IEICE TRANS. INF. & SYST., VOL.E99–D, NO.4 APRIL 2016

LETTER

Object Tracking with Embedded Deformable Parts in Dynamic Conditional Random Fields*

Suofei ZHANG[†a)], **Zhixin SUN**[†], **Xu CHENG**[††], ***Nonmembers***, ***and*** **Lin ZHOU**[††], ***Member***

SUMMARY This work presents an object tracking framework which is based on integration of Deformable Part based Models (DPMs) and Dynamic Conditional Random Fields (DCRF). In this framework, we propose a DCRF based novel way to track an object and its details on multiple resolutions simultaneously. Meanwhile, we tackle drastic variations in target appearance such as pose, view, scale and illumination changes with DPMs. To embed DPMs into DCRF, we design specific temporal potential functions between vertices by explicitly formulating deformation and partial occlusion respectively. Furthermore, temporal transition functions between mixture models bring higher robustness to perspective and pose changes. To evaluate the efficacy of our proposed method, quantitative tests on six challenging video sequences are conducted and the results are analyzed. Experimental results indicate that the method effectively addresses serious problems in object tracking and performs favorably against state-of-the-art trackers.

key words: *visual tracking, conditional random field, deformable part based model, graph model*

1. Introduction

In recent years, visual tracking algorithms have been extensively deployed in various intelligent video surveillance systems. Such a system depends on an object tracking method to trace the position, size and other related values of the target of interest in an efficient way. The resulting information plays a critical role for higher level understanding of video contents like in traffic surveillance, activity analysis, etc. Although many state-of-the-art techniques for visual tracking have been developed [1], it still remains a challenging task for many reasons including viewpoint change, deformation, partial occlusion and cluttered background, etc.

The objective of this study is to build an object tracking framework to handle aforementioned problems simultaneously. For deformation and viewpoint change, our work is inspired by an essential intuition that object tracking by human eyes actually follows the recognition of the target at the first glimpse. The leverage of massive experience in this process brings abundant high-level auxiliary knowledge to handle various problems in tracking. Thus we introduce high performance object detection models, DPMs [2], [3], into our object tracking framework. DPMs exploit Histogram of Oriented Gradients (HOG) features on multiple resolutions to describe object and its details. HOG features can handle scale and illumination variation in practical scenes effectively. On top of that, DPMs consist of a mixture of components to describe different poses and perspectives of target. Tracking target by object recognition and detection methods has been proven as a promising way by researchers [4], [5]. Differing from previous work, our work attempts to introduce prior knowledge of target at the beginning of tracking, while others focus on updating the target model iteratively by on-line learning methods.

To embed DPMs into our tracking frameworks, we treat DPMs as a mixture of star-shaped CRFs as components corresponding to specific poses or views of object. For every component, a DCRF [6] over consecutive frames is composed to predict new position given the last tracking result as prior knowledge. The vertex in this graph model corresponds to a deformable part of object. Pre-defined pairwise potential functions between vertices formulate spatial and temporal deformation of object. Instead of calculating energy of entire graph directly, we successfully restrict the computation into each vertex separately by evaluating the lower bound of energy at each site on current frame.

For partial occlusion in cluttered background, a series of solutions attempt a sparse representation of objects [7], [8] to track parts of target. Differing from these decomposition based methods, our method explicitly describes object as a root part with outline and terminal parts with specific details. When a part is absent from sight, a logistic regression based voting method allows other observed parts still contribute to the final hypothesis as well.

Based on our previous work [9], the main contributions of this work are threefold: (1) we propose some novel potential functions to embed high performance DPMs into DCRF in an efficient way; (2) we propose temporal transition functions between components of DPMs to prompt the robustness and accuracy of our tracker; (3) we implement an unparameterized logistic regression based voting mechanism to handle partial occlusion during tracking. Finally, experiments on challenging video sequences prove the efficacy of our proposed method.

Manuscript received June 22, 2015.
Manuscript revised November 15, 2015.
Manuscript publicized January 19, 2016.

[†]The authors are with the Key Laboratory of Broadband Wireless Communication and Sensor Network Technology, Nanjing University of Posts and Telecommunications, Ministry of Education, Nanjing 210003, China.
[††]The authors are with the School of Information Science and Engineering, Southeast University, Nanjing, Jiangsu, 210096, China.

