A complementary tracking model with multiple features

Peng Gao, Yipeng Ma, Chao Li, Ke Song, Fei Wang, et al.

Peng Gao, Yipeng Ma, Chao Li, Ke Song, Fei Wang, Liyi Xiao, "A complementary tracking model with multiple features," Proc. SPIE 10836, 2018 International Conference on Image and Video Processing, and Artificial Intelligence, 1083618 (29 October 2018); doi: 10.1117/12.2500635

SPIE.

Event: 2018 International Conference on Image, Video Processing and Artificial Intelligence, 2018, Shanghai, China

A Complementary Tracking Model with Multiple Features

Peng Gao, Yipeng Ma, Chao Li, Ke Song, Fei Wang*, and Liyi Xiao

Shenzhen Graduate School, Harbin Institute of Technology, China

ABSTRACT

Discriminative Correlation Filters based tracking algorithms exploiting conventional handcrafted features have achieved impressive results both in terms of accuracy and robustness. In this paper, to achieve an efficient tracking performance, we propose a novel visual tracking algorithm based on a complementary ensemble model with multiple features. Additionally, to improve tracking results and prevent targets drift, we introduce an effective fusion method by exploiting relative entropy to coalesce all basic response maps and get an optimal response. Furthermore, we suggest a simple but efficient update strategy to boost tracking performance. Comprehensive evaluations are conducted on two tracking benchmarks demonstrate and the experimental results demonstrate that our method is competitive with numerous state-of-the-art trackers. Our tracker achieves impressive performance with faster speed on these benchmarks.

Keywords: Object tracking, correlation filter, multiple features, ensemble model, relative entropy

1. INTRODUCTION

Visual object tracking is a basic research problem with various applications in computer vision. The goal of tracking is to estimate the states of an arbitrary target by discriminating between its appearance and that of the surroundings in a video. Object tracking can be applied in many applications, such as video surveillance, assistant driving systems and intelligent traffic control. The most effective trackers should handle all the variations both from background and target itself while track at speeds that far exceed the frame-rate requirement. Despite much progress have been made in recent years, visual object tracking still remains largely unsolved problems due to various challenging factors such as fast motions, background clutters, motion blurs, deformations, etc.[1,2]

In the past decades, Discriminative Correlation Filters (DCF) based approaches[3,4] have drawn a lot of attention from the computer vision community because they are treated as similarity measurements between two image signals in signal processing. Due to DCF enables training and detection with dense sampling strategy and high-dimensional features[3] at frame-rate by efficiently solving a ridge regression problem in the Fourier frequency domain, the time-consuming convolution operations can be straightforwardly avoided. Most state-of-the-art DCF-based trackers employ template handcrafted features such as Histogram of Oriented Gradients (HOGs) and Color Names (CNs) to present the target in a video, they have shown excellent performance on existing visual object tracking benchmarks, but they perform poorly when the appearance of the target changes rapidly such as fast motions and fast deformations. Statistical handcrafted features, such as Color Histograms (CHs), are insensitive to fast target states changing, but they yield inferior performance when illumination variations and background clutters are considered.[5] Some recent tracking algorithms[6,7] concentrate on exploiting all the object representations above-mentioned to train a model, which show favorable performance to target states change and color variations. However, these methods combine the response maps simply, the tracking results are not the optimal.

Motivated by these facts, we propose an appealing complementary ensemble object tracking framework to take advantages of multiple features for visual object tracking. Specifically, we deduce several basic trackers by treating multiple features as independent linear regression problems. Each basic tracker gives a response map of the target position in a new image. Moreover, we observe that the response maps obtained by a single basic tracker are not robust enough to handle more challenge scenarios. Therefore, a continuous fusion technique is proposed, which can significantly improve tracking performance by coalescing the basic response maps using relative entropy, then we can yield more robust and reliable tracking results. Additionally, to further strengthen the proposed tracker to deal with more challenge scenarios, we exploit a robust model update strategy for our tracker. Comprehensive evaluations demonstrate that our proposed complementary ensemble tracking algorithm achieves a considerable performance improvement while running at speeds that far exceed the frame-rate requirement.

