MULTIPLE PERSON AND SPEAKER ACTIVITY TRACKING WITH A PARTICLE FILTER

Neal Checka *Kevin W. Wilson* *Michael R. Siracusa* *Trevor Darrell*

Massachusetts Institute of Technology
Computer Science and Artificial Intelligence Laboratory
Cambridge, MA 02139

ABSTRACT

In this paper, we present a system that combines sound and vision to track multiple people. In a cluttered or noisy scene multi-person tracking estimates have a distinctly non-Gaussian distribution. We apply a particle filter with audio and video state components, and derive observation likelihoods based on both audio and video measurements. Our state includes the number of people present, their positions, and whether each person is talking. We show experiments in an environment with sparse microphones and monocular cameras. Our results show that our system can accurately track the locations and speech activity of a varying number of people.

1. INTRODUCTION

Tracking people in known environments has recently become an active area of research. Robust, multi-person tracking systems have possible use in a wide range of applications, including smart videoconferencing systems, surveillance for security and/or site evaluation, as well as providing location and context features for human-computer interaction. Incorporating audio allows for speech activity detection, and may improve tracking, speech recognition, and source separation.

Previous approaches to tracking multiple people have mostly used only vision or only audio, which is limiting in many real-world scenarios in which both are readily available. We propose a multi-modal tracking architecture with audio and video state components and observations to track location and speech activity.

Kalman filters are commonly used to perform tracking of a single object under Gaussian uncertainty models and linear dynamics; they have been applied extensively in both the audio and video domains [1, 2]. However in a cluttered or noisy scene, Kalman filtering is inadequate because measurements will often have a non-Gaussian distribution.

Particle filtering is an approximation technique for the non-linear and non-Gaussian cases. Many researchers have applied particle filtering to vision problems. Among them, [3] track object contours in dense visual clutter, and [4] track objects by assuming a constant object color distribution over time.

Independently, particle filters have been applied to audio source localization. [5] calculated time delay estimates (TDEs) using cross-correlation and then used a likelihood model to determine the source location based on these TDEs. This method has the advantage that it can cope with spurious peaks in the cross-correlation function caused by reverberations. [6] used a beamformer-based source localization within the particle filter framework. This scheme has the advantage that it does not require intermediate calculation of time-delay estimates.

These single modality trackers have achieved some success; however, we believe that while all sensors have their strengths and weaknesses, there is no single modality for object tracking that always outperforms all others. For example, audio tracking is well-suited for speech activity detection; however, in practice the spatial resolution of audio is worse than the resolution of video. Therefore, it is desirable to integrate the information of various sensor modalities to exploit the benefits of each. [7] combines head tracking and TDOA measurements to detect people in a room and determine whether they are speaking. [8] shows that by using a particle filter, sound and vision can be fused effectively to achieve a more robust tracking of a single object than any of the modalities on their own. To track a speaker, they use a pair of omnidirectional microphones to collect TDOA measurements for their audio observation model and use head contours for their video observation model. Finally, [9] fuses video data obtained from multiple cameras and audio data obtained using microphone arrays to track a single moving object. However, the state representations adopted in these approaches do not explicitly support hypotheses containing a different number of objects.

Extending particle filters to track a varying number of objects presents additional challenges. [10] is an implementation of a particle filter in which a multi-blob likelihood function assigns comparable likelihoods to hypotheses containing different numbers of objects. [11] proposed a hierarchical sampling method where one level tracks the motions of individual objects while the other level handles object addition and deletion.

In this paper, we present a particle filter based multi-person tracker with audio and video state components, including position, height, and whether each user is speaking. Our observations come from two monocular cameras and a microphone array. We introduce the use of a spatiospectral covariance audio model in the particle filtering framework. Our system's strength is that it simultaneously tracks the location and speaking state of a varying number of people.

