Robust Visual Tracking via Basis Matching

Shengping Zhang, *Member, IEEE,* Xiangyuan Lan, *Student Member, IEEE,* Yuankai Qi, Pong C. Yuen, *Senior Member, IEEE*

***Abstract*—Most existing tracking approaches are either based on the tracking by detection framework or the tracking by matching framework. The former needs to learn a discriminative classifier using positive and negative samples, which will cause tracking drift due to unreliable samples. The later usually performs tracking by matching local interest points between a target candidate and the tracked target, which is not robust to target appearance changes over time. In this paper, we propose a novel tracking by matching framework for robust tracking based on basis matching rather than point matching. In particular, we learn the target model from target images using a set of Gabor basis functions, which have large responses on the corresponding spatial positions after a max pooling. During tracking, a target candidate is evaluated by computing the responses of the Gabor basis functions on their corresponding spatial positions. Experimental results on a set of challenging sequences validate that the performance of the proposed tracking method outperforms several state-of-the-art methods.**

***Index Terms*—Visual tracking, Gabor filtering, max pooling, tracking by matching, particle filter**

I. Introduction

VISUAL tracking is one of the fundamental problems in computer vision with numerous applications including intelligent video surveillance [1], traffic monitoring [2], [3], sport video analysis [4], and human-robot interaction [5]. Despite it has attracted increasing interests in recent decades and significant progress have been achieved [6]–[16], visual tracking is still far away from practical use due to unreliable tracking performance caused by large appearance changes such as occlusion, deformation, abrupt motion, illumination variation, and background clutter, etc.

The existing tracking approaches try to improve tracking performance by developing different tracking frameworks [17]–[21]. Two main tracking frameworks are tracking by detection [22]–[26] and tracking by matching [27]–[30]. The former formulates visual tracking as a classification problem and learns a discriminative classifier from the collected positive and negative samples to classify target candidates as the tracked target or background. Due to the unreliable samples collected, the tracker is easy to drift to the background. Different from tracking by detection methods which rely constructing a classifier for tracking, tracking by matching methods track the target by matching the keypoints of interest across consecutive frames. Since key points of interest contain much information of local regions of the tracked target, tracking by matching is capable of tracking the target even if the target suffers appearance variations such as pose changes and partial occlusions. In addition, when some kinds of post-processing methods are used, the tracking performance can be improved further. For example, when clustering the matched points into background pixels and foreground pixels, the matching among background pixels can be excluded, which will further improve tracking performance.

In this paper, we also focus on tracking by matching. However, different from existing tracking by matching methods which only match key points, our idea is to match local regions. Intuitively, matching regions is more reliable than matching pixels especially when the target suffers large appearance changes such as illumination changes or the video quality is not very good with too much noise. In contrast, an image region contains more information than a pixel, which should be more useful to locate the target in subsequent frames. To match regions across frames, we adopt an indirect method rather than directly computing the similarity between the corresponding regions of two frames. In particular, we first learn the template of the tracked target by computing Gabor responses on a set of target template images, then the target template consists of a set of pairs of locations and Gabor kernels. To locate the target in the current frame, the learned target template is used to compute responses of the Gabor kernels on their corresponding positions of the current frame. The tracked target is in the position where the sum of responses of all Gabor kernels is the largest. Our tracking results on a set of fifteen challenging sequences indicate that the proposed method outperforms several state-of-the-art tracking methods.

II. Related Work

In the literature, several appearance models based on sparse coding for visual tracking have been proposed, which usually consist of two steps: local coding and global pooling. After the input image is divided into a set of image patches, each patch is sparsely coded by a linear combination of a set of basis functions. The combination coefficients are used as codes to describe the image patch. The final appearance representation of the input image is obtained by pooling the codes of all patches to form a histogram-like feature vector. Zhang et al. [31] propose to code image patches using independent

This work was supported in part by the National Natural Science Foundation of China (No. 61300111), the China Postdoctoral Science Foundation (No. 2014M550192), and the General Research Fund through the Research Grants Council, Hong Kong, under Grant HKBU 212313.

S. Zhang is with the School of Computer Science and Technology, Harbin Institute of Technology, Weihai 264209, P. R. China. This work was partly done when S. Zhang was a Post-Doctoral Research Fellow with the Department of Computer Science, Hong Kong Baptist University, Hong Kong, P. R. China. E-mail: shengping.zhang@gmail.com.

X. Lan and P.C. Yuen are with the Department of Computer Science, Hong Kong Baptist University, Hong Kong, P. R. China. E-mail: {xylan, pcyuen}@comp.hkbu.edu.hk.

Y. Qi is with Key Laboratory of Intelligent Information Processing of Chinese Academy of Sciences (CAS), Institute of Computing Technology, CAS, Beijing 100190, P. R. China. E-mail: yuankai.qi@vipl.ict.ac.cn.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2016.2539860, IEEE Transactions on Circuits and Systems for Video Technology

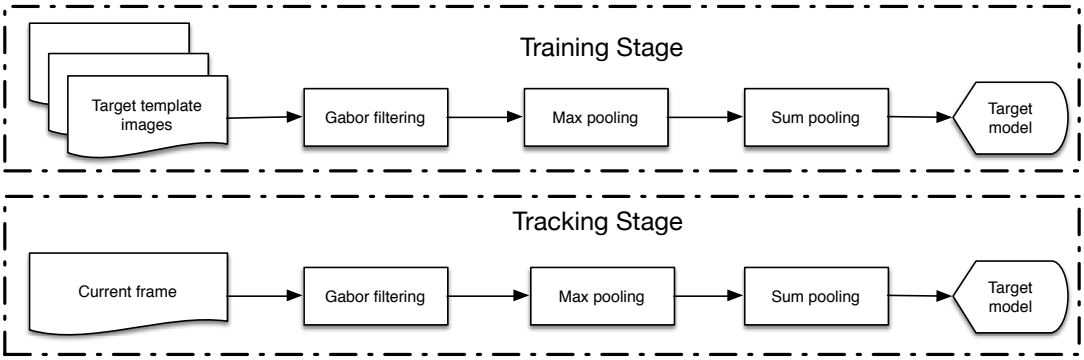


Fig. 1. The conceptual diagram of the proposed tracking method.

component analysis (ICA) basis functions, which are learned from a set of randomly sampled image patches. The average pooling is used to obtain the final appearance representation. In [32], each patch is coded by sparse coding basis functions, which are learned using the sampled patches from the target image. Similar to [31], average pooling is also used to get the final feature representation. Wang et al. [33] combine the sampled patches from the target image and the identity vectors as basis functions to code each patch. All codes are concatenated to form the final feature representation. Zhong et al. [34] learn sparse coding basis functions from the patches sampled from the target image by k-means. The pooling step is similar to [33]. Jia et al. [35] maintain multiple target templates. With a fixed spatial layout, local patches sampled from each template are collected together as basis functions to code patches sampled from the input image with the same spatial layout. They use an alignment pooling operator to preserve the structural relationship among local codes.