*Corresponding author: 1392135844@qq.com (Fei Wang)

2018 International Conference on Image and Video Processing, and Artificial Intelligence,
edited by Ruidan Su, Proc. of SPIE Vol. 10836, 1083618 · © 2018 SPIE
CCC code: 0277-786X/18/$18 · doi: 10.1117/12.2500635

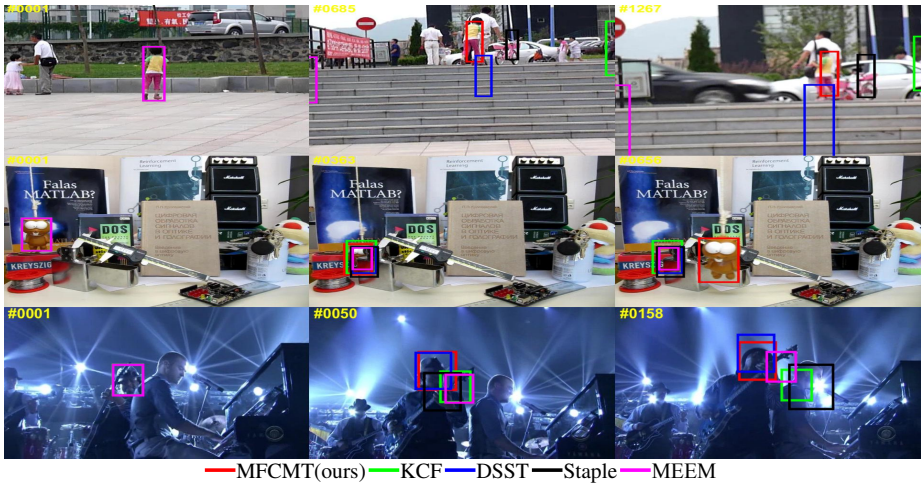


Figure 1. A comparison of our method MFCMT with four state-of-the-art trackers Staple,[6] MEEM,[7] DSST[4] and KCF[3] on three example sequences *Girl2*, *Lemming* and *Shaking* in OTB,[1,2] respectively. Best viewed in color.

2. THE PROPOSED APPROACH

In this section, we introduce the motivation and formulation of our complementary tracking model with multiple features.

2.1 Basic models

In our algorithm, we exploit DCF to construct the template features based basic model due to their lower computational burden. The key point of DCF-based trackers is that the high-dimensional augmentation of negative samples are employed as the train samples to enhance the discriminative ability of the tracking-by-detection framework.[3] Additionally, DCF can quickly find the linear model that provides the best fit to the desired correlation output in the least-squares sense. They are attractive algorithms for tracking, due to their excellent performance and high computational efficiency. We use CHs to construct the statistical features-based basic model. Fig.2 shows the framework of our method. For formulating the basic models, we consider them as several independent ridge regression problems due to the advantage of having a closed-form solution for the optimization problem, which makes it much easier to calculate.

We train several ridge regressions on the L-channel template feature maps $\mathbf{x}_{t,u} \in \mathbb{R}^{M\times N\times L}$, including raw pixels, HOGs and CNs, in frame u, and the desired correlation output $\mathbf{y}_{t,u} \in \mathbb{R}^{M\times N}$ which typically follow a Gaussian function with peak value 1 at the center. Our aim is to find an optimal correlation filter $\mathbf{w}_u$. This can be obtained by minimizing a ridge regression in the Fourier domain as:

$$\min_{\mathbf{w}_u^l} \| \sum_{l=1}^{L} \mathbf{w}_u^l \star \mathbf{x}_{t,u}^l - \mathbf{y}_{t,u}\|^2 + \lambda \sum_{l=1}^{L} \|\mathbf{w}_u^l\|^2 \tag{1}$$

