

Audio-Visual Speaker Localization Using Graphical Models

Akash Kushal Mandar Rahurkar Li Fei-Fei Jean Ponce Thomas Huang
University of Illinois, Urbana Champaign

Abstract

In this work we propose an approach to combine audio and video modalities for person tracking using graphical models. We demonstrate a principled and intuitive framework for combining these modalities to obtain robustness against occlusion and change in appearance. We further exploit the temporal correlations that exist for a moving object between adjacent frames to account for the cases where having both modalities might still not be enough, e.g., when the person being tracked is occluded and not speaking. Improvement in tracking results is shown at each step and compared with manually annotated ground truth.

1 Introduction

Multi-modal information fusion is an important problem in multimedia. The challenge is to combine different modalities to have a synergistic effect. There has been substantial work done in tracking moving objects using video, e.g. [2, 8, 7]. Multiple microphones have also been used to estimate the position of a speaker, e.g. [3]. Depending upon the position of the person, the sound reaches one microphone before the other and thus the signals received by the microphones are displaced by some number of samples τ . However, the problem of using these modalities together is relatively new. Garg et. al [6] address the speaker detection problem by combining multiple features, such as skin color, lip motion etc., using a probabilistic approach. The particle filtering approach of [2] was extended by Vermaak et al. [9] to include audio by modeling the cross correlations between the signals recieved by a microphone array.

Our focus however is using generative graphical models to solve this problem in a systematic manner. This paper builds on the work of Beal et.al [1] on audio-visual tracking and makes the following contributions. First, the video model proposed in [1] models the foreground by a subset of the pixels with constant appearance, translating together against a background of random noise throughout the video sequence. The results shown in [1] have the target person occupying a large portion of the image and the people in the background are moving around which adds randomness to the background appearance. Hence, the model of [1] successfully locks onto the target person. This is not the case in most scenes where the background appearance remains more or less constant. In this case the video model of [1]

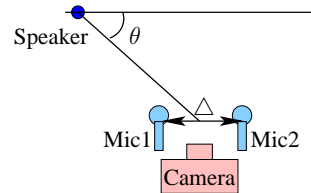


Figure 1. Experimental setup

locks onto the background instead of the target person. We propose an alternate video model that explicitly accounts for the background appearance. The proposed model also explicitly models the intermittent occlusion of the speaker in the video. Second, the model in [1] uses the data as a bag of frames and does not make use of the strong correlation among the position of the moving person in consecutive frames. We model this correlation and propose a dynamic programming algorithm to determine the most likely path of moving person in the video.

2 A Generative Model for Audio-Video Data

Our experimental setup (shown in Fig. 1) is similar to that of [1]. Two microphones with sampling rate $44KHz$ are placed about $30cm$ apart and a camera in the center captures 120×160 video frames at a frame rate of $10fps$. Figure 2 shows the proposed Audio-Visual graphical model. This section describes the audio, video and linkage parameters of the proposed model.

Audio: The variables \mathbf{x}_1 and \mathbf{x}_2 represent the audio signals observed at the microphones 1 and 2 respectively. Both the signals are partitioned into disjoint parts, one for each video frame. The number of samples N in each part is the ratio of the audio sampling rate to the video frame rate (in our case $N = 4400$). \mathcal{A}_f represents the true audio signal corresponding to frame f . The observation \mathbf{x}_1 at the left microphone is generated by adding zero mean random gaussian noise with precision matrix (*inverse covariance matrix*) ν to \mathcal{A}_f . The observed signal \mathbf{x}_2 at the right microphone is generated by shifting \mathcal{A}_f by a discrete sample delay τ , multiplying it by the relative attenuation ratio λ between the two microphones and again adding zero mean random gaussian noise with precision matrix ν to the result. We use circular shift in our implementation to simplify computation. Also, we model ν as a diagonal matrix with a constant diagonal.

