Kernel correlation filters based on feature fusion for visual tracking

Bo Yang1,*, Xiaopeng Hu1 and Fan Wang1

1School of Computer Science and Technology, Dalian University of Technology, No. 2 Linggong Road, Dalian 116024, China.
E-mail: dlutyb@foxmail.com

Abstract. This paper presents a CF-based method for robust visual tracking by multiple features fusion. HOG features and high-level CNN features are used to enhance tracking performance. First, different features are extracted from the target. Secondly, the extracted features are employed to learn the corresponding discriminant models, which are used to generate response maps respectively. Finally, an accurate response map can be obtained by fusing multiple response maps. In addition, we propose a model update strategy to reduce model drift under occlusion. The OTB benchmark is used to test the proposed method. Experimental results show that the proposed method can significantly improve tracking performance in complex situations especially in occlusion and fast motion. In addition, compared with several state-of-the-art methods, the proposed method shows competitive accuracy rate and success rate.

1. Introduction

Object tracking is one of the most popular research topics in the field of computer vision, and has been widely used in many fields. Accurate and robust target tracking is still very challenging due to the effects of illumination variations, background clutter, and occlusion, etc. Discriminative Correlation Filter (DCF) based algorithms have been successfully applied to visual tracking recently. These trackers can learn a correlation filter from a set of training samples and achieve a good balance between tracking accuracy and speed.

KCF [1] is the pioneering work in CF based trackers. KCF learn a correlation filter from a set of training samples as a linear template to distinguish between the target and the surrounding environment. However, KCF still has some issues that need to be addressed. On the one hand, KCF uses single hand-crafted feature, namely the HOG feature, to represent the target. The HOG feature has strong invariance to geometric changes and illumination variation, and is sensitive to fast motion and occlusion. On the other hand, the learning of CF model strongly depends on the tracking results of the previous frame. When the tracking result of a previous frame is inaccurate, model will drift.

In this paper, we propose a CF tracker based on feature fusion to overcome the shortcomings of KCF in occlusion and fast motion situations. As we all know, CNN features have strong discriminative ability, especially in catching semantic information. In this paper, HOG features are used to obtain accurate appearance information, at the same time, high-level CNN features are used to obtain rich semantic information. We use the internal structure of each feature to solve two independent ridge regression problems. First, the target appearance models are established by extracting HOG features and CNN features of candidate regions, respectively. Then, based on the ridge regression, the response maps of each model are obtained by using the polynomial kernel
function to solve the similarity between candidate regions and the target. Finally, the target position is predicted by the linear fusion of the two response maps. Additionally, a model updating strategy is proposed to reduce model drift in occlusion. Experimental results on OTB benchmark show that the proposed method performs better than KCF in terms of accuracy and robustness. The performance of the proposed tracker is significantly improved in complex situations especially in occlusion and fast motion.

2. Related work

2.1. Visual tracking

Visual tracking can be divided into two types: Generative Model [2] and Discriminative Model [3]. The generative model usually looks for the candidate that is most similar to the template as tracking result. Kwon et al. [2] decomposed the appearance model and motion model of the target into several sub-appearance models and sub-motion models, and correlated them to obtain sub-trackers, which can cope with various appearance and motion changes of the target. The discriminant model trains a classifier to distinguish the target from the background, and chooses the candidate sample with the highest confidence as the prediction result. Avdian [4] proposes a support vector tracking algorithm (SVT), which combines SVM [5] with optical flow-based tracking method, uses SVM for detection, and uses optical flow for tracking. In recent years, discriminant method has become the mainstream method in the field of visual tracking and shows promising results on the benchmark, such as OTB.