where $\mathbf{w}_u^l$ refers to the channel l of the DCF $\mathbf{w}_u$ in the frame u, the star symbol $\star$ denotes circular cross-correlation, $\mathbf{x}_{t,u}^l$, $\mathbf{y}_{t,u}$ and $\mathbf{w}_u^l$ are all of size $M \times N$. Following Parseval's theorem and adopting the properties of circulant matrix in KCF,[3] the solution can be gained as:

$$\mathcal{W}_u^l = \frac{\mathcal{Y}_{t,u}^* \odot \mathcal{X}_{t,u}^l}{\sum_{l=1}^{L} (\mathcal{X}_{t,u}^l)^* \odot \mathcal{X}_{t,u}^l + \lambda} \tag{2}$$

Here $\mathcal{W}_u = \mathcal{F}(\mathbf{w}_u)$, and $\mathcal{F}(\cdot)$ denotes the Discrete Fourier Transform (DFT). We using the symbol $*$ to represent the complex conjugation $\mathcal{Y}_u^*$ of a complex number $\mathcal{Y}_u$, the symbol $\odot$ for element-wise multiplication, namely Hadamard product, and $\div$ denotes the element-wise division. Additionally, the regularization coefficient λ alleviates division-by-zero. During detection, given the template features $\mathbf{z}_{t,u+1}$ which are extracted from a basis sample patch centred at the previous position of the target in a new frame $(u + 1)$, the response map can be obtained by the following cross-correlation formulation:

$$\tilde{\mathbf{y}}_{t,u+1} = \mathcal{F}^{-1}(\sum_{l=1}^{L} \mathcal{W}_s^l \odot \mathcal{Z}_{t,u+1}^l) \tag{3}$$

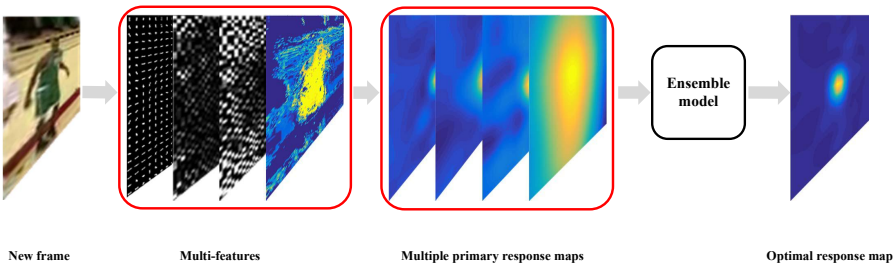


Figure 2. The framework of our proposed method.

where $\tilde{\mathbf{y}}_{t,u+1} \in \mathbb{R}^{M\times N}$ is the response map of predicted states in the new frame, and $\mathcal{F}^{-1}(\cdot)$ denotes the Inverse Discrete Fourier Transform (IDFT).

We also apply a ridge regression to the statistical features $\mathbf{x}_{s,u} \in \mathbb{R}^{M\times N\times K}$ over the foreground and background regions F_u and B_u cropped from an input image patch $\mathrm{I}_u \in \mathbb{R}^{M\times N}$ independently. This can be obtained by minimizing a ridge regression problem as:

$$\min_{\mathbf{v}_u} \sum_{\mathbf{x}_{s,u}\in \mathrm{F}_u} \|\mathbf{v}_u^T \mathbf{x}_{s,u} - \mathbf{y}_{s,u}\|^2 + \sum_{\mathbf{x}_{s,u}\in \mathrm{B}_u} \|\mathbf{v}_u^T \mathbf{x}_{s,u} - \mathbf{y}_{s,u}\|^2 \tag{4}$$

