VISUAL TRACKING VIA STRUCTURAL PATCH-BASED DICTIONARY PAIR LEARNING

Tao Zhou[1], *Fanghui Liu*[1], *Harish Bhaskar*[2], *Jie Yang*[*1], *Lei Chen*[3], *Ping Cai*[4]

[1]Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, China
[2]Dept. of Elec. & Comp. Engg., Khalifa Univ. of Science Technology and Research, Abu Dhabi, U.A.E
[3]School of Computer Science, Nanjing University of Posts and Telecommunications, China
[4]Department of Instrument Science and Engineering, Shanghai Jiao Tong University, China

ABSTRACT

In this paper, a novel visual tracking framework based on Structural Patch-based Dictionary Pair Learning (SPDPL), is proposed. The proposed representation model encapsulates partial and spatial structural variations of the target through a novel dictionary learning scheme. The proposed method facilitates learning a robust and discriminative dictionary by considering all patches from the same part of the target region as one class, thus transforming the tracking problem into a multi-class classification and reconstruction task. Finally, a simple yet effective observation model is designed to obtain the most optimal candidate during tracking. Systems experiments of the proposed tracking algorithm on the Object Tracker Benchmark (OTB) dataset have demonstrated improvements against several other state-of-the-art trackers.

Index Terms— Dictionary learning,Visual tracking, Structural patch-based dictionary pair learning

1. INTRODUCTION

Research in the topic of visual target tracking has evolved due to its applications in behavior analysis, event and activity recognition, video surveillance, and human-computer interaction [1, 2, 3]. Despite increasing interest in target tracking [4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14], factors such as illumination variation, occlusion, fast motion, rotation and deformation (as shown in Fig.1), etc, have continued to challenge the development of efficient, robust and accurate tracking algorithms.

Dictionary learning has been used for the sparse representation of visual signals in different application such as image restoration, object classification and face recognition, among others. In this context, DL has emerged as a critical part of achieving good tracking performance. Some DL based trackers such as [15, 16, 17, 18] have exploited a binary classification formulation of the tracking problem in order to improve tracking performance. However, most available DL-

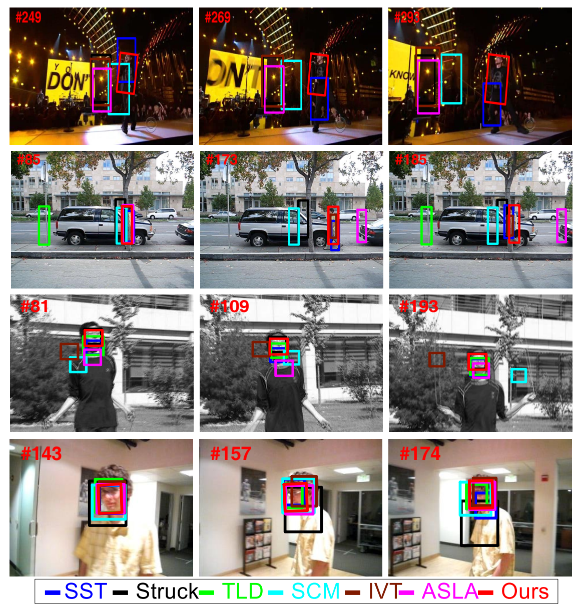

Fig. 1. Tracking results of SST, Struck, TLD, SCM, IVT, ASLA and our tracker on 5 challenging sequences including Skating1 (illumination), David3 (occlusion), Jumping (fast motion) and David (rotation).

based trackers have demonstrated difficulties in discriminating ambiguities between visual patterns, especially, if the target shares geometric and appearance similarities with the clutter in the background. The limitations within DL-based tracking can be attributed to the following inadequacies. First, some existing methods select those patches from the target region to represent the positive dictionary, while patches far away from the target are chosen to represent the negative dictionary [6]. In selecting candidate samples in this manner, the dictionaries fail to faithfully represent the estimated target candidate, particularly when the target is subjected to changes in pose and appearance. In addition, the coefficients obtained over such dictionaries remain non-discriminative of the target from its corresponding background. In addition, it has also been proven that negative samples thus selected, introduce outliers, which forces that tracker to drift away from the target. Further, some tracking algorithms keep the dictionary unchanged throughout the tracking procedure as in [19] or

*Corresponding author: jieyang@sjtu.edu.cn (J. Yang). This research is partly supported by China Postdoctoral Science Foundation (No.2016M601597, 2015M581794), NSFC (No: 61572315, 6151101179), 863 Plan (No. 2015AA042308), and National Science Foundation of Jiangsu Province (BK20161516).