Although these appearance models based on sparse coding have been proposed in the literature, they have some shortcomings: 1) The basis functions of sparse coding are obtained by either learning from random images [31] or directly sampling the patches from the target images [35]. The basis functions are too generative that the resulting sparse codes when coding image patches from the target are not discriminative enough to distinguish them from the background. 2) The sparse codes are pooled by either summing all the coefficient vectors over the entire image which losses the spatial relationship [31], [32] or concatenating all the coefficient vectors which preserves the spatial relationship while degrades the local invariance of the resulting appearance representation [33]. 3) As supported by the biological studies [36], [37] and empirical evaluation [38], [39], non-linear operators such as whitening, rectification and normalization operators can also enhance the representation ability. However, the existing appearance models based on sparse coding only contain the basic coding and pooling operators and ignore the importance of the non-linear operators.

Given the dynamic nature of object tracking, having an online learning mechanism to update the target model is vital to tracking. A large body of work in the literature is dedicated to addressing this issue [22], [27], [40]–[43]. In this part, we will review the significant steps that have been taken to obtain a robust target model.

Some early attempts (*e.g.*, [44], [45]) update the target model by combining the old target model and the tracked current appearance with a proper weighting function. The weighting function is foreseen to adjust the effect of the current appearance on the model. In [40], Jepson et al. propose a probabilistic target model and devise an updating scheme based on the EM algorithm. The probabilistic model takes into account the stable part of the appearance (slowly varying image observations) and the possibility of losing the target due to occlusion, or noise in a unified framework for its decisions.

The concept of feature selection has been widely used for designing robust target models. A pioneer study along this direction is the work of Collins *et al.* where an online and discriminative feature selection scheme is introduced [6]. The idea of online learning for subspaces is developed by Ross *et al.* [42], where a low-dimensional subspace capturing the target appearance is incrementally updated using the past and current tracking results.

For the sparse trackers, the most popular approach for updating the target model is to learn the dictionary in an online fashion. For example, the differences between the current and previous target samples are used in [27] for adapting the dictionary. In [43], an online learning method for creating non-negative dictionaries is proposed. Since solving the optimization problems involving ℓ_1 norms is computationally expensive, Wang et al. in [43] suggest to update the dictionary by gradient descent methods. In [22], a similar idea (gradient descent for updating the dictionary) is utilized albeit authors argued that a discriminative dictionary could be attained by utilizing two disjoint dictionaries to model foreground and background of the target. The interested reader is referred to [46]–[49] for more complete discussions and comprehensive comparisons.

III. Proposed Method

A. Overview

The proposed tracking method consists two stages: training the target template and tracking the target. In the training stage, a set of target template images are used to learn the target appearance model which consists of a set of pairs locations and Gabor basis. In the tracking stage, the learned target model

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2016.2539860, IEEE Transactions on Circuits and Systems for Video Technology

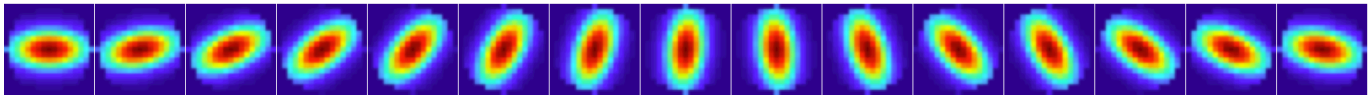


Fig. 2. The visualization of the used Gabor kernels.

is used to locate the target in all possible positions by finding the position with the maximal responses with the basis in the corresponding positions. A conceptual diagram is shown in Figure 1.

B. Target model learning

Given a set of target images, we exploit the active basis model proposed in [51] to learn the target appearance model, which consists a set of pairs of locations and Gabor basis functions such that the Gabor basis functions have the largest responses in the corresponding positions. The reasons why the Gabor basis functions are used are two-folds: First, it is very efficient to compute the responses of Gabor basis functions on local patches of an image. Second, the Gabor basis functions model the local appearance using directional bars which are easily rotated and shifted to capture local appearance changes.

The Gabor basis functions with orientation μ and scale ν are defined as [50]

$$\Phi_{\mu,\nu}(z) = \frac{\|\kappa_{\mu,\nu}\|^2}{\delta^2} e^{(-\|\kappa_{\mu,\nu}\|^2\|z\|^2/2\delta^2)} \left[e^{i\kappa_{\mu,\nu}z - e^{-\delta^2/2}}\right] \quad (1)$$

where $z = (x, y)$ denote the pixel at the position (x, y), and the wave vector $\kappa_{\mu,\nu}$ is defined as $\kappa_{\mu,\nu} = \kappa_\nu^{i\phi_\mu}$ with $\kappa_\nu = \kappa_{max}/f^\nu$ and $\phi_\mu = \pi\mu/8$. κ_{max} is the maximum frequency, f is the spacing factor between kernels in the frequency domain and δ determines the ration of the Gaussian window width to wavelength. After obtaining the Gabor kernels, the responses of an image I with a Gabor kernel $\Phi_{\mu,\nu}$ is computed as $G_{\mu,\nu}(z) = I(z) * \Phi_{\mu,\nu}(z)$. Figure 2 shows 15 Gabor kernels with 15 orientations used in this work.

