2010 International Conference on Pattern Recognition

Multiple Human Tracking Based on Multi-View Upper-Body Detection and Discriminative Learning

Junliang Xing, Haizhou Ai
Department of Computer Science and Technology
Tsinghua University
Beijing 100084, China
Email: ahz@mail.tsinghua.edu.cn

Shihong Lao
Core Technology Center
Omron Corporation
Kyoto 619-0283, Japan
Email: lao@ari.ncl.omron.co.jp

***Abstract*—This paper focuses on the problem of tracking multiple humans in dense environments which is very challenging due to recurring occlusions between different humans. To cope with the difficulties it presents, an offline boosted multi-view upper-body detector is used to automatically initialize a new human trajectory and is capable of dealing with partial human occlusions. What is more, an online learning process is proposed to learn discriminative human observations, including discriminative interest points and color patches, to effectively track each human when even more occlusions occur. The offline and online observation models are neatly integrated into the particle filter framework to robustly track multiple highly interactive humans. Experiments results on CAVIAR dataset as well as many other challenging real-world cases demonstrate the effectiveness of the proposed method.**

***Keywords*-object tracking; object detection; discriminative learning; particle filter;**

I. Introduction

Multiple object tracking in video is of fundamental importance for many applications, such as visual surveillance, traffic safety monitoring, human computer interaction, etc. This could be an easy task when the objects are isolated from each other in a relatively clean background. However, real-world cases often go against this assumption by posing a complex background and serious occlusions among different objects.

To track multiple objects in complex situations, some early methods track motion blobs and regard each individual blob as one human [5], [11]. These methods usually assume the background is fixed and use background subtraction [7] to provide relatively robust object motion blobs. The foreground blob based methods are not discriminative and is likely to fail when the background changes suddenly. Recently, object detection researches have resulted in many promising detectors of particular object classes, e.g., faces [9] and humans or pedestrians [2], [10]. They can provide good observations for detection-based tracking algorithms. By applying object detectors into particle filter [4] framework, impressive results of tracking one single object have been achieved in [6]. In multiple object tracking, detection based methods suffer from the occlusion problem which prevents the detector collecting reliable observations. To cope with this problem, Wu et al. [10] use one full body detector and three part detectors to detect and track partially occluded humans. However, part detector is hard to train and still may fail when object part is not fully visible in more dense situations. What is more, employing more part detectors increases the computation load of the system proportionally. In this paper, only one part detector with a suitable size and discriminative power is used to search for partially occluded objects and a more powerful online discriminative learning process is proposed to deal with much more serious object occlusions.

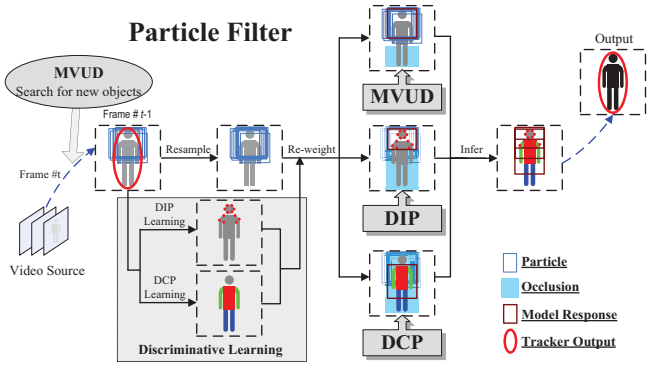

Figure 1. System overview.

The rest of this paper is organized as follows. Section 2 presents the proposed tracking algorithm by describing the learning process of its two components and the implementation in the particle filter framework. Section 3 gives the experimental results and Section 4 concludes the paper.

II. Our Approach

As shown in Figure 1, the proposed multiple human tracking algorithm learns two different kinds of observation models to track humans. The first kind is an offline learned multi-view upper-body detector (MVUD) while the other is online learned discriminative models (including discriminative interest point (DIP) and discriminative color patch (DCP)). These models are neatly coupled together in the particle filter framework to guide the tracking process under different occlusion situations.

1051-4651/10 $26.00 © 2010 IEEE
DOI 10.1109/ICPR.2010.420

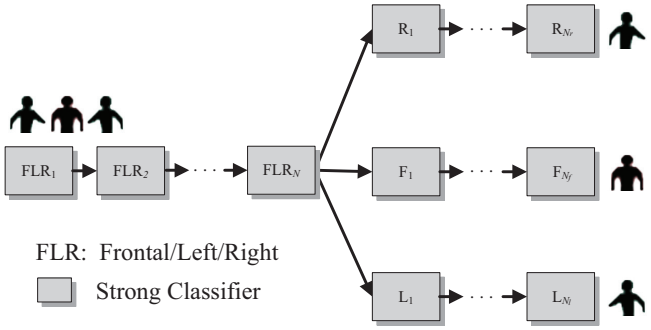

Figure 2. Tree structured multi-view human upper body detector.