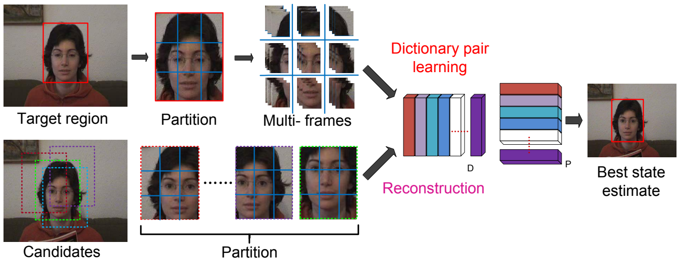

Fig. 2. Tracking framework of the proposed tracking algorithm.

update it directly by only using new observation samples as dictionary items [6, 4]. Without an appropriate update to the learnt dictionary, like in static dictionaries or through heuristic dictionary updates, trackers usually succumb to variations in pose and appearance, that limits the tracking performances and eventually results in drifts.

In this paper, a novel Structural Patch-based Dictionary Pair Learning (SPDPL) tracking algorithm is proposed that encapsulates partial and spatial structural information of the target into the dictionary learning process in order to overcome some of the limitations of existing algorithms highlighted above. According to Fig.2, which illustrates the process flow of the proposed tracking technique, local image patches within the target regions are sampled to form the training sample set. However, those patches which are drawn from the same part of the target region are regarded as one class. A novel dictionary learning scheme that extends the conventional dictionary pair learning model in [20], dubbed as SPDPL, is incorporated to ensure that the dictionary built is discriminative thus enabling improvements in the tracking performance. In addition, an enhanced observation model is designed in order to facilitate optimal candidate estimation.

2. STRUCTURAL PATCH-BASED DICTIONARY PAIR LEARNING

2.1. Problem description

Given a set of target template images $\mathbf{T} = [\mathbf{T}_1, \mathbf{T}_2, \cdots, \mathbf{T}_N]$, each template is divided into M local patches based on a spatial layout constraint (as shown in Fig.2). Each local patch can represent one fixed part of the target region, hence the local patches altogether represent the complete structure of the target. All patches are used as the training samples, i.e. $\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_M] \in \mathbb{R}^{l\times(N\times M)}$, where $\mathbf{X}_m$ denotes the patches collected from $m-$th part of the target, l is the dimension of the image patch, N is the number of target templates and M is the number of local patches within the target region. In the proposed method, these patches from the same part of the target are regarded as one class, hence a multi-class dictionary learning can be described as,

$$\underset{\mathbf{D},\mathbf{C}}{\operatorname{argmin}} \sum_{m=1}^{M} \|\mathbf{X}_m - \mathbf{D}_m\mathbf{C}_m\|_F^2 + \lambda\Psi(\mathbf{C}_m) \tag{1}$$

where $\mathbf{C}$ is the coefficient matrix of $\mathbf{X}$ over $\mathbf{D}$, and $\Psi(\mathbf{C})$ accounts for some regularization terms.

As discussed in [20], if an analysis dictionary $\mathbf{P}$ can be obtained, then the coefficients $\mathbf{C}$ can be directly estimated efficiently using $\mathbf{C} = \mathbf{PX}$. Consider a synthesis dictionary $\mathbf{D} = [\mathbf{D}_1, \cdots, \mathbf{D}_m, \cdots, \mathbf{D}_M]$ and an analysis dictionary $\mathbf{P} = [\mathbf{P}_1; \cdots; \mathbf{P}_m; \cdots; \mathbf{P}_M]$, where $\mathbf{D}_m \in \mathbb{R}^{l\times k}$ and $\mathbf{P}_m \in \mathbb{R}^{k\times l}$ are regarded as sub-dictionary pairs corresponding to class m, and k is the number of atoms in each sub-dictionary. Furthermore, it is always expected that any sub-dictionary $\mathbf{P}_m$ projects a sample from a different class to be the null space, which can effectively preserve the discriminative ability among different classes. Also, samples within each class, $\mathbf{X}_m$ should be well represented by the sub-dictionary pair $\mathbf{D}_m\mathbf{P}_m$, i.e. $\mathbf{X}_m = \mathbf{D}_m\mathbf{P}_m\mathbf{X}_m$. With the above considerations, the following minimization function shall be considered.

