Object Tracking Based on Multi-modality Dictionary Learning

Jing Wang, Hong Zhu(✉), Shan Xue, and Jing Shi

Faculty of Automation and Information Engineering,
Xi'an University of Technology, Xi'an 710048, China
jjing63@hotmail.com, zhuhong@xaut.edu.cn,
xueshanmath@163.com, shijing1003@163.com

Abstract. Sparse representation based methods have been increasingly applied to object tracking. However, complex optimization and a single dictionary limit their deployment during tracking. In this paper, we propose a tracking method based on multi-modality dictionary learning in particle filter framework. First, multi-modality dictionary is formed by background templates and object templates including short-term templates and long-term templates that are updated by K-means clustering. Second, coarse tracking results are achieved by computing the coefficients of object with respect to templates from multi-modality dictionary. Finally, the Local Maximal Occurrence (LOMO) features of coarse tracking results and multi-modality dictionary are compared through observation likelihood function, a candidate result with highest observation score is regarded as the final tracking result. The experimental results demonstrated the effectiveness of our method compared to some state-of-the-art methods.

Keywords: Object tracking · Multi-modality dictionary · Particle filter
K-means clustering

1 Introduction

Object tracking plays an important role in computer vision, which has been widely used in the field of surveillance, intelligent transportation control, medical image and military simulation [1] etc. Even though numerous tracking problems have been studied for decades, and reasonable good results have been achieved, it is still challenging to track general objects in a dynamic environment accurately due to various factors that include noise, occlusion, background cluttering, illumination changes, fast motions, and variations in pose and scale. At present, most of state-of-the-art object tracking methods can be categorized into two types: generative models and discriminative models [2].

The generative methods take the candidate having the best compatibility with the appearance model as the tracked object. For example, Ross et al. proposed an incremental subspace model to adapt object appearance variation [3]. Wang et al. put forward multi-features fusion object model under the guidance of color-feature, and tracking object accurately is realized by the principle of spatial consistency [4]. Wang et al. proposed a probability continuous outlier model to cope with partial occlusion via

Y. Zhao et al. (Eds.): ICIG 2017, Part II, LNCS 10667, pp. 129–138, 2017.
https://doi.org/10.1007/978-3-319-71589-6_12

holistic object template [5]. The latter addresses object tracking as a binary classification to separate the object from background. For example, Kalal et al. first utilized structured unlabeled data and used an online semi-supervised learning method [6], then tracking-learning-detection (TLD) for object tracking in long sequences is proposed subsequently [7]. Babenkon et al. formulated object tracking as an online multiple instance learning [8]. Generally speaking, the former can get more accurate characteristic of object but with high computational complexity. The latter can obtain better tracking accuracy but has to process a large number of training samples. Object needs to be retrained if its appearance changed, and tracking failure can be easily caused by inadequate training samples. In addition, there are some combine both generative and discriminative models [9–11] to get more desirable results.

Recently, sparse representation based methods [9–13] have shown promising results in various tests. Object is represented as a linear combination of a few templates, which are helpful to remove the influences from partial occlusion, illumination and other factors on object based on sparse coding. However, this kind of method is based on solving ℓ_1 minimization that has large computational load, and sparse code is solved by complex optimization. Therefore, a multi-modality dictionary is built in this paper to simplify the sparse coding, and then follow the idea of combination of generative and discriminative to achieve object tracking.

The remainder of this paper is organized as follows. In Sect. 2, particle filter and object representation that are related to our work are reviewed. Section 3 introduces the details of the proposed tracking method. Experimental results and analysis are shown in Sect. 4, and we conclude this paper in Sect. 5.

2 Preliminary

2.1 Particle Filter

Particle filter as the tracking framework in this paper, the object of the next frame is estimated by the observation probability of particles at the current frame [14]. Suppose $Y_t = [y_1, \ldots, y_t]$ are observed images at frames 1 to t, x_t is the state variable that describing object motion parameters at frame t, and follows the following probability distribution:

$$p(x_t \mid Y_t) \propto p(y_t \mid x_t) \int p(x_t \mid x_{t-1}) p(x_{t-1} \mid Y_{t-1}) dx_{t-1} \tag{1}$$

where $p(x_t \mid x_{t-1})$ is state transition distribution, $p(y_t \mid x_t)$ estimates the likelihood of observing y_t at state x_t. Particles are sampled as Gaussian distribution with the center position of previous tracking result. As the number of particles will affect the tracking efficiency, irrelevant particles need to be filtered to reduce the tracking redundancy.

