IEICE TRANS. INF. & SYST., VOL.E93–D, NO.3 MARCH 2010

LETTER

Robust Object Tracking via Combining Observation Models

Fan JIANG[†a)], ***Student Member***, **Guijin WANG**[†b)], ***Member***, **Chang LIU**[†c)], **Xinggang LIN**[†d)], ***and*** **Weiguo WU**[††e)], ***Nonmembers***

SUMMARY Various observation models have been introduced into the object tracking community, and combining them has become a promising direction. This paper proposes a novel approach for estimating the confidences of different observation models, and then effectively combining them in the particle filter framework. In our approach, spatial Likelihood distribution is represented by three simple but efficient parameters, reflecting the overall similarity, distribution sharpness and degree of multi peak. The balance of these three aspects leads to good estimation of confidences, which helps maintain the advantages of each observation model and further increases robustness to partial occlusion. Experiments on challenging video sequences demonstrate the effectiveness of our approach.

key words: *object tracking, combine observation models, feature fusion, particle filter*

1. Introduction

Object tracking is a crucial and basic task in computer vision. It is the foundation of many higher level applications, such as visual surveillance, behavior analysis and human-computer interaction. Tracking aims at distinguishing and locating specific objects from continuously changing backgrounds. The main barriers to robust object tracking are illumination change, shape change, viewpoint change and occlusion.

To handle these visual variations, a variety of features have been introduced for target representation, including color histograms [1], HOG (Histogram of Oriented Gradients) [2] , Local Binary Patterns, Haar-like wavelets, edgelets and so on. Based on these features, different observation models are constructed to evaluate the likelihood between the tracking target and a candidate image window. Generally speaking, observation models can be divided into two categories: Generative models, like Kernel-based tracking [1], try to construct adaptive models in a specific feature space; Discriminative models, like [3], [4], which consider tracking as a classification problem, seek a decision boundary best separating the object and the background.

Because of the complementary characteristics between different features, it is beneficial to integrate different features together. Y. Li [5] proposed to organize three observation models in a cascade manner, in which the last one plays the key role, while the information from the others are not exploited sufficiently. Y. Lei [6] proposed a three-level hierarchy to efficiently combine two parallel observation models in the particle filter framework. However, their approach generally focused on a single object. Without any spatial information, their algorithm is hard to be extended to multiple objects cases. F. Tang [7] proposed a novel algorithm based on semi-supervised learning, which combines two discriminative models according to the accuracy of each classifier. This strategy cannot be extended to generative models, and furthermore, accuracy evaluated on previous samples prevents the observation model from adapting to the latest change.

We propose a novel strategy for combining different observation models. Three simple but efficient parameters, average likelihood, 3db bandwidth and peak-peak ratio, are extracted from the spatial likelihood distribution, which represent the overall similarity, distribution sharpness and degree of multi peak. They are further utilized to evaluate the confidence of an observation model. Observation models are then combined in the particle filter framework, which performs MAP (maximum a posteriori) estimation frame by frame. The novel combination framework based on spatial information maintains the advantages of separate observation models and further makes our algorithm robust to partial occlusion.

The rest of this paper is organized as follows: Sect. 2 introduces our method for confidence evaluation. Section 3 describes how to combine models in the particle filter framework. Experimental results are shown in Sect. 4 and conclusions are drawn in Sect. 5.

2. The Confidences of Observation Models

The idea of combining different observation models aims at maintaining advantages while avoiding disadvantages. Ideally, a good combination take effect when either individual model works. Thus, the most important thing in a combination framework is how to evaluate the confidences of different observation models, indicating which model is more reliable. Figure 1 demonstrates a commonly used combination framework. Likelihood distribution based on color histogram feature and likelihood distribution based on HOG

Manuscript received September 1, 2009.
Manuscript revised November 18, 2009.
[†]The authors are with the Department of Electronic Engineering, Tsinghua University, Beijing 100084, China.
[††]The author is with Sony China Research Laboratory, China.
a) E-mail: jiang_f99@mails.tsinghua.edu.cn
b) E-mail: wangguijin@tsinghua.edu.cn
c) E-mail: chang-liu06@mails.tsinghua.edu.cn
d) E-mail: xglin@tsinghua.edu.cn
e) E-mail: Weiguo.Wu@sony.com.cn
DOI: 10.1587/transinf.E93.D.662

Copyright © 2010 The Institute of Electronics, Information and Communication Engineers

