Deep Object Tracking with Multi-modal Data

Xuezhi Zhang*†, Yuan Yuan*, Xiaoqiang Lu*
*Center for OPTical IMagery Analysis and Learning (OPTIMAL), State Key Laboratory of Transient Optics and Photonics, Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an 710119, Shaanxi, P. R. China.
†University of the Chinese Academy of Sciences, 19A Yuquanlu, Beijing, 100049, P. R. China.

***Abstract*—Object tracking is a challenging topic in the field of computer vision since its performance is easily disturbed by occlusion, illumination change, background clutter, scale variation, *etc*. In this paper, we introduce a robust tracking algorithm that fuses information from both visible images and *infrared* (IR) images. The proposed tracking algorithm not only incorporates convolutional feature maps from the visible channel, but also employs a scale pyramid representation from IR channel. We estimate the target location by fusing multilayer convolutional feature maps, and predict the target scale from a scale pyramid. The pipeline of the proposed method is as follows. First, the hierarchical convolutional feature maps are obtained from visible images using VGG-Nets. Then, the accurate target location is predicted by the maximum response of correlation filters with the visible image feature maps. Finally, we obtain the precise object scale with a scale pyramid from infrared images where the difference between the target and the background is clear. In order to verify the performance of the proposed method, we capture six video sequences under different conditions. These sequences contain both visible channel and IR channel. Ten state-of-the-art tracking algorithms are compared with our method, and the experimental results show the effectiveness of the proposed tracker.**

I. Introduction

Object tracking is a core topic in computer vision [1]. In the typical object tracking setting, given the initial states of an object in the first frame, the task of tracking is to predict the states of the target in the remaining frames [2]. Although large quantities of tracking methods have been proposed, object tracking is still a difficult problem due to the following factors, occlusion, deformation, fast motion, illumination change and scale variation [3].

Most existing tracking algorithms merely use the RGB image. Recently, a few multi-modal visual tracking methods have been developed [4], [5]. Yuan *et al*. [6] propose a robust superpixel tracker fusing both RGB images and depth images. Kumar *et al*. [7] employ thermal infrared sensor colocated with a wide-angle RGB camera to improve the performance of person tracking. Multi-modal information is used in aforementioned research works, however, there is a drawback that all of them use hand-crafted features, which would limit the performance of visual tracking [8].

Recently, deep learning has achieved many breakthroughs in computer vision, such as image classification, object detection, *etc*. The key to the success of deep learning is that its ability of feature representation is powerful. Some tracking methods using deep learning [9] have been proposed recently. For example, *deep learning tracker* (DLT) is proposed by Wang and Yeung [10], and uses the encoder part of the pre-trained *stacked denoising autoencoder network* (SDAE) as a feature extractor. The DeepTrack proposed by Li *et al*. [8], contains a candidate pool of multiple *convolutional neural networks* (CNNs). Both DLT and DeepTrack have a drawback that they only use the outputs of *deep neural network* (DNN) as features. To solve this problem, hierarchical convolutional feature maps are used in [11], but scale variation is not taken into consideration.

In this paper, we propose a robust tracking algorithm to address the aforementioned problems. Firstly, we apply VGG-Nets [12] to compute the hierarchical convolutional feature maps from visible images. The multilayer outputs are used as features, which take both spatial features of shallow layers and semantic features of deep layers into account [11]. Then, we fuse the multilayer convolutional feature maps by resizing them to the same size. Secondly, correlation filters are used to estimate the accurate target location with *fast fourier transform* (FFT). Finally, for the sake of predicting the target scale, we apply correlation filters to a scale pyramid of infrared images where the difference between the target and the background can be distinguished.

The main contributions of the proposed tracking algorithm are as follows:

1) We use dual feature fusions in this work. Our tracker can be more robust by fusing information from both visible images and infrared images, and estimate accurate target location by fusing both spatial features and semantic features from CNN.
2) We integrate a scale pyramid representation of infrared images to our tracker, so the target scale can be estimated precisely and efficiently.
3) Six video sequences are collected as datasets for the experiments. There are many public datasets used for target tracking [2], but these datasets only include visible images or infrared images. The captured video sequences are different from the public datasets, which include both visible images and infrared images for every frame.

II. Related Work

In this section, the tracking algorithms related to our tracker are discussed. First, we introduce some tracking algorithms using deep learning. Second, correlation filters based tracking methods using FFT are presented.

