2018 24th International Conference on Pattern Recognition (ICPR)
Beijing, China, August 20-24, 2018

Segmentation-Guided Tracking with Prior Map Decision

Ding Ma*, Wei Bu†, Yuying Xie‡, Yuehua Cui§ and Xiangqian Wu*
*School of Computer Science and Technology
Harbin Institute of Technology, Harbin, China 150001
Email: xqwu@hit.edu.cn
†Department of New Media Technologies and Arts
Harbin Institute of Technology, Harbin, China 150001
‡Department of Computational Mathematics, Science and Engineering
Michigan State University, East Lansing, USA 48824
§Department of Statistics and Probability
Michigan State University, East Lansing, USA 48824

***Abstract*—For visual tracking, the target object is represented by an appearance model and the location of the target is estimated in each frame. Numerous tracking algorithms model the appearance of the target with a confidence score and rarely take into account the semantic information of the target. In this paper, we propose an efficient tracking algorithm that models the appearance of the target based on semantic segmentation. The overall architecture consists of two parts: the segmentation part and the tracking part. In the segmentation part, an attention model is employed, providing spatial highlights of the candidate region of the target. In the tracking part, the tracker is constructed by an online updated convolutional neural networks to identify the target in subsequent frames, taking advantage of the segmentation information of the target from the segmentation part. To enhance the performance of this architecture, we design an incremental updated prior map taking both the segmentation signal and the tracking signal into consideration. Extensive experiments on two benchmarks including OTB-50, OTB-100, and Temple-Color, show that the proposed method outperforms other trackers.**

I. Introduction

Visual tracking is one of the fundamental application in computer vision and has been widely applied in various fields such as video surveillance, action analysis, robot vision, and human-computer interaction, etc. While for some applications the target is unknown in advance, and the appearance of the target is changed gradually caused by illumination variation, deformation, and scale variation (see Fig. 1), etc.

A common way to solve this problem of unknown target is to train a discriminative classifier during tracking process and to update the classifier to cope with the various variations of the target appearance [1]–[3]. Among these approaches, the visual tracking is formulated as a binary classification problem which utilizes a bounding box to separate the target from background. And the tracker just need to know the initial location of the bounding box without considering the type of the target. Besides, the current prediction is used to update the classifier.

These algorithms achieves high performance with many advance including, improving discriminative ability over time, adapting the appearance variations of the target by online learning and modeling background, etc. And with the powerfulness of the Convolutional Neural Network (CNN) in extracting more discriminative feature representations, these classifier-based tracker [4], [5] further improve performance. However, these approaches have to cope with a rather inaccurate object description by a bounding box. For instance, when the target encounters with non-rigid and deformable, the tracker will be result in drift phenomenon with online learning, since the noise of training data is inevitable during the period of update. The "noise" refers to the background information contained by the target location marked with a bounding

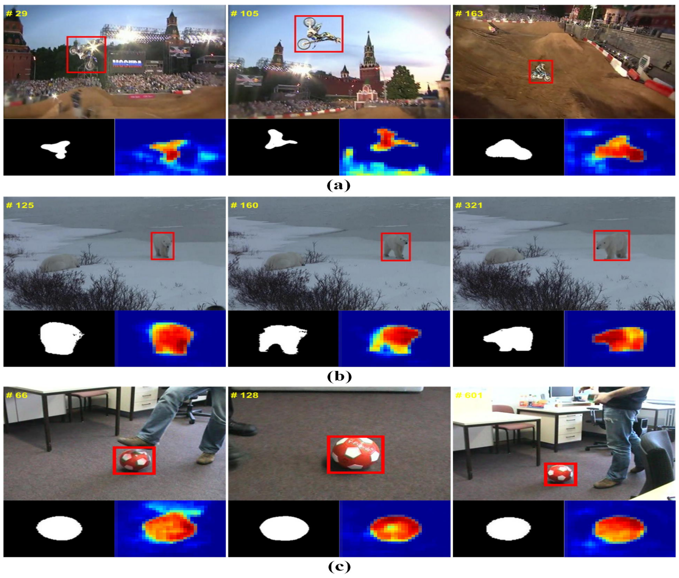


Fig. 1: Tracking in challenging environments including illumination variation (a), deformation (b) and scale variation (c). The bounding box in each frame is the result of our tracker SCPT. The binary image is the segmentation signal of each target. And the heap map is the target objects attention map.