Given a set of N training images containing only the tracked targets as shown in the left column of Figure 3, the Gabor filtering responses of all training images with 15 Gabor kernels can be computed as shown in the right column of Figure 3. Let $G^1_{\mu,\nu}, G^2_{\mu,\nu}, \ldots, G^N_{\mu,\nu}$ be Gabor responses of the training images at scale μ and orientation ν. As we can see that the larger the response is, the more similar the local appearance of the target is with the Gabor kernel. Therefore, the Gabor kernels can be used to model the local appearance of the tracked target. Considering that the target has local deformations among all target template images, we adopt the same strategy as in [51] to shift the Gabor kernels in both local spatial and orientation ranges. In particular, we shift the Gabor kernels in a spatial range with radius of $R_s = 4$ and in an orientation range with radius of $R_o = 1$. An example of shifting with $R_s = 4$ and $R_o = 1$ is shown in Figure 4. Please note that the spatial shift is along the normal vector direction which is perpendicular to the Gabor kernel. Therefore a total of $(2 \times 4 + 1) \times (2 \times 1 + 1) = 27$ shifts are produced. For each spatial position z, its response after shifting can be obtained by a max pooling as

$$G^i_{\mu,\nu}(z) = max\{G^i(\mu, \nu)(z) | \nu \in \Omega_o, z \in \Omega_z\} \quad (2)$$

where $\Omega_o = \{\nu - R_o, \ldots, \nu + R_o\}$ are the shift orientations and $\Omega_z = \{z - R_s, \ldots, z + R_S\}$ are the shift spatial positions.

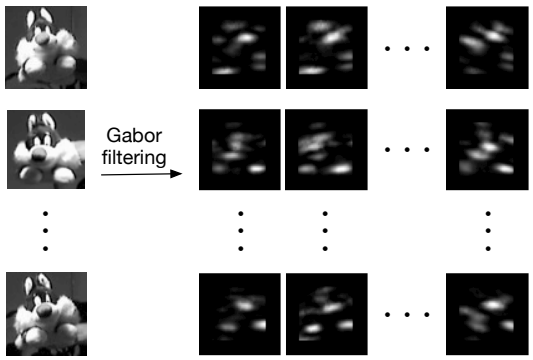


Fig. 3. Examples of the Gabor filtering responses.

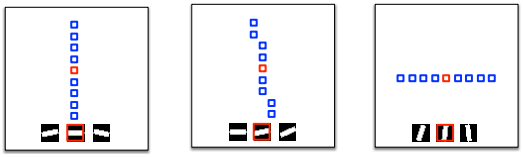


Fig. 4. Examples of shifting three Gabor kernels in both spatial and orientation ranges.

After max pooling, for each spatial position, we sum the responses at the same position cross all training images to get a response map, which can be computed as

$$G_{\mu,\nu}(z) = \sum_{i=1}^{N} G^i_{\mu,\nu}(z) \quad (3)$$

From the response map we can see that which Gabor kernel has the largest response in a spatial position. Based on this response map, we can find a set of $\{z_m, \mu_m, \nu_m\}_{1:M}$, which can be used as the target template. An example is shown in Figure 5 where the most left image is the obtained target template consisting of a set of Gabor kernels on their corresponding positions and the remaining images are the training target images and their corresponding Gabor kernels that match best to the target. As we can see that the target in five target images have some variations and the Gabor kernels obtained from the target template capture well the target in the training images.

Fig. 5. An example of the learned target template on *sylv* sequence. The most left image is the obtained target template consisting of a set of Gabor kernels on their corresponding positions and the remaining images are the training target images and thee corresponding Gabor kernels that match best to the target.

C. Target tracking based on Gabor kernel matching

The particle filter [52] is adopted in our tracking method, which is a popular computation method to recursively approximates the posterior distribution of the state variables characterizing a dynamic system. It consists of two stages: prediction and update. Let $\mathbf{z}_t$ and $\mathbf{I}_t$ be the state variable and the observation at time t, respectively. The posterior distribution of $\mathbf{z}_t$ given all the available observations $\mathbf{I}_{1:t-1} = \{\mathbf{I}_1, \mathbf{I}_2, \ldots, \mathbf{I}_{t-1}\}$ up to time $t-1$ can be predicated using the state transition model $p(\mathbf{z}_t|\mathbf{z}_{t-1})$ as

$$p(\mathbf{z}_t|\mathbf{I}_{1:t-1}) = \int p(\mathbf{z}_t|\mathbf{z}_{t-1})p(\mathbf{z}_{t-1}|\mathbf{I}_{1:t-1})d\mathbf{z}_{t-1} \quad (4)$$

At time t, the observation $\mathbf{I}_t$ is available, the posterior distribution of $\mathbf{z}_t$ is updated using the Bayes rule as

$$p(\mathbf{z}_t|\mathbf{I}_{1:t}) = \frac{p(\mathbf{I}_t|\mathbf{z}_t)p(\mathbf{z}_t|\mathbf{I}_{1:t-1})}{p(\mathbf{I}_t|\mathbf{I}_{1:t-1})} \quad (5)$$

Using a sequential importance sampling technique, we can approximate the posterior distribution $p(\mathbf{z}_t|\mathbf{I}_{1:t})$ by a set of N weighted samples (also called particles) $\{\mathbf{z}_t^i, w_t^i\}_{i=1,\ldots,N}$, where w_t^i are the importance weights of particles $\mathbf{z}_t^i$. Let $q(\mathbf{z}_t|\mathbf{I}_{1:t}, \mathbf{z}_{1:t-1})$ be the importance distribution from which the particles are drawn, the importance weights w_t^i are updated as

$$w_t^i = w_{t-1}^i \frac{p(\mathbf{I}_t|\mathbf{z}_t^i)p(\mathbf{z}_t^i|\mathbf{z}_{t-1}^i)}{q(\mathbf{z}_t|\mathbf{I}_{1:t}, \mathbf{z}_{1:t-1})} \quad (6)$$

To avoid the degeneracy case where the weights of some particles possibly keep increasing, particles are resampled according to their importance weights so as to generate a set of equally weighted particles. In the case of bootstrap filtering [52], the state transition distribution is chosen as the importance distribution $q(\mathbf{z}_t|\mathbf{I}_{1:t}, \mathbf{z}_{1:t-1}) = p(\mathbf{z}_t|\mathbf{z}_{t-1}) \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$ where $\mathbf{\Sigma}$ is a diagonal matrix, the weights are updated as the observation likelihood $w_t^i = p(\mathbf{I}_t|\mathbf{z}_t^i)$.

The key step of applying particle filtering for visual tracking [44], [53] is to compute the weight for each particle using the observation likelihood. In practice, the observation likelihood $p(\mathbf{I}_t|\mathbf{z}_t^i)$ is computed as the similarity between the target template image and the target candidate image parameterized by the particle $\mathbf{z}_t^i$. In the next subsection, we present how to combine our proposed appearance model with a multi-scale pyramid matching to assign an appropriate weight to each particle.