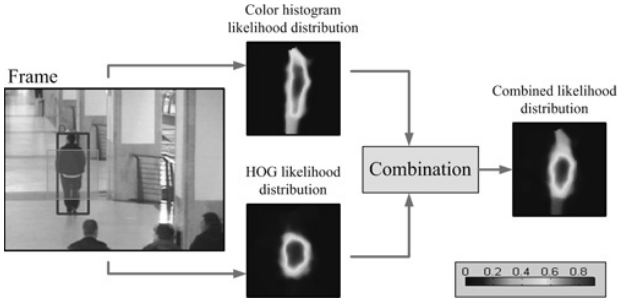

Fig. 1 A commonly used framework for combining different observation models.

feature are calculated separately. They are then combined into one spatial distribution according to their confidence. Our combination is similar to this one, except in the particle filter framework. (Sect. 3)

2.1 Parameters to Represent a Likelihood Distribution

Spatial information is employed to efficiently evaluate the confidence of different observation models. In practice, spatial likelihood distribution is encoded into the following three parameters.

2.1.1 Average Likelihood

Observation likelihood represents the similarity between an observation model and a candidate image win-dow, thus average likelihood shows the overall extent of matching.

Denote likelihood distribution as $d(x, y), (x, y) \in P$, where P is the entire searching area. To avoid noise, average likelihood $\overline{d}$ is calculated in a subset of P, with likelihood exceeds one threshold:

$$\overline{d} = \frac{1}{|R|} \sum_{(x,y)\in R} d(x, y) \tag{1}$$

Where $R = \left\{(x, y)\middle|(x, y) \in P, d(x, y) > th\right\}$ and $|\cdot|$ denotes the number of elements in the set. In practice, threshold th is set to 0.7 times maximum.

2.1.2 3db Bandwidth

3db bandwidth is a widely used measure for distribution sharpness, while distribution sharpness reflects its ability to distinguish target from backgrounds. We adopt the normalized 3db bandwidth in our approach:

$$S = \frac{1}{|P|} \sum_{(x,y)\in P} I\left[d(x, y) > th\right] \tag{2}$$

Where th is a threshold and $I[\cdot]$ is an indicator function.

$$I\left[d(x, y) > th\right] = \begin{cases} 1 & d(x, y) > th \\ 0 & otherwise \end{cases} \tag{3}$$

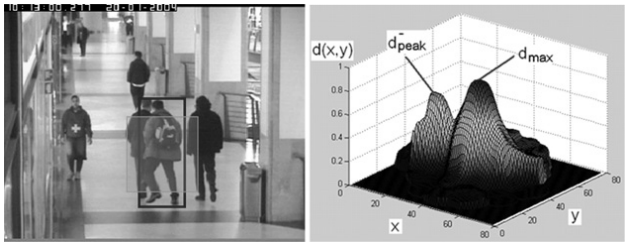

Fig. 2 An example of bimodal distribution.

2.1.3 Peak-Peak Ratio

Peak-peak ratio (γ) is defined as the ratio between the distribution maximum (d_{max}) and the second largest extremum (d^{-}_{peak}).

$$\gamma = d^{-}_{peak}/d_{max} \tag{4}$$

This parameter reflects the degree of multimodal distribution. It decreases when the maximum has larger advantage over the second peak and reaches zero when the distribution is strictly unimodal. In tracking, a unimodal structure of likelihood distribution is preferred because bimodal or multimodal structure generally indicates the existence of interference. Figure 2 gives an example of bimodal distribution when tracking pedestrians using HOG feature.

2.2 Confidence Evaluation

After describing likelihood distribution from three different perspectives, the confidence of an likelihood distribution ($conf$) can be obtained by comparing the current distribution parameters $\left\{\overline{d}, S, \gamma\right\}$ to the initial ones $\left\{\overline{d}_0, S_0, \gamma_0\right\}$. By introducing Gaussian kernels, confidence can be calculated as:

$$conf = \frac{e^{-c\gamma} e^{-k\left[1-min\left(1,\overline{d}/\overline{d}_0\right)\right]}}{S/S_0} \tag{5}$$

Where c and k are constants. In Eq. (5), the first part $e^{-c\gamma}$ corresponds to distribution shape, which will be maximized for a unimodal structure ($\gamma = 0$). The second part $e^{-k\left[1-min\left(1,\overline{d}/\overline{d}_0\right)\right]}$ corresponds to average likelihood. Larger average likelihood will result in higher confidence. An average likelihood larger than the initial value ($\overline{d} > \overline{d}_0$) is considered to be good enough, thus maximizing the second part. The last part S/S_0 in the denominator corresponds to the distribution sharpness. On the whole, Eq. (5) achieves a good balance among the three perspectives. Confidence increases when $\overline{d}$ increases, S decreases and γ decreases.

