DEPTH-WEIGHTED CORRELATION METHOD FOR VISUAL TRACKING WITH OCCLUSION DETECTION

Chenghao Li[1], *Yuan Zhou*[1,2,*], *Bo Cui*[3], *and Chunping Hou*[1]

[1] School of Electrical and Information Engineering, Tianjin University, Tianjin, China
[2] Electrical Engineering Department, Princeton University, Princeton, USA
[3] Institute of Automation, Chinese Academy of Science, Beijing, China
Email: zhouyuan@tju.edu.cn

ABSTRACT

Despite the significant progress, it remains a challenging task for a tracker to distinguish a target from the background when the target is occluded. In this paper, we propose a new tracking method, named as depth-weighted correlation method(DWCM), to handle heavy occlusion. The proposed method uses depth cues as the weights of candidate objects and applies the framework of spatio-temporal context (STC). We also propose a scale update scheme for DWCM, so as to obtain an appropriate scale for the target. Encouraging experimental results show that the proposed tracker obtains state-of-the-art results and handles occlusion better than competing tracking methods.

Index Terms— Visual tracking, occlusion,spatio-temporal context, SVM, spatial weighted map

1. INTRODUCTION

Visual object tracking is one of the core problems in computer vision, with wide-ranging applications such as surveillance, security and auto-control systems. In recent decades, to overcome the challenging problem such as occlusion and scale variation, numerous tracking methods have been proposed in literatures, which can be categorized into generative and discriminative approaches.

Generative methods [1, 2, 3] learn an appearance model to represent the tracked object and search for regions which are the most similar to the object as the predicted target. Different from generative trackers, discriminative approaches [4, 5, 6, 7] formulate object tracking as a classification problem, which aims to find the optimal target location that can best distinguish the target from its background. The methods are relatively more robust in suppressing background clutters than generative methods, but discriminative methods are sensitive to occlusion, leading to tracking failure.

A possible way to deal this occlusion problem in discriminative methods is to utilize depth cue. As an inevitable challenging problem in visual tracking, occlusion could be handled by depth cue to some extent [8]. Hence, the depth cue could improve the tracking performance for the discriminative methods. There exists several tracking methods considering depth cue. In [9], depth cue was fused into the superpixel-based target estimation for robust tracking, however, [9] didn't exploit depth cue for occlusion handling. Gao et al. [10] proposed a layered graph model in image and depth domains for multi-pedestrian tracking. Song and Xiao [11] presented a RGBD tracking method using off the shelf depth sensors.

In this paper, we propose a new visual object tracker based on depth cue, so as to handle occlusion and other challenging factors. The key contributions of the proposed algorithm are summarized as three-fold: (1) We improve the support vector machine (SVM [12]) tracker by including depth cue in its histograms of oriented gradients (HOG) [13] feature, and adopt the improved SVM tracker as the base tracker. (2) When occlusion occurs, the HOG features of tracking target will vary significantly, leading to false classification and therefore tracking failure. A new method is presented to handle heavy occlusion when the occlusion is detected by the proposed occlusion detection mechanism. The method is named as "depth-weighted correlation method"(DWCM), which uses depth cues as the weights of candidate objects and applies the framework of spatio-temporal context (STC) [14]. (3) We also propose a novel scale update scheme for DWCM, so as to obtain an appropriate scale for the target.

2. THE PROPOSED METHOD

2.1. Occlusion detection mechanism

In visual tracking, occlusion is a large challenge for robust tracking. It makes the object appearance change significantly, leading to a tracker losing the target with high probability. Correspondingly, the HOG features of tracking target vary significantly. If the classifier is updated directly when the target is under occlusion, there will be drifting problem, leading to tracking failure. Here we propose an effective occlusion handling mechanism which could actively detects occlusion. To detect the occlusion, we assume that the target

is the closest object in the target bounding box when there is not occlusion. A new object in front of the target inside the bounding box indicates the beginning of occlusion state. Correspondingly, depth histogram inside bounding box is expected to have a newly rising peak with a smaller depth value than target, and a reduction of the number of bins in the histogram around the target depth. Therefore, we can consider it as the direct evidence to detect whether or not the occlusion occurs. If the depth in the $(t-1)-th$ frame has a large change than the depth in the $t-th$ frame, the probability of occlusion occurring is high. If the change is inappreciable, we believe that there is no occlusion in the current frame. The variation of the depth between $(t-1)-th$ frame and $t-th$ frame can be approximated as a Gaussian distribution: $(d_t - d_{t-1}) \sim N(\mu_t, \sigma_t^2)$. We can detect whether or not the occlusion occurs on this basis as:

$$\begin{aligned} O_t^H &= Max(V_t \cdot |\frac{\mu_t - \sigma_t}{\mu_{t-1} - \sigma_{t-1}}|, V_t \cdot |\frac{\mu_{t-1} - \sigma_{t-1}}{\mu_t - \sigma_t}|) \\ V_t &= \frac{\sum_{x,y}^{\mu_t - \sigma_t} \left|(d_t^{x,y} - d_{t-1}^{x,y})\right|}{\sum_{x,y} \left|(d_t^{x,y} - d_{t-1}^{x,y})\right|} \end{aligned} \tag{1}$$

where $d_t^{x,y}$ is the depth value of a pixel located at coordinates (x, y) in the $t-th$ frame, $\sum_{x,y} \left|(d_t^{x,y} - d_{t-1}^{x,y})\right|$ represents the sum of all the difference of depth between $(t-1)-th$ frame and $t-th$ frame, $\mu_t - \sigma_t$ is considered as a threshold, $\sum_{x,y}^{\mu_t - \sigma_t} \left|(d_t^{x,y} - d_{t-1}^{x,y})\right|$ represents the sum of the difference that the value of it is smaller than $\mu_t - \sigma_t$, $|\frac{\mu_t - \sigma_t}{\mu_{t-1} - \sigma_{t-1}}|$ reflects the variation of the depth's statistical properties between $(t-1)-th$ frame and $t-th$ frame, and we believe that the occlusion occurs when the value of it is big enough. A larger O_i^H indicates that an occlusion is more likely.

When the target moves towards or away from the camera, there is some change and the area of the target will be small in consecutive frame, so the value of O_i^H is increased, implying a high probability of occlusion occurring. In fact, there is no occlusion. To address this problem, we introduce the target candidate's confidence into occlusion detection, which can help to reduce ambiguities with depth information. The target candidate's confidence value T_t trained by SVM will be small when occlusion occurs in a frame. T_t has a comparatively higher value when there is no occlusion. We employ a simple but efficient approach to combined depth information and T_t to detect occlusion, that is, we define a likelihood of occlusion in this frame as:

$$\begin{aligned} O_t^T &= \eta_t * e^{(1-T_t)} * O_t^H \\ \eta_t &= \begin{cases} 1, & if \;\; T_t < T_h \\ 0, & if \;\; T_t > T_h \end{cases} \end{aligned} \tag{2}$$

T_h is a threshold for T_t to indicate whether the target candidate is reliable enough. A larger value of O_t for the estimated occlusion means that the occlusion is more likely to occur. It is considered to occur occlusion when the O_t^T is large enough, and the DWCM is proposed to sovle the tracking problem.

2.2. Depth-Weighted Correlation Method

In visual tracking, the position of the tracked target in the next frame is usually close to the position of the tracked target in the current frame. The closer the context location is to the currently tracked target central location, the more important it is to predict the object location in the coming frame, and a larger weight should be set to the context. Furthermore, the depth map is designed for suppressing the background. Then, we generate a spatial weighted map for weighted correlation between target and its surrounding background in the t-th frame:

$$W_t = e^{-\frac{|(x,y)-(x_t^*, y_t^*)|^2}{(\sigma_t^s)^2}} B \tag{3}$$

where σ_t^s denotes a scale parameter, (x_t^*, y_t^*) represents the center location of tracking target, B is a binary map obtained by depth image through an optimized k-means method [15].

When the spatial weighted map is obtained, the DWCM-based tracking process is modeled as a detection problem . The center coordinate of the target in the next frame can be calculated as:

$$(x_{t+1}^*, y_{t+1}^*) = \arg\max_{x,y} \; \mathrm{m}_{t+1}(x, y) \tag{4}$$

$$m_{t+1}(x, y) = F^{-1}(F(H_t(x, y)) \odot F(I_{t+1}(x, y)W_t)) \tag{5}$$

where F denotes the FFT function [16], F^{-1} denotes the inverse FFT function, $\odot$ is the element-wise product, $I(\cdot)$ is image intensity that represents appearance of context which contains target and background regions , $H(\cdot)$ represents a spatio-temporal context model and is described in detail in [14]. The flow chart of the proposed DWCM is illustrated in Fig. 1.