*This work was supported by the Chinese National Natural Science Foundation (Grant No. 61571106, 61170276, 61373135) and Scientific Research Foundation of NJUPT (No. NY213102).

a) E-mail: zhangsuofei@njupt.edu.cn (Corresponding author)

DOI: 10.1587/transinf.2015EDL8139

2. Proposed Tracking Framework

The main flow of our tracking framework is illustrated in Fig. 1. Taking current frame as input to our system, HOG feature over multiple resolutions is extracted from image. Filtered by DPMs filter, a score of hypothesis that interesting part appears at current location can be obtained on each scale. Treated as unary vertex potential function in DCRF framework, the score is regularized by our proposed pairwise potential functions, i.e., previous tracking results as prior knowledge. On top of that, a logistic regression based voting method is used to handle partial occlusion of object, resulting the score of each component in DPMs. Finally, to further prompt the robustness of our system, we propose a component transition function here to re-score each component.

2.1 Embedded DPMs in DCRF

DPMs have been proven as quite effective model to handle challenging object detection tasks. Each component of DPMs can be considered as a star-shaped CRF [10] with dynamic programming based deformation penalty as spatial potential function on current frame. To utilize prior knowledge from previous frames in predicting the status $s_t(x)$ of part x at t-th frame, we propose a novel temporal potential function $V_x(s_{t+1}(x)|s_t(y))$ between corresponding parts over frames to model the consistency between them.

$$V_x(s_{t+1}(x)|s_t(y)) = \mathcal{G}(x-y;\Sigma)\cdot\delta(s_{t+1}(x), s_t(y)) + \frac{1}{1+e^{-\|v_x-v_y\|^2}}(1-\delta(s_{t+1}(x), s_t(y))). \quad (1)$$

Here part x, y are connected parts on different frames, $\delta(\cdot)$ is the Kronecker delta function and $\|\cdot\|$ is the Euclidian distance. If the part x is assumed to be observed at last frame, a three-dimensional normalized Gaussian kernel $\mathcal{G}(x-y;\Sigma)$ is adopted to measure the motions of object. Otherwise if the part is assumed to be occluded, which means the direct prior knowledge about current part from last frame is absent, we keep the temporal connectivity with the difference between part deformation v_x and v_y instead.

Given spatial and temporal pairwise potential functions, we can embed DPMs into DCRF and trace CRFs over consecutive frames. Given observation $o_{1:t+1}$, we can approximate the probability that component s_{t+1} presents by the lower bound of graph energy as

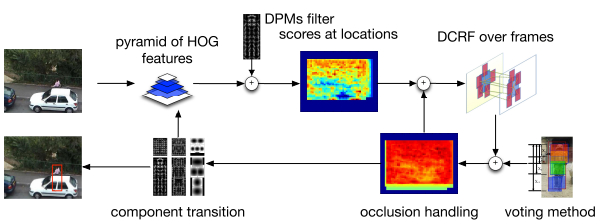

Fig. 1 The workflow of our proposed tracking framework.

$$\prod_{x\in X}\exp\left\{\left[V_x(o_{t+1}(x)|s_{t+1}(x)) + \sum_{y\in N_x} V_{x,y}(s_{t+1}(x), s_{t+1}(y))\right] \cdot \sum_{y\in M_x}\sum_{s_t(y)} V_x(s_{t+1}(x)|s_t(y)p(s_t(y)|o_{1:t}(y)))\right\}. \quad (2)$$

Here $V_{x,y}(\cdot)$ is the spatial pairwise potential function corresponding to deformation in DPMs, N_x means spatial neighbors of x and M_x means temporal neighbors of x. Note that the summation of unary potential $V_x(o_{t+1}(x)|s_{t+1}(x))$ and pairwise spatial potential $V_{x,y}(\cdot)$ in Eq. (2) corresponds to the output of DPMs score on each vertex. Thus the approximation restricts the computation of $p(s_{t+1}|o_{1:t+1})$ into each vertex separately and implements a quite efficient tracking framework based on outputs of DPMs.

