Proceedings of the International Conference on Machine Vision and Machine Learning
Prague, Czech Republic, August 14-15, 2014
Paper No. 60

Robust Tracking via Multi-level Multi-feature Templates

Xueli Bai, Lianyang Ma, Rui Zhang
Institute of Image Communication and Information Processing, Shanghai Jiao Tong University
655 Panyu Rd. Xuhui, Shanghai, China
specially@sjtu.edu.cn, malianyang2009@sjtu.edu.cn, zhang_rui@sjtu.edu.cn

Abstract **-** In this paper, we propose a hierarchical and multi-feature appearance model for object tracking. An object is represented by a hierarchical tree model where the nodes represent object parts. An appearance template is designed for each part and described by a multi-feature set that combines multiple observations, such as colour, intensity, texture and edge. The hierarchical model is represented by a three-level spatial pyramid to characterize the target respectively in high-level, mid-level and low-level representations. We demonstrate the multi-level multi-feature template model (*MMTM*) for object tracking in a generative learning formulation. In this formulation, the weights of different features in each template and the contribution of different templates are learnt to distinguish the object and background. The proposed model is tested on a variety of benchmarks involving object deformation, background cluttering and motion blur. The experimental results demonstrate that our approach can achieve superior performance than previous tracking methods.

Keywords**:** Visual tracking, Multi-feature, Multi-level, Generative learning, Information projection.

1. Introduction

Object tracking is an important research topic in computer vision, with wide applications in video surveillance, human-computer interface, medical imaging and so on (Wu, et al., 2013). Due to the dynamic tracking circumstance in various visual applications, tracking algorithm is required to deal with practical challenges such as illumination change, background cluttering, motion blur, non-rigid object deformation and object pose variation. Therefore, substantial trackers (Zhong, et al., 2012; Yoon et al., 2012) have been proposed to improve tracking accuracy and robustness, which are generally composed of four modules: object initialization, appearance modelling, motion estimation and object localization (Li et al., 2013). In this paper we mainly focus on the appearance modelling since it is significant and effective in improving tracking performance.

There exist wide varieties of appearance models with different visual representations and/or statistical modelling techniques. Literature (Li et al., 2013) summarizes that high-level representation (Hare et al., 2011) is simple, computational efficient but susceptible to global appearance changes, while low-level template representation is effective but may be over-flexible and cause too many detection hypotheses in cluttered image (Lin et al., 2007). Inspired by the demonstration that "deep structure" giving richer descriptions of shape and appearance (Zhu et al., 2010), we argue that high-level, mid-level and low-level representations all contain important discriminative information. Therefore, we utilize all the templates generated from three levels to represent a target object. In addition to using template representations for object tracking, many researchers focus on exploiting multiple features (Yoon et al., 2012). Current approaches usually combine different features by discriminative feature selection, and use only one feature to represent a patch in the template (X. Liu et al., 2011) or to describe a tracker in the tracker set (Kwon et al., 2010). In this paper, we represent each patch or template by multiple features since all of them may contribute to distinguish the target from the background. Furthermore, the contribution of each feature and each template should be measured as according to their discriminative power changes under different circumstances. Therefore, inspired by information projection (Si & Zhu, 2012), we formulate multi-level multi-feature templates for object tracking as a generative learning problem, that is to learn an underlying probability model for the tracked object by maximizing the difference between it and the probability distribution of background images. This formulation is different from previous visual tracking approaches such as Bayesian framework (Yoon, et al., 2012), sparse coding (Jia, et al., 2012), and sparse principle component analysis (Kwon et al., 2010).