For particle i[1], an image block I^i can be obtained using particle parameters. The Gabor response of the image block can be computed and denoted as $I^i_{\mu,\nu}$. The particle weight can be computed as

$$w^i = \sum_{m=1}^{M} I^i_{\mu_m,\nu_m}(z_m) \quad (7)$$

The tracking result in the current frame is the particle with the largest weight whose index can be computed as

$$i^* = \arg \max_{i=1,\ldots,N} w_i \quad (8)$$

IV. Experimental Results

In this section, we present extensive experimental results to evaluate the effectiveness of the proposed method. We first introduce the evaluate protocol in section IV-A including the baseline trackers, test sequences and evaluation criteria. The evaluations are presented in section IV-B and section IV-C, respectively.

A. Evaluation protocol

1) Baseline trackers: We select two groups of trackers from the literature for performance comparison. The first group consists of trackers that are based on sparse coding as reviewed in section II, including the LSA tracker [32], the OLSR tracker [33], the SCM tracker [34] and the SLSA tracker [35]. The second group includes some popular trackers including the fragment-based tracker (Frag) [45], the incremental visual tracker (IVT) [42], the multiple instance learning tracker (MIL) [54] and the online AdaBoost tracker (OAB) [55]. Here we use the source codes downloaded from the authors' websites. Each tracker is run with the optimized parameters.

2) Test sequences: In order to evaluate the performance of the proposed method, we collected fifteen publicly available test sequences including *faceocc2*, *car4*, *david_outdoor*, *david_indoor*, *sylv*, *lemming*, *basketball*, *box*, *CAVIAR*, *bird_2*, *woman*, *singer1*, *face*, *PETS* and *board*[2]. These sequences are recorded either indoor or outdoor and the challenges of these videos include partial occlusion, pose variation, illumination variation, scale change and background clutter. For each sequence, the ground truth for quantitative evaluation is obtained by manually labeling the position and size of the target every five frames.

[1]For simplicity, we ignore the time subscribe.

[2]The *car4*, *david_outdoor*, *david_indoor* and *sylv* can be downloaded from http://www.cs.toronto.edu/~dross/ivt/. The *lemming* and *board* can be downloaded from http://gpu4vision.icg.tugraz.at/subsites/prost. The *CAVIAR* can be downloaded from http://groups.inf.ed.ac.uk/vision/CAVIAR/CAVIARDATA1/. The *faceocc2* can be downloaded from http://vision.ucsd.edu/~bbabenko/project_miltrack.shtml. The *PETS* can be downloaded from http://www.cvg.rdg.ac.uk/PETS2001/. The *face* and *woman* can be downloaded from http://www.cs.technion.ac.il/~amita/fragtrack/fragtrack.htm. The *singer1* and *basketball* can be downloaded from http://cv.snu.ac.kr/research/~vtd/index.html. The *bird_2* can be downloaded from http://ice.dlut.edu.cn/lu/Project/iccv_spt_webpage/iccv_spt.htm

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2016.2539860, IEEE Transactions on Circuits and Systems for Video Technology

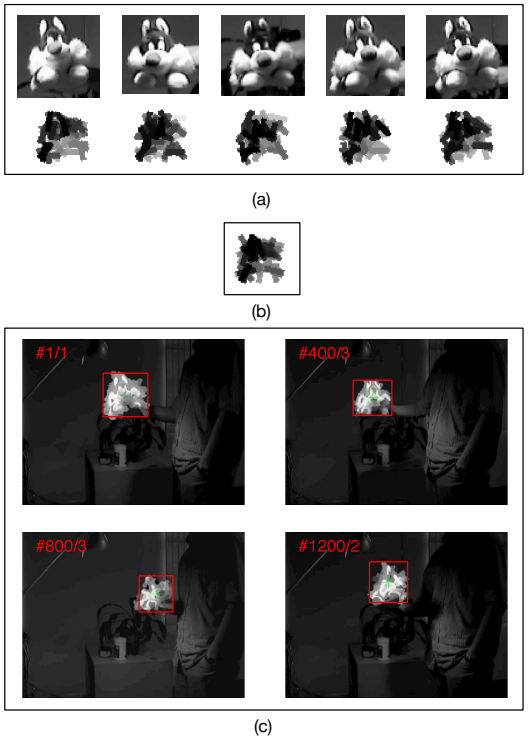


Fig. 7. Intermediate tracking results of the sylv sequence. (a) shows five training images and their corresponding matched models. (b) shows the learned template. (C) shows tracking results of four frames.

3) Evaluation criteria: Two evaluation criteria are employed to quantitatively assess the performance of the selected trackers. The first one is *tracking success rate*, which computes the percentage of correctly tracked frames in a sequence. To evaluate whether the target is correctly tracked or not in a frame, we adopt the PASCAL score [56], which can be computed as $\frac{area(\mathcal{R}^* \cap \mathcal{R}_{gt})}{area(\mathcal{R}^* \cup \mathcal{R}_{gt})}$ where $\mathcal{R}^*$ is the bounding box obtained by a tracker, $\mathcal{R}_{gt}$ is the corresponding ground truth bounding box and $area(\mathcal{R})$ is the area of the bounding box $\mathcal{R}$. The target is correctly tracked in a frame if the score is larger than 0.5. The second one is the *relative center position error*, which computes the distance between the center position of the tracking result and the ground truth relative to the size of the ground truth. At time t, let (x_t, y_t) be the center position of the tracking result, $(\hat{x}_t, \hat{y}_t, \hat{w}_t, \hat{h}_t)$ be the center position and size of the ground truth, the relative center position error at time t can be computed as $\sqrt{(\frac{x_t - \hat{x}_t}{\hat{w}_t})^2 + (\frac{y_t - \hat{y}_t}{\hat{h}_t})^2}$. We also use the average of the relative center position errors over the entire sequence to evaluate the overall performance of a tracker.

B. Quantitative Comparison

We compare the proposed method with the selected eight trackers on all fifteen sequences. The tracking success rates and average relative center position errors are shown in Tables I and II, respectively. As we can see that the proposed tracker achieves the highest tracking success rates on twelve sequences and the lowest average relative center position errors on ten sequences. These results indicate the overall performance of the proposed tracker on all sequences significantly outperforms other trackers. For some sequences, such as *car4* and *PETS*, the proposed tracker still achieves 80% above tracking success rates. In Figure 6, we also show the relative center position errors over time on all sequences. As we can see that the proposed tracker achieves very low relative center errors over time on most sequences.

