A Probabilistic Framework for Multi-modal Multi-Person Tracking

Neal Checka, Kevin Wilson, Vibhav Rangarajan, and Trevor Darrell
Artificial Intelligence Laboratory
Massachusetts Institute of Technology
Cambridge, MA 02139

Abstract

In this paper, we present a probabilistic tracking framework that combines sound and vision to achieve more robust and accurate tracking of multiple objects. In a cluttered or noisy scene, our measurements have a non-Gaussian, multimodal distribution. We apply a particle filter to track multiple people using combined audio and video observations. We have applied our algorithm to the domain of tracking people with a stereo-based visual foreground detection algorithm and audio localization using a beamforming technique. Our model also accurately reflects the number of people present. We test the efficacy of our system on a sequence of multiple people moving and speaking in an indoor environment.

1. Introduction

As the trend to expand computing away from the desktop continues, new research challenges arise. One goal is to make indoor spaces more intelligent, more natural and easier to use by allowing them to use the same visual and audio interface modalities that humans take for granted. Simple things, like presence, posture, gaze, and sounds are extremely important cues in communication between people, and they should be with computers as well. Such pervasive environments would need to determine the location, activity, and identity of its inhabitants. As a result, tracking people in known environments has recently become an active area of research in computer vision. Previous approaches to tracking multiple people have mostly been limited to using solely vision or audio. In this paper, we propose a multimodal tracking architecture to track using audio and video observations.

To achieve optimal performance with multiple sensing modalities, a tracking system must exploit not just the statistics of each modality alone but also relationships between the two. Consider a system that tracks moving objects. Such a system may use video data to track the spatial location of an object. If an object emits sound, the system may use audio data captured by a microphone array to track its location using the time delay of arrival (TDOA) of the audio signals detected at different microphones. A tracker that exploits both these modalities may be more robust and achieve better performance than one that uses either one alone. For example, a tracker using only video data may mistake the background for the object or lose track of the object due to occlusion; whereas, a tracker that also uses audio data could continue tracking the object by following its sound pattern. Conversely, video data could help in situations where an audio tracker alone may fail, such as when the tracked object stops emitting sound or is masked by some background noise.

In a cluttered or noisy scene, measurements will often have a non-Gaussian, multi-modal distribution. Particle filtering is an approximation technique for tracking non-linear and non-Gaussian distributions. Particle filters are sequential Monte Carlo methods based upon point mass representations of probability densities that can be applied to any state space model.

In this paper, we incorporate both audio and video observations within a particle filtering framework for tracking multiple people in an indoor environment. In Section 2, we review related work on different approaches to tracking people. Following that, we introduce our multi-modal dynamic model in Section 3. In Sections 4 and 5, we describe our probabilistic observation models for video and audio, and in Section 6, we show how to combine these modalities in order to track multiple people. In Section 7 we evaluate the performance of our multi-modal tracker. Finally in Section 8, we summarize our contributions and discuss extensions of this work.

2. Previous Work

Much success has been achieved in tracking single objects in a scene using vision alone. One method is to model the foreground object as an ellipse in the image plane. The object is matched from frame to frame either by correlation [16] or by using some statistical properties of the object such as color histograms [1]. To track accurately over time, many systems usually include a model of dynamics. The tracking systems described in [8, 3] use a constant velocity predictor followed by a search for the closest match of a foreground area in the neighborhood of the prediction. The Kidsroom [4] tracker uses color, position, velocity, and size of the blobs to compute distance measures at each frame. The measures are then combined into a match score matrix, which is used to determine object-to-blob correspondence. These tracking methods can be very effective in simple scenes where only one object is present. Some tracking systems build spatio-temporal representations of moving regions in a scene. Kornprobst and Medioni developed an approach that uses a tensor voting methodology to enforce smoothness in space and time of the trajectories of tracked objects [7]. Darrell et al. use an approach where trajectory estimation is implemented by running a connected-components analysis in a spatio-temporal plan-view volume [2].

Kalman filters are commonly used to perform tracking of a single object using a Gaussian uncertainty model and linear dynamics. This approach has been applied in both the audio and video domains. For example, Sturim et al. apply a Kalman filter to spatially smooth raw time delay estimates in the acoustic localization domain [11]. The M2Tracker, a stereo-vision-based vision person tracking system, uses Kalman filters to track object blob centroids over time [9]. However in a cluttered or noisy scene, Kalman filtering is inadequate because measurements will often have a non-Gaussian, multi-modal distribution.