Here, $\mathbf{y}_{s,u}$ is the corresponding regression, i.e. $\mathbf{y}_{s,u} = 1$ for positive samples or $\mathbf{y}_{s,u} = 0$ for negative samples, the parameter vector $\mathbf{v}_u$ can be learned from the color histograms of the previous frames. Thus, the sparse inner product is simply a lookup matrix that $\mathbf{v}_u^T \mathbf{x}_{s,u} = \mathbf{v}_u^{\mathbf{x}_{s,u}}$, where $\mathbf{v}_u^{\mathbf{x}_{s,u}}$ indicates the element of $\mathbf{v}_u$ for which the channel index is non-zero. Then, we adopt color histogram $\mathcal{H}(\mathbf{x}_{s,u})$ to present the object pixels, the solution can be obtained as

$$\mathbf{v}_{u+1} = \begin{cases} \frac{\mathcal{H}_{\mathrm{F}_u}(\mathbf{x}_{s,u})}{\mathcal{H}_{\mathrm{F}_u}(\mathbf{x}_{s,u}) + \mathcal{H}_{\mathrm{B}_u}(\mathbf{x}_{s,u})} & \text{if } \mathbf{x}_{s,u} \in \mathrm{F}_u \cup \mathrm{B}_u \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

Finally, it can be sped up for the detection process by adopting the integral image method to obtain the statistical response map

$$\tilde{\mathbf{y}}_{s,u+1}(i,j) = \mathbf{v}_{u+1}(i,j) + \tilde{\mathbf{y}}_{s,u+1}(i,j-1) + \tilde{\mathbf{y}}_{s,u+1}(i-1,j) - \tilde{\mathbf{y}}_{s,u+1}(i-1,j-1) \tag{6}$$

where $\tilde{\mathbf{y}}_{s,u+1}(i,-1) = 0$ and $\tilde{\mathbf{y}}_{s,u+1}(-1,j) = 0$.

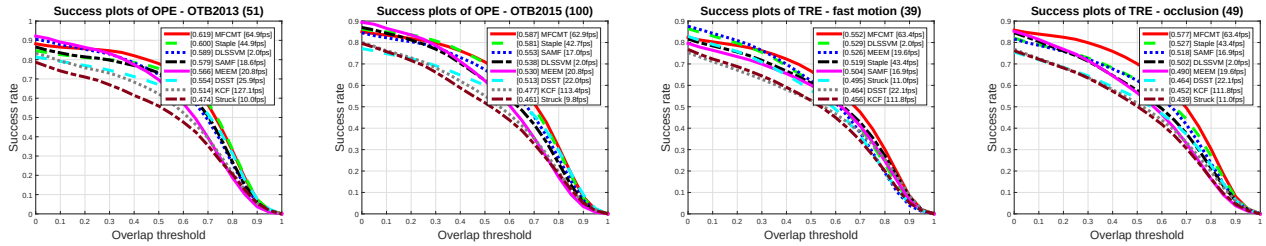
2.2 Ensemble model

In our method, we select the multiple features from both the template and statistical handcrafted features, including raw pixels, HOGs, CNs and CHs. For the i^{th} basic response maps $P_u = \{P^1, P^2, ..., P^i\}$, the single response map $P_u^l \in P_u$, $l = 1, 2, ..., i$, can be considered as a probability map, which consists of a probability distribution $p^l_{(m,n)}$, $(m,n) \in \{1,2,...,M\} \times \{1,2,...,N\}$. The probability distribution is subjected to $\sum p^l_{(m,n)} = 1$ and indicates the probability that position (m,n) is the centroid of the predicted bounding box. Consequently, in order to find the optimal response map Q_u of our ensemble model, we can minimize the relative entropy, i.e. Kullback-Leibler divergence, between each single response map P_u^l and the optimal response map Q_u. We can obtain the optimal response map Q_u by

$$\begin{aligned} &\arg\min_{Q_u} \sum_{l=1}^{i} \sum_{(m,n)} p^l_{(m,n)} \log \frac{p^l_{(m,n)}}{q_{(m,n)}} \\ &s.t. \sum q_{(m,n)} = 1 \end{aligned} \tag{7}$$

where $p_{(m,n)}$ and $q_{(m,n)}$ denote the $(m,n)^{th}$ elements of primary response maps P_u and optimal response map Q_u, respectively. To solve the above equation, we can exploit the Lagrange multiplier method.