2. AN AUDIO-VISUAL STATE-SPACE MODEL

The multiple person tracking problem can be formulated in a state-space estimation framework by associating the locations of all possible configurations of people at time t with an unobserved state vector X_t:

$$X_t = (n_t, \chi_t^1, \ldots, \chi_t^n) \tag{1}$$

where n_t is the number of people, and $\chi_t^i = [x, y, h, s]$ describes an object in the configuration. We track a person's $[x, y]$ floor position and height parameter h. The boolean variable s denotes audio

activity. In our environment, we make the simplifying assumption that localized audio activity implies speech activity.

We use a zeroth order motion model with Gaussian-distributed random excitation forces applied to x, y, and h. We jointly consider the speech activity variable, s, for each person, χ_t^i, as a single bit of the overall "speech activity state" of a particle, X_t. We update this "speech activity state" according to a transition matrix that we defined. This allows us to model the dependence between individuals' speech activity variables as they engage in conversation. This provides a richer model of conversational dynamics than updating each speaker independently.

We apply a prediction model similar to [10] which states that at each time step each object will remain in the scene with probability υ_{remain} and new objects will enter the scene with probability υ_{add}.

3. OBSERVATION MODEL FOR VISION

For a given configuration, the likelihood function, $p(Z_v|X_t)$, measures how well the hypothesized state supports the image data. We consider our video observation Z_v as a collection of i independent visual cues Z_{v_i} where the overall likelihood of our video observation is:

$$p(Z_v|X_t) = \prod_{i=1}^{\#\text{ of cues}} p(Z_{v_i}|X_t) \tag{2}$$

Each visual cue's likelihood $p(Z_{v_i}|X_t)$ is defined as

$$p(Z_{v_i}|X_t) = p(Z_{v_i}^b|X_t) \cdot p(Z_{v_i}^f|X_t) \tag{3}$$

where $p(Z_{v_i}^b|X_t)$ and $p(Z_{v_i}^f|X_t)$ are the likelihoods of observing the background and foreground respectively. We assume that, conditioned on the state, independent processes govern the foreground and background in a visual cue Z_{v_i}. The background and foreground likelihoods are:

$$p(Z_{v_i}^b|X_t) = f_b(\overline{M}_X, I) \tag{4}$$

$$p(Z_{v_i}^f|X_t) = f_f(M_X, I) \tag{5}$$

where I is an image and M_X is a mask we create for each multi-object configuration described by our state X. The mask M_X represents a set of image locations (u, v) which encodes the projection of each person onto each image plane. Based on training data, we selected reasonable functional forms for Equations 4 and 5. The parameters for these functions were found using maximum likelihood estimates based on training data.

Our observations consist of two cues, one based on a background model, Z_{v_1}, and one based on adjacent frame differences Z_{v_2}. A person is modelled as a vertical cylinder in the world coordinate frame. The shape of the cylinder is specified by a fixed radius r and a variable height h as specified in the state. These cylinders are projected onto our image planes using the intrinsic and extrinsic parameters of our calibrated cameras.

We use a background model where we assume pixel-wise and color-channel independence. Each background pixel is modelled with a Gaussian and each foreground pixel is described by a uniform distribution. In a training phase, we compute pixel-wise mean and variance images on a sequence of a few hundred empty background images.

We employ a contrast-normalized difference image I_d for Z_{v_2}. In our difference image, we expect that most of the scene is static so most of the pixels will be zero. Areas of motion will generate large differences between consecutive frames. Therefore, we assume that these foreground and background difference measurements come from exponential distributions with different scale parameters.