2.2 Object Representation

Liao et al. proposed Local Maximal Occurrence (LOMO) feature [15] for the performance of the target in different cameras is inconsistent, which is an effective handmade

feature that can be compared with the characteristics of deep learning network in recent years. The LOMO feature analyzes the horizontal occurrence of local features, and maximizes the occurrence to make a stable representation against viewpoint changes. Specifically, the Retinex algorithm is firstly applied to produce a color image that is consistent to human observation of the scene, then HSV color histogram is used to extract color features of Retinex images, finally, Scale Invariant Local Ternary Pattern (SILTP) descriptor [16] is applied to achieve invariance to intensity scale changes and robustness to image noises.

Since the challenging problems in object tracking and person re-identification are actually the same, in view of the validity of the LOMO feature has been verified, the LOMO feature as the object feature in this paper.

3 Proposed Method

In this paper, object tracking is regarded as the dictionary learning problem. By constructing multi-modality dictionary properly that can describe object precisely, thus the complex optimization can be simplified. The proposed tracking method is presented in Algorithm 1.

Algorithm1: Proposed Tracking Method

Input: image at frame t

Output: tracking result $\hat{x}_t$ of image at frame t

Initialization: construct the multi-modality dictionary by using the first frame of a video

Tracking:

for t=2:end of the video

1. Sample particles based on the tracking result of previous frame, and candidates are filtered by the distance constraint;
2. Solve the coefficients of candidates with respect to multi-modality dictionary using the LARS method;
3. Get the candidate tracking results R according to the coefficients of each candidate (Eq. (6));
4. Compute the observation likelihood score of each candidate result from R by Eq. (8);
5. Candidate with the highest observation score is regarded as the final tracking result;
6. Update the multi-modality dictionary through K-means clustering (Sect. 3.1).

end for

3.1 Multi-modality Dictionary Building and Updating

In general, sparse representation based tracking method usually uses over-complete dictionary to encode the object. Sparse code learning involves two problems: sparse

coding that is to solve the computation of the coefficients to represent the object with the learned dictionary, and dictionary learning that is to solve the problem of constructing the dictionary [17]. With the sparse assumption, a candidate object x_i can be represented as a linear combination of sparse code α_i from dictionary D. The sparse code $\alpha_i \in \mathbb{R}^{n+m}$ corresponding to x_i is calculated by

$$\min_{\alpha_i} \frac{1}{2}\|x_i - D\alpha_i\|_2^2 + \lambda\|\alpha_i\|_1 \tag{2}$$

where over-complete dictionary $D = \left[D_p, D_n\right] \in \mathbb{R}^{d\times(n+m)}$ that is formed by the foreground dictionary $D_p \in \mathbb{R}^{d\times n}$ and background dictionary $D_n \in \mathbb{R}^{d\times m}$. The above problem is also referred to as dictionary learning, which is actually the Lasso regression [18] that can be solved by LARS [19], and then get the sparse code α_i of x_i.

However, the method mentioned above not only requires a large number of templates for over-complete dictionary, but also makes the tracking process more complicated. Actually, for objects in the ideal state without severe external influences, a small number of dictionary templates can distinguish objects from background well; for objects with appearance changed, more dictionary templates will bring many errors. So if a suitable dictionary for the current object can be obtained in real time, there is no need to build over-complete dictionary and experience complex optimization process.