3. Combination Using Particle Filter

Intuitively, likelihood distribution can be calculated pixel-

Table 1 Tracking by combining observation models.

Input:	
	N weighted particles in previous frame: $\left\{X_{t-1}^i, w_{t-1}^i\right\}_{i=1}^N$
<1>	Resample:
	simulate particles $\left\{X_{t-1}^i, \frac{1}{N}\right\}_{i=1}^N$ from $\left\{X_{t-1}^i, w_{t-1}^i\right\}_{i=1}^N$
<2>	Prediction: for $i = 1 \ldots N$
	Simulate $X_t^i \sim p\left(X_t \middle\| X_{t-1}^i\right)$, obtains $\left\{X_t^i\right\}_{i=1}^N$
<3>	Observe: for each particle X_t^i
	Get observations $L_{1,t}^i$ and $L_{2,t}^i$, obtains $\left\{X_t^i, L_{1,t}^i, L_{2,t}^i\right\}_{i=1}^N$
<4>	Estimate likelihood distribution:
	Estimate $d_{1,t}(x, y)$ and $d_{1,t}(x, y)$ by Eq. (6)
<5>	Evaluate confidence $Conf_{1,t}$ and $Conf_{2,t}$ by Eq. (5)
<6>	Update particle weights through combination:
	$w_t^i = \frac{Conf_{1,t}}{Conf_{1,t} + Conf_{2,t}} L_{1,t}^i + \frac{Conf_{2,t}}{Conf_{1,t} + Conf_{2,t}} L_{2,t}^i$
	Obtains weighted particles $\left\{X_t^i, w_t^i\right\}_{i=1}^N$ for the next frame
Output:	
	Target location in the current frame $X_t = \frac{\sum_{i=1}^N X_t^i w_t^i}{\sum_{i=1}^N w_t^i}$

by-pixel using a sliding window in the searching area. However, this is not acceptable due to the computational cost, especially when there is scale change. To deal with this problem, we combine observation models in the particle filter framework by the following steps.

- Estimate likelihood distributions from particles.
- Evaluate model confidences at a single scale.
- Combine models by updating particle weights.

Our tracking system follows the framework of SIR particle filter [8]. Target center (x, y) as well as a scale factor s are chosen as the state vector, denoted as $X_t = (x_t, y_t, s_t)^T$. Width-height ratio is regarded to be constant after initialization. In addition, we assume that targets move at constant velocity.

3.1 Distribution Estimation by Particles

Our distribution is estimated around scale 1, which means selecting particles with approximately the same size to the target in the previous frame. These selected particles are denoted as $\left\{X^i, L^i\right\}_{i=1}^M$, where X^i is the state vector, and L^i is the likelihood. Likelihood distribution can be estimated:

$$d(x, y) = \frac{\sum_{i=1}^M L^i G\left(x, y \middle\| x^i, y^i, \sum\right)}{\sum_{i=1}^M G\left(x, y \middle\| x^i, y^i, \sum\right)} \tag{6}$$

Where $G\left(x, y \middle\| x^i, y^i, \sum\right)$ denotes a Gaussian distribution with mean (x^i, y^i) and covariance $\sum$. Each Gaussian distribution corresponds to one particle. The ones closer to point (x, y) contribute more in the ensemble.

3.2 The Entire Algorithm

Table 1 demonstrates the entire tracking algorithm which combines two observation models.

4. Experimental Results

Basically, pedestrian tracking epitomizes the main problems in object tracking, so we choose pedestrian tracking to evaluate our combination algorithm.

We implement the proposed tracking approach and test it on a large number of video sequences. The evaluation database consists of two parts. Some of them come from the CAVIER dataset [9], with significant illumination change and frequent inter-human occlusion. The others are captured by us, containing situations like pose changing and occlusion by background.

4.1 Implementation Details

For pedestrian tracking, we adopt one generative color histogram model and one discriminative HOG model. HOG feature is robust to illumination change but is sensitive to pose change and dose not distinguish individuals. Color histogram is able to distinguish pedestrians with different clothes and also adapts to slight pose change, but it is seriously affected by illumination change. The natural complementary makes HOG feature and color histogram suitable for combination.