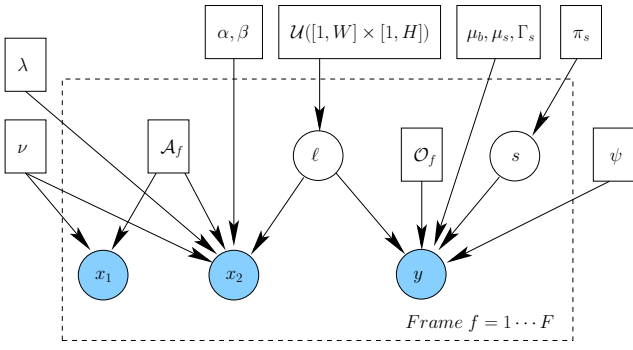


Figure 2. Audio-Video model

Video: The observed video frame \mathbf{y} is a vector with entries y_i for $1 \leq i \leq W \times H$ corresponding to the intensity of pixel i . Here W and H are the width and height of the observed frame. To account for the variability in the appearance of the speaker we allow the speaker’s appearance in each frame to belong to one of S appearance classes. The appearance class s and the corresponding foreground image μ_s and mask Γ_s are picked from the probability distribution π_s . The foreground image μ_s and the mask Γ_s are also represented as vectors of size $W \times H$. A common translation $\ell = (\ell_x, \ell_y)$ for the foreground image and mask is then picked from a uniform distribution. The boolean parameter \mathcal{O}_f for each frame f indicates whether the person is visible or occluded in frame f . If the person is visible (i.e. $\mathcal{O}_f = 0$), the observed frame is generated by first applying the mask Γ_s to μ_s , translating the resulting image by ℓ , combining the masked and shifted foreground image with the background μ_b and finally adding gaussian noise with precision matrix (*inverse covariance matrix*) ψ to the resulting image. On the other hand, if the person is occluded in frame f (i.e. $\mathcal{O}_f = 1$), the observed frame is generated by adding gaussian random noise again with precision matrix ψ to the background μ_b . In this case the video modality does not provide any information about ℓ . Again for computational ease, we model ψ as a diagonal matrix with constant diagonal entries.

Audio-Video Linkage: Assuming that the distance between the microphones Δ is small compared to the distance to the speaker and that the angle θ in figure 1 is close to $\pi/2$, the sample delay τ can be approximated as $\tau = \alpha \ell_x + \alpha' \ell_y + \beta$. Also, if the x -axis of the camera and the line joining the microphones are both horizontal one can assume $\alpha' = 0$ (similar to [1]).

3 Parameter Estimation and Tracking

We use the Expectation Maximization (EM) Algorithm for learning the parameters of the proposed model and obtaining the track estimate. In the E -step of each iteration the

posterior distribution over the hidden variables conditioned on the observed data is updated and the M -step updates the parameter estimates. Let θ denote the set of all the parameters $\{\nu, \lambda, \mathcal{A}_f, \alpha, \beta, \mu_b, \mu_s, \Gamma_s, \mathcal{O}_f, \pi_s, \psi\}$. To track the speaker through the video sequence we need to obtain an estimate of the translation ℓ of the foreground mask in every frame. After obtaining the posterior distribution over ℓ given the data $p(\ell|\mathbf{x}_1, \mathbf{x}_2, \mathbf{y}, \theta)$ we can obtain the most likely position as $\hat{\ell} = \operatorname{argmax}_{\ell} p(\ell|\mathbf{x}_1, \mathbf{x}_2, \mathbf{y}, \theta)$ for each frame.

Notation: Let the diagonal matrices $\nu = \nu \mathbf{I}$ and $\psi = \psi \mathbf{I}$. We will use $\mu_{\ell, s}$ to denote the image obtained by translating μ_s, Γ_s by ℓ and combining with μ_b . $\bar{\Gamma}_s$ denotes the inverted mask. Also, we use q to describe the posterior distribution conditioned on the data. The q notation omits the observed data as well as the parameters for compact representation. Hence, $q(s) = p(s|\mathbf{x}_1, \mathbf{x}_2, \mathbf{y}, \theta)$. The notation $\mathcal{N}(\mu, \nu)$ denotes a gaussian probability distribution with mean μ and precision matrix ν . The transformation G_{ℓ} circularly shifts an image by ℓ . Similarly, L_{τ} circularly shifts an audio signal by τ . Let $\mathbf{x}_{2, \ell}$ be $L_{-\lceil \alpha \ell + \beta \rceil} \mathbf{x}_2$. Finally, $\langle \rangle$ denotes averaging over all frames and $\langle \rangle_p$ denotes averaging over the frames for which the condition p holds.

Figure 3 shows the EM update equations obtained for the Audio-Visual model of figure 2. The audio-video linkage parameters α and β are updated as follows. We create a system of equations, one for each frame, mapping the most probable horizontal position ℓ_x onto the most probable shift τ (the one with highest cross-correlation). That is, we write equations of the form $\tau_f = \alpha \ell_{x, f} + \beta$ for $1 \leq f \leq F$. Since our model does not account for points of no speech explicitly, some of these equations might be erroneous. We use the RANSAC Algorithm [4] to eliminate outliers and obtain a robust estimate for α and β . Similar to [5], the update computations can be represented as convolutions and hence can be performed efficiently using FFT. This optimization reduces the computational complexity of each EM iteration to $O(F S W H \log(W H))$.