2.2 Occlusion Handling

For partial occlusion problem, as shown in Fig. 1, a voting method is implemented based on the vertex result in Eq. (2). Assuming an object is partially occluded, instead of aggregating the scores of all parts $X = x_0, \ldots, x_n$ as in Eq. (2), we attempt to select an optimal subset $X_c^* = \{x_k, \ldots, x_l\}$ from X to maximize the mean of normalized scores of vertices as

$$\psi(X_c^*) = \max_{X_c}\frac{1}{|X_c|}\sum_{j\in X_c} p'(s(x_j)). \quad (3)$$

where $p'(s(x_j))$ is unparameterized logistic regression result of vertex score in Eq. (2), $|X_c|$ means the number of vertices in set X_c. For generic object tracking tasks, a simple greedy search strategy can be employed to add parts into X_c sequentially. The voting method has also been proven empirically as an efficient mechanism with acceptable results as in [11].

2.3 Component Transition

In our experiments we notice that simply taking position of optimal component with highest score as tracking result leads to frequent jitters among components. To prompt the robustness of our system, we propose a component transition function to re-weight the score $\psi(X^*_{c_{t+1}})$ of each component and impose the consistency of component over frames.

$$\phi(c_{t+1}) = \psi(X^*_{c_{t+1}})\sum_{c_t}\phi(c_t)[\alpha_1\delta(c_{t+1}, c_t) + \alpha_2(1-\delta(c_{t+1}, c_t))] \quad (4)$$

Here parameters α_1, α_2 are tuned according to specific scene. This simple transition mechanism yields small but noticeable improvement to the stability as well as accuracy of our system.

3. Experimental Results

In this section, we progressively evaluate the performance of our method with different configurations and compare the final proposed tracker with existing trackers on various tasks. In experiments, we initialize the algorithm with

DPMs trained for PASCAL VOC 2009 [3], which contain six components consisting of nine deformable parts. All experiments are carried out in MATLAB with Intel Core i5 Duo 2.93GHz CPU environment. Since most of computation of our method is spent on HOG features extraction, we implement an adaptive HOG feature extraction mechanism as shown in Fig. 1. The mechanism depends on the scale of previous results, only extracting related features in the pyramid. Therefore the computational cost of our method depends on the ratio of target size and background size rather than absolute size of images in video sequence. For the majority of our experiments, the tracker can process one frame in 0.4s. It is relatively much faster than detecting object with DPMs directly (2.5 second per frame).

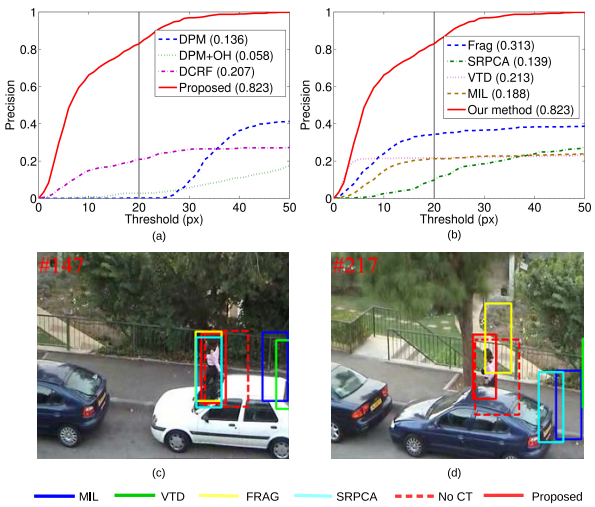

Fig. 2 Quantitative and qualitative evaluations of different methods on "Woman" sequence: (a) performances of our method with different configurations, (b) comparison of other leading methods and our proposed method, (c) and (d) are tracking results of different methods at frame #147 and #217 respectively.

First, we empirically analyze the influence of each proposed mechanism according to tracking results on a challenging video sequence, the "Woman" sequence [12]. Here four different configurations are taken into consideration: detection by DPMs directly (DPM), detection by DPMs and occlusion handling (DPM+OH), tracking by DCRF merely with Gaussian kernel in Eq. (1) (DCRF), and tracking with complete temporal potential function and component transition (Proposed). Since there is no tracking failure problem for detection methods, we follow the evaluation protocol proposed by [4] in Fig. 2 (a) with center errors.