Finally, we can get the centroid (x_u, y_u) of the predicted bounding box at a new frame by find the maximum value of the optimal response map Q_u of our ensemble model.



(a)OPE on OTB-2013 (b)OPE on OTB-2015 (c)fast motion (d)occlusion

Figure 3. Success plots of OPE, TRE and SRE for the top ten trackers in our comparison on the OTB.[1,2] The first value in the legend indicate the scores for each tracker. The last numbers in the bracket denote the speed of the trackers in mean FPS. Best viewed in color.

2.3 Model update

For the online update, we should update the models using the most reliable frames. We consider three criteria, the highest peak values of optimal response maps, template features response maps and statistic features response map, i.e. $\max(Q_u)$, $\max(\tilde{\mathbf{y}}_{t,u})$ and $\max(\tilde{\mathbf{y}}_{s,u})$, respectively. When all these three criteria of the current frame are great than their corresponding average values of the last 10 frames with predefined thresholds γ_q, γ_t and γ_s, the detected result in the current frame is considered to be highly reliable. Then we can update the proposed ensemble model with learning rate parameters for different type of features as

$$\begin{aligned} \mathcal{W}^u_{new} &= (1-\eta_t)\mathcal{W}^{u-1} + \eta_t \mathcal{W}^u \\ \mathcal{H}^u_{new} &= (1-\eta_s)\mathcal{H}^{u-1} + \eta_s \mathcal{H}^u \end{aligned} \tag{8}$$

where η_t and η_s are the learning rate which set to 0.02 and 0.04, respectively.

Additionally, we first estimate the position, then search on the scale similar to DSST[4] by adopting a multi-scale template model at the estimated object position to handle object variation. And the multi-scale template model only executed when the full response maps are more reliable.

3. EXPERIMENTS

We perform all the experiments of our proposed tracker on OTB-2013 and OTB-2015 benchmarks[1,2] which contain 51 sequences and 100 sequences, respectively. The results demonstrate that our algorithm achieves considerable results.

3.1 Implementation details

We exploit raw pixels, HOGs and CNs as template features to construct three different kinds of basic models. The cell size of HOGs is 4×4 and the orientation bin number of HOGs is 9, we also multiple the template features by a Hanning window during tracking. We exploit the distribution of color values in the RGB cube with histograms using 32 bins per channel in the statistic models, namely the color bin number of our color histograms is 32×32×32. For each frame, the search region is cropped twice the last object size around the estimated position, and the cropped search region is resized to a fixed size 150×150. In order to compute the histograms more correctly, we define 85% of the last object size as the foreground surrounded by the background patch to avoid mislabeling as suggested in DAT.[5] The update predefined thresholds γ_F, γ_T and γ_S are set to 0.5, 0.7 and 0.5, respectively. The regularization coefficient in 2 is $\lambda_T = 0.001$. The standard deviation of the desired correlation response output is set to 1/16 of the object size.The proposed tracker is implemented in MATLAB. All experiments are conducted on an Intel i5-4590 CPU at 3.3GHz with 8GB RAM.

3.2 Experiments on OTB

To evaluate our proposed tracker comprehensively, we follow the protocol of OTB.[1,2] We use one-pass evaluation (OPE) as suggested by OTB to verify the performance of our tracker. We also report the speed of the trackers in average frames per second (FPS) over all the sequences. We present both the metrics mentioned above by precision and success plots. We provide a comprehensive comparison of our approach with several state-of-the-art trackers including Staple,[6] SAMF,[8] MEEM,[7] DSST,[4] DLSSVM,[9] KCF[3] and Struck.[10] Here, Staple[6] can be considered as the simple version of our MFCMT

that does not employ neither the proposed ensemble method or the model update strategy, and SAMF[8] is another variation of our method that without none of CHs ,the ensemble method or the model update strategy.