According to Eq.(4), the target location in the current frame is computed by maximizing the confidence map derived from the weighted context region surrounding the previous target location. However in tracking process, the scale of the target often changes over time. Therefore, the scale parameter should be updated accordingly. In realistic scenes, the scale change of the target's width is different with the height. Nevertheless, the target scales of the width and height are updated by the same proportion for STC. Thus, it is not reasonable to update the scale by that way. More importantly, if the scale of a target is not updated correctly, the scale parameter σ_t^s in (Eq.(3)) will not be correctly updated either, and a larger weight may be set for the target's surrounding backgrounds (spatial context), leading to drifting.

To solve this problem, a new method is proposed to estimate the scale, thus enhance the tracking performance. We denote that s_t^x is the estimated target scale of width between

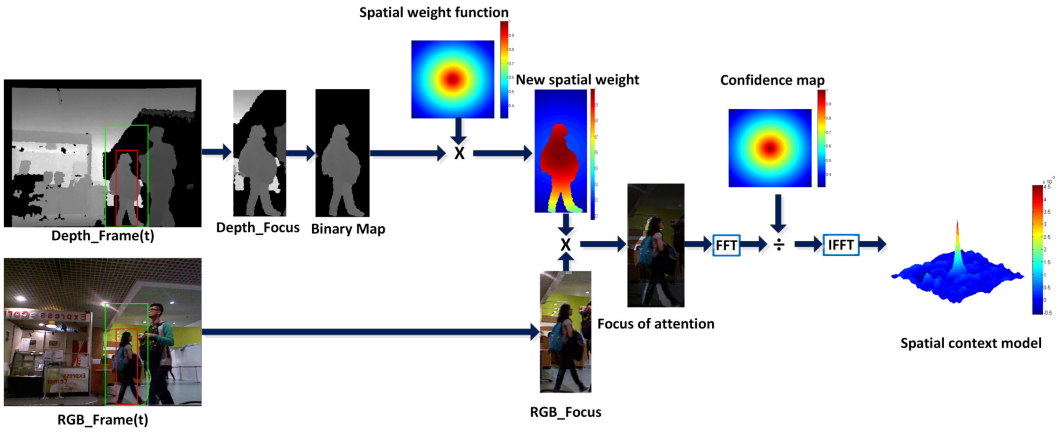

Fig. 1. A system flowchart of the proposed DWCM. Note that the region inside the green rectangle is the context region which includes the target and its surrounding background. The target locations are indicated by the red rectangle. The binary images is combined with the spatial weight map by multiplication. The center of the tracking target is the position of the spatial context model's peak.

two consecutive frames, s_t^y is the estimated target scale of height between two consecutive frames. To avoid over sensitive adaptation and to reduce noise introduced by estimation error, s_{t+1}^x is calculated as the average of the estimated target scale of width from n consecutive frames, s_{t+1}^y conform similar situation. Let $\lambda \in (0,1)$ be a constant filter parameter, which represents the weight of the scale of the previous frame. s_{t+1}^x and s_{t+1}^y in the next frame can be expressed as follows:

$$\begin{aligned} s_{t+1}^x &= \lambda s_t^x + \tfrac{1-\lambda}{n} \textstyle\sum_{i=1}^{n} \left(\sqrt{\tfrac{x_{t-i}^s}{x_{t-1-i}^s}} s_{t-1-i}^x\right) \\ s_{t+1}^y &= \lambda s_t^y + \tfrac{1-\lambda}{n} \textstyle\sum_{i=1}^{n} \left(\sqrt{\tfrac{y_{t-i}^s}{y_{t-1-i}^s}} s_{t-1-i}^y\right) \\ x_t^s &= [x_t^* - \arg\min_x \ |\mathrm{m_t}(\mathrm{x}, \mathrm{y_t^*}) - \zeta \mathrm{m_t}(\mathrm{x_t^*}, \mathrm{y_t^*})|]^{\frac{1}{\kappa}} \\ y_t^s &= [y_t^* - \arg\min_y \ |\mathrm{m_t}(\mathrm{x_t^*}, \mathrm{y}) - \zeta \mathrm{m_t}(\mathrm{x_t^*}, \mathrm{y_t^*})|]^{\frac{1}{\kappa}} \end{aligned} \tag{6}$$

where $\kappa > 1$ is a fixed parameter that is used to suppress the rate of scale variation in order to avoid the influence of partial occlusion and deformation factors, $\zeta \in (0,1)$ is a constant parameters, $m_t(\cdot)$ denotes the confidence map that is computed by Eq.(5). According to Eq.(6), the scale of target is updated. Thus the scale parameter σ_{t+1}^s in Eq.(3) for the spatial weight map should be updated as:

$$\sigma_{t+1}^s = \sqrt{s_t^x s_t^y}\,\sigma_t^s \tag{7}$$

2.3. Adaptive classifier updating

In this section, we establish an adaptive classifier updating scheme for alleviating the potential problem of introducing inaccurate background information during updates. The adaptive classifier updating mechanism is illustrated as follows:

Case 1: No change in state. The model update is straightforward, if the object is neither occluded, nor has undergone any of the other transformations. The SVM is trained by using the new bounding boxes as the positive example and randomly picked bounding boxes that do not overlap with the target as negative examples. And the tracking results are obtained by SVM tracker.