It is easy to observe in Fig. 2 (a) that the proposed tracking method brings significant improvement to DPMs based detection. Note that adding unparameterized occlusion handling to DPMs directly actually leads to worse result. It agrees with the illustration in Fig. 3. The voting method helps observed parts to contribute as well as the whole entirety of target, thus introducing extra noises into detection results. However these noises can be filtered out by prior knowledge in terms of temporal potentials in DCRF so that more robust tracking results are ensured. On the other hand, tracking with DCRF without occlusion handling achieved a desirable result at the beginning of the sequence. However the method failed to follow the target around frame #125, where a long-term partial occlusion occurs, finally leading

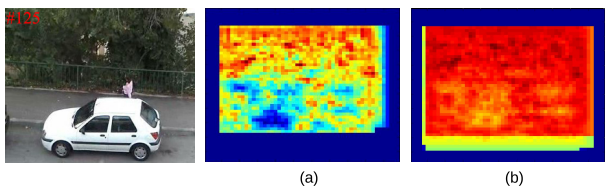

Fig. 3 Resulting heat maps of different detection methods at frame #125 of "Woman" sequence. The red part represents higher possibility of hypothesis that the target occurs at current position. (a) heat map of DPMs, (b) heat map of DPMs with occlusion handling.

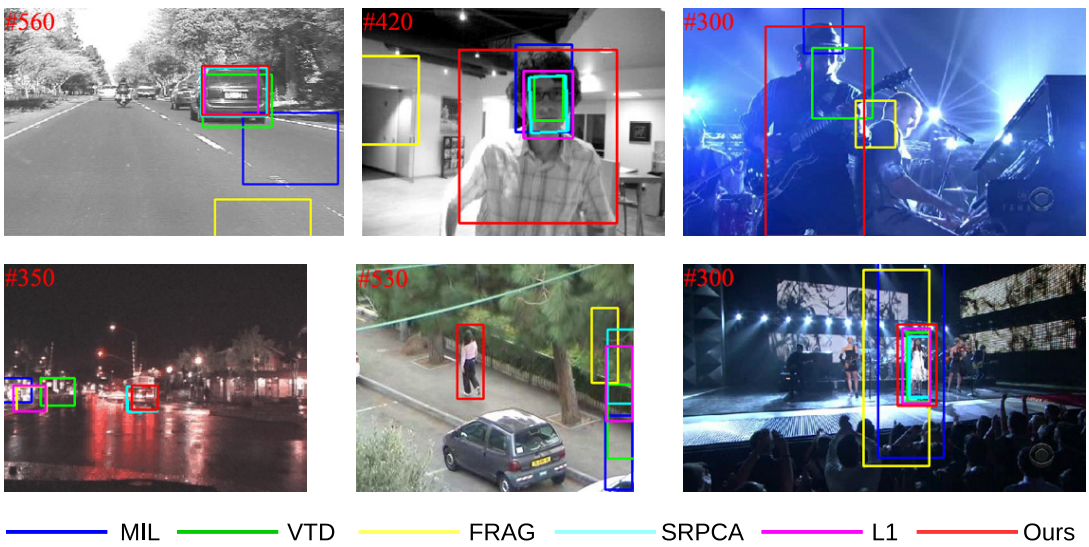

Fig. 4 Sample tracking results on different tasks.

Table 1 Average center error (pixels) for various methods.

Video Clip	MIL	VTD	Frag	SRPCA	L1	proposed
Woman	122.4	186.6	92.6	120.2	-	14.9
Car 4	60.1	12.3	179.8	3.0	4.1	8.6
Car 11	43.5	27.1	63.9	2.3	33.3	2.4
Singer 1	15.2	4.1	22.0	4.7	4.6	4.2
David Indoor	16.1	13.6	76.7	9.2	7.6	49.8
Shaking	11.2	6.1	85.7	-	103.7	102.6
Average	44.8	41.6	86.8	27.9	30.7	30.4

to mitigated result.

We also show quantitative (Fig. 2 (b)) and qualitative (Fig. 2(c-d)) comparison of various methods on "Woman" sequence in Fig. 2. Here four representative tracking methods are considered, i.e., MIL [4], VTD [13], Frag [12], and SRPCA [7]. It can be seen that due to serious partial occlusion, the MIL and VTD trackers fail to trace target around frame #147. Moreover, during frame #147 to #217, there exists an obvious size variation of target. Therefore the FRAG and SRPCA trackers drift from target as well. DPMs based method shows high robustness to practical challenges and is the sole method which can trace target during the whole sequence. Note that the dashed rectangles (No CT) in Fig. 2(c-d) represent tracking results of our method without component transition mechanism described in Sect. 2.3. The jitters occur here since another component with incorrect shape obtains highest score at these frames. The component transition effectively imposes the component change between frames and prompts the robustness of tracking.

