IMPROVED GRAPHICAL MODEL FOR AUDIOVISUAL OBJECT TRACKING

Hao Tang and Thomas S. Huang

Department of Electrical and Computer Engineering
University of Illinois at Urbana-Champaign
Email: {haotang2, huang}@ifp.uiuc.edu

ABSTRACT

Object tracking plays an important role in multimedia surveillance systems, in which the major types of data are video and audio captured by cameras and microphone arrays. In this paper, we describe a systematic approach to audiovisual object tracking, originally proposed by Beal et al, based on graphical models that jointly combine audio and video variables under a single probabilistic framework. We seek to improve this approach through three aspects: First, we introduce background subtraction preprocessing of video data. Second, we modify the video model to exclude the background from being transformed. Third, we extend the joint model to a dynamic Bayes net. These improvements yield satisfactory results on single person tracking in a noisy outdoor environment with far-field background road traffic, and handle situations where the target is lost due to occlusions.

1. INTRODUCTION

The analysis of multimedia data is of paramount importance in today's society. Two major types of multimedia data are video and audio captured by video cameras and microphone arrays. In many surveillance applications, with both video and audio available, a fundamental task is to track an object in the scene. In the literature, there has been extensive work on object tracking based on either video or audio only. In the belief that by fusing video and audio cues one can potentially improve tracking performance, researchers have begun to exploit the correlations between the two modalities. This has led to systems in which video and audio cues are loosely coupled. In those systems, a video tracker and an audio tracker are run independently, and the results of the two trackers are combined at a high level. However, in order to maximize the benefit of joint audiovisual processing, cue fusion must be performed at a low level in a systematic manner.

In 2003, Beal et al proposed a novel approach to audiovisual object tracking based on graphical models that jointly combine audio and video variables under a single probabilistic framework [1]. This approach for the first time models the joint statistical characteristics of the video and audio signals. The joint model exploits the dependency of the time delay between the audio signals received at two microphones on the object position in the scene. As a result, calibration is not required for it is automatically performed by parameter learning in the model.

We seek to improve this approach through three aspects: First, we introduce background subtraction preprocessing of video data to make the foreground object "stable" against the background. Second, we modify the video model to exclude the background from being transformed. Third, we extend the joint model to a dynamic Bayes net (DBN) to characterize the temporal evolution of the object trajectory.

This paper is organized as follows. Section 2 describes the joint audiovisual model proposed by Beal et al. Sections 3-5 elaborate our improvements to the joint model, namely background subtraction preprocessing, the modified video model, and the DBN, respectively. We present our experiment results in Section 6, and conclude the paper in Section 7.

2. JOINT AUDIOVISUAL MODEL

Assuming the experiment setup in Fig. 1(a), the joint audiovisual model proposed by Beal et al is illustrated in Fig. 1(b). It is based on graphical models, or Bayesian networks (BN). The joint model consists of an audio model and a video model constructed in a symmetric fashion. In the audio model, the observed audio signals x_i at mic $i = 1,2$ are described in terms of a latent audio signal a. a is attenuated by a factor λ_i on its way to mic i and is received at mic 2 with a time delay τ relative to mic 1. Additionally, it is contaminated by zero-mean Gaussian noise with precision matrix v_i. a is defined by a mixture model, of which the component r, with a prior probability π_r, is a zero-mean Gaussian template with precision matrix η_r. Mathematically, this generative process is described as

$$p(r) = \pi_r, \; p(a \mid r) = N(a; 0, \eta_r)$$
$$p(x_1 \mid a) = N(x_1; \lambda_1 a, v_1)$$
$$p(x_2 \mid a, \tau) = N(x_2; \lambda_2 L_\tau a, v_2)$$

where L_τ is a delay operator.