2.2. Correlation filter-based tracking

Recently, tracking based on CF has attracted considerable attention and achieved significant progress. CF based trackers learn a correlation filter from a set of training samples, and train a linear template to distinguish between the target and the surrounding environment. The pioneering work of CF based trackers is MOSSE [6], which can learn correlation filters in the frequency domain using a small number of samples. With the introduction of non-linear kernel, scale estimation, maximum edge classifier, spatial regularization and so on, trackers based on correlation filters significantly improve tracking performance. For example, the concepts of circulant matrices and kernel are introduced in CSK [7]. Heriques et al. propose a kernel correlation filter (KCF) [1] by extending CSK. SRDCF [8] adds a spatial regularization function to punish filter coefficients outside the target area, mainly solving the boundary effect brought by circulant matrices. BACF [9] can extract a larger search field to obtain negative samples, which effectively improves the ability of tracker identification. CRSRCF [10] performs Temporal CRSR in temporal domain, and proposes to combine the regularization weight with correlation filters in temporal domain.

3. Proposed method

HOG is a general feature adopted by many CF based methods. However, HOG features are sensitive to occlusion and large deformation. CNN features are powerful in image representations, have rich semantic information, and are not sensitive to deformation. In this paper, we propose a CF based framework that combines hand-crafted appearance features with CNN semantic features to achieve robust target tracking and achieve excellent performance.

3.1. Feature fusion

Feature expression is of great importance to target tracking. Choosing appropriate robust features can greatly improve the tracking performance. KCF uses a single hand-crafted feature, namely HOG feature for visual tracking. However, HOG features can’t effectively deal with complex tracking scenes such as occlusion or fast motion. In recent years, with the development of deep learning, CNN features have been widely used in the field of image representation due to their powerful feature expression capabilities. The high layer of CNN can capture semantic information more effectively. Semantic information can be used to deal with large appearance changes and can make up for the
shortcomings of HOG in handling occlusion or fast motion. In this paper, ImageNet-VGG-Net [11] is used to extract convolution features. As illustrated in Figure 1, we can see the input image, the CNN features of the second layer, the third layer, the fourth layer, and the fifth layer. Cov-5 layer can capture more semantic information. In this paper, HOG features and cov-5 layer features are combined to improve tracking performance.

\[
\begin{align*}
\text{Input} & \quad \text{Conv-2} & \quad \text{Conv-3} & \quad \text{Conv-4} & \quad \text{Conv-5} \\
\end{align*}
\]

\[
\text{Figure 1. Visualization of five layers features of CNN.}
\]

HOG features and the features of cov-5 layer are used to generate discriminative models \( Z^{\text{hog}} \) and \( Z^{\text{cnn}} \) respectively, which can be used to get CNN and HOG response maps. Then the two response maps \( r^{\text{hog}} \) and \( r^{\text{cnn}} \) are fused to predict the target. The fusion strategy is defined as follows:

\[
r = \beta \cdot r^{\text{hog}} + (1 - \beta) \cdot r^{\text{cnn}}
\]

Where \( \beta \) is a weighted coefficient, \( \beta \in [0,1] \), and in this paper, \( \beta = 0.3 \). After obtaining the final response, the target position is determined according to the maximum response.

\[
\begin{align*}
\text{APCE}=11.1822 & \quad \text{APCE}=9.4487 & \quad \text{APCE}=10.5198 \\
\text{APCE}=13.2231 & \quad \text{APCE}=8.0795 & \quad \text{APCE}=8.0997 \\
\text{APCE}=12.6452 & \quad \text{APCE}=10.7186 & \quad \text{APCE}=7.1971 \\
\end{align*}
\]

\[
\text{Figure 3. APCE value of partial video sequences under occlusion.}
\]

\[
\text{Figure 4. The precision and success plots on OTB benchmark.}
\]

Table 1. Precision rate (PR) and success rate (SR) between the proposed tracker and the state-of-the-art trackers. The best and second best results are in red and green colors, respectively.

| Tracker | Precision rate | Success rate |
|---------|---------------|--------------|
| CT      | 0.406         | 0.306        |
| DFT     | 0.496         | 0.389        |
| CSK     | 0.545         | 0.398        |
| TLD     | 0.608         | 0.437        |
| Struck  | 0.656         | 0.474        |
| DSST    | 0.728         | 0.541        |
| KCF     | 0.728         | 0.509        |
| OURS    | 0.773         | 0.559        |
3.2. Update strategy
In the case of occlusion, not only because of the limitation of single hand-craft feature, but also because of the insufficiency of its update strategy, the accuracy of KCF will decrease. KCF updates every frame even in occlusion, so that the wrong features will be learned and the model will drift.