In this paper, object dictionary is formed by short-term templates D_s and long-term templates D_l, the templates of background dictionary D_n are selected randomly from non-object area of video image. Therefore, multi-modality dictionary D is built by the two parts, that is $D = [D^S, D^L]$, where $D^S = [D_s, D_n] \in \mathbb{R}^{d\times(m_s+n_s)}$, $D^L = [D_l, D_n] \in \mathbb{R}^{d\times(m_l+n_l)}$, and the dictionary template is represented by the observed pixel values. More specifically, D_s is initialized by the transformed object templates of the first frame, that is the current object moves 1–2 pixels along four directions (up, down, left and right). D_l is initialized by the clustering center of D_s, that is $D_l = \frac{1}{l_s}\sum_{i=1}^{l_s} D_s^{(i)}$, where l_s represents the number of templates in D_s, here let $l_s = 9$. It should be noted that the observation vector of each template usually constraints its columns to have ℓ_2-norm less than or equal to 1.

For multi-modality dictionary, when object appearance changes little, short-term dictionary can distinguish object from background effectively, and long-term dictionary can reduce errors accumulation; when object appearance changes greatly, short-term dictionary can track object continuously, and long-term dictionary can prevent loss of correct sampled object. Thus, the combination of two modality dictionaries can better balance the adaptability and robustness of ℓ_1 trackers, and it is crucial for updating multi-modality dictionary. D_s is trained and updated using the candidates sampled in the previous frame. D_l is trained and updated using accurate result in all previous frames, and then according to the theory of K-means clustering, the category that the current object belongs to is identified by calculating the Euclidean distance between the current object and the clustering centers of long-term dictionary, as shown in Eq. (3). The long-term dictionary is represented by the cluster center of each category, which reduces the amount of computation effectively.

$$\begin{cases} x^t \in D_l^{c^i}, & d\left(f(x^t), f\left(D_l^{c^i}\right)\right) \in [0, d\_\max + \Delta] \\ x^t \in D_l^{c^{new}}, & d\left(f(x^t), f\left(D_l^{c^i}\right)\right) > d\_\max + \Delta \end{cases} \tag{3}$$

where $D_l^{c^i}$ represents the existed category of long-term dictionary, $D_l^{c^{new}}$ represents the new category of long-term dictionary, $d(\cdot)$ denotes Euclidean distance, $f(\cdot)$ indicates the corresponding LOMO feature, $d\_\max$ represents the maximum value of Euclidean distance between templates $d\left(D_s, D_s^{'}\right)$ in initialized short-term dictionary D_s, and Δ is variable.

3.2 Tracking Based on Multi-modality Dictionary

The proposed method is based on particle filter framework, and all the sampled particles are expressed as $Y = \{y_1, y_2, \ldots, y_N\} \in \mathbb{R}^{d \times N}$. Then irrelevant particles are filtered by the distance constraint that is the distance between the center coordinate of the sampled object $p\left(y_i^t\right)$ and the center coordinate of tracking result of previous frame $p(x_{t-1})$ should meet $\left\|p\left(y_i^t\right) - p(x_{t-1})\right\|_2 \leq \max(w, h)$, where w and h represent the width and height of bounding box of previous tracking result respectively. The candidate samples are expressed as $X = \{x^i \mid i \in [1, q]\} \in \mathbb{R}^{d \times q} (q \ll N)$.

Assuming that the multi-modality dictionary can be adapt to the object appearance changes well, the value of the cost function between the ideal tracking result and the templates of object dictionary should be minimal. The cost function of short-term dictionary and long-term dictionary are expressed as Eqs. (4) and (5) respectively. Then the best coefficients are solved using the LARS method [19].

$$l_S\left(x^i, D^S\right) = \min_{\alpha_s^i} \frac{1}{2} \left\|x^i - D^S \cdot \alpha_s^i\right\|_2^2 + \lambda \left\|\alpha_s^i\right\|_1 \tag{4}$$

$$l_L\left(x^i, D^L\right) = \min_{\alpha_l^i} \frac{1}{2} \left\|x^i - D^L \cdot \alpha_l^i\right\|_2^2 + \lambda \left\|\alpha_l^i\right\|_1 \tag{5}$$