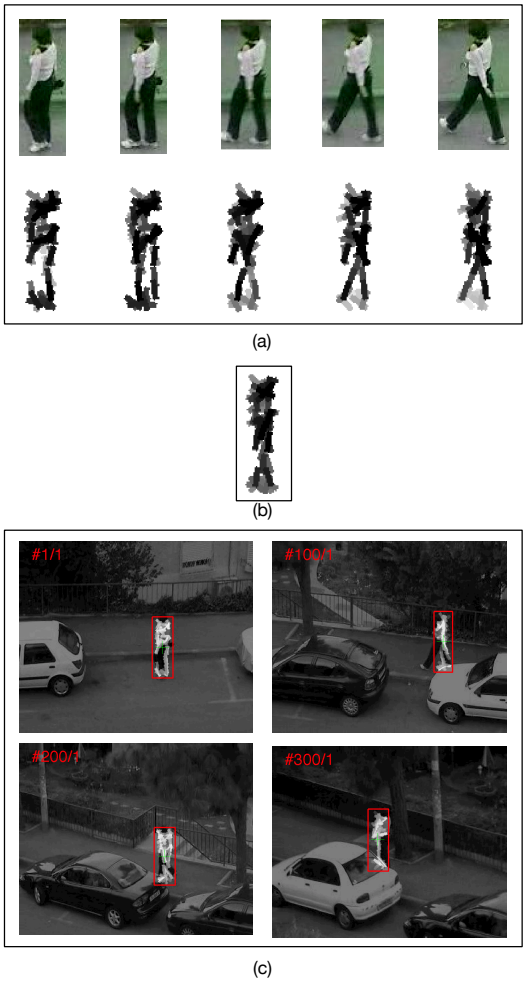


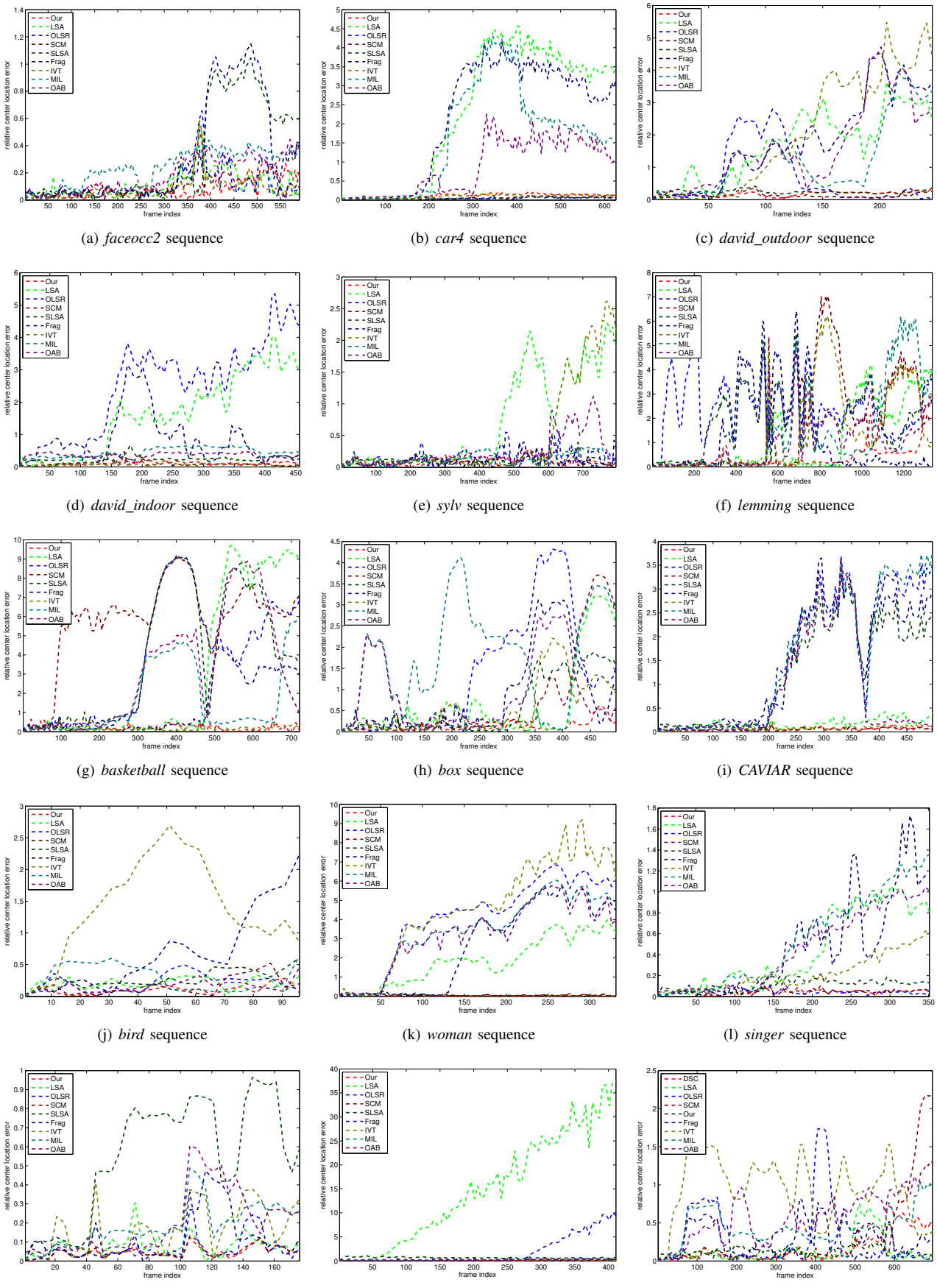
Fig. 8. Intermediate tracking results of the woman sequence. (a) shows five training images and their corresponding matched models. (b) shows the learned template. (C) shows tracking results of four frames.

C. Qualitative Comparison

We first show some intermediate tracking results on the *sylv* and *woman* sequences in Figures 7 and 8, respectively. As we can see that only five images are needed to learn an accuracy target model. The learned target template can accurately match both the training images and the four test images. Two representative tracking results of the LSA, SLSA, IVT, MIL and our trackers on fifteen sequences are shown in Figure 9. In the following, we will compare how they perform when the target undergoes occlusion, illumination changes, pose changes or appear in background clutters.