978-1-5090-0690-8/16/$31.00

A. Deep Learning based Tracking

Due to the fact that DNN is powerful in terms of feature representation, many tracking methods based on deep learning are proposed recently. *Deep learning tracker* (DLT) which is proposed by Wang and Yeung [10], uses the encoder part of the pre-trained *stacked denoising autoencoder network* (SDAE) as a feature extractor. Wang *et al*. [13] study the properties of the hierarchical features of CNN, and discover that different convolutional layers represent objects from different views. According to this discovery, *fully convolutional networks* (FCNT) is developed in [13] using the outputs of two convolutional layers, and uses a feature map selection method to reduce computation complexity. Ma *et al*. [11] propose to use the multilayer outputs of CNN as features, and employ adaptive correlation filters to estimate the target location. All of the above-mentioned trackers are applied to visible images, while our tracking method integrates both visible images and infrared images.

B. Correlation Filters based Tracking

Correlation filters based tracking methods have an advantage in tracking speed, so they are hot topics in the field of object tracking recently. Henriques *et al*. [14] propose a tracking algorithm named CSK using linear correlation filters and kernel trick. CSK is very fast for the reason that the responses of correlation filters can be computed by FFT. A tracking method named DSST is developed by Danelljan *et al*. [15]. It is robust to scale variation for using a scale pyramid. Henriques *et al*. [16] propose *Dual Correlation Filter* (DCF) which extend CSK to multiple channels. All the aforementioned trackers use hand-crafted features, while multilayer convolutional feature maps are used in the our tracking algorithm.

III. Proposed Method

In this section, we first introduce how to extract and fuse the multilayer convolutional features from visible images. Then the theory of correlation filter is described for learning a discriminative classifier. Finally, we will use multilayer convolutional features of visible images, correlation filters and the scale pyramid of infrared images to predict the location and scale of the target.

A. Hierarchical Convolutional Features

In order to compute the hierarchical convolutional features, the VGG-Nets [12] are used in this work. VGG-Nets are pre-trained on large-scale ImageNet dataset. The number of weight layers of the VGG-Nets ranges from 16 to 19, we use VGG-Net-19 to denote the VGG-Net that contains 19 weight layers. The size of convolution filters of VGG-Net-19 is as small as 3×3, to guarantee the depth of the CNN.

Multilayer convolutional features are used in this work. Fig. 1 shows the process of extracting and fusing hierarchical convolutional features. The shallow layer feature maps are low level features, and they are not effective to discriminate the categories of objects. However their size is large, which

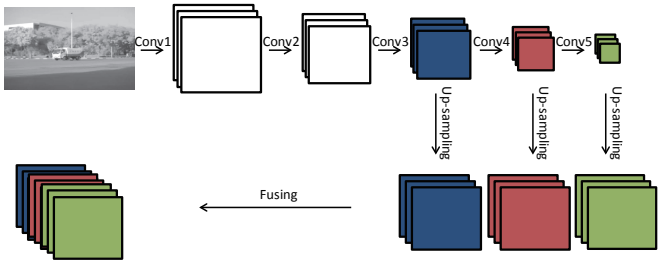

Fig. 1 The flow chart of extracting and fusing hierarchical convolutional features

is very important for tracking to precisely locate the target. Oppositely, the size of deep layer feature maps is so small that we can only approximatively predict the target location. Although the drawback of small size is obvious, the discrimination ability of deep layer feature maps is acknowledged. To achieve the goal of developing a robust tracker, the outputs of both shallow layers and deep layers are used as features. Specifically, we choose conv3-4, conv4-4 and conv5-4 layers as feature representations.

Before the visible images are imported into VGG-Net-19, a series of preprocessing operations should be done, due to the captured visible images are gray. Firstly, we reshape the inputs to 224×224, and then copy and concatenate them to $224 \times 224 \times 3$. Finally, we subtract the average value from the preprocessed images.

After preprocessing, we import visible images into the VGG-Net-19, and the hierarchical convolutional feature maps are obtained. Since the convolution and pooling operations will reduce the size of the feature maps, the size of feature maps from different layers is different. In order to fuse convolutional feature maps, we resize the multilayer feature maps to a constant size using up-sampling. Through a large number of experiments, we find a simple and effective fusion method that concatenates the feature maps of different layers as a rectangular cuboid. In this work, we concatenate the outputs of conv3-4, conv4-4 and conv5-4 layers as features.