Particle filtering is an approximation technique for the non-linear and non-Gaussian cases. In vision, Isard and Blake [5] developed and applied the CONDENSATION algorithm, a type of particle filter, to track curves in dense visual clutter. In the audio domain, Vermaak and Blake [13] proposed a new framework for TDOA source localization based on a particle filtering approach. In their approach, time delay estimates (TDEs) are calculated using cross-correlation and then a likelihood model is used to determine the source location based on the obtained TDEs. By using this multi-hypothesis approach, this method has the advantage that it can cope with spurious peaks in the cross-correlation function caused by reverberations. Ward and Williamson used beamformer-based source localization within the particle filter framework [15]. This scheme has the advantage that it does not require intermediate calculation of time-delay estimates. However, the specific state representation adopted in these approaches does not explicitly support hypotheses containing a different number of objects.

Extending the particle filters to track a varying number of objects presents additional challenges. BraMBLE, a multiple blob tracker, is an implementation of a particle filter in which the number of objects being tracked may vary [6]. The key innovation of this system is a multi-blob likelihood function that assigns comparable likelihoods to hypotheses containing different numbers of objects. Tao et al. [12] proposed a hierarchical sampling method for multiple-object particle filtering. One level tracks the motions of individual objects while the other level handles object addition and deletion.

These unimodal probabilistic trackers have achieved some success; however, we believe that while all sensors have their specific strengths and weaknesses, there is no single modality for object tracking that always outperforms all others. Therefore, it is desirable to integrate the information of various sensor modalities to exploit the benefits of each. Pingali et al. combine head tracking and TDOA measurements in a non-probabilistic framework to detect people in a room and determine whether they are speaking [10]. Vermaak and Blake [14] show 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. Zotkin et al. [17] fuse video data obtained from multiple cameras and audio data obtained using microphone arrays to track a single moving object.

To our knowledge, the system presented in this paper is the first to combine audio-visual person tracking with a multi-object particle filtering framework.

3 The Model

3.1 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. We can denote a state of the world as

$$X_t = (n_t, \chi_t^1, \ldots, \chi_t^n) \quad (1)$$

where n_t is the number of people, and $\chi_t^i = [x, y, d, s]$ describes an object in the configuration. We track a person's $[x, y]$ floor position and size d. The boolean variable s denotes whether or not the person is speaking.

3.2 Dynamic Model

We use a zeroth order motion model with a random excitation force applied to the particles. If $\alpha(t) = (x, y)$ is a 2D vector of the target coordinates at time t then the dynamics can be written as:

$$\alpha(t + \delta t) = \alpha(t) + F\delta t \quad (2)$$

$$d(t + \delta t) = d(t) + G\delta t \quad (3)$$

```
1. For each object $\chi^i_{t-1}$ in a particle $X_{t-1}$
    (a) Predict $\chi^i_t$ by propagating $\chi^i_{t-1}$ according to dynamic model.
    (b) With probability $(1 - \upsilon_{remain})$, delete $\chi^i_t$ and update the person count $n_t$
2. With probability $\upsilon_{add}$ and $n_t < N_{max}$, create new $\chi^{n_t}_t$ according to the new object location prior and update person count $n_t$
```

Figure 1: Algorithm for predicting multiple objects

where F and G are independent random excitation forces that are distributed as a Gaussian with zero mean and variances σ^2_α and σ^2_d respectively. We treat s as a two state Markov chain. Initially, the particles are uniformly distributed.

We apply a prediction model similar to [6] 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}. New object locations are chosen according to a new object location prior distribution. Currently our prior distribution is uniform within a creation zone near the entrances. We only generate new objects at time t if $n_{t-1} < N_{max}$, where N_{max} is the maximum number of tracked objects. The algorithm for generating a new particle hypotheses $\{X^i_t\}$ from the previous set of particles $\{\tilde{X}^i_{t-1}\}$ is shown in Figure 1.