Second, to evaluate the performance of our proposed tracking framework, we compare various leading methods on six tracking tasks. Here we further take L1 [14] tracker into consideration. In Fig. 4, we illustrate some key results of various methods on different sequences. One can see that the proposed method simultaneously tackles challenging problems such as severe occlusion (Woman, Shaking), perspective change (Woman, Singer 1), illumination variation (Car 11, Singer 1, Shaking) and scale variation (Car 4, David Indoor, etc.). Note that for David Indoor and Shacking sequences, our tracker attempts to trace entirety of human body rather than face as other trackers do. This is due to the utilization of pedestrian model from DPMs.

We also list average center errors of various methods in Table 1. The results of our tracker on David Indoor and Shacking sequences are calculated by the distance from center of our results to center of ground truth face data as well. Although our proposed algorithm successfully traces target during the whole entirety of video sequence as shown in Fig. 4, the modification of criterion leads to mitigation of statistical results. However, one can still see from the comparison of average results that our proposed method performs favorably against other state-of-the-art trackers.

4. Conclusion

In this paper, a robust tracking approach is proposed by embedding DPMs into DCRF. We employ high-performance DPMs to model the appearance of object. By introducing abundant prior knowledge into tracking, we can handle various challenging problems in practical scenes. Also, an occlusion handling mechanism is proposed to be complementary to DPMs during tracking. Our proposed tracker represents promising efficacy and efficiency in different tracking tasks.

References

[1] M. Isard and A. Blake, "Condensation-conditional density propagation for visual tracking," International Journal of Computer Vision, vol.29, no.1, pp.5–28, 1998.

[2] P. Felzenszwalb, D. McAllester, and D. Ramanan, "A discriminatively trained, multiscale, deformable part model," 2008 IEEE Conference on Computer Vision and Pattern Recognition, pp.1–8, June 2008.

[3] P.F. Felzenszwalb, R.B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," IEEE Trans. Pattern Anal. Mach. Intell., vol.32, no.9, pp.1627–1645, Sept. 2010.

[4] B. Babenko, M.-H. Yang, and S. Belongie, "Robust object tracking with online multiple instance learning," IEEE Trans. Pattern Anal. Mach. Intell., vol.33, no.8, pp.1619–1632, 2011.

[5] X. Cheng, N. Li, T. Zhou, L. Zhou, and Z. Wu, "Robust superpixel tracking with weighted multiple-instance learning," IEICE Trans. Inf. & Syst., vol.E98, no.4, pp.980–984, 2015.

[6] Y. Wang, K.-F. Loe, and J.-K. Wu, "A dynamic conditional random field model for foreground and shadow segmentation," IEEE Trans. Pattern Anal. Mach. Intell., vol.28, no.2, pp.279–289, Feb. 2006.

[7] D. Wang, H. Lu, and M.-H. Yang, "Online object tracking with sparse prototypes," IEEE Trans. Image Process., vol.22, no.1, pp.314–325, 2013.

[8] X. Mei and H. Ling, "Robust visual tracking and vehicle classification via sparse representation," IEEE Trans. Pattern Anal. Mach. Intell., vol.33, no.11, pp.2259–2272, 2011.

[9] S. Zhang, X. Cheng, H. Guo, L. Zhou, and Z. Wu, "Tracking deformable parts via dynamic conditional random fields," 2014 IEEE International Conference on Image Processing (ICIP), pp.476–480, Oct. 2014.

[10] J. Lafferty, A. McCallum, and F.C. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," 2001.

[11] G. Shu, A. Dehghan, O. Oreifej, E. Hand, and M. Shah, "Part-based multiple-person tracking with partial occlusion handling," 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.1815–1821, 2012.

[12] A. Adam, E. Rivlin, and I. Shimshoni, "Robust fragments-based tracking using the integral histogram," 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp.798–805, 2006.

[13] J. Kwon and K.M. Lee, "Visual tracking decomposition," 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.1269–1276, June 2010.

[14] C. Bao, Y. Wu, H. Ling, and H. Ji, "Real time robust l1 tracker using accelerated proximal gradient approach," 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.1830–1837, June 2012.