978-1-5386-3788-3/18/$31.00 ©2018 IEEE

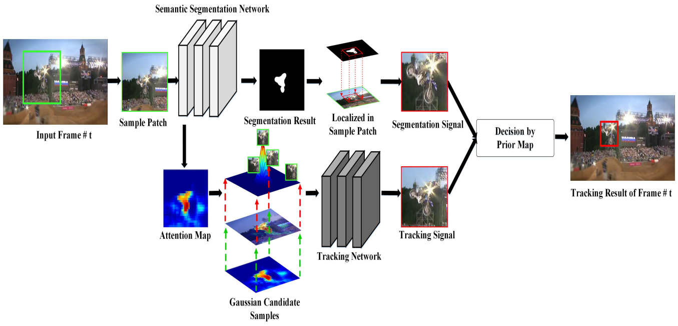


Fig. 2: An overview of the proposed tracking algorithm. The proposed tracker consists of double networks. And the final tracking result is computed by the prior map. The figure is best viewed in color.

box. Thus, the pixel-wise foreground/background labeling in object segmentation is urgently needed to efficiently alleviate the influence from the background information within the bounding box. Nevertheless, because of the shortcomings in object segmentation (sensitive to complex background, fast motion, etc), the approach to utilize the object segmentation directly to track the target may not efficient. Inspired by this, we aim to consider both tracking and segmentation jointly to robustly track the target in difficult scenarios.

In this paper, we propose a simple but effective tracker based on two CNN models, which formulates the appearance of the target by semantic segmentation signal and binary classification signal. First, we employ a semantic segmentation network to get the segmentation signal and the attention map of the target. Second, the tracking signal is obtained by the tracking network processing on samples extracted from the attention map by Gaussian distributed sample method. Third, the final tracking result is computed by a prior map decision strategy. The prior map and tracking network are both updated online with periodic scheme to address the drift problem.

The main contributions of this work are summarized as follows:

(1) We propose a novel and efficient tracking-by-segmentation framework which is robust to cope with the appearance variations of the target by online learning;
(2) We propose a dual CNN model to extract the segmentation signal and the tracking signal. And both signals are selected with a prior map decision strategy to improve the tracking performance;
(3) Quantitative and qualitative evaluations demonstrate the outstanding performance of our tracking algorithm compared to the state-of-the-art techniques on 2 public benchmarks: OTB-50 [6], OTB-100 [7] and Temple-Color [8].

II. Related Work

A. Segmentation-based Tracking

Segmentation-based tracker is efficient to handle non-rigid and deformable objects. And there are some prior works for

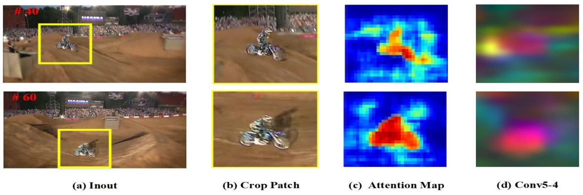


Fig. 3: Visualization of the attention map and convolutional layer (conv5-4) [9]. (a) is the input image; (b) is the crop patch from input image; (c) is the attention map from our semantic segmentation network; (d) is the conv5-4s feature map, which is not clear on target outline.

segmentation-based tracking. Aeschliman et al. [10] propose pixel-level probabilistic models for combining target tracking and segmentation, but this algorithm assumes the environment is relative static. Wang et al. [3] present a robust discriminative appearance model based tracking method using online random forests and mid-level feature (superpixels). [11] extend the idea of Hough Forests to the online domain and couple the voting based detection and back-projection with the GrabCut [12]. A pixel-wise classification is provided by [13] to get soft segmentations of the target. Recently, Hong et al. [14] tracks the target by different configuration across pixel, superpixel and bounding box levels. [15] employs the gradient boosting decision tree to track and segment non-rigid and deformable objects. Although these methods performs well in constrained environments, they have an inherent limitation that they lack the knowledge of the holistic model of the target, which are vulnerable in dynamic situations including complex background, motion blur, etc.