To solve this problem, the reliability of the tracking results needs to be considered when updating the model in the occlusion situation. To determine whether the tracker is experiencing occlusion, we visualize the response map. As shown in Figure 2, when the target encounters occlusion, the response fluctuates sharply and is disturbed by multiple peaks. In this paper, the average peak correlation energy (APCE) is used to measure the variability of the response map. The average peak correlation energy is defined as follows:

$$APCE = \frac{|S_{\text{max}} - S_{\text{min}}|^2}{\text{mean} \left( \sum (S_{i,j} - S_{\text{mean}})^2 \right)}$$  \hspace{1cm} (2)

Where $S_{\text{max}}$ and $S_{\text{min}}$ denote the maximum, minimum of the response. $S_{i,j}$ denotes the $i$th row $j$th column element of the response. As shown in Figure 3, when the target is not occluded, the response fluctuates slightly and APCE will be at a high value; when the target is partially occluded, the response will fluctuate and APCE will decrease; when the target is completely occluded, the response will fluctuate significantly and APCE will be at a low value. Thus, the average peak correlation energy can be used to reflect the fluctuation level of response, thereby adjusting the model update strategy.

4. Experiments
The experimental platform is MATLAB R2017b. The proposed tracker is tested on an Intel I7-4790K, 4.00 GHz CPU with 32 GB RAM desktop computer.

4.1. Datasets and evaluation metrics
The proposed method is evaluated by the standard benchmark datasets OTB [12], which contains several test sequences with ground truth. These test sequences cover 11 attributes. OPE (One-Pass...
Evaluation) is used to evaluate the proposed algorithm. Meanwhile, the precision and success plots are obtained to evaluate the accuracy and success rate of the proposed tracker.

4.2. Results
The proposed tracker is compared with the state-of-the-art trackers including the baseline KCF [1], CT[13], CSK[7], DFT[14], TLD[15], Struck[3] and DSST[16] on OTB benchmark.

4.2.1. Quantitative results.
Figure 4 shows the comparison results of between the proposed tracker and the state-of-the-art trackers. Experimental results show that the proposed tracker shows competitive accuracy and success rates. The precision and success scores are summarized in Table 1 for a more intuitive comparison. The best and second best results are in red and green colors, respectively. Compared with KCF, the precision and success rates of the proposed tracker are improved by 4.5% and 5.0% respectively. It can be seen that the proposed tracker has strong robustness and can improve the performance of tracking.

In addition, the performance of the proposed tracker is compared with the state-of-the-art trackers in complex scenarios. The experimental results of six challenging attributes are shown in Figure 5 and Figure 6, including DEF, OPR, OCC, FM, BC and MB. Compared with KCF, the performance of the proposed tracker is significantly improved in various scenarios such as occlusion, fast motion, etc. For example, in the case of fast motion, the proposed tracker has a significant gain of 9.3% in precision rate and 6.6% in success rate. The experimental results verify the effectiveness of the proposed algorithm in a variety of challenging scenarios, and also verify that the proposed algorithm has excellent accuracy and robustness.