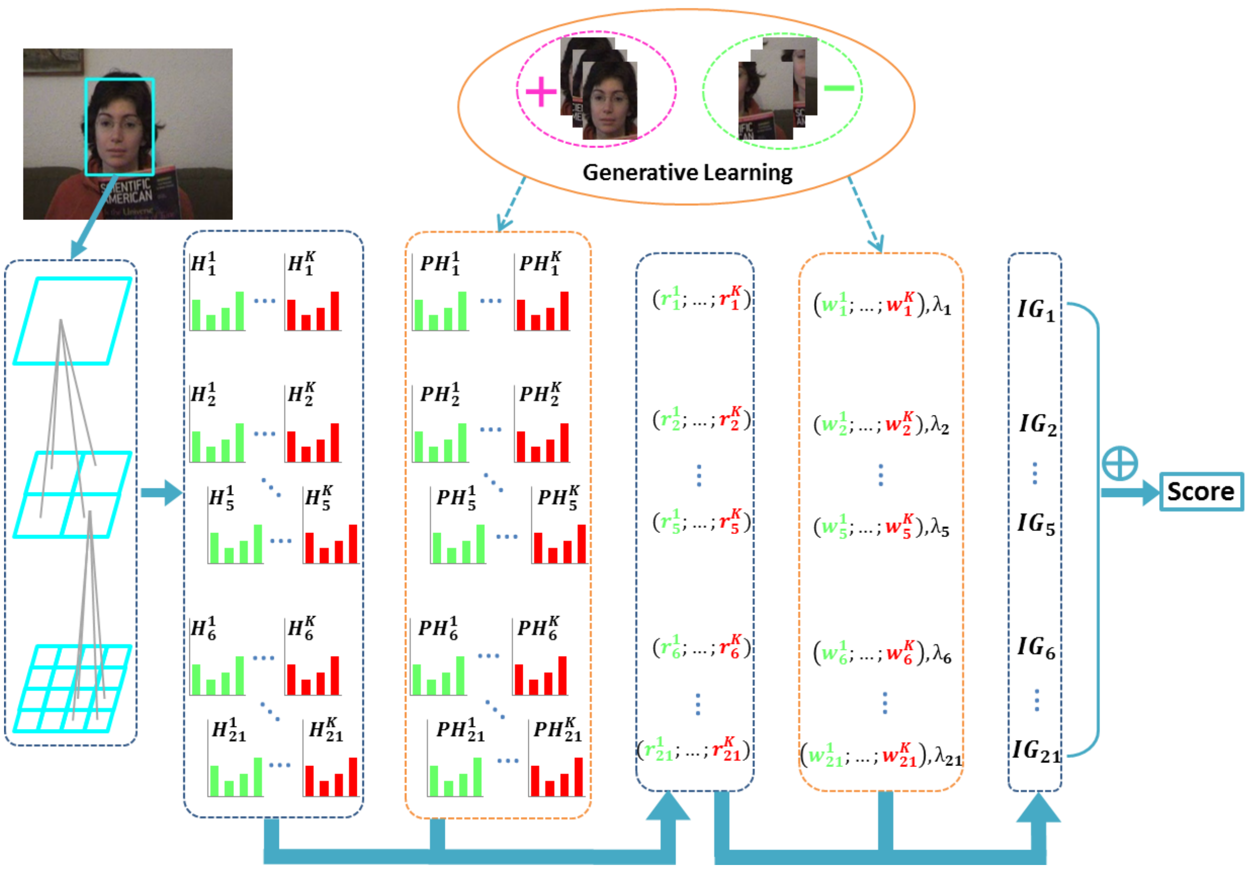


Fig.1 The method overview.

Our contributions are mainly summarized in two aspects: 1) we propose a multi-level multi-feature appearance templates to describe the tracked object. The proposed model is represented by a pyramid structure which encodes high-level, middle level and low level information of the object and contributes to increase tracking accuracy. Besides, the multiple feature representation enables the model to utilize all the visual information for each template, which improves tracking robustness in dynamic conditions. 2) We exploit a generative learning formulation to learn the parameters of the proposed model, i.e. the template weights and feature weights.

2. Proposed Method

The goal of visual tracking is to determine the object position in each frame of video sequence. In this paper, we achieve this goal by finding out the location with highest score in the candidate location set, which is generated from the surroundings of the last observed object position. As shown in Fig.1, we represent each candidate bounding box with a three level spatial pyramid which is a quad-tree like representation. A template is designed for each node. Templates in three scales describe the high-level, mid-level and low-level information respectively. Totally K features in histogram representations $\left(\boldsymbol{H}_i^1; \ldots; \boldsymbol{H}_i^K\right)$ are extracted for i^{th} template T_i to combine different appearance cues including colour, texture, edge and intensity. We formulate the multi-level multi-feature templates for object tracking as a generative learning problem based on the information projection principle (Si & Zhu, 2012). In this formulation, two probability models for positive and negative samples are parameterized by multiple templates and multiple features. The parameters are learnt by maximizing information gain to boost discrimination, i.e. the feature prototypes $\left(\boldsymbol{PH}_i^1;\ \ldots;\boldsymbol{PH}_i^K\right)$, feature weights $\boldsymbol{w}_i$ and template weight λ_i. In the testing phrase, we first calculate the feature responses $\left(r_i^1;\ \ldots;\ r_i^K\right)$ by comparing the extracted features with the feature prototypes on T_i. The score of each template (i.e. IG_i) is computed with the learnt parameters $\left(w_i^1;\ \ldots;w_i^K\right)$ and λ_i. The score of each location is obtained by summing all the template scores $(IG_1;\ \ldots;\ IG_{21})$.

