Real-time Audio-Visual Tracking for Meeting Analysis

David Demirdjian Kevin Wilson Michael Siracusa Trevor Darrell
Computer Science and Artificial Intelligence Laboratory
Massachusetts Institute of Technology
Cambridge, MA 02139, USA
{demirdji,kwilson,siracusa,trevor}@csail.mit.edu

ABSTRACT

We demonstrate an audio-visual tracking system for meeting analysis. A stereo camera and a microphone array are used to track multiple people and their speech activity in real-time. Our system can estimate the location of multiple people, detect the current speaker and build a model of interaction between people in a meeting.

Categories and Subject Descriptors

I.2.10 [**Artificial Intelligence**]: Vision and Scene Understanding—*3D/stereo scene analysis*; I.4.8 [**Image Processing and Computer Vision**]: Scene Analysis—*tracking, sensor fusion*

General Terms

Algorithms

Keywords

tracking, stereo, speaker localization, microphone array

1. INTRODUCTION

Tracking people in known environments has recently become an active area of research. Robust, multi-person tracking systems are useful in a variety of domains, including meeting analysis, smart video conferencing, and human-computer interaction. Incorporating audio allows for speech activity detection, and may improve tracking. Knowledge gathered by such a system can help improve speech recognition and source separation.

2. SYSTEM OVERVIEW

Our audio-visual system consists of a stereo camera and a microphone array. The stereo camera provides 3D data to a multi-person tracker which outputs the location and height of each person in the camera's field of view. Audio signals obtained from the microphone array are used to estimate the probability that a speech event has occurred at any given position relative to the array. Our system combines this information to provide a coherent description of where people are located, and the likelihood that each person is speaking. Finally, a graph is built to model the interaction between people in a meeting. The following sections detail our system.

ICMI'04, October 13–15, 2004, State College, Pennsylvania, USA.
ACM 1-58113-954-3/04/0010.

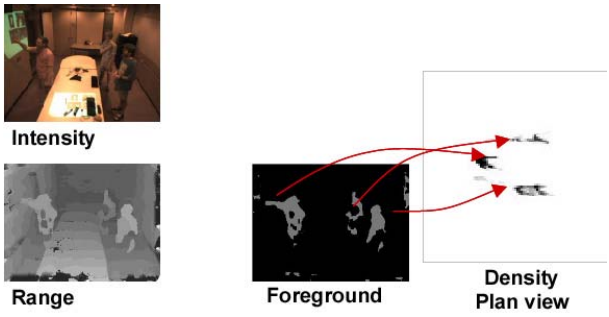

Figure 1: *(left)* **Intensity and range (disparity) image observed by the stereoscopic sensor.** *(center)* **Foreground likelihood density.** *(right)* **"Plan view" foreground likelihood density.**

3. PERSON TRACKER

We have developed a system that can perform dense, fast range-based tracking with modest computational complexity. We apply ordered disparity search techniques to prune most of the computation during foreground detection and disparity estimation. This yields a fast, illumination-insensitive 3D tracking system. Details of the system are presented in [3, 4]. Here we review the components which are relevant to the integration with our audio subsystem.

Our system relies on a stereo background model that is learned online using a technique similar to [6]. Each pixel in a new disparity image is compared to this model yielding a likelihood estimate that that pixel was a result of a foreground object. These likelihoods are projected on a "plan view" image, and a set of potential locations of foreground objects are extracted by searching for high likelihood densities. Then these potential locations and corresponding likelihoods are used in a probabilistic framework, similar to [2] to track people in real-time.

4. MICROPHONE ARRAY AND BACKGROUND NOISE MODEL

The audio system consists of N cardioid, synchronized microphones in a known configuration.

Our current approach [2] uses a learned background noise model. This model [5] encapsulates the relationship between the audio recorded at each microphone in the presence of background noise. The relationship is represented by an N x N covariance matrix for each frequency. We learn this model during a period of time when no speakers are active.

Once the background noise model is learned, a likelihood L of

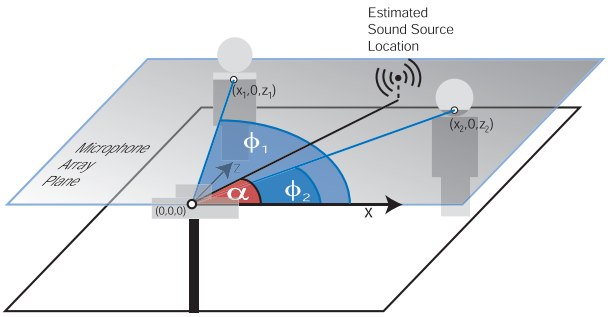

Figure 2: Models for determining audio source localization. The camera and array coordinate systems were aligned and their X,Z plane is parallel to the ground. By positioning the microphone array at approximately mouth level, we can make the simplifying assumption that speech sources emanate from within the microphone array plane.