$$\underset{\mathbf{P},\mathbf{D}}{\operatorname{argmin}} \sum_{m=1}^{M} \|\mathbf{X}_m - \mathbf{D}_m\mathbf{P}_m\mathbf{X}_m\|_F^2 + \lambda\|\mathbf{P}_m\mathbf{X}_m^c\|_F^2 \qquad s.t.\ \|\mathbf{d}_i\|_2^2 \le 1 \tag{2}$$

where $\mathbf{X}_m^c$ denotes the complementary data matrix of $\mathbf{X}_m$ from the entire training sample set $\mathbf{X}$, and λ is the regularization parameter. Although the model in Eq.(2) has introduced the notion of class-specific structure into dictionary pair learning, the discriminative ability cannot be effectively exploited if the coding coefficients have a large within-class variance. Furthermore, the optimization problem in Eq.(2) is generally non-convex, and therefore requires variable matrix $\mathbf{C}$ to be introduced in order to solve it. Thus, the proposed structural patch-based dictionary pair learning (SPDPL) problem is formulated as,

$$\begin{aligned} < \mathbf{P}^*, \mathbf{D}^*, \mathbf{C}^* > = \underset{\mathbf{P},\mathbf{D},\mathbf{C}}{\operatorname{argmin}} \{ \sum_{m=1}^{M} \|\mathbf{X}_m - \mathbf{D}_m\mathbf{C}_m\|_F^2 \\ + \lambda_1\|\mathbf{P}_m\mathbf{X}_m - \mathbf{C}_m\|_F^2 + \lambda_2\|\mathbf{P}_m\mathbf{X}_m^c\|_F^2 \\ + \beta\|\mathbf{C}_m - \bar{\mathbf{C}}_m\|_F^2 \} \quad s.t.\ \|\mathbf{d}_i\|_2^2 \le 1 \end{aligned} \tag{3}$$

where $\bar{\mathbf{C}}_m$ denotes a matrix with each column of it being the mean of all columns in $\mathbf{X}$, λ_1, λ_2 and β are the regularization parameters. By coding $\mathbf{X}_m$ over $\mathbf{D}_m$, the reconstruction term $\|\mathbf{X}_m - \mathbf{D}_m\mathbf{C}_m\|_F^2$ ensures that the representation power of the synthesis dictionary $\mathbf{D}_m$ is preserved. The term $\|\mathbf{P}_m\mathbf{X}_m - \mathbf{C}_m\|_F^2$ ensures that the coding coefficients can be analytically obtained using $\mathbf{C}_m = \mathbf{P}_m\mathbf{X}_m$ with a small reconstruction error. In addition, the minimization term of $\|\mathbf{C}_m - \bar{\mathbf{C}}_m\|_F^2$ is introduced to impose the coding coefficients of all samples in class m to be more closer to their mean, thereby reducing the variation of the coding coefficients of each class. Overall, according to the proposed SPDPL algorithm, the synthesis dictionary $\mathbf{D}$ and the analysis dictionary $\mathbf{P}$ can ensure that the training samples have lager between-class distance and smaller within-class variations thereby improving its discriminative ability. In addition, the representation coefficients can be analytically optimized in an efficient manner as detailed in the following section.

2.2. Optimization

The objective function described in Eq.(3) concerning all variables $\mathbf{D}$, $\mathbf{P}$ and $\mathbf{C}$ has no closed-form solution, however, it can be efficiently solved by iteratively estimating each variable by keeping the other fixed. Thus the following iterative optimization scheme is proposed to determine the local optimum for each variable in Eq.(3).