B. Classification-based Tracking

Classification-based tracking aims to learn a classifier that discriminates the target from background. Numerous classification algorithms [1], [2], [16] have been used in tracking methods. [2] address the adaptive appearance model within a multiple instance learning framework, which degrades the influence of incorrectly labeled training samples to some extent. Hare et al. [16] propose a kernelized Structured SVM (SSVM) for visual tracking. The SSVM formulates the tracking problem as a structured output prediction problem that admits a consistent target representation for both learning and detection. Furthermore, [1] propose a dual linear structured SVM tracker (DLSSVM), which approximates intersection kernel with an explicit feature map. With the explicit feature map, the DLSSVM can extract high dimensional linear features to better represents the target than the SSVM.

Recently, automatic feature extraction using Convolutional Neural Network (CNN) brings significant performance improvements in tracking area. The performance of the classification-based tracking methods have improved further. Hong et al. [5] take outputs from hidden layers of a pre-trained CNN (from imagenet) as feature descriptors, and the features are fed into an online SVM to learn discriminative target appearance models. This approach transfer CNN pre-trained

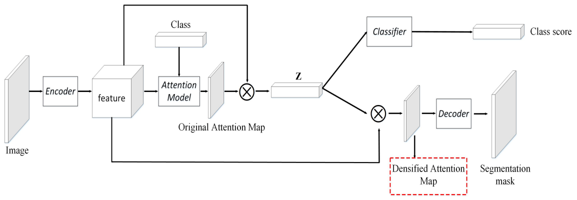


Fig. 4: Overall architecture of the TransferNet. Dotted red rectangular box denotes the input of tracking network.

on a large-scale dataset constructed for image classification to visual tracking task, but the representation may not be very effective due to the fundamental difference between classification and tracking tasks. In order to solve this problem, Bohyung Han [17] train a shallow CNN with limited tracking sequences, and has achieved the state-of-the-art performance by updating the network online. In order to improve ability of the inter-class classification, Heng Fan [18] utilizes recurrent neural network (RNN) to model the structure of the object, and combines it into CNN to improve its robustness in face of similar distractors. In addition, Bohyung Han [19] proposes an ensemble based tracker which aims to learn more robust appearance model of the target. Contrary to the existing approaches, our algorithm takes advantage of the pixel-wise labeled signal from semantic segmentation network and trains a tracking network by online updating.

III. The Proposed Algorithm

In this section, we present the details of our tracker (see Fig. 2). First, a semantic segmentation network is used to get the segmentation signal and the attention map of the target. Second, the tracking signal is obtained by the tracking network processing on samples extracted from the attention map by Gaussian distributed sample method. Third, the final tracking result is computed by a prior map decision strategy. Fourth, the tracking network is updated online with periodic and stochastic scheme to address the drift problem.

A. Semantic Segmentation Network

Some prior works [5], [9] transfer CNN pre-trained on a large-scale dataset constructed for image classification task to visual tracking task and utilize deeper features to obtain the semantic information of the target, which may not be effective due to the difference of the training data between image classification and visual tracking. In this work, the semantic information is extracted by training the attention model directly on data of the tracking domain. Results show that the semantic information got from the attention model has more apparent outlines than that got from deeper convolution layer (see Fig. 3). Furthermore, in amount of visual tracking algorithms, the candidate region is simply manually set as a rectangular region which is a little bit larger than the bounding box of the frame before during tracking process. Employing this kind of strategy cannot take fully advantage of the information of image. Instead, setting the attention map learned from the image as the guide to sample the candidates is able to make full use of the information provided.