B. Discriminative Correlation Filters

In this work, the one-dimensional correlation filters are used as scale filters for predicting the target scale, and the two-dimensional correlation filters are used as translation filters for estimating the target location. Since both one-dimensional filters and two-dimensional filters are used, the general theory of discriminative correlation filters will be introduced in the following.

The correlation filters based trackers learn a discriminative classifier from a single patch x centred around the target for predicting the attributes of the target. Note that x refers to the hierarchical convolutional feature maps when estimating the target location, and x refers to the HOG feature vectors extracted from infrared images when predicting the target scale. All cyclic shifts x_n of patch x are used as training samples, the corresponding labels y_n are generated by a

gaussian function, and n is the location of discrete signal. The cost function of the correlation filter is

$$\varepsilon = \sum_{n} \|h \cdot x_n - y_n\|^2 + \lambda \|h\|_2^2, \tag{1}$$

where h denotes the correlation filter, λ is a constant parameter of the regularization term. The regularization parameter λ can avoid the problem of division by zero. In order to find the optimal correlation filter h^*, we need to minimize the cost function as follows:

$$h^* = \arg\min_{h} \sum_{n} \|h \cdot x_n - y_n\|^2 + \lambda \|h\|_2^2, \tag{2}$$

where $\cdot$ denotes the inner product. The corresponding solution in the frequency domain of (2) is:

$$H_t^d = \frac{A_t^d}{B_t^d} = \frac{Y \odot \overline{X_t^d}}{\sum_{d=1}^{D} X_t^d \odot \overline{X_t^d} + \lambda}. \tag{3}$$

The capital letters denote the solutions in the frequency domain of the corresponding lowercase letters, for example, Y is the FFT solution of y. t is the frame number of sequences, and H_t^d is the fourier transform of the optimal correlation filter over d-channel features of t-th frame image, $\overline{X_t^d}$ is the complex conjugation of X_t^d.

In order to obtain a robust approximation, we update the numerator A_t^d and denominator B_t^d of the correlation filter H_t^d in (3) separately as:

$$A_t^d = (1-\eta)A_{t-1}^d + \eta Y \odot \overline{X_t^d}; \tag{4}$$

$$B_t^d = (1-\eta)B_{t-1}^d + \eta \sum\nolimits_{d=1}^{D} X_t^d \odot \overline{X_t^d}; \tag{5}$$

$$H_t^d = \frac{A_t^d}{B_t^d + \lambda}, \tag{6}$$

where η denotes the learning rate. The response of the correlation filter, namely the correlation score map, is computed as follows:

$$f_t = \mathscr{F}^{-1}(\sum_{d} H_t^d \odot \overline{X_t^d}), \tag{7}$$

where f_t is the correlation score map, and $\mathscr{F}^{-1}$ denotes the *Inverse Fast Fourier Transform* (IFFT) operator. Through finding the max response of the correlation score map f_t, we can predict the location and scale of the target.

C. Location and Scale Estimation

Our tracking algorithm can be divided into two steps successively. Firstly, the target location is predicted by translation correlation filters with multilayer convolutional features from visible images. Secondly, the target scale is estimated by scale correlation filters with the HOG features of scale pyramids from infrared images. Both visible images and infrared images are integrated into our tracking method.

We need to locate the target before estimating its scale. As described in subsection III-A, we firstly import the preprocessed visible images into VGG-Net-19, and then concatenate the feature maps from conv3-4, conv4-4 and conv5-4 layers for fusing the multilayer convolutional features. The concatenated feature maps are denoted by x. Finally, we utilize x to train and update the two-dimensional multi-channel translation correlation filters h_{trans}, and the target location is estimated by finding the max response of the pixel-wise sum of the translation correlation filters.

Fig. 2 Example images captured by the IR Thermal Imaging System. From top to bottom, they are Car, Head, HeadRotate, HeadCard, Roper and Book respectively. Left: visible images. Right: infrared images.

After the target location obtained, the target scale is estimated by following steps. Firstly, the scale pyramid of infrared images should be constructed. Let S and a denote the size of the scale filter and the scale factor between feature layers, respectively. The size of target in current frame is $M \times N$. For all $n \in \{-\lfloor \frac{S-1}{2} \rfloor, \cdots, \lfloor \frac{S-1}{2} \rfloor\}$, we utilize the patches of size $a^n M \times a^n N$ centred around the predicted location to construct the scale pyramid of infrared images. Then we reshape the patches in the scale pyramid to a constant size, and compute the HOG feature vectors from these patches. Note that x denotes the HOG feature vectors. Finally, the one-dimensional multi-channel scale correlation filters h_{scale} are trained and updated using the HOG feature vectors x, and we obtain the target scale from the max response of pixel-wise sum of scale correlation filters.