4.2.2. Qualitative results.
To qualitatively evaluate the proposed tracker, the tracking results of certain frames of three test sequences are shown in Figure 7, including jogging, motorRolling and jumping. As can be seen from Figure 7, in several challenging scenarios, the proposed tracker has better accuracy and robustness, while some other trackers will drift. For example, in jogging, in the case of occlusion, KCF drifts. Because that HOG is sensitive to occlusion. Besides, the tracking strategy of KCF makes the tracker learn the wrong features all the time and lead to error accumulation. In motorRolling, the prediction result of KCF gradually deviate from the real target because of the fast motion. The proposed tracker significantly improves the ability of the tracker to distinguish between the target foreground and the semantic background by fusing deep semantic features and HOG features. In a word, semantic
information is of great significance for classification. By using the strategy of fusing CNN features and HOG features, the proposed tracker obtain a stronger ability of target feature expression, which can improve the robustness and accuracy of tracking.

5. Conclusion
In this paper, a CF based method for visual tracking by feature fusion is proposed. HOG appearance feature and high-level CNN features are fused to improve tracking performance in complex situations especially in fast motion and occlusion. In addition, we propose an update strategy based on APCE to alleviate the tracking drift problem. To verify the effectiveness of the proposed method, we compared it with some state-of-the-art methods on OTB benchmark. Experimental results show that the accuracy and robustness of the proposed method are significantly improved compared to the baseline KCF. The proposed method shows excellent performance in complex situations especially in occlusion and fast motion.

Acknowledgments
The authors declare that they have no conflict of interest.

References
[1] Henriques J F, Caseiro R, Martins P, et al. High-speed tracking with kernelized correlation filters. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(3): 583-596.
[2] Kwon J, Lee KM. Visual tracking decomposition. In: Proceedings of CVPR, 2010: 1269-1276.
[3] Hare S, Golodetz S, Saffari A, et al. Struck: structured output tracking with kernels. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016, 38(10): 2096-2109.
[4] Avidan S. Support vector tracking. IEEE transactions on pattern analysis and machine intelligence, 2004, 26(8): 1064-1072.
[5] Burges C J C. A tutorial on support vector machines for pattern recognition. Data mining and knowledge discovery, 1998, 2(2): 121-167.
[6] Bolme D S, Beveridge J R, Draper B A, et al. Visual object tracking using adaptive correlation filters. In: Proceedings of CVPR, 2010: 2544-2550.
[7] Henriques J F, Caseiro R, Martins P, et al. Exploiting the circulant structure of tracking-by-detection with kernels. In: Proceedings of ECCV, 2012(7575): 702-715.
[8] Danelljan M, Hager G, Khan F S, et al. Learning spatially regularized correlation filters for visual tracking. In: Proceedings of ICCV, 2015: 4310-4318.
[9] Galoogahi H K, Fagg A, Lucey S. Learning background-aware correlation filters for visual tracking. In: Proceedings of ICCV, 2017: 1135-1143.
[10] Han R, Guo Q, and Feng W. Content-related spatial regularization for visual object tracking. In: Proceedings of the IEEE International Conference on Multimedia and Expo, 2018:1-6.
[11] Nguyen K, Fookes C, Sridharan S. Improving deep convolutional neural networks with unsupervised feature learning. In: Proceedings of the IEEE International Conference on Image Processing (ICIP), Canada, 2015: 2270-2274.
[12] Wu Y, Lim J, Yang M H. Online object tracking: A benchmark. In: Proceedings of CVPR, 2013: 2411-2418.
[13] Zhang K H, Zhang L, Yang M H. Real-time compressive tracking. In: Proceedings of ECCV, 2012(7574): 864-877.
[14] Sevilla L L, Learned M E. Distribution fields for tracking. In: Proceedings of CVPR, 2012: 1910-1917.
[15] Kalal Z, Mikolajczyk K, Matas J. Tracking-learning-detection. IEEE transactions on pattern analysis and machine intelligence, 2012, 34(7): 1409-1422.
[16] Danelljan M, Häger G, Khan F, et al. Accurate scale estimation for robust visual tracking. In: Proceedings of the British Machine Vision Conference, 2014.