Occlusion: Figures 9(a)–9(c) present tracking results on the *face*, *faceocc2* and *woman* sequences which mainly involve partial occlusion. The SLSA tracker fails to track the face in

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2016.2539860, IEEE Transactions on Circuits and Systems for Video Technology



(a) *faceocc2* sequence (b) *car4* sequence (c) *david_outdoor* sequence

(d) *david_indoor* sequence (e) *sylv* sequence (f) *lemming* sequence

(g) *basketball* sequence (h) *box* sequence (i) *CAVIAR* sequence

(j) *bird* sequence (k) *woman* sequence (l) *singer* sequence

(m) *face* sequence (n) *PETS* sequence (o) *board* sequence

Fig. 6. The relative center position error plots on fifteen test sequences.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2016.2539860, IEEE Transactions on Circuits and Systems for Video Technology

TABLE I
THE TRACKING SUCCESS RATES ON FIFTEEN TEST SEQUENCES BY THE PROPOSED TRACKER AND OTHER BASELINE TRACKERS. FOR EACH SEQUENCE, THE BEST RESULT IS SHOWN IN RED FONT.

	Our	LSA	OLSR	SCM	SLSA	Frag	IVT	MIL	OAB
faceocc2	**0.99**	0.95	0.86	0.97	0.65	0.68	0.97	0.54	0.83
car4	0.89	0.32	**1.00**	0.98	0.96	0.27	0.93	0.38	0.57
david_outdoor	0.95	0.20	0.59	**0.97**	0.79	0.25	0.32	0.20	0.14
david_indoor	**1.00**	0.29	0.23	0.96	0.71	0.21	0.91	0.19	0.11
sylv	**1.00**	0.54	0.90	**1.00**	0.98	0.91	0.76	0.91	0.80
lemming	**0.63**	0.48	0.13	0.48	0.18	0.50	0.40	0.38	0.36
basketball	**0.76**	0.35	0.22	0.18	0.16	0.45	0.43	0.28	0.30
box	**0.95**	0.53	0.44	0.63	0.61	0.47	0.57	0.17	0.30
CAVIAR	**1.00**	0.79	0.40	0.99	0.40	0.35	1.00	0.39	0.36
bird_2	**0.96**	0.63	0.65	0.79	0.58	0.34	0.17	0.51	0.90
woman	**1.00**	0.15	0.14	0.99	0.97	0.38	0.14	0.25	0.31
singer1	**1.00**	0.26	0.98	0.99	0.66	0.25	0.27	0.25	0.27
face	**1.00**	0.97	0.98	0.93	0.25	0.85	0.79	0.82	0.75
PETS	0.84	0.11	0.62	0.95	0.56	0.50	**0.99**	0.53	0.58
board	**0.96**	0.88	0.73	0.74	0.16	0.55	0.14	0.26	0.21

TABLE II
THE AVERAGE RELATIVE CENTER ERRORS ON FIFTEEN TEST SEQUENCES BY THE PROPOSED TRACKER AND OTHER BASELINE TRACKERS. FOR EACH SEQUENCE, THE BEST RESULT IS SHOWN IN RED FONT.

	Our	LSA	OLSR	SCM	SLSA	Frag	IVT	MIL	OAB
faceocc2	**0.08**	0.11	0.11	0.10	0.31	0.31	0.09	0.24	0.18
car4	0.11	2.28	**0.05**	0.05	0.05	2.14	0.09	1.62	0.88
david_outdoor	**0.13**	1.69	0.68	0.18	0.25	1.85	2.08	1.30	1.64
david_indoor	0.11	1.64	2.41	0.10	0.26	0.84	**0.07**	0.44	0.35
sylv	**0.09**	0.71	0.14	0.09	0.12	0.11	0.48	0.14	0.24
lemming	**0.39**	1.15	2.50	1.75	1.96	1.24	1.48	1.47	1.01
basketball	0.21	3.01	3.74	5.76	3.88	1.28	**0.14**	1.61	3.15
box	**0.13**	0.65	1.34	0.74	0.60	0.84	0.55	1.80	0.90
CAVIAR	**0.05**	0.17	1.55	0.05	1.28	1.45	0.08	1.55	0.80
bird_2	**0.11**	0.20	0.22	0.17	0.29	0.71	1.43	0.31	0.17
woman	0.05	1.86	4.13	**0.04**	0.05	2.62	4.63	3.45	3.25
singer1	**0.04**	0.44	0.04	0.04	0.11	0.48	0.22	0.55	0.45
face	**0.05**	0.08	0.07	0.05	0.55	0.12	0.17	0.19	0.20
PETS	0.13	14.97	1.84	0.09	0.67	0.23	**0.07**	0.16	0.13
board	**0.13**	0.15	0.29	0.31	0.87	0.30	0.95	0.45	0.57

Figure 9(a) where the book has different color with the face. In Figure 9(c), when the real windshield of the car has similar color with the woman's legs, the SLSA tracker is able to track the woman accurately. The LSA tracker is capable of handling partial occlusion when the target has strong contrast with the background (e.g. in *face* and *faceocc2* sequences). However, when a background region is similar to the target as shown in Figure 9(c), the LSA tracker drifts to the background region. The IVT and Our trackers perform well in these sequences.