4 Observation Model for Vision

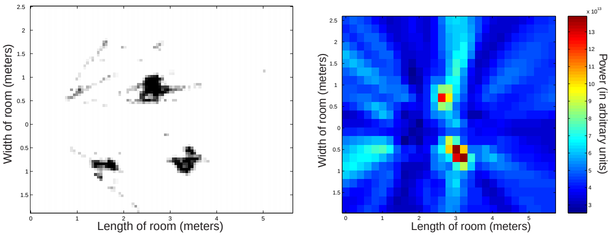
Our video observations z_v consist of plan-view images of detected foreground points as shown in Figure 2(a). These foreground points are detected using a range-based background model as described in [2].

4.1 Likelihood Model

At each time step t, we observe a binary plan-view image of the foreground points I. For each multi-object configuration X, we create a mask M_X which represents a set of image locations (x, y) where the hypothesized objects should generate foreground points.

$$M_X = \bigcup_{i=1}^{n} M_{\chi^i} \tag{4}$$

where M_{χ^i} is a square mask of size d centered on the hypothesized object location. For a given configuration, the likelihood function, $p_v(z_v|X_t)$, measures how well the hypothesized state supports the image data. From our observations, we calculate the following measurements:



(a) (b)

Figure 2: (a) Video observations consist of binary plan-view images of foreground points. (b) 2D map of audio power in the microphone array coordinate system. Brighter colors correspond to locations of high power.

$$z_f = \frac{|I \cap M_X|}{|M_X|} \tag{5}$$

$$z_b = \frac{|I \cap \overline{M}_X|}{|\overline{M}_X|} \tag{6}$$

where z_f is the fraction of the mask area covered by foreground points and z_b is the fraction of the non-mask area covered by foreground points. In a training phase, we empirically determined the distributions of these measurements. We model the distribution of z_f with a Gaussian whose parameters, σ_f and μ_f, were determined from the training data. This yields the likelihood function:

$$p_f(z_f|X_t) = \frac{1}{\sqrt{2\pi}\sigma_f} e^{\frac{-(z_f - \mu_f)^2}{2\sigma_f^2}} \tag{7}$$

We model the distribution of z_b as an exponential distribution:

$$p_b(z_b|X_t) = \lambda_v e^{-\lambda_v z_b} \tag{8}$$

The above equations implicitly depend on X_t since it was used to calculate z_f and z_b. Finally, the overall likelihood of a configuration is given by,

$$p_v(z_v|X_t) = p_f(z_f|X_t) \cdot p_b(z_b|X_t) \tag{9}$$

5 Observation Model for Sound

The sound measurement system consists of M omnidirectional microphones that are synchronized in time. These microphones form a steerable array that can be used to localize sound sources in the room. To use the array for source localization, we steer it to a fixed set of locations in the room and calculate the response power at each location. The response power is calculated for the delay-and-sum beamformer described by:

$$Y(t) = \sum_{m=1}^{M} \alpha_m X_m(t - \Delta_m) \tag{10}$$

where X_m is the signal for the mth microphone, α_m is a scaling factor, and Δ_m is the delay between the source signal and microphone m. In practice, the array signals are bandpass filtered before beamforming to ignore frequencies that contain more noise than speech energy. The beam power is then calculated as the integral over a time period of the square of the beamformed signal:

$$\bar{Y}(\tau) = \int_{\tau}^{\tau+T} |Y(t)|^2 dt \tag{11}$$

where T is a fixed-length time window. If the beamformer were steered toward the true source location, then one would expect the beamformer output power to be large. Figure 2(b) is a 2D visualization of an audio observation.

5.1 Likelihood Model

The likelihood function $p_a(z_a|X_t)$ measures how well the state X_t supports the audio data at time t. In our system, z_a is a 2-dimensional response map denoted by $B_{obs}(x, y)$. We compare z_a to a response pattern, $B_X(x, y)$, that we synthesize based on a simple anechoic sound propagation model. We synthesize this response pattern by generating synthetic signals with a spectrum similar to that of speech and delaying these signals appropriately for each hypothesized speaker location. These signals are then used to calculate array response powers as described above. Our response map representation has the advantage of making explicit the spatial distribution of audio power while taking into account sidelobe patterns. These features would be difficult to incorporate into a TDE formulation of the problem.