A. Multi-View Upper-Body Detector

Part detectors have been proved to be an effective way to detect and track partially occluded objects [10]. Generally speaking, the smaller a part is, the larger probability it will be fully visible. From this prospective, a smaller object part has a larger traceability. However, smaller object part detector becomes harder to learn since it provides less information for learning. Although employing multiple part detectors of different size will remedy this problem [10], the computation cost will increase simultaneously. In this paper, we only train one part detector covering the upper-body area which is the most informative region of the human body. To deal with the view variances of the upper-body, the training samples are divided into three different views, i.e., frontal-rear, left profile and right profile, and trained using the method in [3]. The multi-view upper-body detector provides a very discriminative model. Figure 2 shows the structure of the detector. For details about the training process, please see [3].

B. Online Discriminative Learning

Although part detector could detect many partially occluded humans, it is likely to fail when more serious occlusions happen which prevent the part region from fully visible. What is more, the general human detector tends to drift when humans are close to each other due to its congenital deficiency at distinguishing different humans. To address these problems, an online learning process is proposed to effective collect the discriminative features of each human and be used to track a human under more serious dense situations.

During the online discriminative learning process, two different types of features are explored, the discriminative interest points and the discriminative color patch. The interest points are those have an expressive texture in their respective localities which provide the local information of one object and could be visible in very dense situation, while the color patch could be one salient image region (e.g. the clothes region) which provides the global information of one object and could be used to re-track the object after long time full occlusion.

(a) Learn discriminative interest points from candidate points.

(b) Learn discriminative color patch among different objects.

Figure 3. Discriminative learning process.

A DIP is assumed to have these properties: (1) it is an interest point and could be easily tracked; (2) it belongs to one object only; (3) and its motion coincides with the object's motion. To get a certain number of discriminative interest points that meet these requirements, we first generate a large pool of interest points using the KLT algorithm [8] within the bounding box of the object, and then learn the discriminative ones among them by filtering with the above properties in a greedy strategy. Denote the interest point set generated by KLT algorithm as $\mathbf{I} = \{I_l\}_{l=1}^{L}$ and the discriminative interest point set as $\mathbf{I}^d = \{I_l^d\}_{l=1}^{L_d}$. First, the learning process selects one interest point with the highest traceability (denoted as w_l which can be obtained by KLT algorithm) from $\mathbf{I}$, and then checks whether the location of the point lies on the object and its velocity has the same direction with the object it belongs. If the checking passed, this point is regarded as a discriminative point and added to $\mathbf{I}^d$; else it is removed from $\mathbf{I}$ and turns to the next point in $\mathbf{I}$ with the highest traceability. This process iterates until enough discriminative points have been selected (e.g., $L_d = 30$). Figure 3 (a) gives a typical case of the DIP learning process.

As for a DCP, it is supposed to have the ability to re-track one object after it has been fully occluded by other objects. So it should have a distinguished color distribution covering a relative small range of color values. Usually one object has multiple color modes and can be represented by a set of color patches as $\mathbf{C} = \{C_m\}_{m=1}^{M}$ and $\mathbf{C}^d = \{C_m^d\}_{m=1}^{M_d}$ corresponds to the discriminative color patch set. The DCM learning process learns $\mathbf{C}^d$ from $\mathbf{C}$ when the object is detected and isolated from other object and then updates $\mathbf{C}^d$ during the tracking process. For a human in most visual surveillance scenarios, the color patch covering the cloth region is most likely to be the discriminative one that differs from other humans. Figure 3 (b) gives a typical case of the DCP learning process.

C. Particle Filter Implementation

We couple the offline trained multi-view upper-body detector and the online learned DIP and DCP model in the particle filter [4] framework which has been widely used in object tracking. Denoting the object state sequence as $\mathbf{s}_{1:t} = \{\mathbf{s}_1, \ldots, \mathbf{s}_t\}$ and the observation sequence as $\mathbf{o}_{1:t} = \{\mathbf{o}_1, \ldots, \mathbf{o}_t\}$, object tracking is formalized as a sequential Bayesian estimation problem by a two-step recursion of Prediction (P) and Update (U):

$$\mathrm{P}: p(\mathbf{s}_t|\mathbf{o}_{1:t-1}) = \int D(\mathbf{s}_t|\mathbf{s}_{t-1}) p(\mathbf{s}_{t-1}|\mathbf{o}_{1:t-1}) d\mathbf{s}_{t-1} \quad (1)$$