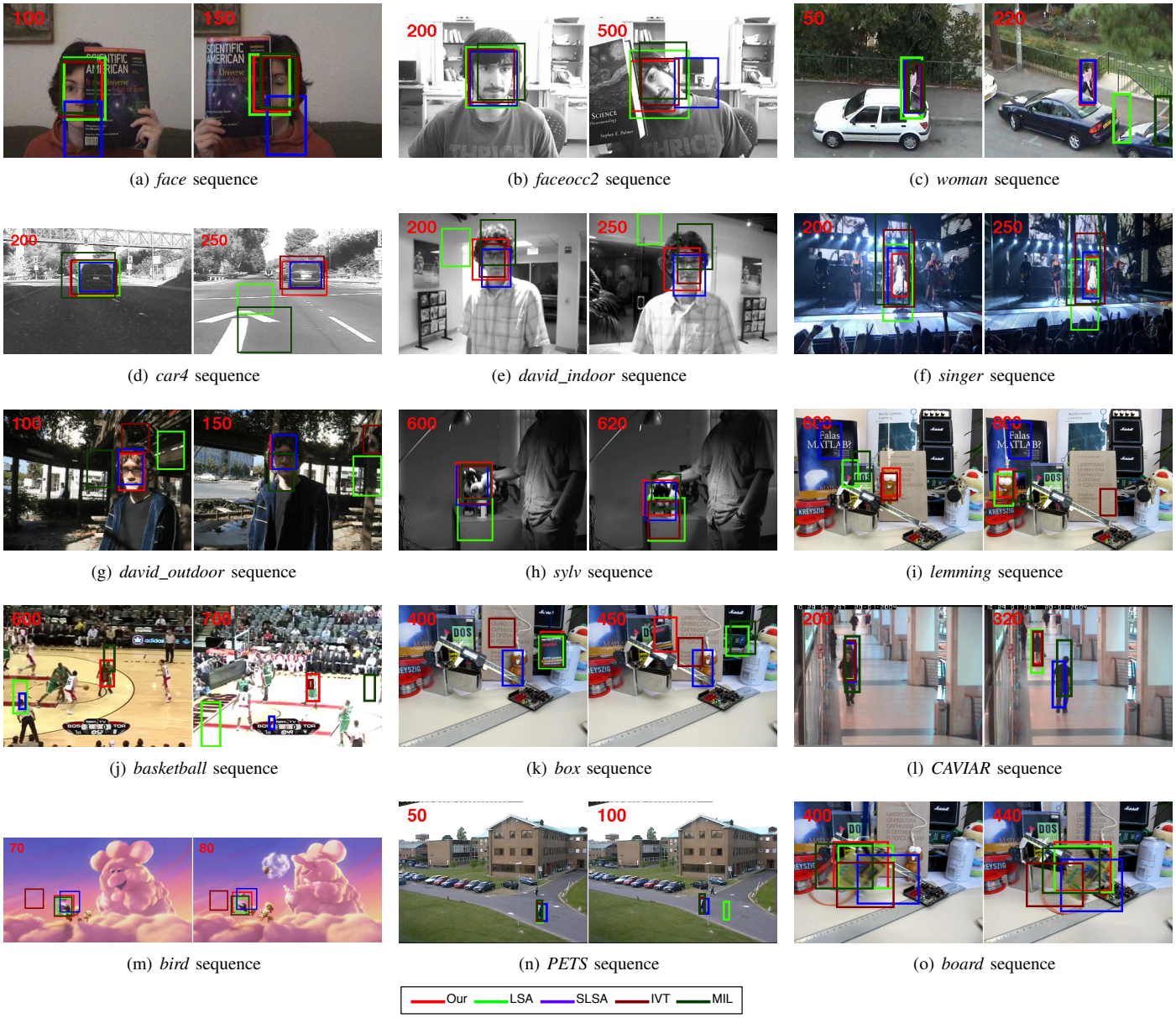
Illumination changes: The *car4*, *david_indoor*, *singer* and *david_outdoor* sequences suffer from severe illumination changes as shown in Figures 9(d)–9(g). Both our tracker and the SLSA tracker successfully track the target on these sequences. The LSA tracker drifts to the background far away from the target in *car4*, *david_indoor* and *david_outdoor*. In *singer1*, it also has small drift away from the target. Similarly, the MIL and IVT trackers also drift to the background due to the several illumination changes. Our tracker uses the pre-processing and normalization operators and is able to handle illumination changes even if the stage light changes drastically as seen from Figure 9(f).

Pose change: The *sylv*, *basketball* and *bird* involve pose changes. It should be noted that the rectangle region of the tracked target in the *bird* sequence also contains a large number of background pixels, which increases the difficulty of discriminating the target from the background. As seen from Figures 9(h), 9(j) and 9(m), our tracker accurately tracks the target. Even if the SLSA tracker achieves acceptable performance on *sylv* and *bird* sequences, it fully losses the target on the *basketball* sequence. The LSA, IVT and MIL trackers achieve worse performance on these sequences.

Background clutter: The *lemming*, *box* and *board* sequences involve complex background. The *CAVIAR* and *PETS* sequences have simple background but they contain disruptors which are similar to the tracked target. As we can see from Figures 9(i), 9(k) and 9(o), although our tracker has small drift away from tracked target, its performance is still superior to other trackers. As shown in Figures 9(l) and 9(n), when the disruptors appear, our tracker is capable of accurately tracking the target. However, other trackers drift to either the disruptors or the background.

V. CONCLUSIONS AND FUTURE WORK

In this paper, we propose a novel framework of tracking by matching. Different from existing methods that solve tracking by matching pixels across frames, we propose to match local patch across frames by pooling Gabor responses. Since Gabor kernels are capable of capturing local structure features,



(a) *face* sequence (b) *faceocc2* sequence (c) *woman* sequence

(d) *car4* sequence (e) *david_indoor* sequence (f) *singer* sequence

(g) *david_outdoor* sequence (h) *sylv* sequence (i) *lemming* sequence

(j) *basketball* sequence (k) *box* sequence (l) *CAVIAR* sequence

(m) *bird* sequence (n) *PETS* sequence (o) *board* sequence

Fig. 9. Sample tracking results of the compared algorithms on fifteen test sequences.

their responses have the ability of locating local features of the tracked target. By using multiple layer pooling of the Gabor responses, the target can be accurately tracked. Our experimental results on a set of challenging sequences validate that the proposed method outperforms several state-of-the-art methods.

In the future, we will consider to use learning based methods such as sparse coding to replace the Gabor filtering, which may achieves better tracking performance. In addition, the tracking speed of the currently proposed method still needs to be improved for real-time applications.