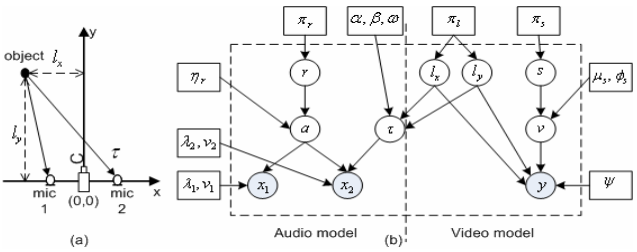

Fig. 1. (a) Camera and microphone setup. (b) The joint audiovisual model proposed by Beal et al.

Likewise, in the video model, the observed image y is described in terms of a latent image v. v is first spatially shifted by $l=(l_x,l_y)$ pixels (drawn from a uniform distribution) and then contaminated by zero-mean Gaussian noise with precision matrix ψ. v is defined by a mixture model, of which the component s, with a prior probability π_s, is a Gaussian template with mean μ_s and precision matrix ϕ_s. Mathematically, this generative process is described as

$$p(s)=\pi_s,\ p(v\mid s)=N(v;\mu_s,\phi_s)$$
$$p(l)=\pi_l,\ p(y\mid v,l)=N(y;G_l v,\psi)$$

where G_l denotes a spatial shift operator.

In the setup in Fig. 1(a), the dependency of the time delay τ on the object position l is modeled by a linear mapping, varied by zero-mean Gaussian noise with precision ω

$$p(\tau\mid l)=N(\tau;\alpha l_x+\alpha' l_y+\beta,\omega)$$

Finally, the BN in Fig. 1(b) gives a factorized form of the joint distribution over all random variables

$$p(x_1,x_2,y,a,\tau,r,v,l,s)=p(x_1\mid a)p(x_2\mid a,\tau)p(a\mid r)p(r)$$
$$p(y\mid v,l)p(v\mid s)p(s)p(\tau\mid l)p(l)$$

An efficient E-M algorithm is derived to learn and infer in the model [1]. Tracking is done by evaluating

$$\hat{l}=\arg\max_l p(l\mid x_1,x_2,y)$$

The above audio and video models both use a technique termed transformed mixture of Gaussians (TMG) [2]. In the audio model the transformations are time delays and in the video model spatial shifts. An analysis of TMG reveals a drawback of the video model. TMG is a transformation-invariant clustering technique that tends to capture the most typical or stable part of an image sequence. In practice, the background of a scene is usually deemed as more stable than the foreground object. In this case, TMG tends to track the background instead of the desired foreground object. Although the audio model is fighting against this trend, it is however less effective. Due to the high level of noise present in the audio signals, the time delay is difficult to estimate. The much stronger visual cues cause a bias toward the video model. An example of failure of tracking is demonstrated in Fig. 6.

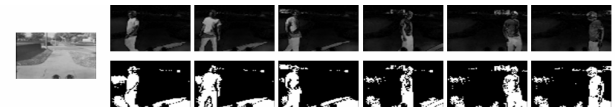

Fig. 2. Background subtraction preprocessing. (Left) the extracted background. (Top) background subtracted video frames. (Bottom) binarized video frames.

3. BACKGROUND SUBTRACTION PREPROCESSING

Since TMG tends to capture the most stable part of an image sequence, a natural idea would be to force to make the foreground object "stable" relative to the background. This is achieved by a background subtraction preprocessing procedure. First, we extract the background of the scene by averaging all the video frames across time. Next, we subtract the obtained background from every video frame. The preprocessed video data, with the foreground object made stable, is then used to train the joint audiovisual model.

The foreground object can be made even more stable by further binarizing the preprocessed video data. We choose an intensity threshold θ so that all pixels greater than θ are set to 1 and the rest 0. See Fig. 2.

4. MODIFIED VIDEO MODEL

In the video model, the background is spatially shifted along with the foreground object. An alternative to background subtraction preprocessing is that we modify the video model to exclude the background from being transformed. The new BN is shown in Fig. 3.