IV. Experimental Evaluation

In this section, we first introduce the datasets and compared algorithms used in the experiments. Then the tracking evaluation metrics used in this paper are described. Finally, we evaluate the proposed tracking method quantitatively.

A. Datasets and Compared Algorithms

For the sake of evaluating the proposed tracker, we capture six challenging video sequences. All the captured sequences

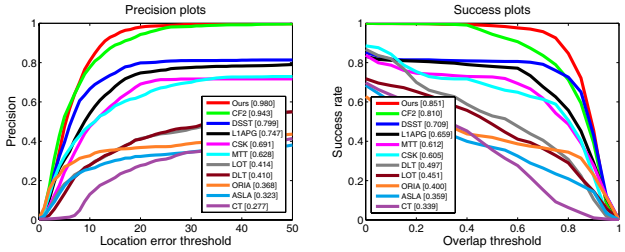

Fig. 3 Precision plot (left) and success plot (right) of all trackers, using one-pass evaluation (OPE).

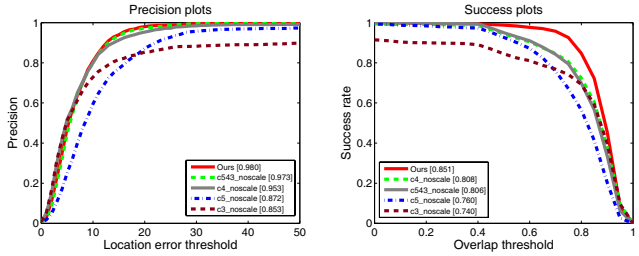

Fig. 4 Precision plot (left) and success plot (right) of the trackers using convolutional features, using one-pass evaluation.

are collected by the IR Thermal Imaging System in different time (e.g. day and night) and places (e.g. indoor and outdoor). These sequences contain two channels, one is the gray visible channel, the other is IR channel. Fig. 2. shows some representative images of both visible channel and the corresponding infrared channel. The challenges of the captured video sequences include scale variation, occlusion,in-plane rotation, out-of-plane rotation, fast motion and pose variation. The six video sequences are named Car, Head, HeadRotate, HeadCard, Roper and Book. The lengths of these sequences are 81, 464, 296, 462, 241 and 353, respectively. The frame rate of the captured videos is 8 fps.

In order to evaluate the tracking performance, ten state-of-the-art trackers are selected to compare with our algorithm. They are CT [17], ASLA [18], L1APG [19], LOT [20], ORIA [21], MTT [22], CSK [14], DSST [15], DLT [10] and CF2 [11]. The codes for the first seven tracking algorithms are from [2], and the others can be downloaded from the author's respective homepages.

B. Evaluation Metrics

In order to evaluate the proposed method, three objective metrics are used in this paper. They are precision, success rate and *average overlap ratio* (AOR). Precision denotes the percentage of frames whose predicted location is within a certain level distance from the ground truth. Success rate means the proportion of successful frames whose overlap between the tracked bounding box and the ground truth bounding box is larger than the threshold. AOR is the mean of overlap ratio of all the frames. The bigger AOR is, the better the tracker is.

C. Quantitative Evaluation

Fig. 3 shows overall performances for all trackers using both precision plot and success rate plot under *one-pass evaluation* (OPE). In general, the performance of our tracking algorithm is the best among all the trackers, and CF2 tracker is the second. On the one hand, because our tracker integrates both the spatial features and the semantic features by fusing multilayer convolutional features from visible images, the target location estimated by our tracker is more precise. Although CF2 uses the strategy of coarse-to-fine and hierarchical convolutional features, its distance precision is lower than ours, for the reason that the strategy of coarse-to-fine is easily disturbed by noise. However, our tracker reduce the effect of noise by concatenating convolutional feature maps. On the other hand, the success ratio of our tracker is also better than CF2, due to using a scale pyramid from infrared images.