3. The Multi-scale Multi-feature Templates

3. 1. The Multi-scale Templates and Object Representation

In this paper, a hierarchical tree-structure model for object tracking is proposed to cover the object configuration information in different level. Specifically, a three level quad-tree is designed for the object representation. As shown in Fig.1, the root node in the first level describes the high-level information of the entire object. In the second level, 4 child nodes segment the object to 2×2 grid and each of them represents one part of the object, which describe the mid-level information. Similarly, there are 16 nodes in a 4 by 4 grid layout in the third level, which represent the low-level information. Thus, there are totally 21 nodes to represent the target in different levels and locations. Each node in the tree is described by a template. We formulate the template set *B* to represent the object by assembling all templates:

$$B=\left(T_1,T_2,\ldots,T_{21}\right). \tag{1}$$

It is noted that this hierarchical model will degenerate to the high-level representation by only using the first level, which may ignore local information. And if we merely use the third level in this model, it will become a low-level template representation, which may be hard to capture global information. Generally speaking, the deeper hierarchical model may encode more information. However, it also requires extra computational cost. Thus, a three level hierarchical model is adopted in this work as a trade-off between effectiveness and efficiency.

3. 2. Multi-Feature Appearance Model for a Template

In addition to the spatial pyramid, which specifies multi-level spatial configuration information, we construct an appearance model for each template in the template set. For a template T_i, its appearance model ρ_i is represented by two terms as:

$$\rho_i=\left(\boldsymbol{PH}_i,\ \boldsymbol{w}_i\right), \tag{2}$$

$$\boldsymbol{PH}_i=\left(\boldsymbol{PH}_i^1;\ldots;\boldsymbol{PH}_i^K\right), \tag{3}$$

$$\boldsymbol{w}_i=\left(w_i^1;\ldots;w_i^K\right), \tag{4}$$

where vector $\boldsymbol{PH}_i$ denotes the prototype set containing *K* feature prototypes of this template, and vector $\boldsymbol{w}_i$ denotes the weights of each feature in computing the template response. Specifically, the feature prototype corresponds to a certain typical parametric appearance pattern (e.g., texture, colour, etc.) of each template in our model. In general, we can use any off-the-shelf visual features for a template, e.g. HOG (Dalal & Triggs, 2005), colour histogram, local binary pattern (LBP) (Ojala, et al., 2002) and intensity.

In this paper, since all the features are represented in histogram format, we compute a scalar-valued feature response r_i^k ($k\in\{1,\ldots,K\}$) of an incoming feature vector $\boldsymbol{H}_{\mathrm{i}}^k$ to the pro-type $\boldsymbol{PH}_{\mathrm{i}}^k$ by measuring histogram correlation between them: $r_i^k = C\left(\boldsymbol{H}_i^k, \boldsymbol{PH}_i^k\right)$, which is shown as Eq.(5)[1],

$$C(\boldsymbol{H}_1, \boldsymbol{H}_2) = \frac{\sum_{j=1}^{J}\left(\boldsymbol{H}_1(j) - \overline{\boldsymbol{H}_1}\right)\sum_{j=1}^{J}\left(\boldsymbol{H}_2(j) - \overline{\boldsymbol{H}_2}\right)}{\sqrt{\sum_{j=1}^{J}\left(\boldsymbol{H}_1(j) - \overline{\boldsymbol{H}_1}\right)^2 + \sum_{j=1}^{J}\left(\boldsymbol{H}_2(j) - \overline{\boldsymbol{H}_2}\right)^2}}$$
$$\overline{\boldsymbol{H}_1} = \frac{1}{J}\sum_{j=1}^{J}\boldsymbol{H}_1(j),\ \overline{\boldsymbol{H}_2} = \frac{1}{J}\sum_{j=1}^{J}\boldsymbol{H}_2(j), \tag{5}$$

where $\boldsymbol{H}_{\text{i}}(j)$ and J refer to the value of j^{th} bin and the dimension of $\boldsymbol{H}_{\text{i}}$ respectively, while $\overline{\boldsymbol{H}_{\text{i}}}$ is the mean value of the histograms. As we use multiple features to represent a single template, the feature response R_i of whole template T_i is the weighted sum of all the individual feature responses $\boldsymbol{r}_i = \left(r_i^1;\ \ldots; r_i^K\right)$:

$$R_i = \sum_{k=1}^{K} w_i^k r_i^k = \boldsymbol{w}_i^T \boldsymbol{r}_i$$
$$s.t.\quad w_i^k \geq 0,\ \sum_{k=1}^{K} w_i^k = 1, \tag{6}$$

where $\boldsymbol{w}_i = \left(w_i^1;\ \ldots; w_i^K\right)$ is the weights of feature responses for T_i, and all the weights in this vector sum to one while the value of them is non-negative.