Generally, an image observation of a "good" object candidate is effectively represented by the object templates and not the background templates, thereby, leading to a sparse representation. Likewise, an image observation of a "bad" object candidate can be more sparsely represented by a dictionary of background templates. Therefore, for ideal sampled object, the difference between the $\ell_1 - norm$ of coefficients of object templates and background templates should be larger. Then the candidate tracking results R are formed by the first p samples satisfying the condition, as shown in Eqs. (6) and (7).

$$R = [R_S, R_L] = \left[I_S^i|_p, I_L^i|_p\right] \quad (i \in [1, q]) \tag{6}$$

$$\begin{aligned} I_S^i &= \max\left(\left\|\alpha_{s^+}^i\right\|_1 - \left\|\alpha_{s^-}^i\right\|_1\right) \\ I_L^i &= \max\left(\left\|\alpha_{l^+}^i\right\|_1 - \left\|\alpha_{l^-}^i\right\|_1\right) \end{aligned} \tag{7}$$

where α^i_{s+} and α^i_{s-} represent the coefficients of object templates and background templates in short-term dictionary, α^i_{l+} and α^i_{l-} represent the coefficients of object templates and background templates in long-term dictionary.

Eventually,the observation likelihood function is built for each candidate tracking result, then the candidate tracking result with the highest similarity is regarded as the final tracking result $\hat{x}$, as shown in Eqs. (8–9).

$$\hat{x} = \arg\max(\omega_s \cdot s_s + \omega_l \cdot s_l)\ s_l = sim\big(f\big(R^j_L\big), f(D_L)\big) \tag{8}$$

$$\begin{aligned} s_s &= sim\big(f\big(R^j_s\big), f(D_S)\big) \\ s_l &= sim\big(f\big(R^j_L\big), f(D_L)\big) \end{aligned} \quad j \in [1, p] \tag{9}$$

where $sim(\cdot)$ represent the similarity that is calculated by Bhattacharyya distance, $f(\cdot)$ is the LOMO feature of the corresponding image area. $\omega_s = s_s/(s_s + s_l)$ and $\omega_s = s_s/(s_s + s_l)$ are the weights.

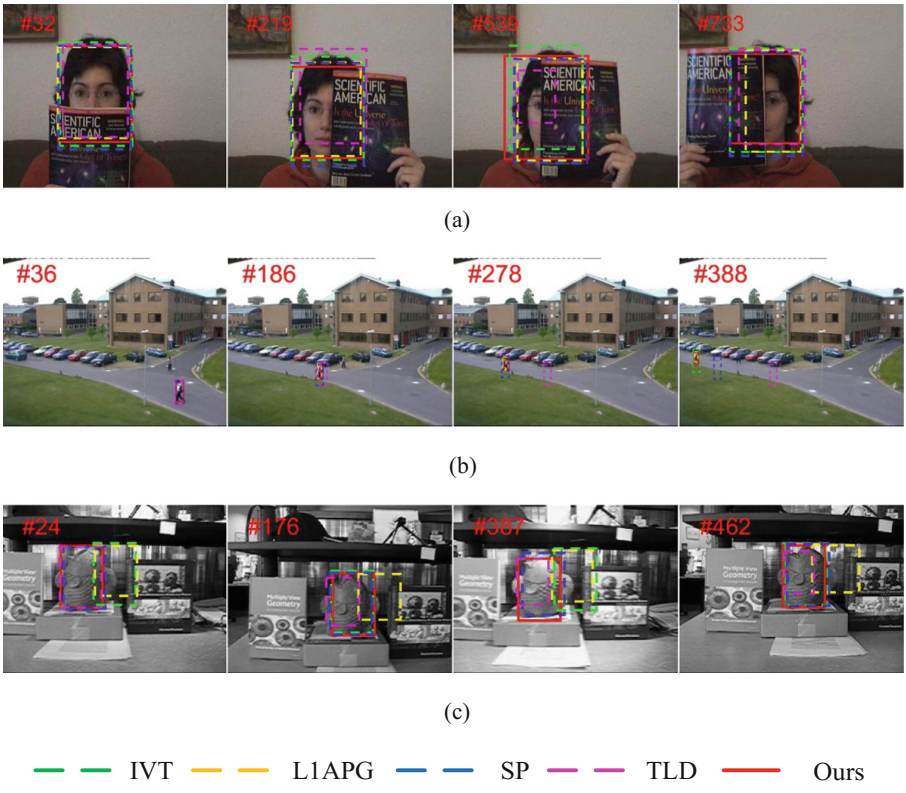

Fig. 1. Some representative results of test sequences. (a) *FaceOcc1*; (b) *Walking*; (c) *Fish*.