1) **Update of C**. In order to estimate $\mathbf{C}$ while keeping the other variables fixed, the objective function in Eq.(3) is equivalent to minimizing,

$$\min_{\mathbf{C}} \sum_{m=1}^{M} \|\mathbf{X}_m - \mathbf{D}_m\mathbf{C}_m\|_F^2 + \lambda_1\|\mathbf{P}_m\mathbf{X}_m - \mathbf{C}_m\|_F^2 + \beta\|\mathbf{C}_m - \bar{\mathbf{C}}_m\|_F^2 \quad (4)$$

The objective function in Eq.(4) is a conventional least squares problem and therefore can be optimized using,

$$\mathbf{C}_m^* = (\mathbf{D}_m^T\mathbf{D}_m + (\lambda_1 + \beta)\mathbf{I})^{-1} (\mathbf{D}_m^T\mathbf{X}_m + \lambda_1\mathbf{P}_m^T\mathbf{X}_m + \beta\bar{\mathbf{C}}_m) \quad (5)$$

where $\mathbf{I}$ is an identity matrix. In the first iteration, $\bar{\mathbf{C}}_m$ is initialized to zero. In the following iteration, $\bar{\mathbf{C}}_m$ is computed as the mean of the updated coefficient matrix $\mathbf{C}_m$.

2) **Update of P**. $\mathbf{P}$ can be optimized by transforming the objective function in Eq.(3) into,

$$\min_{\mathbf{P}} \sum_{m=1}^{M} \lambda_1\|\mathbf{P}_m\mathbf{X}_m - \mathbf{C}_m\|_F^2 + \lambda_2\|\mathbf{P}_m\mathbf{X}_m^c\|_F^2 \quad (6)$$

Thus, a closed-form solution of $\mathbf{P}$ can be obtained as:

$$\mathbf{P}_m^* = \lambda_1\mathbf{C}_m\mathbf{X}_m^T(\lambda_1\mathbf{X}_m\mathbf{X}_m^T + \lambda_2\mathbf{X}_m^c\mathbf{X}_m^{c^T})^{-1} \quad (7)$$

3) **Update of D**. With the assumption that all variables except $\mathbf{D}$ remain fixed, the objective function in Eq.(3) can be re-written as,

$$\min_{\mathbf{D}} \sum_{m=1}^{M} \|\mathbf{X}_m - \mathbf{D}_m\mathbf{C}_m\|_F^2 \quad s.t. \ \|\mathbf{d}_i\|_2^2 \leq 1 \quad (8)$$

The process of updating $\mathbf{D}_m$ is the same as detailed in [20, 21]. The optimization rule for minimizing the objective function in Eq.(3) is non-convex against three unknown variables. Therefore, an iterative optimization algorithm can alone provide a local minimum. In addition, it also can be proven that the above objective function is non-increasing under the aforementioned optimization steps.

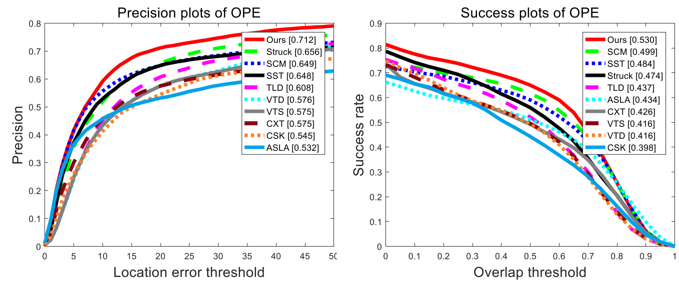

Fig. 3. Overall performance comparison of OPE using precision and success plots. The performance score of each tracker is shown in the legend.

