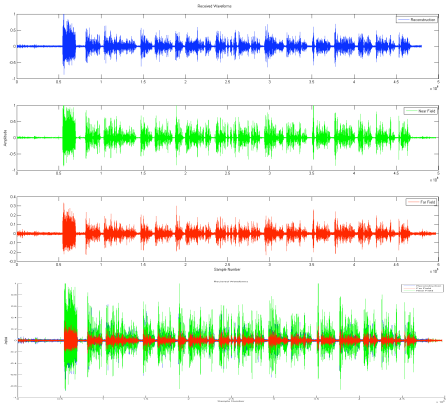

Figure 7. The beamformer output, far field and near field speech waveforms. In the last column the three are plotted on top of one another to illustrate the improvement in SNR obtained by beamforming - the average SNR of a farfield mic was 11dB. The SNR at the output of the beamformed mic was 16dB. The near field microphone has a SNR of 21dB.

[8] J. H. DiBiase, H. F. Silverman, and M. S. Branstein. Robust localization in reverberant rooms. *Microphone Arrays: Signal Processing Techniques and Applications*, 2001. 2

[9] 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. 1

[10] T. Gustafsson, B. D. Rao, and M. M. Trivedi. Source localization in reverberant environments: Modeling and statistical analysis. *IEEE Transactions on Speech and Audio Processing*, 11(6), Nov. 2003. 4

[11] K. S. Huang and M. M. Trivedi. Integrated detection, tracking, and recognition of faces with omnivideo array in intelligent environments. *EURASIP Journal on Image and Video Processing*, 2008(1), 2008. 4

[12] H. K. Maganti, D. Gatica-Perez, and I. McCowan. Speech enhancement and recognition in meetings with an audio-visual sensor array. *IEEE Trans. on Audio, Speech, and Language Processing*, 15(11):2257–2269, Nov 2007. 2, 4

[13] E. Murphy-Chutorian and M. M. Trivedi. 3d tracking and dynamic analysis of human head movements and attentional targets. *Distributed Smart Cameras, 2008. ICDSC 2008. Second ACM/IEEE International Conference on*, pages 1–8, Sept. 2008. 4

[14] E. Murphy-Chutorian and M. M. Trivedi. Head pose estimation in computer vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 99(1), 5555. 4

[15] 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. 1

[16] 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. 1

[17] 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. 2

[18] S. T. Shivappa, B. D. Rao, and M. 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*, 2008. 1

[19] S. T. Shivappa, B. D. Rao, and M. M. Trivedi. Multimodal information fusion using the iterative decoding algorithm and its application to audio-visual speech recognition. In *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, Las Vegas, Nevada, USA, 2008. 1

[20] S. T. Shivappa, B. D. Rao, and M. M. Trivedi. Role of head pose estimation in speech acquisition from distant microphones. In *Accepted at Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, Taiwan, 2009. 2, 4

[21] S. T. Shivappa, M. M. Trivedi, and B. D. Rao. Person tracking with audio-visual cues using the iterative decoding framework. In *5th IEEE International Conference On Advanced Video and Signal Based Surveillance*, Santa Fe, New Mexico, USA, [Best Paper Award], 2008. 1, 3, 5

[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. 1

[23] Z. Zeng, J. Tu, B. M. Pianfetti, and T. S. Huang. Audiovisual affective expression recognition through multistream fused hmm. *Multimedia, IEEE Transactions on*, 10(4):570–577, June 2008. 1

[24] D. Zhang, D. Gatica-Perez, S. Bengio, I. McCowan, and G. Lathoud. Modeling individual and group actions in meetings: A two-layer hmm framework. *Computer Vision and Pattern Recognition Workshop, 2004 Conference on*, pages 117–117, June 2004. 2