SMART ROOM: PARTICIPANT AND SPEAKER LOCALIZATION AND IDENTIFICATION

Carlos Busso, Sergi Hernanz, Chi-Wei Chu*, Soon-il Kwon, Sung Lee*,
Panayiotis G. Georgiou, Isaac Cohen*, Shrikanth Narayanan

Integrated Media Systems Center, Department of Electrical Engineering,
*Department of Computer Science
Viterbi School of Engineering, University of Southern California, Los Angeles

ABSTRACT

Our long-term objective is to create Smart Room Technologies that are aware of the users presence and their behavior and can become an active, but not an intrusive, part of the interaction. In this work, we present a multimodal approach for estimating and tracking the location and identity of the participants including the active speaker. Our smart room design contains three user-monitoring systems: four CCD cameras, an omnidirectional camera and a 16 channel microphone array. The various sensory modalities are processed both individually and jointly and it is shown that the multimodal approach results in significantly improved performance in spatial localization, identification and speech activity detection of the participants.

1. INTRODUCTION

New developments in communications technologies have brought to light a number of exciting and challenging applications that promise to change the way people communicate and interact. An application that has recently gained significant attention in the literature is the development of multimodal, unobtrusive *Smart Room Technologies* (SRT): monitor and infer important clues about users in specific environments such as their spatial position, identities and behavior. This is a challenging multidisciplinary application that involves research in diverse topics including object tracking, speaker activity detection, speaker identification, human action recognition and user behavior modeling.

One of the well-studied areas in SRT is the *detection and tracking of user locations*. Two important sources of information are the visual and the acoustic modality. Within a multimodal framework, these two sources have been used to track a single active speaker using methods such as *Sequential Monte Carlo* [1] [2], *Kalman filtering* [3] and *Dynamic Bayesian Networks* (DBN) [4], taking advantage of the complementary information represented by these two modalities.

Recently, [5] and [6] extended these approaches to track multiple speakers using particle filtering, while at the same time achieving *active speaker detection*, which is another important aspect of smart room technologies. In [7], visual clues were used to track users and a microphone array to select the active speaker by computing the distance between the visual and acoustic results.

Another important aspect of SRT is *speaker identification* (SID), in which the identity of the user is detected. There are several additional possible biometric systems for smart room applications (*e.g.,* retina, fingerprint), although most of them are impractical due to their invasive nature. One feasible option is to classify the user according to acoustic speech features [8] or through face recognition.

In this paper, we propose a real time multimodal approach to determine the spatial position of the user, detect speaker activity, and additionally determine the speaker's identity aimed at applications such as remote video-conferencing and audio-video indexing and retrieval for tasks such as meetings.

Our conference room contains three user monitoring systems: four synchronized cameras located in the corners of the room, a full-circle 360 degree camera located at the center of the table, and an array of sixteen microphones located at the end of the table. The location of each user is computed based on (i) the 3D polygon surface model from 4 synchronized cameras and (ii) a face detection technique using a full-circle 360 degree camera. Subsequently a dynamic model, under the Gaussian distribution assumption, is used with a moving window to combine the above information and *localize the participants*. The *Speaker ID*, operating on far-field sound obtained from the microphone array, algorithm employs a standard Gaussian Mixture model based on MFCCs. Finally, the active speaker's *identity and location* is estimated by fusing all the information channels.

The long-term objective of this project is to create a system which is cognizant of the users and can become an active but non-intrusive member of the ir interaction. The specific goal of this paper is to present a smart room design suitable for real time multi-speaker remote video-conferencing, with augmented information channels containing speaker IDs and relative location of the participants and the active speaker. Moreover, the extracted information can be used in a number of other applications such as video indexing and retrieval, human posture inference [9], modeling of human behavior, and as the device technologies further mature, for applications such as audio-visual speech and emotion recognition [10].

2. THE SMART ROOM

The present initial design primarily comprises microphones and cameras for activity sensing. The microphone array consists of 16 omnidirectional microphones that process sound at 48kHz sampling frequency. Fourteen microphones are distributed on a square frame of 50×50cm and two microphones are raised in the middle of the frame to allow for vertical plane localization. The room is acoustically treated on three walls and has a full-wall glass window on the other side, and has ceiling panels and carpeting on the floor.

The 3D camera system consists of 4 firewire CCD cameras near the corners on the ceiling that overlook the meeting area around the main table and capture the image sequences of the meeting from multiple angles. Each camera provides 1024×768 images at 15 frames per second, but we scale them to 320×240 for real time processing. The room is lighted with halogen lights.

At the center of the meeting table, a full-circle omnidirectional (360°) camera captures the faces of all participants. The size of the original omnidirectional image is 1280×960.

The next subsection describes the algorithms used to process each of these raw information sources.

2.1. Microphone array

One modality of localization is the sound source localization using a microphone array. The principle of sound source localization is based on the *Time Difference of Arrival* (TDOA) of the sound to the various sensors and the geometric inference of the source location

Fig. 1. Four firewire CCD cameras

from this TDOA. In this microphone array implementation we first estimate the pair-wise delays [11] and then employ a least-squares estimation procedure for the source localization[12].

Georgiou *et al*[11] have demonstrated the impulsive nature of audio signals and introduced a time delay estimation approach to mitigate its effects. The algorithm called *Fractional Lower Order Statistics-Phase Transform Method* (FLOS-PHAT) is based on a signed-non linearity on the input signal that reduces the detrimental effects of outliers.

As is common practice, this implementation of the FLOS-PHAT algorithm employs memory in order to approximate the expectation in the lower order statistics, and additionally the memory varies as a function of time to mitigate temporal propagation of errors.

Subsequently, based on the TDOA estimates, a computationally simple algorithm presented by Huang *et al*[12], called *One Step Least Squares* (OSLS), can be used to spatially locate the souce using these pairs of delays.

The resulting localization algorithm is quite robust, but as expected, not very accurate in range due to the small aperture of the array. We expect, however, that this shortcoming will be countered by the visual modalities, which have higher accuracy in the horizontal plane.

2.2. Speaker ID

Speaker identification was implemented by analyzing the short-time spectrum (through mel frequency cepstral coeeficients, MFCCs) of the spoken phrases. In speaker recognition, the *Gaussian Mixture Model* (GMM), a weighted sum of Gaussian distributions, has been found to be good to capture the speaker information in MFCCs, and hence a GMM with 16 mixtures was used as a speaker model. Model training was accomplished by the standard *Expectation-Maximization* (EM) algorithm. All frames were initially divided into 16 clusters. An initial model was obtained by parameter estimation (for mean and covariance matrices), which were estimated from the vectors in each cluster. The prior weights of GMM can be simply set by the proportion of feature vectors in each cluster. Next, the feature vectors are clustered by the *Maximum Likelihood* (ML) method using the previously estimated model. This process is iteratively executed until the model parameters converged. Additionally, we have created a silence/background noise model.

The speech signal was obtained through beamforming from the microphone array (see Fig. 4). The result of the speaker identification was in terms of pairs, (S_i, P_i), where P_i refers to the probability of speech activity of speaker S_i given for all speakers i. This information is evaluated and transmitted to the fusion algorithm every 1 second.

We should note that the acoustic signal processed is a reverberant, far-field signal corrupted by noise, and so the performance of this method is expected to be lower compared to a case when clean signal from a close-talking microphone is to be used.

Fig. 2. Omnidirectional image from 360°camera and its panoramic transform

2.3. Video detection

The goal of visual tracking is to detect and track the 3D locations of the participants in the meeting room using video streams acquired by multiple synchronized cameras.

We use a Gaussian background-learning model to segment moving regions in the scene. When large variations from the learned Gaussian models are detected the foreground pixels are extracted. These pixel changes are then merged into regions. However, this method will segment actual people as well as their shadows and reflections. In our indoor setting, the shadow regions cast by diffused light do not have strong boundaries. We eliminate the shadows by combining the foreground pixels detection and the edge features detection [9] for segmenting into moving regions and corresponding cast shadows. The resulting regions are the silhouettes of the moving objects in the room.

The detected silhouettes across the views are integrated for inferring the 3D visual hulls of people in the room [13]. The silhouette contour is converted to a polygon approximation and a visual hull with polyhedral representation is then computed directly from these polygons [14]. This polygonal 3D approximation of the shapes is fast and is done in real-time. In detecting the locations of the people in the meeting room, we only need an estimation of general location of blobs of shapes instead of a precise reconstruction. Furthermore, we want the detection to cover an area as large as possible given a limited number of cameras. For this purpose we use a variation of the visual hull method proposed by [15]: the polyhedral visual hull is required to be the integration of only a subset (at least 3 out of 4) of the silhouettes instead of all of them. The resulting visual hull shape is less accurate, but the 3D shape of all people in the room can be approximated.

The computed visual hull is in a polygonal representation. We randomly sample points on the polygon surface and construct a height map of those points. This map assumes the XY plane in the Cartesian space is the meeting room floor and the Z coordinate represents the height. The local maximums of the height are then detected and considered as heads of the meeting participants. In this process some thresholds are applied to eliminate small regions such as moving chairs.

2.4. Full-circle 360-degree camera

We have added an omnidirectional (360°) camera on the meeting table to capture faces of all participants in order to get thumbnail representation of "who's talking". The image of the omnidirectional (360°) camera is the result of the projection of the surrounding scene into a hemisphere. We can unroll the captured original image and project it back onto a cylinder as in Fig. 2.
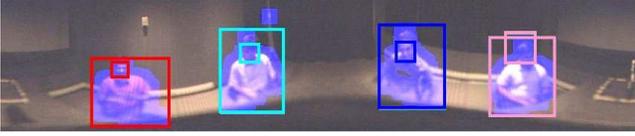
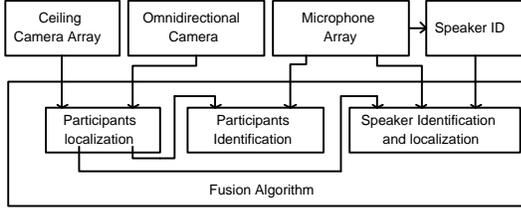**Fig. 3**. Detection of participants' faces with the 360°camera

Fig. 4. The system is distributed running over TCP, with information exchange as depicted above.

To detect the foreground region, we use adaptive Gaussian background-learning model. All pixels in a new frame are compared to the current color distribution in order to detect moving blobs prior to capturing the faces. Morphological operators are used to group detected pixels into foreground regions, and small regions are eliminated. Pixel color distributions are updated in these regions for adapting the background model to slow variations. In these moving regions, we perform face detection. The face detector is based on Haar-like features and is implemented using Intel's open source computer vision library [16]. To accurately detect faces under low light level conditions, the color histogram of detected regions is normalized beforehand. Detected regions are then tracked using a graph-based tracking approach [17]. These regions correspond to the upper body of the meeting participants. Spatial and temporal information of tracking regions are combined as a graphical structure where nodes represent the detected moving regions and edges represent the relationship between two moving regions detected in two separate frames. Each newly processed frame generates a set of regions corresponding to the detected moving objects. The size of original omnidirectional image is 1280×960 and the panoramic image resolution is 848×180. The average size of detected faces is approximately 30×30. The faces are detected and tracked at approximately 13 FPS in a 2.8 GHz Pentium4 PC. In Fig. 3 we show an example of detection and tracking of the participants' faces during a meeting.

2.5. Synchronization

Each modality was initially processed independently and asynchronously. Therefore, the estimated 3D coordinates from the polygonal representation (X_v) and from the microphones array (X_{MA}), the angles of the faces detected (X_θ), and the speaker information from the acoustic analysis (S_i, P_i) are sent to the fusion algorithm for integration. Although the results are received in an asynchronous manner, they are transformed and processed in a synchronous fashion.

3. MULTIMODAL INTEGRATION

The various modalities are subsequently received and processed by a fusion algorithm for the purpose of finding and tracking the participants' spatial locations and identifying where and who the current active speaker is. Fig. 4 shows the information flow between the various modules, and what information is used for each decision.
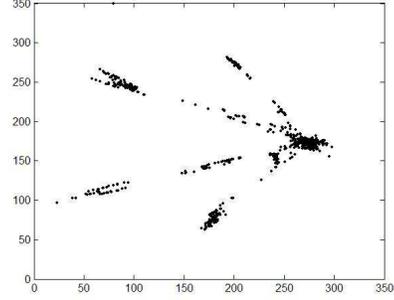

Fig. 5. Microphone Array Distribution

3.1. Participant localization

It is well known that visual tracking algorithms have better spatial resolution than acoustic localization techniques [7, 6]. Hence, our algorithm for localization of all the participants' location employs a dynamic visual approach that uses only information obtained by cameras $X = (X_v, X_\theta)$. Based on the distribution of the samples X, we model the position of each speaker as a multidimensional Gaussian distribution.

A single distribution with covariance K of a significant spread and mean M is initialized at the center of the room. As data are obtained, the variance and mean converge to the detected object's location. When information is received for a location scoring below a certain threshold of belonging to the existing distribution, a new multidimensional Gaussian is initialized at (M, K). The process continues sequentially until all the speakers are detected, with new data points either spawning new participant models or adapting the existing ones. In addition, temporal filtering ensures that false participant detections are identified and removed. This procedure allows us to determinate not only the spatial positions of the participants (X_P), but also the number of participants in the room (N_P).

3.2. Participant Identification

The spatial location of the current speaker (X_{MA+P}) as obtained from the microphone array (X_{MA}) and participants' location information (X_P), as well as the speaker ID from the GMM algorithm (S_i, P_i) are used to determine the identities of the participants. The goal is to detect who the participants are and also correlate their identity with their location in space (derive the "seating arrangement").

Fig. 5 shows a sample scatter plot of the raw microphone array localization X_{MA}, and as can be observed, the range information is highly noisy. For simplicity, we model $P(C_i|X_{MA})$, the probability that the acoustic source comes from cluster i given X_{MA}, as a multidimensional Gaussian distribution centered at the locations X_P and with a large variance in range and smaller variance in the other two dimensions.

Using (S, P), the probabilistic identity of the participant along with $P(C|X_{MA})$, the probabilistic location of the current speaker, over time and with physical constraints[1] we estimate the participants seating arrangement (L).

3.3. Speaker Identification and Localization

We compute activity speaker detection by employing all modalities: X_{MA+P}, which is derived from the visual modality and the microphone array, and (S, P) obtained from the acoustic analysis of the signal. The information is fused as described in (1), where r_{ij} is the correlation measure between the probabilities of the current speaker belonging in cluster j and being speaker i.

---

[1]Such that a participant can only be at one point in space at a time, and one position can only be occupied by one participant at a time

		Session	Strong Decision	Weak Decision	Decision
A	Speaker ID (GMM based)	1	58.30%	60.70%	ID
		2	56.70%	58.40%	ID
B	Microphone Array + Video	1	68.10%	69.50%	Loc
		2	71.00%	72.00%	Loc
C	Microphone Array + Video + Speaker ID (Assumes known seating arrangement L)	1	73.20%	76.50%	Loc+ID
		2	77.90%	79.50%	Loc+ID
D	Microphone Array + Video + Speaker ID (Participant location (L) learned through data)	1	73.80%	77.60%	Loc+ID
		2	74.90%	76.70%	Loc+ID
E	Speaker-location learned through data (L)	1	93.30%		L
		2	94.90%		L

Fig. 6. All of the above results are obtained in real time, and include the whole length of the meeting, with *no* time given for initial convergence. **A**: Speaker ID as obtained purely from the speech signal using a GMM; **B**: Localization obtained by the two visual information channels and the microphone array; **C**: Speaker Identification & Localization based on all information channels. Assumes perfect knowledge of L, the seating arrangement of the participants; **D**: As C, but the mapping of speaker-location, L, is continuously estimated from the data; **E**: Speaker Location mapping, L.

$$P(S_i) = P_i \cdot \sum_i^n r_{ij} \cdot P(C_j | X_{MA}) \tag{1}$$

4. RESULTS & DISCUSSION

The experiments were performed using two meetings (each 5 minute long) with four participants, processed in real time. Off-line computations were also performed later for comparison purposes. The conversation in the meetings was casual with many interuptions, overlaps and short utternaces, making this an extremely challenging task for both the microphone array and the Speaker ID. We used two criteria: strong decision, in which the detection is considered correct if the speaker was active at least 50% of the time interval, and weak decision, in which the detection is considered correct if the speaker was active in any part of the time interval.

The participants localization algorithm takes about 3 seconds per participant to converge during the start of the meeting. As can be observed from the results in Fig. 6 (rows C & D), the speaker identification and localization based on all the modalities is fairly robust, achieving about 70% performance. This is a significant improvement of about 30% compared to the speaker ID based purely on the speech signal as shown in row A, which suffers from the far field and noisy nature of the data.

Similarly, there is a significant improvement in the accuracy of localization (row B) as contrasted to the performance based purely on the microphone array. The microphone array as a single modality is very unreliable when it comes to the range of the speaker (as can be observed from Fig. 5), both due to the noisy environment as well as the range errors due to the aperture of the array. The multimodal localization accuracy is also further improved by the acoustic speaker ID modality, as the correlation between active speaker and sound source location is providing additional information. This results in about 10% improvement when comparing all modalities (row C & D) versus the visual and microphone array only (row B).

Finally, the identification of the participants' spatial arrangement (row E) is extremely accurate, a fact that explains the very close results observed in rows C & D.

