Person Tracking Using Audio-Video Sensor Fusion

Neal Checka NEALC@AI.MIT.EDU
Kevin Wilson KWILSON@AI.MIT.EDU
MIT Artificial Intelligence Laboratory, 200 Technology Square, Cambridge MA, 02139 USA

1. Introduction

Audio and video signals originating from the same source tend to be related. To achieve optimal performance, 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, such a system may use audio data captured by a microphone array to track its location using the time delay of arrival of the audio signals at different microphones. A tracker that exploits both these modalities may be more robust and achieve better performance than one which uses either one alone. Each modality may compensate for weaknesses of the other one. 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 where an audio tracker alone may fail to track the object as it stops emitting sound or is masked by some background noise.

2. Related Work

Tracking people in known environments has recently become an active area of research in computer vision. Several person tracking systems have been developed to detect the number of people present as well as their 3D position over time. These systems generally use a combination of foreground/background classification, clustering of novel points, and trajectory estimation in one or more camera views [1]. [5] uses a Kalman filter for tracking multiple talkers using a microphone array. There has been less work done in the audio-visual tracking domain. [4] showed 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.

Our test environment is a conference room equipped with 32 omnidirectional microphones spread across the ceiling and 2 stereo cameras on the adjacent wall. In this paper, we first describe our existing vision tracking and microphone array processing algorithms, and then present a framework in which the two modalities can be combined to track multiple objects.

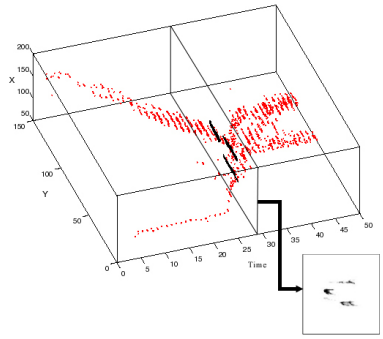

Figure 1. $(X, Y, Time)$ plot of plan-view foreground densities over time for a sequence of 3 moving people.

3. Vision Tracking

Fast foreground detection is necessary to detect rapidly moving objects and dynamic activity patterns. 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 disparity search computation during foreground detection and disparity estimation, yielding a fast, illumination-insensitive 3D tracking system. Details of our vision system are presented in [3].

When tracking multiple people, we have found that rendering an orthographic vertical projection of detected foreground pixels is a useful representation[1]. A “plan view” image facilitates correspondence in time since only 2D search is required. Previous systems would segment foreground data into regions prior to projecting into a plan-view, followed by region-level tracking and integration, potentially leading to sub-optimal segmentation and/or object fragmentation. Instead, we developed a technique that altogether avoids any early segmentation of foreground data. We merge the plan-view images from each view and estimate over time a set of trajectories that best represents the integrated foreground density. Trajectory estimation is performed by finding connected components in a spatio-

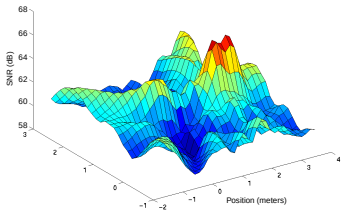

Figure 2. Array power response as a function of position (three speakers).

temporal filtered volume.

4. Microphone Array Processing

Our system uses the position estimate from the vision tracker as the initial guess from which to begin a gradient ascent search for a local maximum in beam output power (Figure 2). Positional standard deviation of our tracker was measured to be 10cm for a target moving along a straight line in the region of the room in which our experiments were performed. Gradient ascent to the nearest local maximum can therefore be expected to converge to the location of the speaker of interest when no other speakers are very close by. After localizing the target, we use the location estimate to calculate delays and shading coefficients for a delay-and-sum beamformer used to separate the target signal from background noise. Details of our microphone array processing algorithms are presented in [2].

5. Probabilistic Audio-Visual Tracking Framework

Our framework uses a probabilistic model to describe the observed data. Our problem can be formulated in a state-space estimation framework by expressing the likelihood that a hypothesized configuration $X_t = (x^1, ..., x^n)$ of n objects gave rise to an observation z at time t, where $x^i = [x, y, h, f]$ is a state vector of an object. A person is modelled as vertical cylinder with fixed radius where $[x, y]$ specifies floor position, h is the height of the cylinder, and f is the voicing frequency. The problem is then to estimate the state given measurements from the microphone array signals and the vision tracker. In our system, the video observation is a foreground disparity image and the audio observation is the output power of a delay-and-sum beamformer steered toward the hypothesized location. The Kalman filter is the optimal solution to the tracking problem when the posterior density at every time step is unimodal (Gaussian) and the dynamics of the system are linear. In a cluttered or noisy scene, our measurements have a non-Gaussian, multi-modal distribution (Figures 1 and 2). Particle filtering is an approximation technique for the non-linear and non-Gaussian cases. Particle filters are sequential Monte Carlo methods based upon point mass representations of probability densities, which can be applied to any state space model. The key idea is to represent the posterior density function by a set of random samples with associated weights and to compute estimates based on these samples and weights. As the number of samples becomes very large, this representation becomes a good approximation to the true posterior pdf. In order to apply particle filters to this problem, a state transition density $p(x_t|x_{t-1})$, or a model of how the states propagate, is required. We are currently investigating a variety of state dynamic models in order to reflect the different kinds and degrees of variability that are appropriate to person tracking. The other requirement is a likelihood function of the tracking and microphone data, i.e., $p(z_t|x_t)$ where z_t represents the data observed at time t.

5.1 Likelihood Models

For a given object, the likelihood $L(z|x_i)$ measures how well the observations support a hypothesis of a single object. The likelihood is computed as a product

$$L(z|x_i) = L(z_a|x_i) * L(z_v|x_i) \quad (1)$$

where $L(z_a|x_i)$ and $L(z_v|x_i)$ measure how well the audio and video data support the hypothesis x_i, respectively. For a configuration of multiple objects X, the total likelihood $L_T(z|X)$ is computed as

$$L_T(z|X) = \Phi(L(z|x_1), ..., L(z|x_m) \quad (2)$$

where $L_T(z|X)$ is a function of the likelihood, calculated according to equation 2, of the m objects.

References

[1] D.J. Beymer and K. Konolige. Real-time tracking of multiple people using stereo. In *Frame-Rate99*, 1999.

[2] K. Wilson, N. Checka, D. Demirdjian, and T. Darrell. Audio-video array source localization for perceptual user interfaces. In *PUI*, 2001.

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

[4] J. Vermaak, M. Gangnet, A. Blake, and P. Perez. Sequential Monte Carlo fusion of sound and vision for speaker tracking. In *ICCV*, 2001.

[5] D.E. Sturim, M.S. Brandstein, and H.F. Silverman. Tracking multiple talkers using microphone-array measurements. In *ICASSP*, 1997.