3. 3. Generative Learning for Template Parameter

To build an adaptive and informative appearance model for tracking the object in the dynamic and cluttered background, we adopt a generative learning algorithm to determine the parameters for the templates and their feature weights, which is generalized from the *information projection principle* (Si & Zhu, 2012). Let $D^+ = \{I_1^+, I_2^+, \ldots, I_N^+\}$ denotes the positive samples, which are composed of multiple tracking results from previous frames to alleviate over-fitting of the learned model, and $D^- = \{I_1^-, I_2^-, \ldots, I_M^-\}$ indicates the negative samples, consisting of the training images obtained from the object surroundings in previous frame. Let $f(I)$ be the underlying probability distribution for the tracking target, and $q(I)$ be the reference (background) distribution, where D^+ is sampled from $f(I)$ and D^- sampled from $q(I)$. Our objective is to learn a probability model $p(I; MMTM)$ that approach $f(I)$ starting from $q(I)$. To simplify notation in the following discussion, we denote the template response of T_i on D^+ and D^- by $\{\boldsymbol{w}_i^T \boldsymbol{r}_{i,n}^+\}_{n=1}^{N}$ and $\{\boldsymbol{w}_i^T \boldsymbol{r}_{i,m}^-\}_{m=1}^{M}$, respectively.

For the *MMTM* of the target object, the model space $\Omega(MMTM)$ is defined as: $\Omega(MMTM) = \{p(I; MMTM) | E_p[R_i] = E_f[R_i], \forall i\}$, where $E_p[r_i] = E_f[r_i]$ implies the constraint that the model expectation of each template response from the positives samples is expected to match the empirical statistics. Based on the max entropy principle that $\hat{p} = argmin_{p\in\Omega_p(MMT)} KL(p\|q)$, a factorized log-linear model (Si & Zhu, 2012) is derived as in Eq. (7) because there is no overlapping of templates in each pyramid level:

$$\hat{p}(I; MMTM) = q(I)\prod_{i=1}^{I_T}\left[\frac{1}{z_i}\exp(\lambda_i R_i)\right], \tag{7}$$

where λ_i and z_i denote the parameters of template weight and normalizing factor for T_i in *MMTM*,

[1] This equation is defined in http://www.opencv.org.cn/opencvdoc/2.3.2/html.

respectively, and I_T (equals to 21) refers to the number of templates in the model.

We formulate the learning objective as a regularized information gain of *MMTM* similar to (Si & Zhu, 2012), shown as Eq. (8):

$$IG(MMTM) = KL(f \parallel q) - KL(f \parallel \hat{p}) - M(MMTM)$$
$$= \sum_{i=1}^{I_T} \left\{ \lambda_i E_f \left[\boldsymbol{w}_i^T \boldsymbol{r}_i \right] - \log z_i - \frac{1}{2} \alpha \boldsymbol{w}_i^T \boldsymbol{w}_i - \frac{1}{2} \beta \lambda_i^2 \right\} \tag{8}$$
$$s.t.\ w_i^k \geq 0,\ \sum_{k=1}^{K} w_i^k = 1$$

where $[KL(f \| q) - KL(f \| \hat{p})]$ is the difference of Kullback-Leibler divergence of the learned model $\hat{p}(I; MMTM)$ approaching $f(I)$ relative to $q(I)$, which measures information-theoretical improvement. $M(MMTM)$ accounts for the regularization term on model complexity, in which α and β denote the trade-off parameters on shrinking the weight λ_i and punishing extreme heterogeneity of $\boldsymbol{w}_i$, respectively. Thus, the optimal object appearance model is learned by maximizing each template's information gain IG_i.

Similar to (Si & Zhu, 2012), we calculate the feature prototype of each T_i by the average of feature histograms extracted from all the positive samples on the same template. Then, for each template, as there are two parameters to calculate: λ_i and $\boldsymbol{w}_i$, and neither of them has analytic solution, we gain the optimal values by alternating optimization, during which we need to get the solution of λ_i with a fixed $\boldsymbol{w}_i$, as well as the solution of $\boldsymbol{w}_i$ with a fixed λ_i.