4. OBSERVATION MODEL FOR SOUND

The sound measurement system consists of N omnidirectional microphones that are synchronized in time. These microphones form a steerable array that can be used to localize sound sources in the room. Our audio observations, Z_a, consist of an N_f point short-time Fourier transforms (STFTs) of each microphone's signal. We represent this as N_f complex vectors $\mathbf{w}(k)$, each with dimension N. We assume that these STFT coefficient vectors are jointly Gaussian and are independent of the coefficients in all other frequency bins. We evaluate the likelihood of the coefficients according to a hypothesized N by N spatiospectral covariance matrix, $\mathbf{R_h}(\mathbf{k})$, where k represents frequency bin number. Each particle has its own $\mathbf{R_h}(\mathbf{k})$, with likelihood:

$$p(Z_a|X_t) = \prod_{k=1}^{N_f} \mathcal{N}(\mathbf{w}(k); \mathbf{0}, \mathbf{R_h}(\mathbf{k})) \tag{6}$$

The hypothesized covariance matrix consists of three additive components:

$$\mathbf{R_h}(\mathbf{k}) = \mathbf{R_b}(\mathbf{k}) + \sum_{i=1}^{n} \mathbf{R_{s_i}}(\mathbf{k}) + \lambda \mathbf{I} \tag{7}$$

The first component, $\mathbf{R_b}(\mathbf{k})$, models the background noise in the environment. We assume a constant level of background noise and estimate $\mathbf{R_b}(\mathbf{k})$ from a period of time when no speakers are active; however, this does not preclude the inclusion of a time-varying background noise model. The second component is the sum of one outer product, $\mathbf{R_{s_i}}(\mathbf{k})$ for each active speaker:

$$\mathbf{R_{s_i}}(\mathbf{k}) = s_i \mathbf{v_i}(\mathbf{k}) \mathbf{v_i}(\mathbf{k})^\top \tag{8}$$

This models each speaker as a point source emitting from its hypothesized location in an anechoic environment. In Equation 8, v_i represents the propagation vector from that location and s_i is the speech activity bit.

We have found that hypothesizing additional speakers often incorrectly increases the likelihood of an observation according to the first two terms. This is due to the fact that our anechoic propagation model is only approximate. By increasing the overall variance, the second term may increase the likelihood of an observation produced by a signal from an unrelated direction. This can be seen as decreasing the specificity of our model. To balance this effect we incorporate a third term, $\lambda\mathbf{I}$, where $\mathbf{I}$ is the $N \times N$ identity matrix with:

$$\lambda = \mu_0 - \mu_1 m \tag{9}$$

where m is the number of active speakers and μ_0 and μ_1 are empirically determined scalar constants. By adding more energy as

the number of active speakers decreases, this term partially counteracts the additional variance introduced by the second term. μ_1 was chosen to add as much energy to $\mathbf{R_h}(\mathbf{k})$ as would a typical speaker, and μ_0 was chosen to keep $\mathbf{R_h}(\mathbf{k})$ positive definite in typical scenarios with a small number of active speakers. We have found this third term to substantially improve the performance of our system.

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 and fits well into the particle filtering framework.

5. COMBINED AUDIO AND VIDEO PROBABILITY

We use Equations 2 and 6 to determine the weight w of each particle in our particle filter. In addition to the above likelihood functions, we employ a prior on the number of people in the scene to penalize unnecessarily complicated explanations of the observation. We use a geometric distribution on the number of people. This term is used when calculating the particle weights in the update stage of the filter.

The combined probability $p(Z|X_t)$ for both audio and video data is obtained by multiplying the corresponding likelihoods from the audio and video source.

Our particle filter provides a probability distribution on the state of the world. To decide on a single consistent explanation of the scene, we estimate the number of people $\hat{n}$ and their most likely state $\hat{X}_t$ using the following algorithm.

1. Calculate the marginal distribution of the number of people and use it to find a maximum-likelihood estimate for the number of people:

$$\hat{n} = \arg\max_{i \in S} \sum_{n^j = i} w^j \tag{10}$$

   where $S = \{1, \ldots, N_{max}\}$

2. Estimate the tracked position as the weighted sum of the person locations:

$$\hat{X}_t = \frac{\sum_{j \in \hat{n} = n^j} w_t^j X_t^j}{\sum_{j \in \hat{n} = n^j} w_t^j} \tag{11}$$