The subplot (a) and (b) in Figure 3 illustrates the success plots of the evaluation participated trackers on both OTB-2013 and OTB-2015. Our complementary ensemble trackers (MFCMT) performs best with all evaluation metrics in the two benchmarks. The results of TRE and SRE show the robustness of our ensemble model. Our tracker leads to 27% gains in success plots on OTB-2015 compared with Struck which is the best performance tracker in original benchmark.[1] Staple and MEEM are trackers which developed based on multiple trackers, our tracker significantly improves them by 4% and 10%, respectively. KCF is the basis of our template features primary model, and DSST provides the scale estimation method, but our tracker performs favorably over them while by 20% and 13%, respectively. As for tracking speed, our tracker also run at a significantly high speed of 63 FPS, which superior to other up-to-date trackers, including MEEM, DLSSVM and Staple.

Additionally, video sequences contained in both the OTB-2013 and OTB-2015 benchmarks are annotated with 11 different attributes, such as fast motion, motion blur, occlusion and scale variation. For detailed experiments, we also evaluate our method on various challenging attributes in OTB-2015. The complete comparison on challenging scenarios of fast motion and occlusion are illustrated as the subplot (c) and (d) in Fig.3. It is clear that our approach can deal with most challenging factors at speeds that far exceed the frame-rate requirement.

4. CONCLUSIONS

In this paper, we propose a novel complementary ensemble object tracking method with multiple features. We observe that the response map obtained by a single model cannot deal with more challenges. Therefore, we adopt a relative entropy-based continuously fusion technique to further enhance the proposed tracker to deal with more challenge scenarios and improve tracking performance. Response maps obtained from multiple features-based models can be coalesced using our ensemble method. Furthermore, we also present a simple and effective model update strategy to improve both the accuracy and robust of tracking performance. We conduct evaluations on modern online tracking benchmarks. The evaluation results demonstrate that the proposed method is more effective and faster than several state-of-the-art trackers.

ACKNOWLEDGMENTS

This work is supported by the Science and Technology Planning Project of Guangdong Province, China (Grant No. 2016B090918047).

REFERENCES

[1] Wu, Y., Lim, J., and Yang, M.-H., "Online object tracking: A benchmark," in [*CVPR*], 2411–2418, IEEE (2013).
[2] Wu, Y., Lim, J., and Yang, M.-H., "Object tracking benchmark," *IEEE TPAMI* **37**, 1834–1848 (September 2015).
[3] Henriques, J. F., Caseiro, R., Martins, P., and Batista, J., "High-speed tracking with kernelized correlation filters," *IEEE TPAMI* **37**, 583–596 (March 2015).
[4] Danelljan, M., Häger, G., Khan, F. S., and Felsberg, M., "Discriminative scale space tracking," *IEEE transactions on pattern analysis and machine intelligence* **39**, 1561–1575 (August 2017).
[5] Possegger, H., Mauthner, T., and Bischof, H., "In defense of color-based model-free tracking," in [*CVPR*], 2113–2120, IEEE (2015).
[6] Bertinetto, L., Valmadre, J., Golodetz, S., Miksik, O., and Torr, P. H., "Staple: Complementary learners for real-time tracking," in [*CVPR*], 1401–1409, IEEE (2016).
[7] Zhang, J., Ma, S., and Sclaroff, S., "Meem: robust tracking via multiple experts using entropy minimization," in [*ECCV*], 188–203, Springer (2014).
[8] Li, Y. and Zhu, J., "A scale adaptive kernel correlation filter tracker with feature integration.," in [*ECCV*], 254–265, Springer (2014).
[9] Ning, J., Yang, J., Jiang, S., Zhang, L., and Yang, M.-H., "Object tracking via dual linear structured svm and explicit feature map," in [*CVPR*], 4266–4274, IEEE (2016).
[10] Hare, S., Golodetz, S., Saffari, A., Vineet, V., Cheng, M.-M., Hicks, S. L., and Torr, P. H., "Struck: Structured output tracking with kernels," *IEEE TPAMI* **38**, 2096–2109 (October 2016).