Motivated by this fact, we employ the TransferNet [20] which can not only get more precise semantic information, but also provide spatial highlights of the candidate region of the target by attention maps. In the architecture of the TransferNet (see Fig. 4), given a feature descriptor extracted from the encoder and its associated class label, the attention model aims to learn a set of positive weights, where each weight represents the relevance between the feature of each location and a specific category. Then we aggregate the feature and the category information by element-wise product and get a category-specific feature $\mathbf{z}$. The feature $\mathbf{z}$ feeds into two different ways. One way is a classifier which helps the attention model effectively exhibit the highlight spatial regions as the attention model and the classifier are optimized under a same classification objective function. The other way is to get the ultimate segmentation mask. To collect more useful information for segmentation, the densified attention map is obtained by fusing the feature $\mathbf{z}$ and the original feature, since the feature information loses while flowing. And then the segmentation mask is finally got by the Decoder processing on the densified attention maps. More details about it can be found in reference [20].

As we know, the data of the TransferNet [20] includes VOC [21] and COCO [22], the target objects of which are often located in the center of the image or other conspicuous regions, while the target objects of data in visual tracking databases are more unobvious. Besides, objects in visual tracking databases often suffer from motion blur, occlusion and sharply appearance variation, which rarely appear in VOC and COCO. In addition, the segmentation masks are rarely available in most tracking datasets. Hence, it is necessary to annotate the segmentation information within the tracking dataset. In this work, AVOL300++ dataset [23] is selected to annotate the segmentation information. The ALOV300++ dataset contains 314 videos with 89364 frames in total. For efficient training, we select 17000 frames approximately as our training set.

In order to get more precise segmentation masks, we select an interactive segmentation method. Given a raw image, firstly, we crop the image patch, size of which is 1.5 times as big as the ground-truth bounding box; Secondly, for each image patch, we extract superpixel segmentation map by means of ERS; Thirdly, with the guidance of superpixel segmentation map, we indicate the location and region of the object and background by using strokes, which are called markers. And the maximal-similarity based region merging mechanism is to guide the merging process with the help of markers. The merged regions are initially segmented by mean shift segmentation, and then the object contour is extracted by labelling all the non-marker as either background or object.

B. Tracking Network

The tracking network is constructed with the help of the Gaussian distributed sample method on attention maps. Atten-

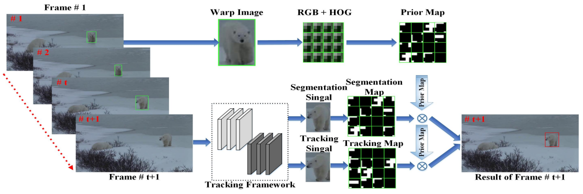


Fig. 5: The flowchart of prior map decision.

tion maps are severed as the candidate regions which highlight the saliency region of a specific target. Given a specific target, candidate samples are extracted from original image with different responses of attention maps. Stronger the response is, more candidate patches are sampled, and vice versa.

In this work, a shallow tracking network is designed. The tracking network has three hidden layers including two convolutional layers (conv 1-2) and one fully connected layer (fc 3). The convolutional layers are initialized with VGG network [24]. The fully connected layer is combined with ReLU and dropout. Doubts may come if the shallow tracking network is powerful enough to locate the object, since features at deeper layers have the ability to capture more high-level semantic information for discriminating objects from background. The answer is yes, as the tracking network gets the semantic information from the segmentation network as the reference. Besides, the deeper tracking network is too time-consuming for online update.

Training & Testing The tracking network is pre-trained on ALOV300++ dataset [23] to obtain effective representation. The fully connected layer is followed by a binary classification layer, which is responsible for distinguishing the target from background. When testing our tracking network, the whole tracking network is fine-tuned online corresponding to the input sequence.

C. Prior Map Decision Strategy

In order to obtain an effective and robust object appearance model, we design a decision strategy measuring the similarities between the prior map and the two signals from two networks, the semantic segmentation network and the tracking network. And the prior map plays a role of a regularization to enforce the two networks to focus on the target object (see Fig. 5). In the rest of this section, we first illustrate the construction of prior map. Then, we will show the procession of prior map decision strategy.