For optimizing λ_i with $\boldsymbol{w}_i$ fixed, the optimal value (λ_i^*, z_i^*) is obtained by solving $\frac{\partial IG}{\partial \lambda_i} = 0$ as Eq. (9)

$$\left(\lambda_i^*, z_i^*\right):\ E_f \left[\boldsymbol{w}_i^T \boldsymbol{r}_i \right] - E_{\hat{p}} \left[\boldsymbol{w}_i^T \boldsymbol{r}_i \right] = \beta \lambda_i, \tag{9}$$

where $E_f[\boldsymbol{w}_i^T \boldsymbol{r}_i] \approx \frac{1}{N} \sum_{n=1}^{N} \boldsymbol{w}_i^T \boldsymbol{r}_{i,n}^+$ is empirical expectation by the mean template response on positive samples. The term of $E_{\hat{p}}[\boldsymbol{w}_i^T \boldsymbol{r}_i]$ is approximately estimated using the response on D^-: $E_{\hat{p}}[\boldsymbol{w}_i^T \boldsymbol{r}_i] = E_q \left[\frac{1}{z_i} exp(\lambda_i \boldsymbol{w}_i^T \boldsymbol{r}_i) \boldsymbol{w}_i^T \boldsymbol{r}_i \right] \approx \frac{1}{M} \sum_{m=1}^{M} \left[\frac{1}{z_i} exp\left(\lambda_i \boldsymbol{w}_i^T \boldsymbol{r}_{i,m}^-\right) \boldsymbol{w}_i^T \boldsymbol{r}_{i,m}^- \right]$, where $z_i \approx \frac{1}{M} \sum_{m=1}^{M} exp\left(\lambda_i \boldsymbol{w}_i^T \boldsymbol{r}_{i,m}^-\right)$. On computation, we can solve Eq.(9) by Newton method (Si & Zhu, 2012).

For optimizing $\boldsymbol{w}_i$with a fixed λ_i, we select two elements in $\boldsymbol{w}_i$ to be updated in each iteration while the others are fixed. The optimal value is obtained by iteratively traversing over all pairs of elements in $\boldsymbol{w}_i$ and we adopt the gradient projection method to optimize these two elements until the objective in Eq. (8) does not increase.

In the testing phrase, we exploit these parameters to compute the location score S_L of a certain candidate area A_c, as $S_L(A_c) = \sum_{i=1}^{21} \left(\lambda_i \boldsymbol{w}_i^T \boldsymbol{r}_{i,A_c} - \log z_i\right)$, where λ_i, $\boldsymbol{w}_i$, and z_i are acquired in the generative learning progress, and $\boldsymbol{r}_{i,A_c}$ is the feature response of i^{th} template in A_c. The target is located by the candidate area with highest score.

4. Experimental Results

4. 1. Implementation Details

In our method, we use multiple features to represent the appearance model for each template including the histogram on the colour plane (RGB and YUV), original intensity, LBP and HOG. We empirically set the value of regularization terms α and β in Eq. (8) to be 3.0 and 0.1 respectively. The positive sample set is updated similar to the template set updating method in ASLA (Jia, et al., 2012); the negative samples are collected only in the latest frame. The search region is defined by enlarging the last detected bounding box by one and a half of the object size in each direction.

4. 2. Quantitative Results

We evaluate our proposed method on the latest video benchmark (Wu, et al., 2013), where each video sequence contains one or multiple challenges. We also compare our method with five different trackers estimated in the benchmark: SCM (Zhong, et al., 2012), ASLA (Jia, et al., 2012), MTT (Zhang, et al., 2012), Struck (Hare, et al., 2011) and LSK (B. Liu, et al., 2011). The evaluation metric we use is the same as introduced in the benchmark (Wu, et al., 2013): the overlap ratio of bounding boxes. Specifically, the score of the ground-truth bounding box b_a and the tracked bounding box b_t is defined as $S = \frac{|b_a \cap b_t|}{|b_a \cup b_t|}$, where ∩ and ∪ denote the intersection and union regions of the them, and $|\cdot|$ stands for the number of pixels in the region. The measurement of performance on a sequence of frames is given by the number of successfully tracked frames whose overlap S is greater than a certain threshold. The ratio of successful frames at different thresholds varied from 0 to 1 is given by the success plot. We compare our method with the state-of-arts on the overall dataset and the non-rigid object deformations (DEF), motion blur (MB), background cluttering (BC), out-of-plane rotation (OPR) and fast motion (FM) sub-datasets, which are labelled by the benchmark. The plots are shown in Fig.2.