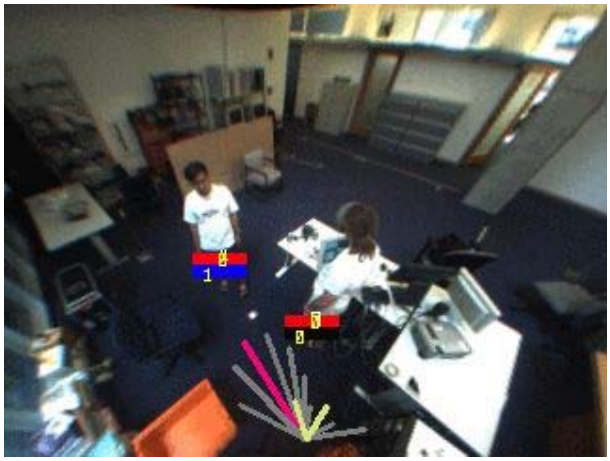

Figure 3: Extract of a video showing a 2-person meeting. Location of detected people is shown by a red square displayed at the feet of the persons. The speaker is shown with a blue square. The yellow arrows show the likelihood of audio source location for different directions.

observing some audio signal emitted by a non-background source can be derived from audio observations and the noise model (assuming an anechoic environment). In practice, we bandpass filter the array signals to emphasize the frequencies most useful to speech source localization. Although our environment is fairly reverberant, we have found that our model allows for reasonably accurate localization.

5. GROUP INTERACTION MODEL

Estimation from the person tracker and the audio subsystem are combined to estimate speaker locations. More precisely, the likelihood L is filtered temporally and spatially in the vicinity of each tracked person in order to estimate a likelihood that they are actually talking.

The speaker localization and identification is then used over time to build a model of the interaction between the tracked persons. Our approach is similar to [1] and models the group interaction using a graphical model (Markov chain) that describes the probability of a person talking after another one. Such a model is useful to show social dominance of a person or sub-group in a group, the

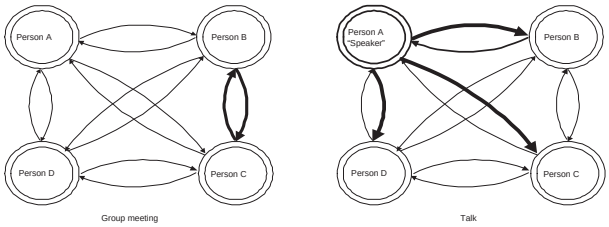

Figure 4: Example of such models in the case of "group meeting" and "talk". These models show very specific structures for "group meeting" (graph very homogeneous) and "talks" (interaction mainly between the speaker and the rest of the group, but almost no interaction between the rest of the group).

degree of interaction between pairs of people as well as analyzing the meeting type (e.g. discussion, talk, ...).

Example of such models in the case of "group meeting" and "talk" are shown Figure 4. These models show very specific structures for "group meeting" (graph very homogeneous) and "talks" (interaction mainly between the speaker and the rest of the group, but almost no interaction between the rest of the group).

Videos showing the audio-visual tracking system can be seen at:
`http://www.ai.mit.edu/~demirdji/movie/AVtracking/`

6. REFERENCES

[1] S. Basu, T. Choudhury, B. Clarkson, and A. Pentland. Towards measuring human interactions in conversational settings. In *IEEE Int'l Workshop on Cues in Communication, Kauai, Hawaii*, 2001.

[2] N. Checka, K. Wilson, M. Siracusa, and T. Darrell. Multiple person and speaker activity tracking with a particle filter. 2004.

[3] T. Darrell, D. Demirdjian, N. Checka, and P. Felzenszwalb. Plan-view trajectory estimation with dense stereo background models. In *International Conference on Computer Vision*, 2001.

[4] D. Demirdjian, K. Tollmar, K. Koile, N. Checka, and T. Darrell. Activity maps for location-aware computing. In *Proceedings of IEEE Workshop on Applications of Computer Vision (WACV'02)*, pages 70–75, 2002.

[5] D. Johnson and D. Dudgeon. *Array signal processing: Concepts and Techniques*. Prentice Hall, 1993.

[6] C. Stauffer and W. Grimson. Learning patterns of activity using real-time tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8), 2000.