3. THE PROPOSED TRACKING FRAMEWORK

3.1. Tracking framework

Considering that $\mathbf{y}_{1:t} = \{\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_t\}$ represent observation samples at the t-th frame, the aim of target tracking is to estimate the state of the hidden variables using,

$$p(\mathbf{z}_t|\mathbf{y}_{1:t}) \propto p(\mathbf{y}_t|\mathbf{z}_t)\int p(\mathbf{z}_t|\mathbf{z}_{t-1})p(\mathbf{z}_{t-1}|\mathbf{y}_{1:t-1})d\mathbf{z}_{t-1} \quad (9)$$

where the target motion motion in temporally consecutive states is represented as $p(\mathbf{z}_t|\mathbf{z}_{t-1})$, and $p(\mathbf{y}_t|\mathbf{z}_t)$ denotes the observation model that estimates the likelihood of observing $\mathbf{y}_t$ at the current state $\mathbf{z}_t$. The inference for the optimal target state is estimated as the maximum a posteriori estimation (MAP) in the t-th frame using,

$$\hat{\mathbf{z}}_t = \underset{\mathbf{z}_t^i}{\operatorname{argmax}}\, p(\mathbf{y}_t^i|\mathbf{z}_t^i)p(\mathbf{z}_t^i|\mathbf{z}_{t-1}) \quad (10)$$

where $\mathbf{z}_t^i$ indicates the i-th candidate of the state $\mathbf{z}_t$ and $\mathbf{y}_t^i$ denotes the i-th candidate predicted by $\mathbf{z}_t^i$.

3.2. Observation model

Each candidate sample $\mathbf{y}_t^i$ can be divided into M local patches (e.g., $\{\mathbf{y}_t^{i1}, \cdots, \mathbf{y}_t^{ij}, \cdots, \mathbf{y}_t^{iM}\}$). Thus, the reconstruction residual of each local patch is obtained using,

$$e_i^j = \|\mathbf{y}_t^{ij} - \mathbf{D}_j\mathbf{P}_j\mathbf{y}_t^{ij}\|_2^2 \quad (11)$$

where $\mathbf{y}_t^{ij}$ denotes the j-th patch in i-th candidate, $\mathbf{D}_j$ and $\mathbf{P}_j$ are the dictionary pairs that correspond to the j-patch (j class). For each each candidate sample $\mathbf{y}_t^i$, the observation likelihood can be measured using,

$$p(\mathbf{y}_t^i|\mathbf{z}_t^i) \propto \exp\{-\sum_j^M e_i^j\} \quad (12)$$

In Eq. (12), the reconstruction residual of each patch is obtained using the corresponding sub-dictionary pair, and these reconstruction residuals of all patches are combined to form the final reconstruction residual of each candidate.

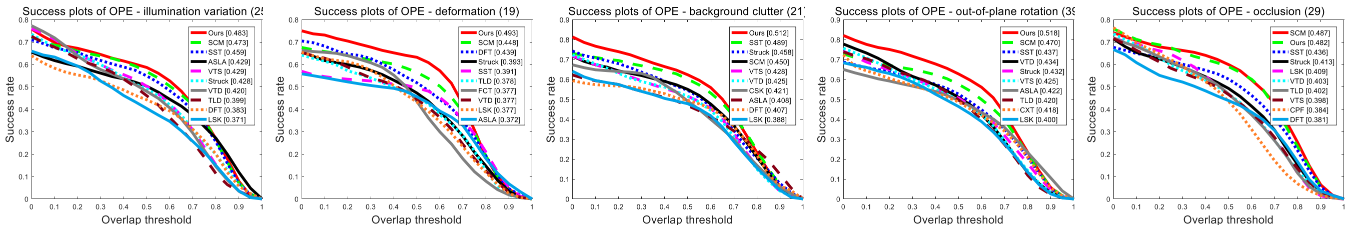

Fig. 4. Overlap success plots over five tracking challenges of illumination variation, occlusion, background clutter, scale variation and in-plane rotation.

3.3. Dictionary update

The limitation of maintaining a fixed dictionary during tracking is that it fails in dynamic scenes as the appearance changes due to factors such as illumination, deformation and pose. However, if the dictionary is directly updated using new observation samples, errors are likely to accumulate and the tracker will eventually drift to the background. To deal with the dictionary update issue, dictionary pair $\mathbf{D}$ and $\mathbf{P}$ requires to be updated online after obtaining the new tracking result at the t-th frame. Within this proposed technique, in the current target template set $\mathbf{T} = [\mathbf{T}_1, \mathbf{T}_2, \cdots, \mathbf{T}_N]$, the update scheme proposed in [6] is adopted to update the target template set using new observation samples thereby producing the set $\mathbf{T}^{'}$.