4 Experiments and Analysis

The test video sequences (*FaceOcc1, Walking* and *Fish*) are selected from object tracking benchmark [1]. We test four state-of-the-art methods on the same video sequences for comparison. They are IVT [3], L1APG [13], TLD [7] and SP [5]. The code of all those trackers are public available, and we keep the parameter settings provided by authors for all the test sequences. In this paper, we use the error rate (*error*) and overlap rate (*overlap*) to evaluate the tracking performance of each tracking method. *error* is the Euclidean distance between the center coordinate obtained from tracking method and tracking ground truth, which means the smaller the value, the more accurate position the method tracks. *overlap* is the overlap ratio between the tracking window of the method and the ideal tracking window, which means the larger the value, the more suitable window the method has. Figure 1 shows some representative results of test sequences.

Tracking error plots and tracking overlap plots for all the test sequences are shown in Figs. 2 and 3. The main tracking problem in *FaceOcc1* is that the object is occluded in large area for a long time. TLD fails to track when object is occluded in large area,

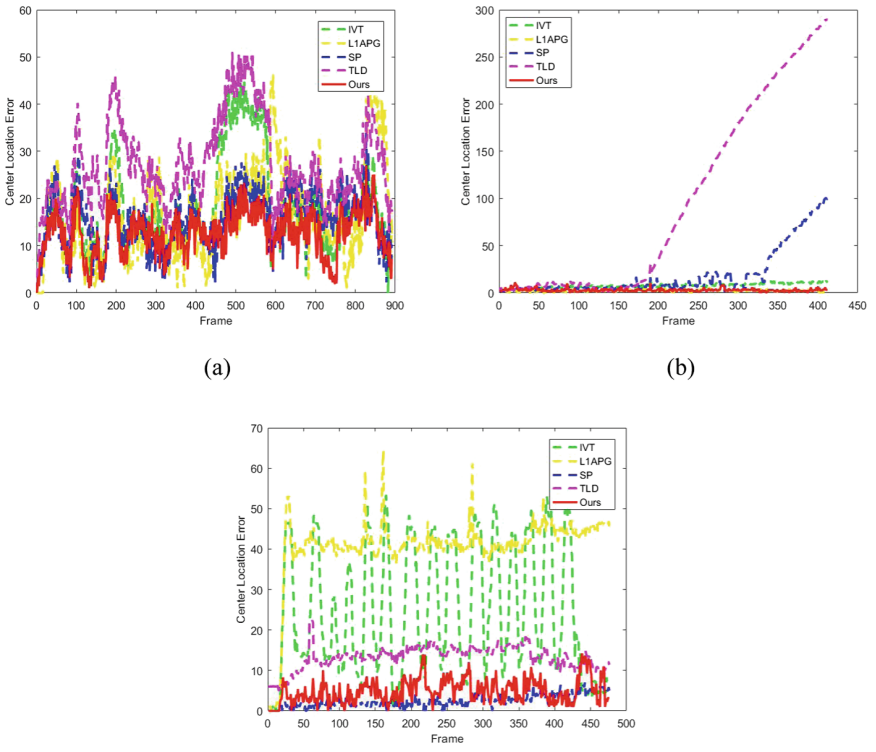

Fig. 2. Tracking error plots for all test sequences. (a) *FaceOcc1*; (b) *Walking*; (c) *Fish*.