Case 2: Occlusion. When the object is in a (partial or fully) occluded state, the classifier is not updated. And the tracking results are calculated by the proposed DWCM. For a robust tracker, it is very crucial that how to recover from occlusion in the challenge videos. A list of possible target candidates are identified around the position of the tracking results calculated by DWCM and the initial position of the occlusion occurring. By examining the list of possible target candidates, the tracker interprets target recovery when at least one candidate's score evaluated by the SVM classifier is high. The occlusion subroutine ends if the target is recovered from occlusion.

3. EXPERIMENTS

To evaluate the performance of the proposed tracker, we compile a set of challenge tracking sequences from [11]. These video sequences have many challenging factors including illumination changes, heavy occlusion, scale variation, etc. We compare our approach with the state-of-the-art methods, including Kernelized Correlation Filters (KCF) [17], structured output tracker (Struck) [4], spatio-temporal context tracker (STC) [14], complex cell tracker (CCT) [18], compressive tracker (CT) [19], circulant structure tracker with kernels

Fig. 2. Comparison of tracking results on 9 challenge sequences (from left to right and top to down are *basketball, bear, child, face1, walking1, face2, car, walking2* and *cup*, respectively).

(CSK) [5], adaptive color tracker (CN) [20] and tracking revisited using RGBD (RGBD) [11]. All the parameters of the proposed algorithm are fixed for the experiments. For other trackers, we used the original implementation provided by the respective authors.

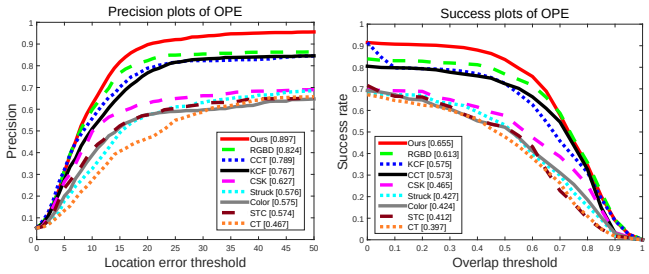

Fig. 3. The performance score for each tracker is shown in the legend.

To quantitatively compare all the trackers, Fig. 2 exhibits a subjective comparison with selected trackers on several representative frames.It can be obviously observed that our proposed tracking method can maintain a precise trail with smaller errors as compared to the other trackers, some of which falsely detect the tracking target when there are occlusion and other challenging factors. Furthermore, our tracker gives the most accurate bounding box, due to the adaptive classifier updating and depth weighted correlation method used in our model, which prevent the tracker from drifting off to the background.

The experimental results for the trackers is based on two different metrics: the precision plot and success plot. The ranking of trackers is based on the Area Under Curve (AUC) score. For more details about the adopted metrics we refer readers to [21]. We show the precision plot and success plot in Fig. 3, which indicates that our approach achieves overall the best performance using both the metrics and significantly outperforms the second best tracker RGBD with 7% performance gain using the metric of AUC score.

4. CONCLUSION

In this paper, we propose depth-weighted correlation method for visual tracking with occlusion detection. By exploiting the depth cue for occlusion detection and correlation between the the target and its surrounding regions, our algorithm is robust to certain condition of occlusion . The proposed update strategy reduces the drifting problem caused by model update without noisy samples through the adaptive classifier updating scheme. Experiments on challenging video sequences demonstrate that the proposed approach performs better than several state-of-the-art approaches.

5. ACKNOWLEDGEMENT

This work is supported by the National Natural Science Foundation of China (No.61571326, 61471262, 61520106002) and National Natural Science Foundation of Tianjin (No.16JCQNJC00900).

6. REFERENCES

[1] Junseok Kwon and Kyoung Mu Lee, "Visual tracking decomposition," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 1269–1276.

[2] Tianzhu Zhang, Si Liu, Narendra Ahuja, Ming-Hsuan Yang, and Bernard Ghanem, "Robust visual tracking via consistent low-rank sparse learning," *International Journal of Computer Vision*, vol. 111, no. 2, pp. 171–190, 2015.