In the modified video model, the latent image v represents only the foreground object. v is spatially shifted by l pixels before it is combined with the background b to form the observed image y. In this model, b is a deterministic variable whose value is obtained as described in the previous section. The observed image y is then drawn from

$$p(y\mid v,l)=N(y;G_l v+b,\psi)$$

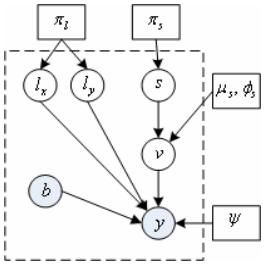

Fig. 3. The modified video model. In this model, the background is excluded from being transformed.

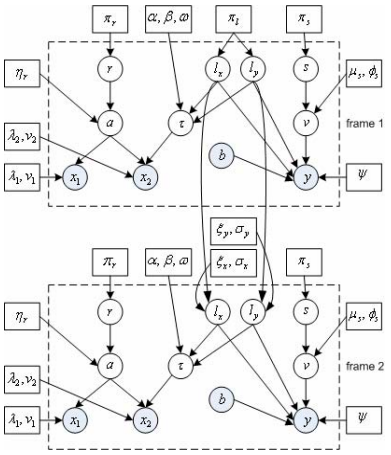

Fig. 4. The dynamic Bayes net. For more slices, unroll the DBN.

5. MODELING TEMPORAL DYNAMICS

The joint audiovisual model described above assumes all random variables are i.i.d. across time, without mentioning any inter-frame dependencies. Thus, it is called the "static" model. In practice, however, an object is observed to change its position slowly over time. In order to characterize the temporal evolution of the object trajectory, we extend the static model to a dynamic Bayes net (DBN) [3], as shown in Fig. 4. In the DBN, the change of object position l between two successive time frames is bounded by a Gaussian with mean ξ and precision σ

$$l^t = l^{t-1} + (\Delta l)^t, \left|(\Delta l)^t\right| \le N(\xi, \sigma)$$

The task of tracking in the DBN is then to evaluate

$$\hat{l}^{1:t} = \arg\max_{l^{1:t}} p(l^{1:t} \mid x_1^{1:t}, x_2^{1:t}, y^{1:t})$$

It would be useful to assume a predefined ξ and $\sigma = \infty$. This leads to an approximated realization of the DBN using the Viterbi algorithm [4]. First, we run the static model on the data to generate for each time t an N-best candidate list of the position estimates $\hat{l}_i^t$, with each $\hat{l}_i^t$ associated with a target probability

$$P_{\text{target}}(\hat{l}_i^t) = p(\hat{l}_i^t \mid x_1^{(t)}, x_2^{(t)}, y^{(t)})$$

We form a trellis of $\hat{l}_i^t$ as shown in Fig. 5. Define the link probability between the candidate i at time t and the candidate j at time $t+1$ as

$$P_{\text{link}}(l_i^t, l_j^{t+1}) = \begin{cases} 1 & \left|l_j^{t+1} - l_i^t\right| \le \xi \\ 0 & \text{otherwise} \end{cases}$$

The total probability of a short path connecting two candidates in successive time frames in the trellis is defined as the weighted sum of the target and link probabilities

$$P_{\text{total}}(i, j) = w_1 * P_{\text{target}}(l_i^t) + w_2 * P_{\text{link}}(l_i^t, l_j^{t+1})$$

where the weights w_1 and w_2 are set empirically.

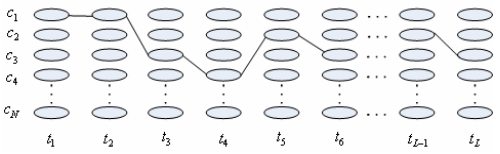

Fig. 5. Viterbi search for the path with highest total probability.

The Viterbi algorithm finds the "best" path in the trellis of which the total probability P_{total} is maximum. The candidates $\hat{l}_i^t$ along the "best" path are then our optimal position estimates of the object for the corresponding time frames.