In order to demonstrate the effectiveness of both the scale pyramid from infrared images and the fusion of multilayer convolutional feature maps, we compare our tracker with other tracking methods using convolutional features. Fig. 4 shows the results using both precision plot and success rate plot under one-pass evaluation. By comparing our method (Ours) with our proposed method without scale estimation (c543_noscale), we can draw a conclusion that the effect of the scale pyramid representation from infrared images is obvious. The precision plots are similar, but the success rate plot of our method with a scale pyramid representation is better than the other. By comparing with our method using the single-layer convolutional feature maps without scale estimation (c3_noscale, c4_noscale and c5_noscale), we find that the fusion of multilayer convolutional feature maps is superior to single-layer convolutional feature maps.

Table I presents the quantitative comparisons of average overlap ratio, using the red, blue and green font to denote the top three results. On the whole, drift occurs in almost all the trackers, except for CF2 and our tracker, it is because that the captured sequences contain much noise. The average overlap ratios of our method on all the sequences are in the top three. The scale pyramid representation is also used in DSST, but the average overlap ratio of DSST is worse than our method. The reasons may lie in two aspects. Firstly, the hand-crafted features are used in DSST, which causes the drift phenomenon in the Roper dataset. However, our tracker uses the fusion of multilayer convolutional features from VGG-Net-19. Secondly, DSST uses a scale pyramid of visible images to estimate the target scale, but the proposed tracker employs infrared images to construct the scale pyramid, since the appearance difference between the target and the background in infrared image is more obvious than that in gray image.

Fig. 5 shows both precision plot and success rate plot under one-pass evaluation for four challenge factors (scale variation, occlusion, deformation and fast motion) respectively. Our tracker is more robust to scale variation than other trackers, due to the use of the scale pyramid from infrared images. The difference between the target and the background in infrared images can be easily distinguished, which is useful for estimating the target scale, and the scale pyramid ensures that the proposed tracker can estimate the precise target scale. In addition, our tracker is also the best among all the trackers in

TABLE I Average overlap ratio. Red: best, Blue: second, Green: third.

	CT	ASLA	L1APG	ORIA	MTT	CSK	DLT	DSST	CF2	LOT	Ours
Car	0.6	0.75	0.75	0.9	0.73	0.66	0.74	0.89	0.67	0.8	0.88
Head	0.18	0.08	0.9	0.17	0.86	0.84	0.32	0.9	0.9	0.26	0.92
HeadRotate	0.23	0.25	0.86	0.28	0.83	0.87	0.6	0.83	0.87	0.26	0.88
HeadCard	0.68	0.77	0.92	0.83	0.91	0.9	0.84	0.93	0.92	0.83	0.92
Roper	0.31	0.21	0.04	0.21	0.32	0.25	0.39	0.31	0.81	0.51	0.86
Book	0.03	0.1	0.56	0.03	0.07	0.16	0.11	0.46	0.78	0.08	0.76
Average	0.34	0.36	0.67	0.4	0.62	0.61	0.5	0.72	0.83	0.45	0.87

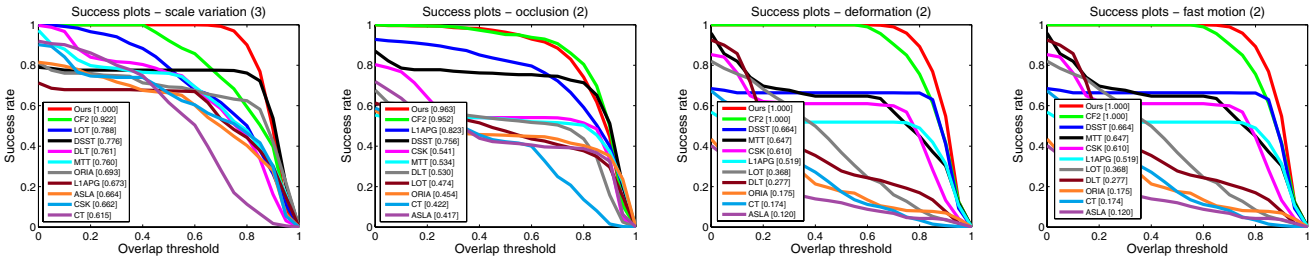

Fig. 5 Precision plot (up) and success plot (down) over four challenges (scale variation, occlusion, deformation and fast motion). One-pass evaluation used.

the aspects of deformation and fast motion for using the fusion of hierarchical convolutional features. However, the proposed tracker is slightly poor in the aspect of occlusion, due to the use of holistic feature representation and correlation filter which isn't robust to occlusion for a long time.