[3] David A Ross, Jongwoo Lim, Ruei-Sung Lin, and Ming-Hsuan Yang, "Incremental learning for robust visual tracking," *International Journal of Computer Vision*, vol. 77, no. 1-3, pp. 125–141, 2008.

[4] Sam Hare, Amir Saffari, and Philip HS Torr, "Struck: Structured output tracking with kernels," in *Proceedings of the 2011 International Conference on Computer Vision*. IEEE Computer Society, 2011, pp. 263–270.

[5] João F Henriques, Rui Caseiro, Pedro Martins, and Jorge Batista, "Exploiting the circulant structure of tracking-by-detection with kernels," in *Computer Vision–ECCV 2012*, pp. 702–715. Springer, 2012.

[6] Si Liu, Tianzhu Zhang, Xiaochun Cao, and Changsheng Xu, "Structural correlation filter for robust visual tracking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4312–4320.

[7] Fanyi Xiao and Yong Jae Lee, "Track and segment: An iterative unsupervised approach for video object proposals," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 933–942.

[8] Darrel Greenhill, J Renno, James Orwell, and Graeme A Jones, "Occlusion analysis: Learning and utilising depth maps in object tracking," *Image and Vision Computing*, vol. 26, no. 3, pp. 430–441, 2008.

[9] Yuan Yuan, Jianwu Fang, and Qi Wang, "Robust superpixel tracking via depth fusion," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 24, no. 1, pp. 15–26, 2014.

[10] Shan Gao, Zhenjun Han, Ce Li, Qixiang Ye, and Jianbin Jiao, "Real-time multipedestrian tracking in traffic scenes via an rgb-d-based layered graph model," *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 5, pp. 2814–2825, 2015.

[11] Shuran Song and Jianxiong Xiao, "Tracking revisited using rgbd camera: Unified benchmark and baselines," in *Computer Vision (ICCV), 2013 IEEE International Conference on*. IEEE, 2013, pp. 233–240.

[12] Olivier Chapelle, "Training a support vector machine in the primal," *Neural computation*, vol. 19, no. 5, pp. 1155–1178, 2007.

[13] Pedro F Felzenszwalb, Ross B Girshick, David McAllester, and Deva Ramanan, "Object detection with discriminatively trained part-based models," *IEEE transactions on pattern analysis and machine intelligence*, vol. 32, no. 9, pp. 1627–1645, 2010.

[14] Kaihua Zhang, Lei Zhang, Qingshan Liu, David Zhang, and Ming-Hsuan Yang, "Fast visual tracking via dense spatio-temporal context learning," in *Computer Vision–ECCV 2014*. 2014, pp. 127–141, Springer.

[15] Jianfeng Wang, Jingdong Wang, Jingkuan Song, Xin-Shun Xu, Heng Tao Shen, and Shipeng Li, "Optimized cartesian k-means," *IEEE Transactions on Knowledge and Data Engineering*, vol. 27, no. 1, pp. 180–192, 2015.

[16] Peter Welch, "The use of fast fourier transform for the estimation of power spectra: a method based on time averaging over short, modified periodograms," *IEEE Transactions on audio and electroacoustics*, vol. 15, no. 2, pp. 70–73, 1967.

[17] João F. Henriques, Caseiro Rui, Pedro Martins, and Jorge Batista, "High-speed tracking with kernelized correlation filters," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 3, pp. 583–596, 2015.

[18] Dapeng Chen, Zejian Yuan, Yang Wu, Geng Zhang, and Nanning Zheng, "Constructing adaptive complex cells for robust visual tracking," in *Computer Vision (ICCV), 2013 IEEE International Conference on*. IEEE, 2013, pp. 1113–1120.

[19] Kaihua Zhang, Lei Zhang, and Ming-Hsuan Yang, "Real-time compressive tracking," in *Computer Vision–ECCV 2012*, pp. 864–877. Springer, 2012.

[20] Martin Danelljan, Fahad Shahbaz Khan, Michael Felsberg, and Joost van de Weijer, "Adaptive color attributes for real-time visual tracking," in *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*. IEEE, 2014, pp. 1090–1097.

[21] Yi Wu, Jongwoo Lim, and Ming-Hsuan Yang, "Online object tracking: A benchmark," in *Computer vision and pattern recognition (CVPR), 2013 IEEE Conference on*. IEEE, 2013, pp. 2411–2418.