6. EXPERIMENTS AND RESULTS

We carried out a series of experiments based on our in-house data, captured by an off-the-shelf camera and two microphones, as shown in Fig. 1(a). The data include the video of a person walking in a noisy outdoor environment with far-field background road traffic and his speech. We used 420 audiovisual frames (about 30s) of the data for parameter learning, and tracking is part of inference in the model. The parameters are randomly initialized except that μ_s is set to be the first frame of the video data. All the precision matrices are assumed to be diagonal. For the DBN, we set $\xi = 5$, $w_1 = 1$, and $w_2 = 10$. Usually, the E-M algorithm converges after 5 iterations.

Fig. 7 is the result of background subtraction preprocessing and Fig. 8 the result of the modified video model. In both results the person in the scene is accurately tracked. Fig. 9 shows the result of the DBN. Here, the result of the DBN is identical to that of the modified video model. A reasonable explanation is that the video cues (and probably the audio cues) have provided sufficient information for the static model to accurately track the person, thus leaving no improvement space for the DBN. However, in situations where the person is nearly totally occluded, the advantage of the DBN becomes obvious.

Fig. 10 shows the result of the static model with an artificial occluding bar. When the person walks behind the bar, visual cues disappear and the static model loses the target. However, the DBN overcomes this drawback. It can accurately track the person even he is totally buried behind the bar, as shown in Fig. 11.

Exp.	Total frames	Correct frames	Accuracy
1	420	420	100%
2	420	420	100%
3	420	420	100%
4	420	400	95.24%
5	420	420	100%

Table 1. A comparison of tracking performance of various experiments. 1. Background subtraction preprocessing. 2. Modified video model. 3. DBN. 4. Modified video model with occlusion. 5. DBN with occlusion.

Fig. 6. Tracking result for the joint audiovisual model with raw video input

Fig. 7. Tracking result for the joint audiovisual model with background subtraction preprocessing

Fig. 8. Tracking result for the joint audiovisual model with the modified video model

Fig. 9. Tracking result for the DBN

Fig. 10. Tracking result for the static model with an artificial occluding bar

Fig. 11. Tracking result for the DBN with an artificial occluding bar

Table 1 is a summary of tracking performance of the above various experiments. One should note that the DBN demonstrates great potential for object tracking in those scenes where possible occlusions exist.

7. CONCLUSION

In this paper, we describe a novel approach to audiovisual object tracking, originally proposed by Beal et al, which is based on graphical models that jointly combine audio and video variables under a single probabilistic framework. We have improved this approach through three aspects: First, we introduce background subtraction preprocessing of video data to make the foreground object stable against the background. Second, we modify the video model to exclude the background from being transformed. Third, we extend the joint model to a dynamic Bayes net to characterize the temporal evolution of the object trajectory. A series of our experiments have shown that both background subtraction preprocessing and the modified video model yield satisfactory results on single person tracking in a noisy outdoor environment with far-field background road traffic, and the DBN can further handle situations where the target is lost due to occlusions.

The DBN is a promising technique for modeling temporal data. It is potentially useful for simultaneously tracking multiple objects in a clutter background with occlusions. Our future work will be focused on applying the DBN technique to multiple audiovisual object tracking in a complex scene in which the objects might be occluded by one another.

8. ACKNOWLEDGMENTS

The author would like to thank Mandar Rahurkar and Amit Sethi for providing the audiovisual database used in the experiments in this paper.

9. REFERENCES

[1] M. Beal, N. Jojic and H. Attias, "A graphical model for audiovisual object tracking," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 25:7, pp. 828-836, July 2003.

[2] B. Frey and N. Jojic, "Transformation-invariant clustering using the EM algorithm," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 25:1, pp. 1-17, January 2003.

[3] K. Murphy, *Dynamic Bayesian Networks: Representation, Inference and Learning*. PhD Thesis, UC Berkeley, Computer Science Division, July 2002.

[4] G. D. Forney, "The Viterbi algorithm," *Proceedings of the IEEE* 61(3), pp. 268–278, March 1973.