1) Prior Map: Colors and edges are most sensitive to human visual system [25], which provide useful cues to discriminate the foreground object from the background. In this work, we use color features in the RGB color space and HOG features to construct the prior map. We first normalize the initial bounding box P provided by the first frame to a fixed size 32×32. Then we sample n local image patches inside the warped bounding box $\tilde{P}$ sequentially (see Fig. 5). Next, transform the local image patch p into the RGB color space, denoted as $f^{RGB}(p)$. Furthermore, we extract the HOG features to capture edge orientation information of the target, denoted as $f^{HOG}(p)$. Then, we normalize both $f^{RGB}(p)$ and $f^{HOG}(p)$ to $[0,1]$, and concatenate $f^{RGB}(p)$ and $f^{HOG}(p)$ to form a union feature vector $F^{RGB}_{HOG}(p)$. The union feature vector is rescaled to $[0,1]$ by:

$$F^{RGB}_{HOG}(p) \leftarrow \frac{F^{RGB}_{HOG}(p) - \min(F^{RGB}_{HOG}(p))}{\max(F^{RGB}_{HOG}(p)) - \min(F^{RGB}_{HOG}(p))} \tag{1}$$

where $\max(\cdot)$ and $\min(\cdot)$ denotes the maximal and minimal operators, respectively.

Next, $F^{RGB}_{HOG}(p)$ in Eq. 1 is encoded into a set of vectorized Boolean maps $B(P) = \{b_i(p)\}_{i=1}^{n}$ by:

$$b_i(p) = \begin{cases} 1 & F^{RGB}_{HOG}(p) \geq \theta \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

where θ is a threshold to control the information within the $b_i(p)$. And $B(P)$ is the initial prior map for tracking.

2) Decision Strategy: The decision strategy is formulated by:

$$D_i = \begin{cases} 1 & R_S \geq R_T \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

$$B_i(P) = \begin{cases} (1-\rho)\text{map}_i^T + \rho B_{i-1}(P) & D_i = 0 \\ (1-\rho)\text{map}_i^S + \rho B_{i-1}(P) & D_i = 1 \end{cases} \tag{4}$$

$$R_{S,T} = B_i(P) \otimes \text{map}_i^T \tag{5}$$

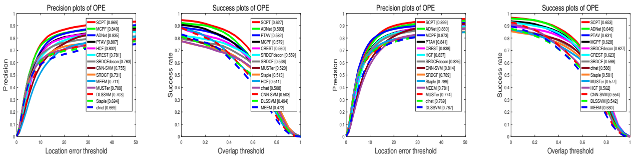
where R_S is the relevance score between the prior map and the segmentation signal (the generative model), R_T is the relevance score between the prior map and the tracking signal (the discriminative model). D_i is the decision parameter controlling the selection of both signals. $B_{i-1}(P)$ is the *Boolean map* at $(i-1)$-th frame. ρ is a learning parameter that controls the update frequency of $B_{i-1}(P)$, which copes with rapid appearance variations and also alleviates the drift problem thanks to retaining the information in the first frame. The operator $\otimes$ is the convolution operator. And map_i^* is the *Boolean map* for either the segmentation signal or tracking signal.

At last, the bounding box regression technique which is popular in object detection [26] is employed to improve target localization accuracy in the tracking task.

D. Updating Model with Periodic and Stochastic Schemes

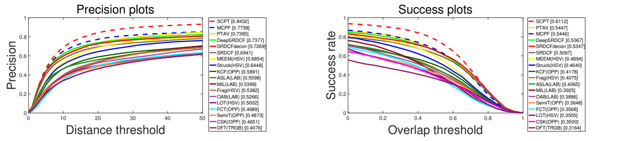
In this work, both periodic and stochastic update mechanisms are presented to handle occlusion and drift during online tracking.

1) Periodic: We accumulate a number of key frames to update the tracking network. According to the relevance score in Eq. 5, if the relevance score is larger than the threshold f^{θ}, the corresponding frame is denoted a key frame. When the number of key frames reaches a threshold, the tracking network is updated.



(a) (b) (c) (d)

Fig. 6: (a) and (b) are the precision and success plots on OTB50, respectively. (c) and (d) are the precision and success plots on OTB100, respectively.



(a) (b)

Fig. 7: (a) and (b) are the precision and success plots on TC128, respectively.

2) Stochastic: When the target comes across scale variation and deformation, the segmentation signal is more reliable than the tracking signal to some extent. This is because the generative model is more robust than the discriminative model in processing the shape variation of the target. When the decision strategy mentioned above selects the segmentation signal as the final tracking result, the tracking network is updated immediately at current frame.