Our simple propagation model yields plausible response patterns; however, the peaks in the synthesized response patterns tend to be narrower than peaks in actual observations due to unmodelled noise and reverberation in the environment. To compensate for these narrow peaks, we blur each synthesized response pattern with a Gaussian kernel of standard deviation 28 centimeters.

In practice, we synthesize these response patterns by precomputing a separate response pattern, $B_{x',y'}(x, y)$, for each possible speaker location (x', y'). We generate multi-speaker response patterns by taking linear combinations of our precomputed patterns:

$$B_X(x, y) = \sum_{i \in V} B_{x_{\chi^i}, y_{\chi^i}}(x, y) \tag{12}$$

where V is the set of people in X who are speaking ($s_{\chi^i} = 1$).

Form an initial set of particles $\{X_0^{(i)}, i = 1, \ldots, N_s\}$ and give them uniform weights $\{w_0^{(i)} = 1/N_s\}, i = 1, \ldots, N_s$. As each new frame of data is received:

1. Resample the particles from the previous frame $\{X_{t-1}^{(i)}\}$ according to their weights $\{w_{t-1}^{(i)}\}$ to form a resampled set of particles $\{\tilde{X}_0^{(i)}, i = 1, \ldots, N_s\}$
2. Predict a new set of particles $\{X_t^{(i)}\}$ by propagating $\{\tilde{X}_{t-1}^{(i)}\}$ according to dynamic model described in Figure 1.
3. For each particle X_t, calculate the video likelihood according to the function
   $$p_v(z_v|X_t) = p_f(z_f|X_t) \cdot p_b(z_b|X_t)$$
   and each audio likelihood according to the function
   $$p_a(z_a|\chi_t) = \lambda_a e^{-\lambda_a C}$$
   The combined audio and video likelihood is computed by
   $$p_{total}(Z|X_t) = p_v(z_v|X_t) \cdot p_a(z_a|X_t)$$
4. Weight the new particles according to the total likelihood function
   $$w_t^{(i)} = p_{total}(Z|X_t^{(i)}) \cdot p_n(n)$$
   and normalize so that $\sum_i w_t^{(i)} = 1$
5. Store the particles and their weights $\{X_t^{(i)}, w_t^{(i)}, i = 1, \ldots, N_s\}$

Figure 3: Algorithm for multi-modal multiple person particle filter

To compare our synthesized response pattern to the observed pattern, we first normalize each pattern to lie in the interval $[0, 1]$. We then compute the L2 norm of the difference between the two patterns and model the audio likelihood as an exponential function of this norm:

$$C = \|\tilde{B}_X(x, y) - \tilde{B}_{obs}(x, y)\|^2 \tag{13}$$
$$p_a(z_a|X_t) = \lambda_a e^{-\lambda_a C} \tag{14}$$

where $\tilde{B}$ are normalized response patterns.

6 Combined Audio and Video Probability

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. In particular, configurations with more objects tend to have more foreground pixels. We model this by:

$$p_n(n) = e^{\frac{-(n-|I|/\kappa)^2}{2\sigma_n^2}} \tag{15}$$

where κ and σ_n^2 are the expected number of pixels per person and the variance on number of people respectively. This term is used when calculating the particle weights in the update stage of the filter. Without this term, a hypothesis that includes an additional person hiding behind the true location of an occupant would be just as likely as the hypothesis without the hidden person.

The combined probability $p_{total}(Z|X_t)$ for both audio and video data is obtained by multiplying the corresponding likelihoods from the audio and video models.

$$p_{total}(Z|X_t) = p_v(z_v|X_t) \cdot p_a(z_a|X_t) \tag{16}$$

We use this likelihood function and prior in a particle filtering framework to track multiple objects. Figure 3 is an overview of our algorithm.

Our particle filter provides a probabilistic model of the state of the world. To decide on a single consistent explanation of the scene, we use 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:

   $$\hat{m} = \arg\max_{i \in S} \sum_{m^j = i} w^j \tag{17}$$

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

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

   $$E\{X_t\} = \sum_{i \in \hat{m}} w_t^{(i)} \chi_t^{(i)} \tag{18}$$

7 Experiments

Our test environment, depicted in Figure 4, is a conference room equipped with 15 omnidirectional microphones spread across the ceiling and one stereo camera on the wall. Our system can easily be extended to incorporate video data from multiple stereo cameras as in [2]. The audio and video subsystems were calibrated independently, and for our experiments, we performed a joint calibration by finding the least-squares best-fit alignment between the two coordinate systems.