4 Temporal Constraints

We now describe a modification to the Audio-Video model proposed above that allows us to model the correlation in the position of the person among consecutive frames in the video. The modified model (Fig. 4), has edges connecting the hidden variables ℓ_f across consecutive frames. The variables ℓ_{f+1} are now conditioned on the variables ℓ_f and have a gaussian distribution with mean ℓ_f and precision parameter η . Applying the EM algorithm directly on this modified model becomes intractable due to the huge number of hidden variables. Hence, we use a two step process to estimate the trajectory of the person. First, we use the original model of (Fig. 2) to esti-

<p>E-Step:</p> $q(l, s) = \frac{1}{Z} \mathcal{N}(\mathbf{x}_2 \lambda L_{[\alpha\ell + \beta]} \mathcal{A}_f, \nu) [\pi_s \mathcal{N}(\mathbf{y} \boldsymbol{\mu}_{\ell, s}, \boldsymbol{\psi})]^{1 - \mathcal{O}_f}$ <p>M-Step:</p> $\boldsymbol{\mu}_b = \frac{\langle [\sum_{\ell, s} (q(\ell, s) G_\ell \bar{\boldsymbol{\Gamma}}_s)^{1 - \mathcal{O}_f} \mathbf{y}] \rangle}{\langle [\sum_{\ell, s} q(\ell, s) G_\ell \bar{\boldsymbol{\Gamma}}_s]^{1 - \mathcal{O}_f} \rangle}$ $\boldsymbol{\mu}_s = \frac{\langle \sum_{\ell} q(\ell, s) G_\ell^{-1} \mathbf{y} \rangle_{\mathcal{O}_f=0}}{\langle \sum_{\ell} q(\ell, s) \rangle_{\mathcal{O}_f=0}}$ $\boldsymbol{\psi} = \frac{WH}{\langle [\sum_{\ell, s} q(\ell, s) \mathbf{y} - \boldsymbol{\mu}_{\ell, s} ^2]^{1 - \mathcal{O}_f} [\mathbf{y} - \boldsymbol{\mu}_b ^2]_{\mathcal{O}_f} \rangle}$ $\pi_s = \langle \sum_{\ell} q(\ell, s) \rangle_{\mathcal{O}_f=0}$ $\nu = \frac{2N}{\langle \mathcal{A}_f - \mathbf{x}_1 ^2 + \sum_{\ell} q(\ell) \lambda \mathcal{A}_f - \mathbf{x}_2, \ell ^2 \rangle}$ $\lambda = \langle \sum_{\ell} q(\ell) \frac{\mathbf{x}_2, \ell^T \mathcal{A}_f}{\mathcal{A}_f^T \mathcal{A}_f} \rangle$	$q(\ell) = \sum_s q(\ell, s)$ $\boldsymbol{\Gamma}_s(i) = \begin{cases} 1 & \mathcal{E}_0(i) > \mathcal{E}_1(i) \\ 0 & o/w \end{cases}$ <p>where $\mathcal{E}_0 = \langle \sum_{\ell} (q(\ell, s) G_\ell^{-1} (\mathbf{y} - \boldsymbol{\mu}_b)^2) \rangle_{\mathcal{O}_f=0}$</p> $\mathcal{E}_1 = \langle \sum_{\ell} (q(\ell, s) (G_\ell^{-1} (\mathbf{y} - \boldsymbol{\mu}_s)^2)) \rangle_{\mathcal{O}_f=0}$ $\mathcal{O}_f = \begin{cases} 1 & \mathcal{E}_{bg} < \mathcal{E}_{fg} \\ 0 & o/w \end{cases}$ <p>where $\mathcal{E}_{bg} = \mathbf{y} - \boldsymbol{\mu}_b ^2$</p> $\mathcal{E}_{fg} = \sum_{\ell, s} q(\ell, s) \mathbf{y} - \boldsymbol{\mu}_{\ell, s} ^2$ $\mathcal{A}_f = \sum_{\ell} q(\ell) \frac{\mathbf{x}_1 + \lambda \mathbf{x}_2, \ell}{1 + \lambda^2}$
---	--