4. IMPLEMENTATION AND EXPERIMENTS

4.1. Parameter setting

In order to experiment with the proposed methodology, each target region is normalized to a 32×32 size, and is divided into $M = 9$ local patches. In the first N frames, it is assumed that the location of target has been obtained using the IVT method [9], and N is set to 10. The regularization parameters λ_1, λ_2 and β are set to 0.05, 0.05 and 0.001 respectively. The proposed tracker is updated every 5 frames to learn a new dictionary pair which eventually results in obtaining a good balance between effectiveness and computational demand. All the parameters are maintained fixed across all experiments. To compare the computational demand of the proposed tracker with the others, all trackers with MATLAB codes are tested using MATLAB on an i3 3.20 GHz machine with 4 GB RAM. Table 1 reports the average Frames Per Second (FPS) of the partial top trackers.

4.2. Results on Object Tracker Benchmark (OTB)

Evaluation Metrics. Two evaluation metrics: precision plot and success plot are chosen for performance benchmarking [22]. According to [22] the results of 29 different trackers are compared using the OTB. In addition to these baseline strategies, two other recent tracking techniques: FCT [23] and SST [24] have also been included into this experiment.

Overall Performance: In order to benchmark the tracking algorithms, the One Pass Evaluation (OPE) curve of both precision and success plots has been used. The evaluation of the overall performance of the top 10 trackers is shown in Fig.3.

Table 1. FPS comparison of the partial top trackers.

Algorithm	SCM	SST	ASLA	IVT	ℓ_1	Ours
FPS	0.5	2.1	1.1	16.5	0.3	7.5

From Fig.3, it can be seen that the proposed method achieves the best performance among all compared trackers. According to the precision and success plots, the proposed tracker also obtains improvements than all compared trackers. The superior performance of the tracker can be attributed to the following reasons. First, the proposed model regards local image patches within a spatial layout of the target region as a class. In this manner, only target information is considered for representation and thus the impact of the background is reduced. In addition, each local patch represents one fixed part of the target region, hence the local patches altogether exploit the complete structure of the target thus improving the robustness of the tracker to occlusion. Second, the proposed dictionary pair learning algorithm introduced into the tracking framework, not only reduces the computation complexity for obtaining coefficients directly, but is also helpful towards improving tracking performance.

Attribute-based evaluation. To further analyze the overall performance of the trackers across all sequences, 11 subset experiments corresponding to different attributes under specific challenging conditions are conducted to benchmark tracking performance. In Fig.4, results for five different challenging attributes are displayed with the top 10 trackers. The experimental results of the proposed approach on sequences under illumination variation, occlusion, background clutter, scale variation, in-plane rotation and other challenges validate its efficiency, effectiveness and robustness.

5. CONCLUSION

This paper proposes a novel visual tracking framework based on structural patch-based dictionary pair learning. According to the proposed model, tracking is transformed into a multi-class classification problem by regarding all patches from the same part of the target as one class. Because the training set only considers target information, the effect of background can be reduced thus resulting in more accurate tracking performance. Besides, the proposed SPDPL algorithm is used to obtain a more robust and discriminative dictionary. Finally, a simple but effective observation model is designed to obtain an optimal candidate. Comparison results on OTB database have verified the effectiveness of the proposed tracker.

6. REFERENCES

[1] T. Zhou, H. Bhaskar, and et al., "Online learning of multi-feature weights for robust object tracking," in *ICIP*. IEEE, 2015, pp. 725–729.

[2] S. Zhang, H. Yao, X. Sun, and X. Lu, "Sparse coding based visual tracking: review and experimental comparison," *Pattern Recognition*, vol. 46, no. 7, pp. 1772–1788, 2013.

[3] F. Liu, T. Zhou, K. Fu, I. Gu, and J. Yang, "Robust visual tracking via inverse nonnegative matrix factorization," in *ICASSP*. IEEE, 2016, pp. 1491–1495.

[4] W. Zhong, H. Lu, and M.-H. Yang, "Robust object tracking via sparsity-based collaborative model," in *CVPR*. IEEE, 2012, pp. 1838–1845.