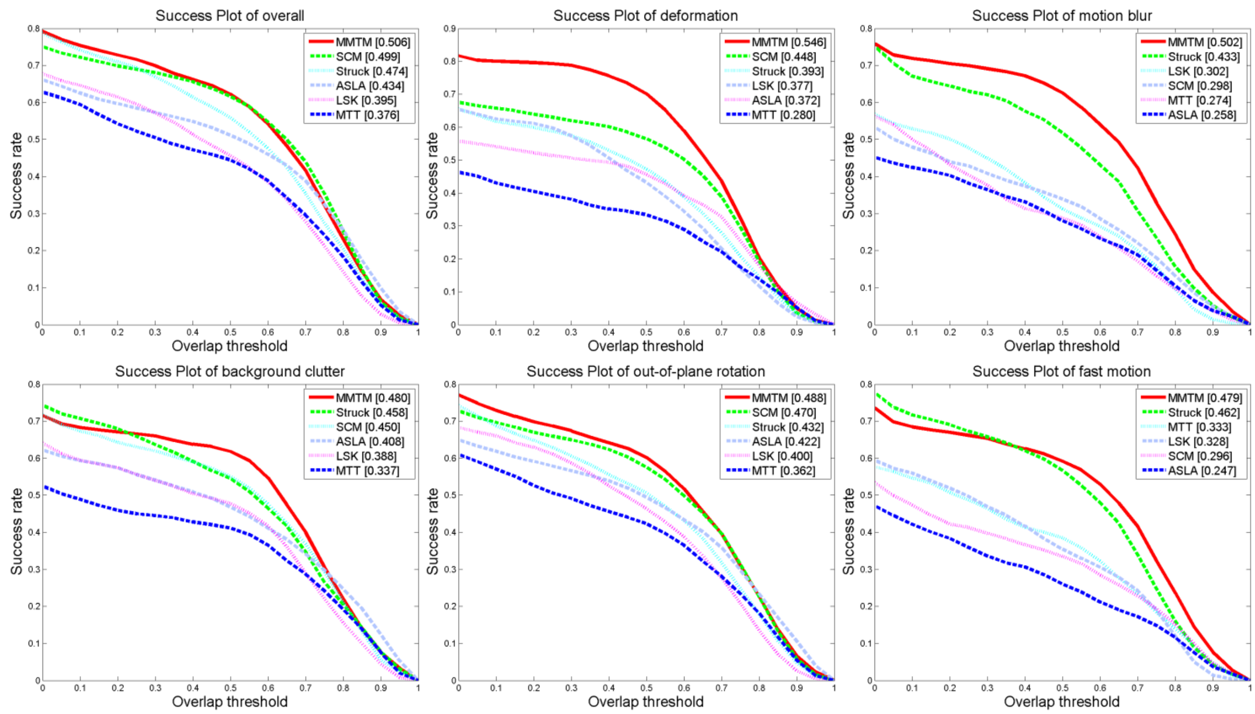


Fig.2 Comparison of success plots on our method (i.e., *MMTM*) and previous trackers. Each panel from left to right and up to bottom corresponds to the results on overall dataset and DEF, MB, BC, OPR, FM sub-datasets.

Table. 1. Results on the dataset and its sub-datasets.

	MTT	LSK	ASLA	Struck	SCM	Ours
Overall	0.376	0.395	0.434	0.474	0.499	**0.506**
DEF	0.280	0.377	0.372	0.393	0.448	**0.546**
MB	0.274	0.302	0.258	0.433	0.298	**0.502**
BC	0.337	0.388	0.408	0.458	0.450	**0.480**
OPR	0.362	0.400	0.422	0.432	0.470	**0.488**
FM	0.333	0.328	0.247	0.462	0.296	**0.479**