Figure 3. EM update equations

mate the parameters using the EM-algorithm as described in the previous section. After the model parameters have been learned and the estimates of the posterior distributions $q(\ell_f)$ have been computed for each frame f , the most likely trajectory is estimated using the modified model. Let $q(\ell_1, \ell_2, \dots, \ell_F)$ be the posterior probability of the trajectory $\ell_1, \ell_2, \dots, \ell_F$ using the modified model. One can show that $q(\ell_1, \ell_2, \dots, \ell_F) \propto \prod_{f=1}^{F-1} p(\ell_{f+1} | \ell_f) \prod_{f=1}^F q(\ell_f)$. Hence, the most likely trajectory maximizes the product on the right hand side. Taking logs turns this product into a sum of log probabilities. Let \mathcal{P}_f be the negative log probability table corresponding to $-\log(p(\ell_f | \ell_{f-1}))$ and let \mathcal{Q}_f be the negative log probability table corresponding to $-\log(q(\ell_f))$. The best trajectory minimizes the total *penalty* (sum of the entries of \mathcal{P}_f and \mathcal{Q}_f corresponding to ℓ_f for all the frames) and can be efficiently computed by dynamic programming as shown in Algorithm 1. The working of the algorithm can be visualized using Fig. 6(a). The background shows the log probabilities \mathcal{Q}_f for each frame and the blue points show the selected trajectory. The algorithm finds a path (values for ℓ_f) from the left to right in Fig. 6(a) that minimizes the *penalty*. The \mathcal{Q}_f term in the *penalty* accounts for the probabilities assigned to each value of ℓ_f by the original model whereas the \mathcal{P}_f term depends on η and penalizes paths with sudden changes in the values of ℓ_f .

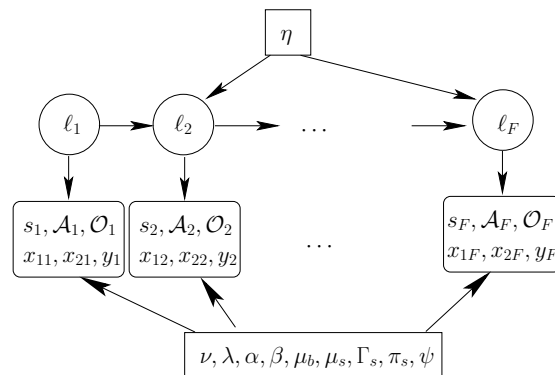


Figure 4. Modeling temporal correlation.

5 Results and Analysis

We tested our tracking algorithm on audio-video sequences captured by the setup in figure 1. The video consists of a person moving from left to right and back in front of the setup. The background $\boldsymbol{\mu}_b$ was initialized using the mean of all the frames in the video. The number of classes for the foreground and mask was kept at 2. The foreground images for the 2 classes were initialized with random images from the video and the mask was initialized using simple background subtraction. Our implementation of the EM algorithm converges in about 5 iterations and processes about 0.5 frames per class, per iteration, per second. Figures 5(a) and 5(b) show the estimated means and masks for

```

Input: Negative log probability tables  $\mathcal{P}_f$  and  $\mathcal{Q}_f, \forall f$ 
Output: Trajectory  $S = [\ell_1 \dots \ell_f]$  minimizing penalty
 $bestVal \leftarrow \mathcal{Q}_1$ 
for  $f = 2$  to  $F$  do
  for all assignment  $x_j$  to  $\ell_f$  do
     $newBest(x_j) \leftarrow INFINITY$ 
    for all assignment  $x_k$  to  $\ell_{f-1}$  do
      if  $bestVal(x_k) + \mathcal{P}_f(x_j|x_k) + \mathcal{Q}_f(x_j) < newBest(x_j)$ 
        then
           $newBest(x_j) \leftarrow bestVal(x_k) + \mathcal{P}_f(x_j|x_k) + \mathcal{Q}_f(x_j)$ 
           $bestPrev(f-1, x_j) \leftarrow x_k$ 
        end if
      end for
    end for
   $bestVal \leftarrow newBest$ 
end for
 $y \leftarrow indexOfMinimumValue(bestVal)$ 
 $S(F) \leftarrow y$ 
for  $f = F-1$  to  $1$  do
   $S(f) \leftarrow bestPrev(f, y)$ 
   $y \leftarrow bestPrev(f, y)$ 
end for

```

Algorithm 1: Selecting the most likely trajectory

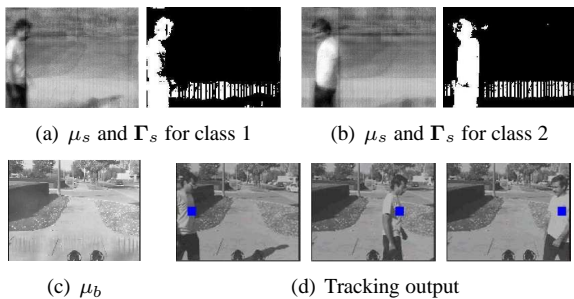


Figure 5. Video model parameter estimates

the two classes. The two appearance classes have locked onto the person moving in different directions. Fig. 5(c) shows the estimated background and the blue square in Fig. 5(d) shows the tracking output.