The multi-modal multiple person tracker is a prototype system that was developed in Matlab. Experiments were conducted offline on synchronized audio and video feeds. We expect a real-time system to be possible with an optimized implementation.

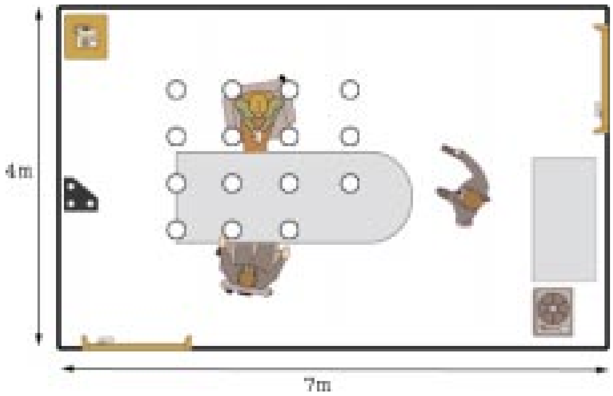


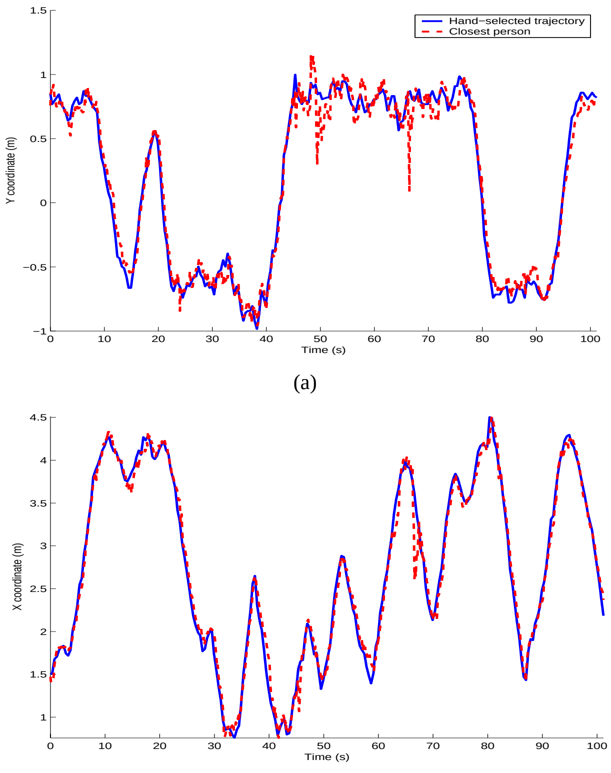
Figure 4: The test environment. A schematic view of the environment with stereo cameras represented by black triangles and microphones represented by empty circles.

7.1 Overview

We illustrate our system on a test sequence of three people interacting in our test environment. This sequence tests the performance of our tracker in tracking multiple people. It consists of 700 frames, captured at 12 frames per second, of one to three people walking around the room. In this sequence, the subjects were carrying on a conversation in which typically only one person was speaking. In this experiment, the particle filter was run with 750 particles and the parameter values listed in Table 1.

7.2 Analysis

Figure 6 presents key frames of the sequence being tracked using audio and video. The tracking result is visualized as foreground points superimposed on the audio response pattern observed by the microphone array. The squares represent the state of the world as estimated by the algorithm described in Section 6. The colors of the squares range from white to orange, where the amount of orange represents our estimate of whether the person is talking. In the audio response map "hotter" colors denote higher audio power. After the first few frames, the particles are able to lock onto the initial occupant of the room and track his location as shown in Figure 6(a). At frame 50, our estimation algorithm chooses an incorrect configuration in which it describes the scene with a large silent person and a small talking person. However, by frame 100 the tracker recovers to the correct configuration as shown in Figure 6(b). By frame 150, another person has entered and the system responds appropriately. In frame 200, the audio response map indicates that a third person has entered the room. Although the third person is not yet in the field of view of the camera, the tracker has used audio to detect him. In frames 250 to 500, as the



(a)

(b)

Figure 5: These plots show one person's trajectory in the x and y directions as a function of time. The solid blue line is the hand labeled trajectory ("ground truth"). The dotted green line is the trajectory of χ_1, the first person in the state. Since we are not tracking identity, the tracker cannot guarantee that χ_1 is always the same person, but the tracker dynamics usually preserve continuity in the scene. The dashed red line is the trajectory of the closest χ_i to the ground truth reported by the tracker.

moving occupants take turns talking, the tracker fuses audio and video to track them. Shortly before frame 550, one of the non-speaking subjects has moved out of the field of view of the camera. As a result, our estimation algorithm reports only two people. By frame 600, the tracker has recovered.