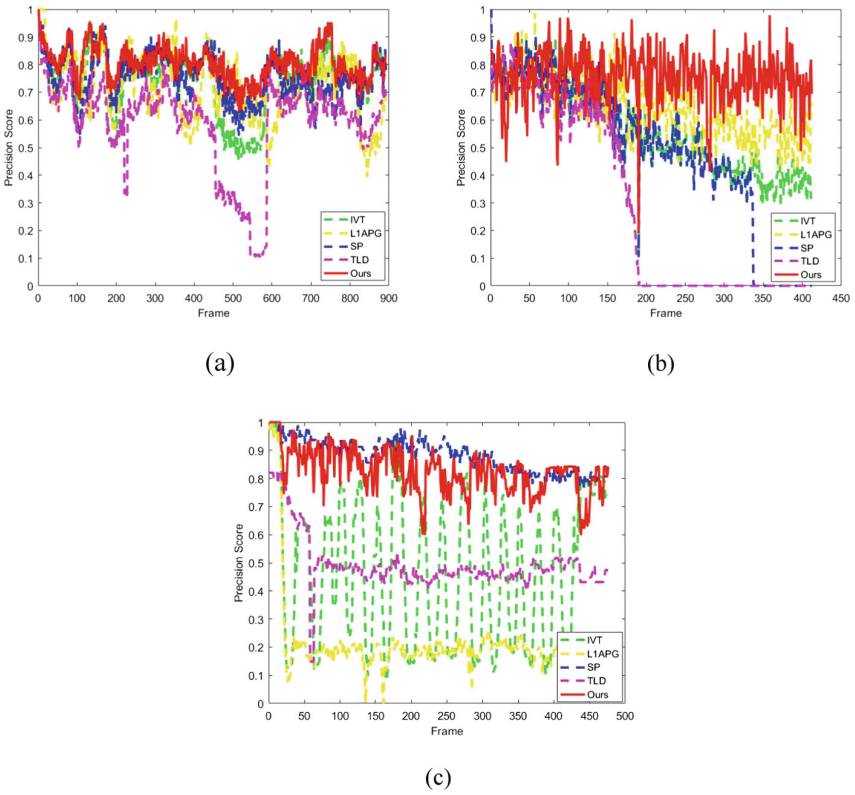

Fig. 3. Tracking overlap plots for all test sequences. (a) *FaceOcc1*; (b) *Walking*; (c) *Fish*.

but it can back to track object well when object remains in the normal state. The performance of IVT, L1APG, SP and our method can maintain low tracking error and high overlap rate, in which, our method performs the best. The main tracking problem in *Walking* is partially occlusion and object scales variation. When object scale becomes small, object cannot be distinguished from background clearly, so TLD and SP lose object. IVT, L1APG and our method can track object continuously, but the tracking bounding box of IVT is too large to fit the object size. Our method performs the best and L1APG performs the second best. The main tracking problem in *Fish* is the illumination changes. As the object is affected by the illumination and camera shake, IVT and L1APG start to drift. TLD can track object roughly, but the tracking bounding box is small. SP and our method show the promising performance, in which SP is the best tracker, and there is a slightly difference between SP and our method.

Table 1 shows the mean of tracking error and tracking overlap rate, in which bold fonts indicate the best performance while the Italic underlined fonts indicate the second best ones. From these data, we can conclude that the proposed method has good performance on occlusions, illumination and object scale variation, etc.

Table 1. The mean of tracking error and tracking overlap rate.

Method	*FaceOcc1*		*Walking*		*Fish*	
	error (pixel)	*overlap* (%)	*error* (pixel)	*overlap* (%)	*error* (pixel)	*overlap* (%)
IVT	17.92	74.92	7.62	56.57	24.42	47.38
L1APG	16.99	71.57	*2.92*	*65.34*	40.70	21.43
SP	*14.61*	*75.13*	19.32	48.88	**2.48**	**88.38**
TLD	27.38	59.69	95.89	29.06	13.22	49.54
Ours	**13.73**	**80.46**	**2.23**	**75.29**	*4.08*	*82.62*

5 Conclusions

In this paper, object tracking method based on multi-modality dictionary is proposed, which addresses object tracking as a problem of learning a dictionary that can represent object accurately. Under the particle filter framework, a multi-modality dictionary is built and updated by clustering, which makes the candidate tracking result can be obtained easily by comparing the coefficients difference with respect to multi-modality dictionary. And then the final tracking result is determined by calculating observation function precisely through employing LOMO feature. By applying some benchmark videos, the experimental results show that the proposed method is more robust against occlusions, illumination changes and background interference.