REFERENCES

[1] H. Zhou, A. Wallace, and P. Green, "Efficient tracking and ego-motion recovery using gait analysis," *Signal Processing*, vol. 89, no. 12, pp. 2367–2384, 2009. 1

[2] S. Gao, Z. Han, C. Li, Q. Ye, and J. Jiao, "Real-time multipedestrian tracking in traffic scenes via an RGB-D-based layered graph model," *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 5, pp. 2814–2825, 2015. 1

[3] K.-H. Lee, J.-N. Hwang, and S.-I. Chen, "Model-based vehicle localization based on 3-D constrained multiple-kernel tracking," *IEEE Trans. Circuits Syst. Video Techn.*, vol. 25, no. 1, pp. 38–50, 2015. 1

[4] X. Zhou, L. Xie, Q. Huang, S. J. Cox, and Y. Zhang, "Tennis ball tracking using a two-layered data association approach," *IEEE Transactions on Multimedia*, vol. 17, no. 2, pp. 145–156, 2015. 1

[5] H. Zhou, M. Fei, A. Sadka, Y. Zhang, and X. Li, "Adaptive fusion of particle filtering and spatio-temporal motion energy for human tracking," *Pattern Recognition*, vol. 47, no. 11, pp. 3552–3567, 2014. 1

[6] R. Collins, Y. Liu, and M. Leordeanu, "On-line selection of discriminative tracking features," *TPAMI*, vol. 27, no. 10, pp. 1631–1643, 2005. 1, 2

[7] Z. Han, J. Jiao, B. Zhang, Q. Ye, and J. Liu, "Visual object tracking via sample-based adaptive sparse representation (AdaSR)," *Pattern Recognition*, vol. 44, no. 9, pp. 2170–2183, 2011. 1

[8] S. Zhang, H. Yao, X. Sun, and S. Liu, "Robust visual tracking using an effective appearance model based on sparse coding," *ACM Transactions*

on Intelligent Systems and Technology, vol. 3, no. 3, pp. 1–18, 2012. 1

[9] L. Zhang, W. Wu, T. Chen, N. Strobel, and D. Comaniciu, "Robust object tracking using semi-supervised appearance dictionary learning," *Pattern Recognition Letters*, vol. 62, pp. 17–23, 2015. 1

[10] Y. Yuan, H. Yang, Y. Fang, and W. Lin, "Visual object tracking by structure complexity coefficients," *IEEE Transactions on Multimedia*, vol. 17, no. 8, pp. 1125–1136, 2015. 1

[11] F. Poiesi and A. Cavallaro, "Tracking multiple high-density homogeneous targets," *IEEE Trans. Circuits Syst. Video Techn.*, vol. 25, no. 4, pp. 623 – 637, 2015. 1

[12] Y. Sui, S. Zhang, and L. Zhang, "Robust visual tracking via sparsity-induced subspace learning," *IEEE Transactions on Image Processing*, vol. 24, no. 12, pp. 4686–4700, 2015. 1

[13] L. Zhao, X. Gao, D. Tao, and X. Li, "Tracking human pose using max-margin markov models," *IEEE Transaction on Image Processing*, vol. 24, no. 12, pp. 5274–5287, 2015. 1

[14] X. Sun, H. Yao, S. Zhang, and D. Li, "Non-rigid object contour tracking via a novel supervised level set model," *IEEE Transaction on Image Processing*, vol. 24, no. 11, pp. 3386–3399, 2015. 1

[15] L. Zhao, X. Gao, D. Tao, and X. Li, "Learning a tracking and estimation integrated graphical model for human pose tracking," *IEEE Trans. Neural Netw. Learning Syst.*, vol. 26, no. 12, pp. 3176–3186, 2015. 1

[16] Z. Hong, Z. Chen, C. Wang, X. Mei, D. V. Prokhorov, and D. Tao, "MUlti-Store Tracker (MUSTer): A cognitive psychology inspired approach to object tracking," *Proceedings of International Conference on Computer Vision and Pattern Recognition*, pp. 749–758, 2015. 1

[17] 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. 1

[18] Y. Cong, B. Fan, J. Liu, J. Luo, and H. Yu, "Speeded up low-rank online metric learning for object tracking," *IEEE Trans. Circuits Syst. Video Techn.*, vol. 25, no. 6, pp. 922 – 934, 2015. 1

[19] T. A. Biresaw, A. Cavallaro, and C. S. Regazzoni, "Tracker-level fusion for robust bayesian visual tracking," *IEEE Trans. Circuits Syst. Video Techn.*, vol. 25, no. 5, pp. 776 – 789, 2015. 1

[20] M.-C. Chuang, J.-N. Hwang, K. Williams, and R. Towler, "Tracking live fish from low-contrast and low-frame-rate stereo videos," *IEEE Trans. Circuits Syst. Video Techn.*, vol. 25, no. 1, pp. 167 – 179, 2015. 1

[21] X. Mei, Z. Hong, D. V. Prokhorov, and D. Tao, "Robust multitask multiview tracking in videos," *IEEE Trans. Neural Netw. Learning Syst.*, vol. 26, no. 11, pp. 2874–2890, 2015. 1

[22] C. Gong, K. Fu, A. Loza, Q. Wu, J. Liu, and J. Yang, "Discriminative object tracking via sparse representation and online dictionary learning," *IEEE Transactions on Cybernetics*, vol. 44, no. 4, pp. 539–553, 2014. 1, 2

[23] S. Zhang, X. Yu, Y. Sui, S. Zhao, and L. Zhang, "Object tracking with multi-view support vector machines," *IEEE Transactions on Multimedia*, vol. 17, no. 3, pp. 265 – 278, 2015. 1

[24] W. Liu, A. B. Chan, R. W. H. Lau, and D. Manocha, "Leveraging long-term predictions and online learning in agent-based multiple person tracking," *IEEE Trans. Circuits Syst. Video Techn.*, vol. 25, no. 3, pp. 399 – 410, 2015. 1

[25] Q. Wang, F. Chen, W. Xu, and M.-H. Yang, "Object tracking with joint optimization of representation and classification," *IEEE Trans. Circuits Syst. Video Techn.*, vol. 25, no. 4, pp. 638 – 650, 2015. 1

[26] B. Ma, J. Shen, Y. Liu, H. Hu, L. Shao, and X. Li, "Visual tracking using strong classifier and structural local sparse descriptors," *IEEE Transactions on Multimedia*, vol. 17, no. 10, pp. 1818–1828, 2015. 1

[27] S. Zhang, H. Yao, H. Zhou, X. Sun, and S. Liu, "Robust visual tracking based on online learning sparse representation," *Neurocomputing*, vol. 100, no. 1, pp. 31–40, 2013. 1, 2

[28] X. Lan, A. J. Ma, and P. C. Yuen, "Multi-cue visual tracking using robust feature-level fusion based on joint sparse representation," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1194–1201. 1

[29] S. Zhang, H. Zhou, F. Jiang, and X. Li, "Robust visual tracking using structurally random projection and weighted least squares," *IEEE Trans. Circuits Syst. Video Techn.*, vol. 25, no. 11, pp. 1749 – 1760, 2015. 1

[30] X. Lan, A. J. Ma, P. C. