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

- 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: (1) four CCD cameras, (2) an omnidirectional camera and (3) a 16 channel microphone array
- Significant gains in speaker identification and localization can be achieved by employing multimodal approach results

Introduction

Objectives and Goals

- Long term goal is to create a system which is cognizant of the users and can become an active but non-intrusive member of the interaction
- Multidisciplinary approach that involves research in topics including object tracking, speaker activity detection and identification, human action recognition and user behavior modeling
- The present initial design primarily comprises microphones and cameras for activity sensing
- This work focuses on speaker identification & localization Example applications include:
  - real-time multispeaker remote video conferencing, where low resource participants can receive augmented information channels containing speaker ID's and relative location of the active and inactive participants
  - augmented summarization with information such who the active speaker is, who he is addressing and who are the other participants of the interaction and where they are all located
- The extracted information can be used in a number of other applications such as video indexing and retrieval, human posture inference [Cohen,03], modeling of human behavior, and as the device technologies further mature, for applications such as audio-visual speech and emotion recognition [Busso,04]

Related Work

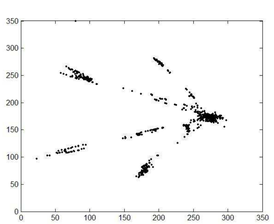
- Detection and tracking of single user locations: *Sequential Monte Carlo* [Vermaak,01] [Zotkin,01], *Kalmar filter* [Spors,01], *Dynamic Bayesian Network* [Pavlovic,00]
- Multiple speaker tracking [Checka,04], [Gatica-Perez,03]
- Speaker identification[Kwon,03]

The Smart Room

	Configuration	Sample Rate	Location
Microphone Array	16 microphones	48 KHz	one size of the table
Calibrated Cameras	4 firewire CCD Camera	15 fps (1024×768)	near corners of the ceiling
Omnidirectional Camera	1 full-circle camera	13 fps (1280×960)	center of the table

Microphone array

- TDE based localization employing FLOS-PHAT variant [Georgiou,99]. Employs lower order statistics based on α-Stable theory
- OSLS algorithm employed for geometric mapping [Huang,99] (Equivalent to spherical interpolation)



Microphone Array Localization Distribution

- The localization algorithm is robust, but not very accurate in range due to small aperture of the array

Speaker ID

- Speaker identification using 16-mixture GMMs for speakers and silence/background noise model
- The Speaker identification results is given probabilistically (S_i, P_i)
- Speech signal obtained through beamforming

Video detection

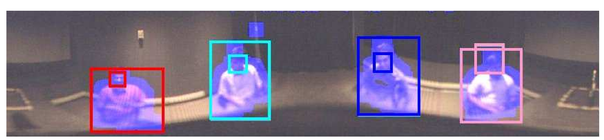
- Gaussian background-learning for segmenting moving regions
- Shadows are removed by combining foreground pixels and edge features detection [Cohen,03] resulting in silhouettes
- Silhouettes converted to a polygon approximations and a visual hull is then computed [Matusik,01]
- A height map is constructed and the local maxima of the height are considered as heads of the participants
- Thresholds are applied to eliminate small regions such as moving chairs

Four firewire CCD cameras, silhouettes and height map

Full-circle 360-degree camera

- Image of the omnidirectional camera, which is formed by projecting the surrounding scene into a hemisphere, is unrolled and projected it back onto a cylinder

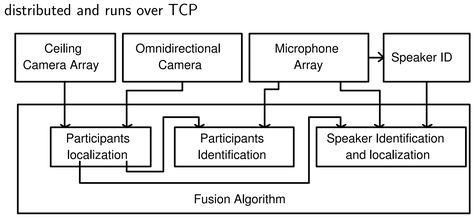
- Moving blobs are detected with Gaussian background model and morphological operators
- In these moving regions, we perform face detection, which is based on Haar-like features and is implemented using Intel's open source computer vision library [Kuranov, 02]
- Detected regions (upper bodies) are tracked using a graph-based approach [Cohen,99]



Detection of participants' faces with the 360°camera

Multimodal Integration

- Each modality was processed independently and asynchronously
- The fusion algorithm receives: 3D coordinates from the polygonal representation (X_v) and from the microphones array (X_{MA}), the angles of the detected faces (X_θ), and the speaker information from the acoustic analysis (S_i, P_i)
- The system is distributed and runs over TCP



LOC - Participants localization

- *Identify the spatial location of all participants*
- The position of each speaker is modeled as a multidimensional Gaussian distribution (M,K), whose parameters are obtained from the samples $X = (X_v, X_\theta)$
- The speakers are detected sequentially, by using a Gaussian model whose parameter (M,K) are continually adapted to fit the received data
- If new data scores low on the multidimensional densities of existing speakers a new participant is added at the center
- Temporal filtering ensures that false participant detections are identified and removed
- As result, the number (N_P) and the spatial positions of the participants (X_P) is estimated

L - Participant Identification

- *Detect the identity and the spatial location of the participants in the room*
- The probability that the acoustic source comes from cluster i given X_{MA}, $(P(C_i|X_{MA}))$, is modeled as a multidimensional Gaussian distribution centered at the locations X_P with large variance in range and smaller variance in the other two dimensions
- $P(C|X_{MA})$ and (S_i, P_i) are used to determine the identities of the participants and the seating arrangement (L), over time and with physical constraints

ID - Speaker Identification and Localization

- *Identify current speaker*
- The active speaker is found by fusing all the modalities
- The ID is found by employing a correlation measure r_{ij} between the probabilities of the current speaker belonging spatially in cluster j and being speaker i

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

Results and Discussion

- Two meetings (5 minute) with four participants, processed in real time. Casual conversation with many interruptions, overlaps, short utterances and with *no* time given for initial convergence
- Strong decision: Speaker was active at least 50% of the time interval
- Weak decision: Speaker was active in any part of the time interval
- **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 and Localization based on all modalities. (L) is assumed known
- **D**: As C, but the seating arrangement, (L), is continuously estimated from the data
- **E**: Speaker seating arrangement performance,(L)

		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

- Localization algorithm takes about 3 sec. per participant to converge during initialization
- The speaker identification and localization based on all the modalities is fairly robust, achieving over 75% performance, about 30% better than unimodal speaker ID
- There is a significant improvement in the accuracy of localization (B) as contrasted to the performance based purely on the microphone array
- The multimodal localization accuracy is also further improved by the acoustic speaker ID modality (row C and D) (10% improvement)
- The identification of the participants' spatial arrangement(row E) is extremely accurate

Future work

- Investigate further integrated recognition technologies including face recognition, gesture recognition and head pose estimation
- Collect and share a multimodal data corpus from this testbed