Our color histogram model is a 512 ($8 \times 8 \times 8$) dimensional vector in RGB space and updated online by IPCA [10]. Its likelihood is evaluated by Bhattacharyya coefficients [1]. The discriminative model is an offline SVM classifier [2], without any online update. Even with a discriminative model not updated online, our combination framework is proved to work effectively.

4.2 Results

Among 199 pedestrians in 86 video sequences (over 50,000 frames), after manually assigning the initial positions, our tracking algorithm successfully tracks 169 pedestrians, achieving a correction rate of 84.9%. Correction rates for only HOG model and only color histogram model are 67.4% and 60.8%.

Figure 3 is an example on sequence ThreePastShop2cor in CAVIAR dataset. There are some shops on the left hand side, introducing significant illumination change. Meanwhile, occlusion occurs frequently, further increasing the difficulty of tracking. HOG observation model tracks to a wrong pedestrian after occlusion (Fig. 3 (c) #475). Color histogram observation model leads to the drift problem (Fig. 3 (b) #475, #671) and lost its target after partial occlusion (Fig. 3 (b) #1173). Our algorithm (Fig. 3 (a)) maintains the advantage of both models and tracks to the end of the sequence. In addition, we compare our results to the recently proposed MIL tracker [11], the code for which is publicly available. Although they have developed an effective online updating algorithm, using only haar-like features result in their failure after occlusion (Fig. 3 (d)).

Figure 4 is an example of occlusion by background.

(a) The proposed algorithm with combined observation models

(b) Tracking results using only color histogram observation model

(c) Tracking results using only HOG observation model

(d) Tracking results by MIL Tracker [11]

Fig. 3 Tracking results in the CAVIAR ThreePastShop2cor video sequence. Figures with label '2x' means its size doubles the origin.

Fig. 4 Tracking results during partial occlusion.

This video is captured by us. When the pedestrian becomes occluded by the bushes, the confidence of both HOG channel and color histogram channel decrease. Nevertheless, color histogram is more reliable comparatively and can still be used for tracking.

5. Conclusion

In this paper, we propose a novel human tracking algorithm, which combines different observation models. Experimental results on pedestrians have validated its robustness to partial occlusion and its ability to maintain advantages. Although our experiments are based on the combinative usage of HOG model and color histogram model, our combination framework can be easily extended to other observation models.

References

[1] D. Comaniciu, V. Ramesh, and P. Meer, "Kernel-based object tracking," IEEE Trans. Pattern. Anal. Mach. Intell., vol.25, no.5, pp.564–577, 2003.

[2] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," Proc. IEEE Int. Conf. Comp. Vision Pattern Recognition, vol.1, pp.886–893, 2005.

[3] S. Avidan, "Support vector tracking," IEEE Trans. Pattern. Anal. Mach. Intell., vol.26, no.8, pp.1064–1072, 2004.

[4] H. Grabner and H. Bischof, "On-line boosting and vision," Proc. IEEE Int. Conf. Comp. Vision Pattern Recognition, vol.1, pp.17–22, 2006.

[5] Y. Li, H. Ai, T. Yamashita, S. Lao, and M. Kawade, "Tracking in low frame rate video: A cascade particle filter with discriminative observers of different lifespans," IEEE Trans. Pattern. Anal. Mach. Intell., vol.30, no.10, pp.1728–1740, 2008.

[6] Y. Lei, X. Ding, and S. Wang, "Visual tracker using sequential bayesian learning: Discriminative, generative, and hybrid," IEEE Trans. Syst. Man.& Cybern., vol.38, vol.6, pp.1578–1591, 2008.

[7] F. Tang, S. Brennan, Q. Zhao, and H. Tao, "Co-tracking using semi-supervised support vector machines," Proc. IEEE Int. Conf. Computer Vision, vol.1, pp.1–8, 2007.

[8] M.S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp, "A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking," IEEE Trans. Signal Process., vol.50, no.2, pp.174–188, 2002.

[9] CAVIAR, http://homepages.inf.ed.ac.uk/rbf/CAVIAR/

[10] Y. Li, "On incremental and robust subspace learning," Pattern Recognit., vol.37, no.7, pp.1509–1518, 2004.

[11] B. Babenko, M. Yang, and S. Belongie, "Visual tracking with online multiple instance learning," Proc. IEEE Int. Conf. Comp. Vision Pattern Recognition, pp.983–990, 2009.