6. EXPERIMENTS

Our tracking system consists of two widely spaced (approximately 5.2m) monocular cameras and a microphone array. The monocular cameras provide us with 320x240 images at 20 frames per second. The microphone array consists of four linear sub-arrays, each consisting of four omnidirectional microphones. Each microphone provides a separate 16kHz audio stream. The test environment is a 6m by 6m open room with stationary background noise. A schematic of our test environment is shown in Figure 1. Using an audio-visual calibration target, we calibrate the audio subsystem to determine the locations of the sources and the microphones. We then calibrate the video subsystem using these source locations. For these experiments, data was processed offline on synchronized audio and video feeds. Each of our experimental sequences involved two or three people moving around our test

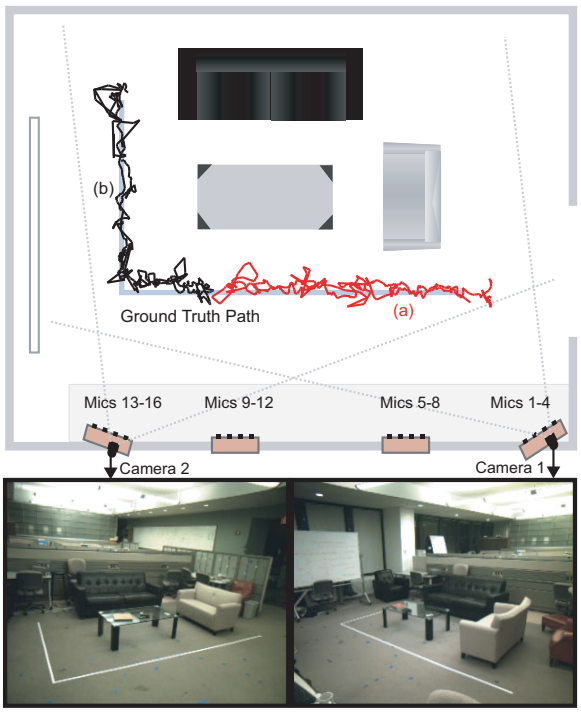


Fig. 1. Test environment consisting of two monocular cameras and four linear microphone sub-arrays. (a) Trajectory of person 1. (b) Trajectory of person 2.

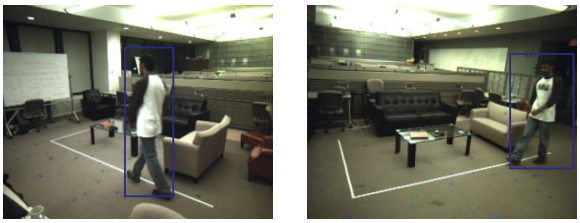
	Person 1	Person 2	Neither
Person 1	90%	3%	7%
Person 2	6%	84%	10%
Neither	30%	5%	65%

Table 1. Speaker confusion matrix (% Accuracy) . The row is our system's hypothesis, and the column is the ground truth.

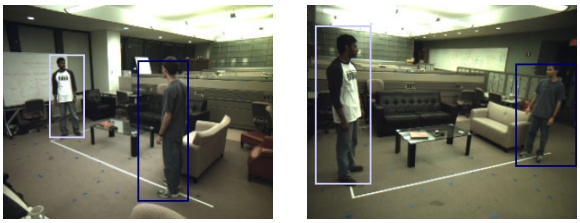
environment and conversing with each other. In the first sequence, two individuals follow a known path in the room and their conversation was restricted to turn-taking dialog with no simultaneous speech. In this experiment, the particle filter was run with 100 manually initialized particles. The speech activity output of the system was compared to hand-labelled ground truth and the results are shown in Table 1. Overall, our system was approximately 80% accurate when determining who is speaking. The signal-to-noise ratio ranged from -4dB to 10dB as people moved about the room. As shown in Figure 1, one person started at each end of an "L"-shaped path (2m x 4m) and walked along the path until they reached the center. The RMS position error for this sequence was 9.2cm.