Acknowledgment. This work is supported in part by National Natural Science Foundation of China (No. 61673318), Natural Science Basic Research Plan in Shaanxi Province of China (No. 2016JM6045) and Scientist Research Program Funded by Shaanxi Provincial Education Department (No. 16JK1571).

References

1. Wu, Y., Lim, J., Yang, M.-H.: Online object tracking: a benchmark. In: Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 2411–2418 (2013)
2. Li, X., Hu, W., Shen, C., Zhang, Z., Dick, A., Hengel, A.V.D.: A survey of appearance models in visual object tracking. ACM Trans. Intell. Syst. Technol. **4**(4), 478–488 (2013)
3. Ross, D., Lim, J., Lin, R.-S., Yang, M.-H.: Incremental learning for robust visual tracking. Int. J. Comput. Vis. **77**, 125–141 (2008)
4. Wang, J., Zhu, H., Yu, S., Fan, C.: Object tracking using color-feature guided network generalization and tailored feature fusion. Neurocomputing **238**, 387–398 (2017)
5. Wang, D., Lu, H., Yang, M.-H.: Online object tracking with sparse prototypes. IEEE Trans. Image Process. **22**, 314–325 (2013)
6. Kalal, Z., Matas, J., Mikolajczyk, K.: P-N learning: bootstrapping binary classifiers by structural constraints. Comput. Vis. Pattern Recognit. **238**(6), 49–56 (2010)
7. Kalal, Z., Mikolajczyk, K., Matas, J.: Tracking-learning-detection. IEEE Trans. Pattern Anal. Mach. Intell. **34**, 1409–1422 (2012)

8. Babenko, B., Yang, M.-H., Belongie, S.: Robust object tracking with online multiple instance learning. IEEE Trans. Pattern Anal. Mach. Intell. **33**, 1619–1632 (2011)
9. Zhong, W., Lu, H., Yang, M.-H.: Robust object tracking via sparsity-based collaborative model. In: Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 1838–1845 (2012)
10. Wang, N., Wang, J., Yeung, D.Y.: Online robust non-negative dictionary learning for visual tracking. In: Proceedings of IEEE Computer Society Conference on Computer Vision, pp. 657–664 (2013)
11. Xing, J., Gao, J., Li, B., Hu, W., Yan, S.: Robust object tracking with online multi-lifespan dictionary learning. In: Proceedings of IEEE Computer Society Conference on Computer Vision, pp. 665–672 (2013)
12. Mei, X., Ling, H.: Robust visual tracking and vehicle classification via sparse representation. IEEE Trans. Pattern Anal. Mach. Intell. **33**, 2259–2272 (2011)
13. Bao, C., Wu, Y., Ling, H., Ji, H.: Real time robust L1 tracker using accelerated proximal gradient approach. In: Proceedings of IEEE Computer Society Conference on Computer Vision, pp. 1830–1837 (2012)
14. Chang, C., Ansari, R.: Kernel particle filter for visual tracking. IEEE Sig. Process. Lett. **12**, 242–245 (2005)
15. Liao, S., Hu, Y., Zhu, X., Li, S.Z.: Person re-identification by local maximal occurrence representation and metric learning. In: Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 2197–2206 (2015)
16. Liao, S., Zhao, G., Kellokumpu, V., Pietikäinen, M., Li, S.Z.: Modeling pixel process with scale invariant local patterns for background subtraction in complex scenes. In: Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 1301–1306 (2010)
17. Aharon, M., Elad, M., Bruckstein, A.: K-SVD: an algorithm for designing overcomplete dictionaries for sparse representation. IEEE Trans. Sig. Process. **54**, 4311–4322 (2006)
18. Tibshirani, R.: Regression shrinkage and selection via the lasso. J. Roy. Stat. Soc. **73**, 273–282 (2011)
19. Rosset, S., Zhu, J.: Least angle regression. Ann. Stat. **32**, 407–499 (2004)