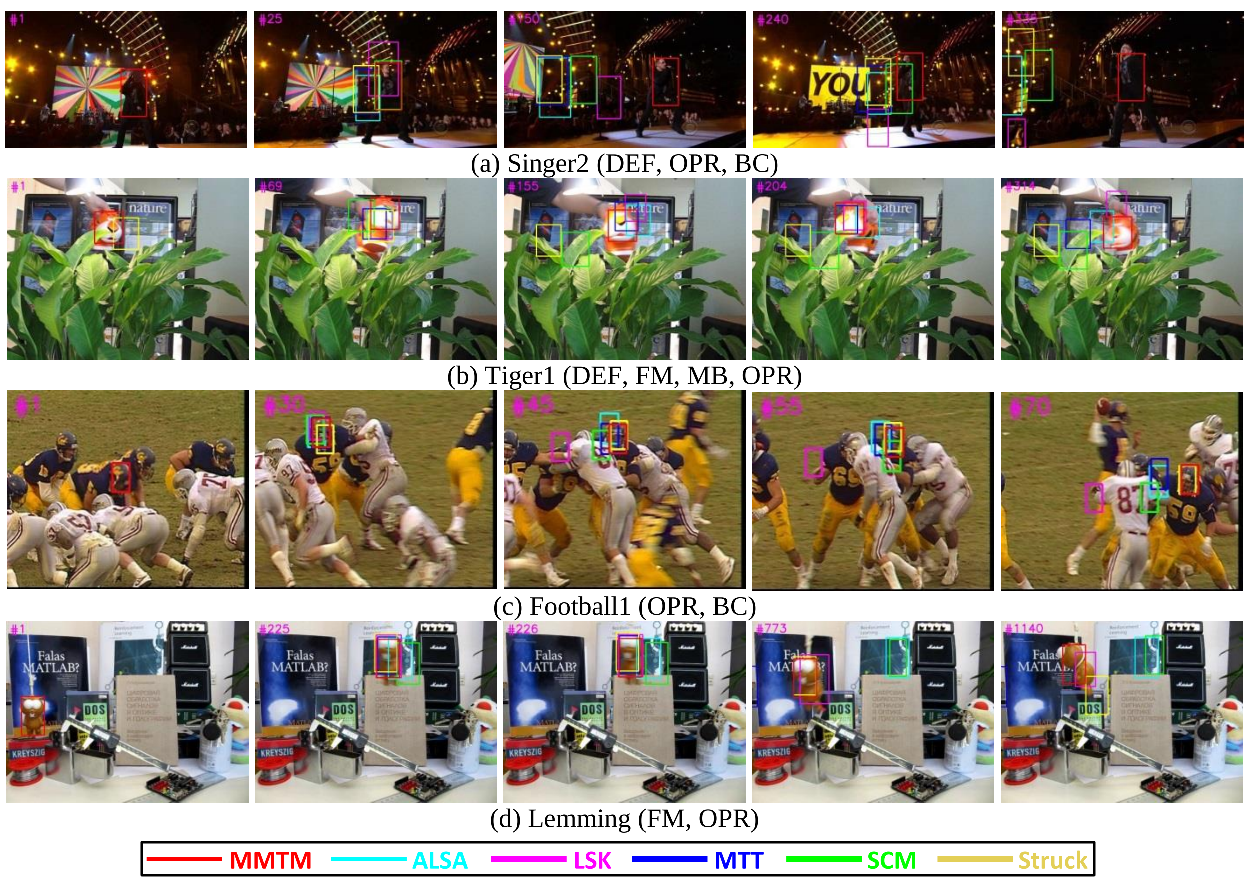
To avoid the unfairness caused by using one success rate at a certain threshold for evaluating all trackers, the benchmark (Wu, et al., 2013) uses the area under curve (AUC) of each plot to represent the performance of a tracker on the dataset. The AUC scores of our tracker and the reference trackers are reported in Table. 1. It is clear that our method outperforms previous trackers on the whole dataset and its sub-datasets. These sub-datasets are selected due to their commonness in real-world tracking problem and they correspond to different challenges in visual object tracking.

4. 3. Qualitative Results and Analysis

Deformation: As shown in Fig.3 (a) and (b), both tracking progresses encounter object non-rigid deformation, where many other trackers fail in tracking the target. The possible reason for our method performs well during these situations is that it employs the appearance information in multiple levels. The high-level template defines the object silhouette, while the middle-level and low-level templates match the discriminative parts, and combing them can detect the target precisely.

Motion Blur and Fast Motion: Generally, motion blur accompanies with fast motion, as illustrated in Fig.3 (b) and (d). Due to the complementary characteristic of multiple feature representation for each template, the proposed method is likely to have enough distinctive power for tracking when the target is blurred, while other methods might endure drift problem.

Background cluttering: Shown in Fig.3 (c), the background includes many cluttering objects with similar appearance to the tracking target. While other methods use the motion estimation to eliminate these disturbing candidates, our appearance model is robust enough to track the object accurately with distinguishing information gathered from the generative learning against these negative samples.