IV. Experiment

To evaluate our approach SCPT, we perform comprehensive experiments on two benchmark datasets: OTB-50 [6], OTB-100 [7] and Temple-Color [8]. The OTB-50 and OTB-100 dataset include 50 and 100 sequences respectively tagged with 11 attributes. The Temple-Color dataset contains 128 color sequences with challenge factor annotations.

A. Experimental Setting

The proposed SCPT runs on a PC with an Intel(R) Core(TM) i7-4790k CPU and an NVIDIA Tesla K40c GPU.

1) Semantic Segmentation Network: The architecture of our semantic segmentation network is similar to [20]. And we employ post-processing based on fully-connected CRF [27].

2) Tracking Network: The architecture of our tracking network consists of a input ($3@107 \times 107$), conv1 ($96@51 \times 51$), conv2 ($256@11 \times 11$), and fc3 (256). We set Gaussian distribution model to generate target candidates in each frame, the number of samples is set to $N = 300$. We typically extract 5 positive samples and 50 negative samples for initialization. We collect 5 positive samples and 20 negative samples for every key frame, where positive and negative examples have ≥ 0.7 and ≤ 0.3 IoU overlap ratios with the estimated target bounding boxes respectively. For offline training, we collect 50 positive and 100 negative samples from every frame, where positive and negative examples have ≥ 0.8 and ≤ 0.4 IoU overlap ratios with ground-truth bounding boxes, respectively. We exploit 1200 training examples for bounding box regression.

3) Prior Map: The extracted HOG features are with cell size 8 pixels and 9 orientation bins. For grayscale sequences, we extract raw intensity and HOG features. And for color sequences, the raw color features from RGB color space and HOG features are extracted. The learning rate ρ, the threshold θ, f_θ are set to 0.95, 0.25 and 10 respectively.

B. Experiments on OTB-50 and OTB-100 Dataset

We compare the proposed SCPT tracker on both OTB-50 and OTB-100 datasets with the following recently published 13 trackers: ADNet [28], MCPF [4], PTAV [29], CREST [30], HCF [9], SRDCFdecon [31], CNN-SVM [5], SRDCF [32], Staple [33], MEEM [34], MUSTer [35], cfnet [36], and DLSSVM [1]. The results of one-pass-evaluation (OPE) are shown in Fig. 6. The SCPT tracker achieves the best performance among the state-of-the-arts on both datasets. The AUCs of the precision plot and the success plot are 0.869 and 0.627 on OTB-50, and 0.899 and 0.816 on OTB-100, respectively.

C. Experiments on Temple-Color Dataset

We perform experiments on the Temple-Color dataset [8] with 128 videos. Among the compared methods, our approach improves the state-of-the-art on this dataset with an error score of 0.8432 (see Fig. 7(a)). Fig. 7(b) shows the success plot on the Temple-Color dataset. Our tracker achieves a success score of 0.6112 outperforming state-of-the-art approaches. And Qualitative results of the proposed SCPT are shown in Fig. 8.

V. Conclusion

In this paper, an online tracking algorithm based on semantic segmentation and prior map decision is proposed. Our tracking algorithm learns the pixel-wise labeled knowledge from the semantic segmentation network with the attention model. With the guidance of the attention model, an online updated tracking network is used to separate the target from the background. The prior map decision strategy is proposed to regulate both networks focusing on the target object. Extensive experimental results on two benchmark databases demonstrate the effectiveness of the proposed algorithm against the state-of-the-art methods for visual tracking.

Acknowledgment

This work was supported in part by the Natural Science Foundation of China under Grant 61472102, and 61672194, by the Distinguished Youth Science Foundation of Heilongjiang Province of China under Grant JC2018021, by the Shandong Provincial Natural Science Foundation, China under Grant ZR2016FM04, and by the Humanity and Social Science Youth foundation of Ministry of Education of China under Grant 14YJC760001.