$$\mathrm{U}: p(\mathbf{s}_t|\mathbf{o}_{1:t}) \propto L(\mathbf{o}_t|\mathbf{s}_t) p(\mathbf{s}_t|\mathbf{o}_{1:t-1}), \quad (2)$$

where $D(\mathbf{s}_t|\mathbf{s}_{t-1})$ is the dynamic model and $L(\mathbf{o}_t|\mathbf{s}_t)$ is the observation model that gives a likelihood of one observation in the state space. The filter distribution $p(\mathbf{s}_t|\mathbf{o}_{1:t})$ usually is complicated which leads to analytical intractability, particle filter provides a neat way to approximate it by a set of weighted particles:

$$p(\mathbf{s}_t|\mathbf{o}_{1:t}) = \sum_{n=1}^{N} \pi_t^n \delta_{\mathbf{s}_t^n}(\mathbf{s}_t), \quad (3)$$

where N is the number of particles and $\delta_s(\cdot)$ denotes the delta-Dirac function at position s.

In dense environments, it is hard to predict and update the object state using only one observation model since the recurring occlusions often fail the observation model (e.g. detector model often drifts when two objects are close to each other). In this paper, we are more confident to overcome this problem because we have got several observation models at hand, including one offline learned detector model and the online learned DIP and DCP models, which could deal with different occlusion situations. Represent the three models as $L_{\mathrm{D}}(\mathbf{o}_t|\mathbf{s}_t)$, $L_{\mathrm{I}}(\mathbf{o}_t|\mathbf{s}_t)$ and $L_{\mathrm{C}}(\mathbf{o}_t|\mathbf{s}_t)$ respectively, the tracking algorithm dynamically employs the suitable model according to the occlusion state of the object.

To find the occlusion status of one object, a visible score is calculated for each object which is defined as the quotient between the number of visible pixels and the number of total pixels within the elliptical object bounding box. When two objects are overlapped, the visible one is decided by the DCP model by calculating the histogram distance between the occlusion region and DCP model. Based on the object visible score, the algorithm switches to the best observation model to track it: if the upper-body region is visible, the MVUD model is used to track the object; if the upper-body region is not visible but a certain number of interest points are visible, the DIP model is used to track the object; if the upper-body region is not visible and not enough interest points are visible, the DCP model is used to track the object.

In the particle filter framework, the observation model needs to give a confidence reflecting the human likelihood when evaluating a particle. Considering the characteristic of a boosted object detector, the likelihood given by the detector is calculated as:

$$L_{\mathrm{D}}(\mathbf{o}_t|\mathbf{s}_t) = \frac{\eta \exp(S)}{\exp(L_T - L_P)}, \quad (4)$$

where L_T is total layer number of the cascade detector, L_P is the layer number the observation passed, S is sum of the marginal distance when passing one layer, and η is a normalization factor which makes the likelihood to be a distribution. For the DIP model, the human likelihood is modeled by calculating the weighted track ratio of the interest points, which can be represented as:

$$L_{\mathrm{I}}(\mathbf{o}_t|\mathbf{s}_t) = \frac{\sum_{l=1}^{L_d} w_l I_l^d}{\sum_{l=1}^{L} w_l I_l}. \quad (5)$$

And for the DCP model, the human likelihood is modeled as the Bhattacharyya coefficient between the particle region and DCP region:

$$L_{\mathrm{C}}(\mathbf{o}_t|\mathbf{s}_t) = \sum_{b=1}^{B} H(b) H'(b), \quad (6)$$

where B is the bin number of the histogram, $H(b)$ is the b-th bin value of histogram in the particle region and $H'(b)$ is the b-th bin value in the DCP region.

Table I gives the overall flowchart of the proposed tracking algorithm.

Table I
HUMAN TRACKING ALGORITHM.

For each tracked object with the particle set $\{\mathbf{s}_{t-1}^{(n)}, \pi_{t-1}^{(n)}\}_{n=1}^{N}$ at the previous time step $t-1$, proceed at time t:

- Resample: simulate $\alpha_n \sim \{\pi_{t-1}^{(n)}\}_{n=1}^{N}$, and replace $\{\mathbf{s}_{t-1}^{(n)}, \pi_{t-1}^{(n)}\}_{n=1}^{N}$ with $\{\mathbf{s}_t^{(\alpha_n)}, 1/N\}_{n=1}^{N}$
- Predict: $\mathbf{s}_t^{(n)} \sim D(\mathbf{s}_t|\mathbf{s}_{t-1}^{(n)})$
- Update: set particle weight according to object occlusion state:
  - If the MVUD region is visible, $\pi_t^{(n)} = L_{\mathrm{D}}(\mathbf{o}_t|\mathbf{s}_t)$
  - Else If enough points are visible, $\pi_t^{(n)} = L_{\mathrm{I}}(\mathbf{o}_t|\mathbf{s}_t)$
  - Else, $\pi_t^{(n)} = L_{\mathrm{C}}(\mathbf{o}_t|\mathbf{s}_t)$
- Output: $\hat{\mathbf{s}}_t \leftarrow \sum_{i=1}^{N} \pi_t^{(n)} \cdot \mathbf{s}_t^{(n)}$

III. EXPERIMENTS

Experiments are carried out on a public dataset CAVIAR [1] and some more challenging real-world video data collected with a hand-held camera.