5. CONCLUSION

In this paper, we have presented the first results from the smart room that we are developing at USC. We have demonstrated that complementary modalities can increase the general participant identification and localization (without any prior knowledge of the number of participants) including the active speaker identification and localization.

Our goal of increasing the system's awareness of the users in the space has many more challenges ahead. In our future work we propose to investigate further integrated recognition technologies including face recognition, gesture recognition and head pose estimation. Additionally we plan to collect and share with the research community a multimodal data corpus from this testbed.

6. REFERENCES

[1] J. Vermaak, M. Gangnet, A. Blake, and P. Perez, "Sequential monte carlo fusion of sound and vision for speaker tracking," in *International Conference on Computer Vision*, 2001, vol. I, pp. 741–46.

[2] D. Zotkin, R. Duraiswami, and L.S. Davis, "Multimodal 3-d tracking and event detection via the particle filter," in *IEEE Workshop on Detection and Recognition of Events in Video*, 2001, pp. 20 – 27.

[3] S. Spors, R.Rabenstein, and N.Strobe, "Joint audio-video object localization and tracking," *IEEE Signal Processing Magazine*, vol. 18, no. 1, pp. 22 31, Jan 2001.

[4] V. Pavlovic, A. Garg, J.M. Rehg, and T.S. Huang, "Multimodal speaker detection using error feedback dynamic bayesian networks," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2000, vol. II, pp. 34 – 41.

[5] N. Checka, K. Wilson, M. Siracusa, and T. Darrell, "Multiple person and speaker activity tracking with a particle filter," in *International Conference on Acoustics, Speech, and Signal Processing*, 2004, vol. V, pp. 881–84.

[6] D. Gatica-Perez, G. Lathoud, I. McCowan, J.-M.Odobez, and D. Moore, "Audio-visual speaker tracking with importance particle filters," in *International Conference on Image Processing*, 2003, vol. III, pp. 25–8.

[7] G. Pingali, G. Tunali, and I. Carlbom, "Audio-visual tracking for natural interactivity," in *Proceedings of the seventh ACM international conference on Multimedia*, Orlando, Fl, 1999, p. 373 382.

[8] S. Kwon and S. Narayanan, "A study of generic models for unsupervised on-line speaker indexing," in *Proceedings of IEEE Automatic Speech Recognition and Understanding Workshop*, 2003, pp. 423–28.

[9] I. Cohen and H. Li, "Inference of human postures by classification of 3d human body shape," in *IEEE International Workshop on Analysis and Modeling of Faces and Gestures*, 2003, pp. 74–81.

[10] C. Busso, Z. Deng, S. Yildirim, M. Bulut, C.M. Lee, A. Kazemzadeh, S. Lee, U. Neumann, and S. Narayanan, "Analysis of emotion recognition using facial expressions, speech and multimodal information," in *Sixth International Conference on Multimodal Interfaces ICMI 2004*, State College, PA, 2004.

[11] P. G. Georgiou, P. Tsakalides, and C. Kyriakakis, "Alpha-stable modeling of noise and robust time-delay estimation in the presence of impulsive noise," *IEEE Transactions on Multimedia*, vol. 1, no. 3, pp. 291–301, September 1999.

[12] Y. Huang, J. Benest, and G. Elko, "Adaptive eigenvalue decomposition algorithm for realtime acoustic source localization system," in *International Conference on Acoustics, Speech, and Signal Processing*, Phoenix, AZ, 1999, vol. II, pp. 937–40.

[13] A. Laurentini, "The visual hull concept for silhouette-based image understanding," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 16, no. 2, pp. 150162, Feb 1994.

[14] W. Matusik, C. Buehler, and L. McMillan, "Polyhedral visual hulls for real-time rendering," in *In Proceedings of Eurographics Workshop on Rendering*, 2001.

[15] E. Boyer and J.-S. Franco, "A hybrid approach for computing visual hulls of complex objects," in *Computer Vision and Pattern Recognition (CVPR'03)*, 2003, vol. I, pp. 695–701.

[16] A. Kuranov, R. Leinhart, and V. Pisarevsky, "An empirical analysis of boosting algorithms for rapid objects with an extended set of haar-like features," in *Intel Technical Report MRL-TR-July02-01*, 2002.

[17] I. Cohen and G. Medioni, "Detecting and tracking moving objects for video surveillance," in *Computer Vision and Pattern Recognition (CVPR'99)*, 1999, vol. II, pp. 319–325.