(a) Singer2 (DEF, OPR, BC)

(b) Tiger1 (DEF, FM, MB, OPR)

(c) Football1 (OPR, BC)

(d) Lemming (FM, OPR)

Fig. 3. Tracking results of different algorithms (best viewed in colour).

Out-of-plane rotation: As shown in Fig.3 (b) and (d), our method overcomes the object OPR challenge while other methods encounter result drift or even target lose. This is possibly achieved by the persistence of distinctive information due to the multi-level multi-feature appearance model, and the tolerance of rotation benefits from the learning progress in each frame.

4. Conclusion

In this paper, we propose a novel multi-level multi-feature template based appearance model for object tracking. A tracked object is represented by combining the templates in different levels. For each template in the model, we characterize it by using multiple features simultaneously to exploit their complementary characteristics. Furthermore, we demonstrate a generative learning formulation to learn the parameters of each template and different features. The proposed method is able to consider both the changes of tracking object and its background, and thus tracks the object accurately and robustly. The experimental results demonstrate that the proposed tracker is robust to various challenges and outperforms previous tracking methods.

Acknowledgements

We would like to thank Dr. Jun Zhu for the valuable discussion and the help in proof reading. This work is supported by NSFC (61071155） and 973 Program (2010CB731401, 2010CB731406).

References

Dalal, N., & Triggs, B. (2005). Histograms of oriented gradients for human detection. In CVPR 2005. IEEE Computer Society Conference on (Vol. 1, pp. 886-893). IEEE.

Hare, S., Saffari, A., & Torr, P. H. (2011). Struck: Structured output tracking with kernels. In Computer Vision (ICCV), 2011 IEEE International Conference on (pp. 263-270). IEEE.

Jia, X., Lu, H., & Yang, M. H. (2012). Visual tracking via adaptive structural local sparse appearance model. In CVPR, 2012 IEEE Conference on (pp. 1822-1829). IEEE.

Kwon, J., & Lee, K. M. (2010, June). Visual tracking decomposition. InComputer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on(pp. 1269-1276). IEEE.

Li, X., Hu, W., Shen, C., Zhang, Z., Dick, A., & Hengel, A. V. D. (2013). A survey of appearance models in visual object tracking. ACM Transactions on Intelligent Systems and Technology (TIST), 4(4), 58.

Lin, Z., Davis, L. S., Doermann, D., & DeMenthon, D. (2007). Hierarchical part-template matching for human detection and segmentation. InComputer Vision, 2007. ICCV 2007. IEEE 11th International Conference on (pp. 1-8). IEEE.

Liu, B., Huang, J., Yang, L., & Kulikowsk, C. (2011). Robust tracking using local sparse appearance model and k-selection. In CVPR, 2011 IEEE Conference on (pp. 1313-1320). IEEE.

Liu, X., Lin, L., Yan, S., Jin, H., & Jiang, W. (2011). Adaptive object tracking by learning hybrid template online. Circuits and Systems for Video Technology, IEEE Transactions on, 21(11), 1588-1599.

Ojala, T., Pietikainen, M., & Maenpaa, T. (2002). Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. TPAMI, IEEE Transactions on, 24(7), 971-987.

Si, Z., & Zhu, S. C. (2012). Learning hybrid image templates (hit) by information projection. Pattern Analysis and Machine Intelligence, IEEE Transactions on,34(7), 1354-1367.

Wu, Y., Lim, J., & Yang, M. H. (2013, June). Online object tracking: A benchmark. In CVPR, 2013 IEEE Conference on (pp. 2411-2418). IEEE.

Yoon, J. H., Kim, D. Y., & Yoon, K. J. (2012). Visual tracking via adaptive tracker selection with multiple features. In Computer Vision–ECCV 2012 (pp. 28-41). Springer Berlin Heidelberg.

Zhong, W., Lu, H., & Yang, M. H. (2012). Robust object tracking via sparsity-based collaborative model. In CVPR, 2012 IEEE Conference on (pp. 1838-1845). IEEE.

Zhu, L., Chen, Y., Yuille, A., & Freeman, W. (2010). Latent hierarchical structural learning for object detection. In CVPR, 2010 IEEE Conference on (pp. 1062-1069). IEEE.

Zhang, T., Ghanem, B., Liu, S., & Ahuja, N. (2012). Robust visual tracking via multi-task sparse learning. In CVPR, 2012 IEEE Conference on (pp. 2042-2049). IEEE.