The second sequence consists of 1836 frames and shows one to three people walking around the room and conversing. In this experiment, the particle filter was run with 200 particles. Figure 2 shows some key frames of this sequence. The rectangles represent the state of the world as determined by our estimation algorithm. The darker rectangles denote active speakers, while the lighter rectangles represent non-speakers. In frame 17, one person enters the room as shown. By frame 211, a second person has entered and has been added to the state of the system. By frame 539,

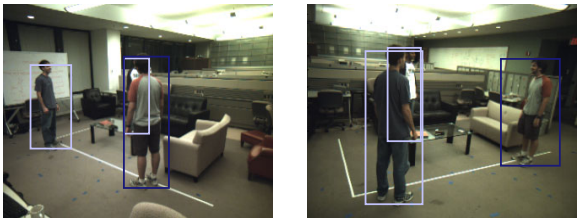
there are now three people in the room, and one person is partially occluded in both views. In spite of the occlusion, the system maintains track of the occluded individual. By frame 1367, one of the three people has left and they have been removed from the system state. [1]



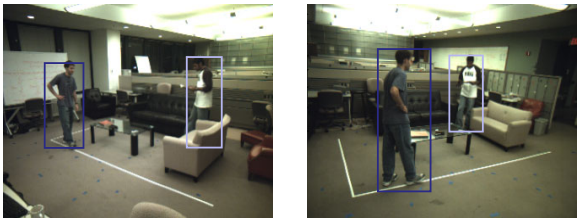
(a) Frame 18



(b) Frame 211



(c) Frame 539



(d) Frame 1367

Fig. 2. Sequence of multiple people being tracked. The squares represent the state of the world as determined by our estimation algorithm. Dark blue represents an active speaker and light blue represents a non-speaker.

7. CONCLUSION AND FUTURE WORK

This paper describes a multi-modal tracking architecture that uses both audio and video observations. The contribution of our work is the development of a multiple person tracking framework with integrated audio and visual state and observation likelihood components. We apply a particle filter to track multiple people using a foreground detection, image-differencing, and spatiospectral covariance matrices. Also, our model accurately reflects the number of people present and their speech activity. This type of system provides information that may be useful for source separation and speech enhancement.

There are many interesting extensions to the work presented in this paper. For example, the incorporation of additional modalities such as color distributions might lead to increased system robustness. Color distributions, as well as audio features like pitch, might be helpful in maintaining the identity of a tracked object. In the audio domain, we intend to explore more sophisticated acoustic models. We also intend to relax the independence assumption between frequency bins since speech is not stationary.

These extensions should lead to more robust a system that can better determine the location and activity of its inhabitants in a pervasive computing environment.

[1] Complete movie sequence at http://www.ai.mit.edu/projects/vip/av/

8. REFERENCES

[1] D.E. Sturim, M.S. Brandstein, and H.F. Silverman, "Tracking multiple talkers using microphone-array measurements," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1997.

[2] A. Mittal and L.S. Davis, "M2tracker: A multi-view approach to segmenting and tracking people in a cluttered scene using region-based stereo," in *European Conference on Computer Vision*, 2002.

[3] M. Isard and A. Blake, "Condensation - conditional density propagation for visual tracking," *International Journal on Computer Vision*, , no. 28(1), pp. 5–28, 1998.

[4] P. Perez, C. Hue, J. Vermaak, and M. Gangnet, "Color-based probabilistic tracking," in *ECCV02*, 2002.

[5] J. Vermaak and A. Blake, "Nonlinear filtering for speaker tracking in noisy and reverberant environments," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2001.

[6] D.B. Ward and R.C. Williamson, "Particle filter beamforming for acoustic source localization in a reverberant environment," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2002.

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

[8] 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.

[9] 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 Event in Video*, 2001.

[10] M. Isard and J.MacCormick, "Bramble: A bayesian multiple-blob tracker," in *International Conference on Computer Vision*, 2001.

[11] H. Tao, H.S. Sawhney, and R. Kumar, "A sampling algorithm for tracking multiple objects," in *Workshop on Vision Algorithms*, 1999, pp. 53–68.