A. Experiment Settings

The multi-view upper body detector is trained from 7504 front-rear, 6986 left profile and 6986 right profile samples with the normalized size 24×24. For the DIP model, 50 best KLT interest points per each object are detected for discriminative learning and if less than 5 features are visible, the DIP model is stopped. And for the DCP model, the color patch is represented by a $32 \times 32 \times 32$ color histogram in RGB color space.

Table II
TRACKING COMPARISON ON CAVIAR.

Algorithm	GT	MT	ML	Frmt	FAT	IDS
Wu et al.[10]	189	140	8	40	4	19
Proposed	189	152	6	37	6	16

GT: ground-truth; MT: mostly tracked; ML: mostly lost; Fgmt: trajectory fragment; FAT: false alarm trajectory; IDS: ID switch.

B. Evaluation Metrics

We adopt the same metrics for evaluating tracking performance as in [10] which are defined as:

– Number of "mostly tracked" trajectories (more than 80% of the trajectory is tracked);
– Number of "mostly lost" trajectories (more than 80% of the trajectory is lost);
– Number of "fragments" trajectories (a result trajectory which is less than 80% of a ground-truth trajectory);
– Number of "false trajectories" (a result trajectory corresponding to no real object);
– The frequency of "identity switches" (identity exchanges between a pair of result trajectories).

C. Results

The CAVIAR dataset consists of 26 sequences with overall 36,292 frames in the size 384×288. The sequences contain intensive inter-object occlusion and frequent interactions between humans. We evaluate our algorithm and compare it with the method in [10]. Table II gives the comparison results from which we can see that our algorithm obtain an improvement on most of the metrics yet only employing one part detector. Some typical tracking results on the sequence *OneStopMoveEnter1cor.mpg* are showed in Figure 4 (a).

The real-world videos contain very complex background with serious occlusions between different objects which are much more complex than those in CAVIAR dataset. As examples, some tracking results on two typical sequences are shown in Figure 4 (b).

IV. CONCLUSION

In this paper, we propose a robust multiple occluded human tracking algorithm in common visual surveillance environments. Observations collected from both an offline boosted multi-view upper-body detector and online learned discriminative features are tightly integrated into the particle filter framework to track humans of different occlusion degrees. Experiment results on public dataset and challenging real-world video data demonstrate the effectiveness of our method. Future work could be focused on mining more discriminative features, e.g. object texture and motion features, to increase the robustness and adaptability of the system.

(a) Tracking results on CAVIAR sequence *OneStopMoveEnter1cor.mpg*.

(b) Tracking results on two real-world video sequences.

Figure 4. Typical tracking results. Points: DIP model; rectangle: DCP model; ellipse: final result (zoom in for a better view).

ACKNOWLEDGMENT

This work is supported in part by National Basic Research Program of China (2006CB303102), Beijing Educational Committee Program (YB20081000303), and it is also supported by a grant from Omron Corporation.

REFERENCES

[1] CAVIAR dataset. http://homepages.inf.ed.ac.uk/rbf/CAVIAR/.

[2] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005.

[3] C. Hou, H. Ai, and S. Lao. Multiview pedestrian detection based on vector boosting. In *ACCV*, 2007.

[4] M. Isard and A. Blake. Condensation – conditional density propagation for visual tracking. *IJCV*, 28(1):5–28, 1998.

[5] M. Isard and J. MacCormick. Bramble: a bayesian multiple-blob tracker. In *ICCV*, 2001.

[6] 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. In *CVPR*, 2007.

[7] C. Stauffer and W. E. L. Grimson. Learning patterns of activity using real-time tracking. *PAMI*, 22(8):747–757, 2000.

[8] C. Tomasi and T. Kanade. Detection and tracking of point features. Technical Report CMU-CS-91-132, Carnegie Mellon University, 1991.

[9] P. Viola and M. J. Jones. Robust real-time face detection. *IJCV*, 57(2):137–154, 2004.

[10] B. Wu and R. Nevatia. Detection and tracking of multiple, partially occluded humans by bayesian combination of edgelet based part detectors. *IJCV*, 75(2):247–266, 2007.

[11] T. Zhao and R. Nevatia. Tracking multiple humans in complex situations. *PAMI*, 26(9):1208–1221, 2004.