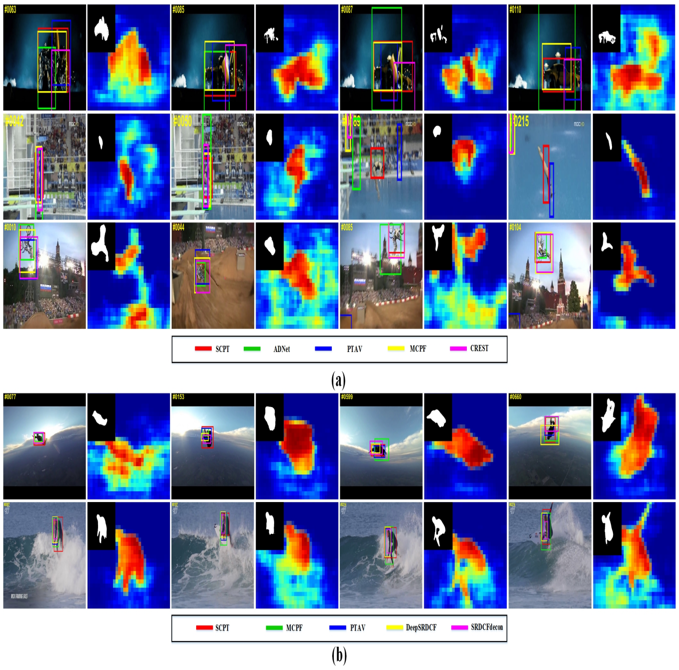


Fig. 8: Some results of the proposed SCPT tracker on a subset of challenging sequences. (a) is the comparison results on OTB dataset, and (b) is the comparison results on Temple-Color dataset.

References

[1] J. Ning, J. Yang, S. Jiang, L. Zhang, and M. H. Yang, "Object tracking via dual linear structured svm and explicit feature map," in *Computer Vision and Pattern Recognition*, 2016, pp. 4266–4274.

[2] B. Babenko, M. H. Yang, and S. Belongie, "Visual tracking with online multiple instance learning," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, 2009, pp. 983–990.

[3] W. Wang, C. Wang, S. Liu, T. Zhang, and X. Cao, "Robust target tracking by online random forests and superpixels," *IEEE Transactions on Circuits & Systems for Video Technology*, vol. PP, no. 99, pp. 1–1, 2017.

[4] T. Zhang, C. Xu, and M. H. Yang, "Multi-task correlation particle filter for robust object tracking," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 4819–4827.

[5] S. Hong, T. You, S. Kwak, and B. Han, "Online tracking by learning discriminative saliency map with convolutional neural network," pp. 597–606, 2015.

[6] Y. Wu, J. Lim, and M. H. Yang, "Online object tracking: A benchmark," in *Computer Vision and Pattern Recognition*, 2013, pp. 2411–2418.

[7] Y. Wu, J. Lim, and M. H. Yang, "Object tracking benchmark," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 37, no. 9, pp. 1834–1848, 2015.

[8] P. Liang, E. Blasch, and H. Ling, "Encoding color information for visual tracking: Algorithms and benchmark," *IEEE Trans Image Process*, vol. 24, no. 12, pp. 5630–5644, 2015.

[9] C. Ma, J. B. Huang, X. Yang, and M. H. Yang, "Hierarchical convolutional features for visual tracking," in *IEEE International Conference on Computer Vision*, 2016, pp. 3074–3082.

[10] C. Aeschliman, J. Park, and A. C. Kak, "A probabilistic framework for joint segmentation and tracking," in *Computer Vision and Pattern Recognition*, 2010, pp. 1371–1378.

[11] M. Godec, P. M. Roth, and H. Bischof, "Hough-based tracking of non-rigid objects," in *IEEE International Conference on Computer Vision*, 2011, pp. 81–88.

[12] C. Rother, V. Kolmogorov, and A. Blake, ""grabcut": interactive foreground extraction using iterated graph cuts," in *ACM SIGGRAPH*, 2004, pp. 309–314.

[13] S. Duffner and C. Garcia, "Pixeltrack: A fast adaptive algorithm for tracking non-rigid objects," in *IEEE International Conference on Computer Vision*, 2014, pp. 2480–2487.