Figure 5 shows one person's trajectory in the x and y directions as a function of time. The solid blue line is the hand labeled trajectory ("ground truth") which was obtained by hand segmenting the location of one of the occupants in every fifth frame. The dashed red line is the trajectory of the closest χ_i to the ground truth reported by the tracker. On this sequence, the subject moved about in our 4m by 7m room, the RMS error was 16cm.

Symbol	**Meaning**	**Value**
σ_α	Speed std dev	2.8m/s
σ_d	Person size std dev	4.3cm
v_{remain}	Object survival probability	0.99
v_{add}	New object arrival probability	0.02
N_{max}	Maximum number of people	4
σ_f^2	Foreground likelihood variance	0.79
μ_f	Foreground likelihood mean	0.3304
λ_v	Scale parameter for background	357
T	Time window for beam power	1s
λ_a	Scale parameter for audio	0.35
κ	Expected number of pixels per person	330
σ_n^2	Variance on the number of people	2

Table 1: Parameter values used for experiments

8 Conclusion and Future Work

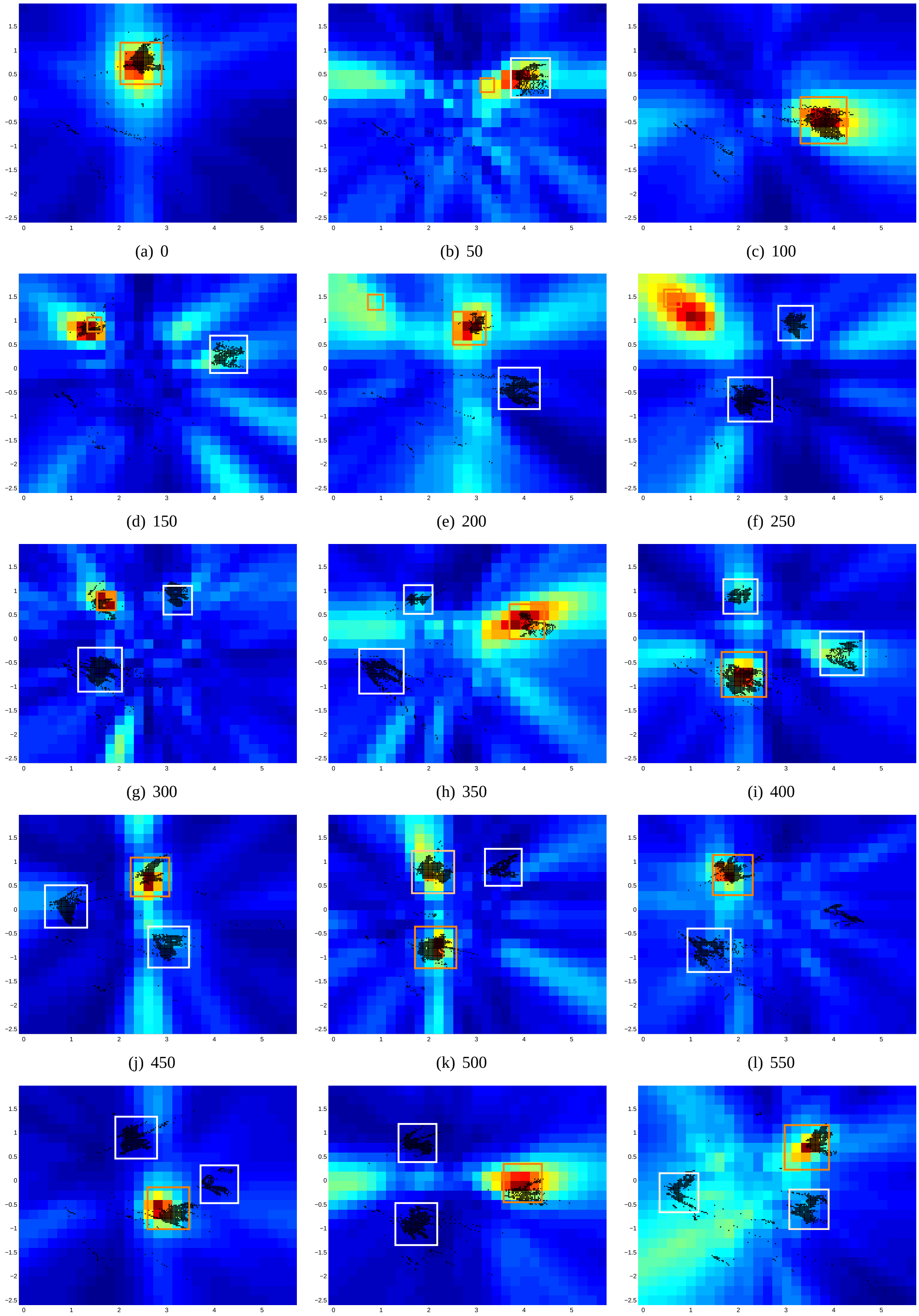
This paper describes a multi-modal tracking architecture to track using audio and video observations. We apply a particle filter to track multiple people using a stereo-based visual foreground detection algorithm and spatial audio responses based on delay-and-sum beamforming. Also, our model accurately reflects the number of people present.

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.

Currently, our system tracks objects on a ground plane. We would like to explore different observational models. For example, in vision, rather than track foreground densities on a ground plane, foreground disparities could be tracked. This might result in better occlusion reasoning since the full 3D information about objects is available. By tracking in 3D, the sampling space is larger so more complicated sampling techniques are required.

Also, different audio source localization techniques should be explored to improve the audio subsystem. More complicated beamforming techniques such as minimum variance distortionless response beamforming could be incorporated.

These extensions and a real-time implementation will hopefully lead to a system that can determine the location and activity of its inhabitants in a pervasive computing environment.



(a) 0

(b) 50

(c) 100

(d) 150

(e) 200

(f) 250

(g) 300

(h) 350

(i) 400

(j) 450

(k) 500

(l) 550

(m) 600

(n) 650

(o) 700

Figure 6: Sequence of multiple people. Foreground points are superimposed on the audio response pattern observed by the microphone array. In the audio response map “hotter” colors denote higher audio power. The squares represent the state of the world as determined by our estimation algorithm. The colors of the squares range from white to orange, where the amount of orange represents our estimate of whether the person is talking.

References

[1] D. Comaniciu, V. Ramesh, and P. Meer. Real-time tracking of non-rigid objects using mean shift. In *Conference on Computer Vision and Pattern Recognition*, 2000.

[2] 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.

[3] I. Haritaoglu, D. D. Harwood, and L.S. Davis. W4: Real-time surveillance of people and their activities. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):809–830, August 2000.

[4] S.S. Intille, J.W. Davis, and A.F. Bobick. Real-time closed-world tracking. In *Conference on Computer Vision and Pattern Recognition*, 1996.

[5] M. Isard and A. Blake. Condensation - conditional density propagation for visual tracking. *International Journal on Computer Vision*, (28(1)):5–28, 1998.

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

[7] P. Kornprobst and G. Medioni. Tracking segmented objects using tensor voting. In *Conference on Computer Vision and Pattern Recognition*, 2000.

[8] J. Krumm, S. Harris, B. Meyers, B. Brummit, M. Hale, and S. Shafer. Multi-camera multi-person tracking for easyliving. In *3rd IEEE Workshop on Visual Surveillance*, 2000.

[9] 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.

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

[11] 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.

[12] H. Tao, H.S. Sawhney, and R. Kumar. A sampling algorithm for tracking multiple objects. In *Workshop on Vision Algorithms*, pages 53–68, 1999.

[13] 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.

[14] 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.

[15] 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.

[16] C.R. Wren, A. Azarbayejani, T.J. Darrell, and A.P. Pentland. Pfinder: Real-time tracking of the human body. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):780–785, July 1997.

[17] 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.