To test robustness against occlusion, we introduced a bar in the video which partially covers the image and hence intermittently occludes the speaker. The video model alone, with \mathcal{O}_f forced to be 0 (no occlusion) for all f , loses track of the moving person when he moves through the artificial occlusion as shown in Fig. 7(a). The combined audio-video model provides a better estimate as shown in Fig. 7(b). Finally, the trajectory (Fig. 7(c)) obtained after applying the temporal constraints is very close to the manually annotated ground truth as can be seen from Fig. 6(b). The tracking results can be found in the accompanying video.

6 Conclusion and Future Work

We propose a new generative model for audio-visual object tracking that improves upon the weaknesses of [1]. The algorithm performs calibration automatically as part of EM.

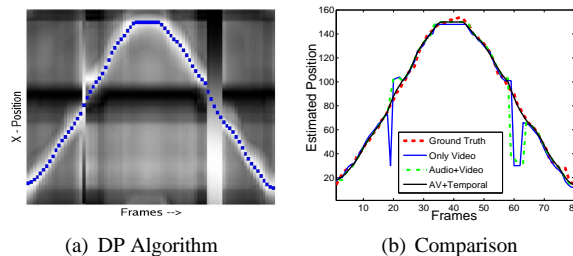


Figure 6. Temporal Constraints

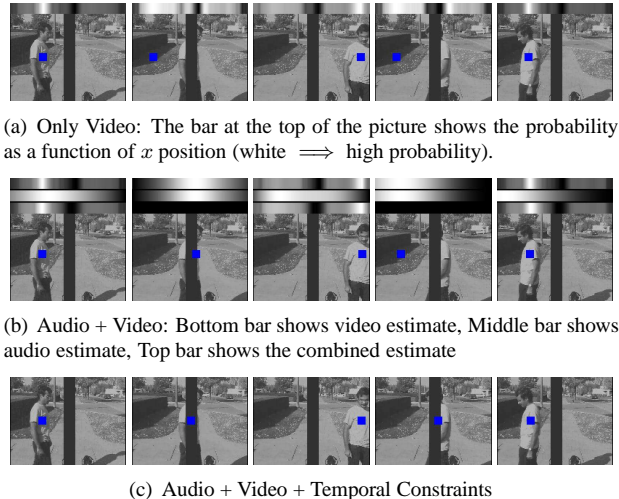


Figure 7. Tracking Results

We also capitalize on the correlation in the position of the person in adjacent frames to obtain a more robust estimate of the person's position. At this point there are multiple directions for future work, including extending the approach to multiple moving persons and allowing for camera motion.

References

- [1] M. J. Beal, N. Jojic, and H. Attias. Graphical model for audiovisual object tracking. *IEEE Trans on PAMI*, 25:828–836, July 2003.
- [2] A. Blake and M. Isard. *Active Contours*. Springer, 1998.
- [3] M. Brandstein and D. Ward. *Microphone Arrays*. Springer, 2001.
- [4] M. A. Fischler and R. C. Bolles. Random sample consensus. *Communications ACM*, 24(6):381–395, June 1981.
- [5] B. J. Frey and N. Jojic. Fast, large-scale transformation-invariant clustering. In *NIPS*, pages 721–727, 2001.
- [6] A. Garg, V. Pavlovic, and J. Rehg. Audio visual speaker detection using dynamic bayesian networks. *Proc. IEEE Conf. Automatic Face and Gesture Recogniton*, 2000.
- [7] N. Jojic and B. Frey. Learning flexible sprites in video layers. *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2001.

- [8] N. Jovic, N. Petrovic, B. Frey, and T. Huang. Transformed hidden markov models: Estimating mixture models of images and inferring spatial transformations in video sequences. *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, June 2000.
- [9] J. Vermaak, M. Gangnet, A. Blake, and P. Perez. Sequential monte carlo fusion of sound and vision for speaker tracking. *Proc. Int'l Conf. Computer Vision*, June 2000.