[14] Z. Hong, C. Wang, X. Mei, D. Prokhorov, and D. Tao, *Tracking Using Multilevel Quantizations*, 2014.

[15] J. Son, I. Jung, K. Park, and B. Han, "Tracking-by-segmentation with online gradient boosting decision tree," in *IEEE International Conference on Computer Vision*, 2016, pp. 3056–3064.

[16] S. Hare, A. Saffari, and P. H. S. Torr, "Struck: Structured output tracking with kernels," in *IEEE International Conference on Computer Vision*, 2012, pp. 263–270.

[17] H. Nam and B. Han, "Learning multi-domain convolutional neural networks for visual tracking," in *Computer Vision and Pattern Recognition*, 2016, pp. 4293–4302.

[18] H. Fan and H. Ling, "Sanet: Structure-aware network for visual tracking," pp. 2217–2224, 2016.

[19] B. Han, J. Sim, and H. Adam, "Branchout: Regularization for online ensemble tracking with convolutional neural networks," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 521–530.

[20] S. Hong, J. Oh, H. Lee, and B. Han, "Learning transferrable knowledge for semantic segmentation with deep convolutional neural network," pp. 3204–3212, 2016.

[21] M. Everingham, L. Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International Journal of Computer Vision*, vol. 88, no. 2, pp. 303–338, 2010.

[22] T. Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollr, and C. L. Zitnick, "Microsoft coco: Common objects in context," vol. 8693, pp. 740–755, 2014.

[23] A. W. M. Smeulders, D. M. Chu, R. Cucchiara, S. Calderara, A. Dehghan, and M. Shah, "Visual tracking: An experimental survey," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 36, no. 7, pp. 1442–68, 2014.

[24] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, "Return of the devil in the details: Delving deep into convolutional nets," *Computer Science*, 2014.

[25] M. S. Livingstone and D. H. Hubel, "Anatomy and physiology of a color system in the primate visual cortex," *Journal of Neuroscience the Official Journal of the Society for Neuroscience*, vol. 4, no. 1, pp. 309–56, 1984.

[26] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 580–587.

[27] P. Krahenbhl and V. Koltun, "Efficient inference in fully connected crfs with gaussian edge potentials," pp. 109–117, 2012.

[28] S. Yun, J. Choi, Y. Yoo, K. Yun, and Y. C. Jin, "Action-decision networks for visual tracking with deep reinforcement learning," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1349–1358.

[29] H. Fan and H. Ling, "Parallel tracking and verifying: A framework for real-time and high accuracy visual tracking," pp. 5487–5495, 2017.

[30] Y. Song, C. Ma, L. Gong, J. Zhang, R. W. H. Lau, and M. H. Yang, "Crest: Convolutional residual learning for visual tracking," in *IEEE International Conference on Computer Vision*, 2017, pp. 2574–2583.

[31] M. Danelljan, G. Hger, F. S. Khan, and M. Felsberg, "Adaptive decontamination of the training set: A unified formulation for discriminative visual tracking," pp. 1430–1438, 2016.

[32] M. Danelljan, G. Hager, F. S. Khan, and M. Felsberg, "Learning spatially regularized correlation filters for visual tracking," in *IEEE International Conference on Computer Vision*, 2015, pp. 4310–4318.

[33] L. Bertinetto, J. Valmadre, S. Golodetz, O. Miksik, and P. Torr, "Staple: Complementary learners for real-time tracking," vol. 38, no. 2, pp. 1401–1409, 2015.

[34] J. Zhang, S. Ma, and S. Sclaroff, "Meem: Robust tracking via multiple experts using entropy minimization," vol. 8694, pp. 188–203, 2014.

[35] Z. Hong, Z. Chen, C. Wang, X. Mei, D. Prokhorov, and D. Tao, "Multi-store tracker (muster): A cognitive psychology inspired approach to object tracking," in *Computer Vision and Pattern Recognition*, 2015, pp. 749–758.

[36] J. Valmadre, L. Bertinetto, J. Henriques, A. Vedaldi, and P. H. S. Torr, "End-to-end representation learning for correlation filter based tracking," pp. 5000–5008, 2017.