V. Conclusion

We develop a robust tracking algorithm based on hierarchical convolutional features, correlation filters, scale pyramid representation and dual fusion. Both the spatial features of shallow layers and the semantic features of deep layers are integrated to estimate the precise target location. The scale pyramid of infrared images and discriminative correlation filters are used to predict the target scale precisely and efficiently. Both visible images and infrared images are integrated to our tracker to improve the tracking performance. Experimental results show that the proposed tracker is better than all the compared tracking algorithms.

References

[1] Q. Wang, J. Fang, and Y. Yuan, "Multi-cue based tracking," *Neurocomputing*, vol. 131, pp. 227–236, 2014.

[2] Y. Wu, J. Lim, and M. Yang, "Online object tracking: A benchmark," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 2411–2418.

[3] X. Lu, Y. Yuan, and P. Yan, "Robust visual tracking with discriminative sparse learning," *Pattern Recognition*, vol. 46, no. 7, pp. 1762–1771, 2013.

[4] R. A. Newcombe, D. Fox, and S. M. Seitz, "Dynamicfusion: Reconstruction and tracking of non-rigid scenes in real-time," in *Proc. IEEE International Conference on Computer Vision and Pattern Recognition*, 2015, pp. 343–352.

[5] L. Ma, J. Lu, J. Feng, and J. Zhou, "Multiple feature fusion via weighted entropy for visual tracking," in *Proc. IEEE International Conference on Computer Vision*, 2015, pp. 3128–3136.

[6] Y. Yuan, J. Fang, and Q. Wang, "Robust superpixel tracking via depth fusion," *IEEE Trans. Circuits Syst. Video Techn.*, vol. 24, no. 1, pp. 15–26, 2014.

[7] S. Kumar, T. K. Marks, and M. J. Jones, "Improving person tracking using an inexpensive thermal infrared sensor," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 217–224.

[8] H. Li, Y. Li, and F. Porikli, "Deeptrack: Learning discriminative feature representations by convolutional neural networks for visual tracking," in *Proc. British Machine Vision Conference*, 2014.

[9] L. Wang, N. T. Pham, T.-T. Ng, G. Wang, K. L. Chan, and K. Leman, "Learning deep features for multiple object tracking by using a multi-task learning strategy," in *Proc. IEEE International Conference on Image Processing*. IEEE, 2014, pp. 838–842.

[10] N. Wang and D. Yeung, "Learning a deep compact image representation for visual tracking," in *Proc. Annual Conference on Neural Information Processing Systems*, 2013, pp. 809–817.

[11] C. Ma, J. Huang, X. Yang, and M. H. Yang, "Hierarchical convolutional features for visual tracking," in *Proc. IEEE International Conference on Computer Vision*, 2015, pp. 3074–3082.

[12] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. International Conference on Learning Representations*, 2015.

[13] L. Wang, W. Ouyang, X. Wang, and H. Lu, "Visual tracking with fully convolutional networks," in *Proc. IEEE International Conference on Computer Vision*, 2015, pp. 3119–3127.

[14] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista, "Exploiting the circulant structure of tracking-by-detection with kernels," in *Proc. European Conference on Computer Vision*, 2012, pp. 702–715.

[15] M. Danelljan, G. Häger, F. S. Khan, and M. Felsberg, "Accurate scale estimation for robust visual tracking," in *Proc. British Machine Vision Conference*, 2014.

[16] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista, "High-speed tracking with kernelized correlation filters," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 3, pp. 583–596, 2015.

[17] K. Zhang, L. Zhang, and M. Yang, "Real-time compressive tracking," in *Proc. European Conference on Computer Vision*, 2012, pp. 864–877.

[18] X. Jia, H. Lu, and M. Yang, "Visual tracking via adaptive structural local sparse appearance model," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 1822–1829.

[19] C. Bao, Y. Wu, H. Ling, and H. Ji, "Real time robust L1 tracker using accelerated proximal gradient approach," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 1830–1837.

[20] S. Oron, A. Bar-Hillel, D. Levi, and S. Avidan, "Locally orderless tracking," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 1940–1947.

[21] Y. Wu, B. Shen, and H. Ling, "Online robust image alignment via iterative convex optimization," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 1808–1814.

[22] T. Zhang, B. Ghanem, S. Liu, and N. Ahuja, "Robust visual tracking via multi-task sparse learning," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 2042–2049.