[5] S. Hare, A. Saffari, and P. Torr, "Struck: Structured output tracking with kernels," in *ICCV*. IEEE, 2011, pp. 263–270.

[6] X. Jia, H. Lu, and M.-H. Yang, "Visual tracking via adaptive structural local sparse appearance model," in *CVPR*, 2012, pp. 1822–1829.

[7] Z. Kalal, K. Mikolajczyk, and J. Matas, "Tracking-learning-detection," *IEEE Trans. on Pattern Anal. and Mach. Intell.*,, vol. 34, no. 7, pp. 1409–1422, 2012.

[8] J. Kwon and K. M. Lee, "Visual tracking decomposition," in *CVPR*. IEEE, 2010, pp. 1269–1276.

[9] D. A. Ross, J. Lim, R.-S. Lin, and M.-H. Yang, "Incremental learning for robust visual tracking," *Int. J. Comp. Vis.*, vol. 77, no. 1-3, pp. 125–141, 2008.

[10] B. Ma, J. Shen, Y. Liu, H. Hu, L. Shao, and X. Li, "Visual tracking using strong classifier and structural local sparse descriptors," *IEEE Trans. Multimedia*, vol. 17, no. 10, pp. 1818–1828, 2015.

[11] Q. Wang, F. Chen, J. Yang, W. Xu, and M.-H. Yang, "Transferring visual prior for online object tracking," *IEEE Trans. Image Proc.*, vol. 21, no. 7, pp. 3296–3305, 2012.

[12] X. Mei and H. Ling, "Robust visual tracking using ℓ_1 minimization," in *ICCV*. IEEE, 2009, pp. 1436–1443.

[13] T. Zhou, H. Bhaskar, F. Liu, and J. Yang, "Graph regularized and locality-constrained coding for robust visual tracking," *IEEE Transactions on Circuits and Systems for Video Technology*, 2016.

[14] H. Zhang, S. Hu, X. Zhang, and L. Luo, "Visual tracking via constrained incremental non-negative matrix factorization," *IEEE Signal Processing Letters*, vol. 22, no. 9, pp. 1350–1353, 2015.

[15] N. Wang, J. Wang, and D.-Y. Yeung, "Online robust non-negative dictionary learning for visual tracking," in *ICCV*. IEEE, 2013, pp. 657–664.

[16] J. Xing, J. Gao, B. Li, W. Hu, and S. Yan, "Robust object tracking with online multi-lifespan dictionary learning," in *ICCV*. IEEE, 2013, pp. 665–672.

[17] Y. Xie, W. Zhang, C. Li, S. Lin, Y. Qu, and Y. Zhang, "Discriminative object tracking via sparse representation and online dictionary learning," *IEEE Trans. Cybern.*, vol. 44, no. 4, pp. 539–553, 2014.

[18] F. Yang, Z. Jiang, and L. Davis, "Online discriminative dictionary learning for visual tracking," in *WACV*. IEEE, 2014, pp. 854–861.

[19] B. Liu, J. Huang, C. Kulikowski, and L. Yang, "Robust visual tracking using local sparse appearance model and k-selection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 12, pp. 2968–2981, 2013.

[20] S. Gu, L. Zhang, W. Zuo, and X. Feng, "Projective dictionary pair learning for pattern classification," in *NIPS*, 2014, pp. 793–801.

[21] M. Yang, W. Luo, and L. Shen, "Dictionary pair learning with block-diagonal structure for image classification," in *International Conference on Intelligent Science and Big Data Engineering*. Springer, 2015, pp. 288–299.

[22] Y. Wu, J. Lim, and M.-H. Yang, "Online object tracking: A benchmark," in *CVPR*. IEEE, 2013, pp. 2411–2418.

[23] K. Zhang, L. Zhang, and M.-H. Yang, "Fast compressive tracking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 10, pp. 2002–2015, 2014.

[24] T. Zhang, S. Liu, C. Xu, S. Yan, B. Ghanem, N. Ahuja, and M.-H. Yang, "Structural sparse tracking," in *CVPR*. IEEE, 2015, pp. 150–158.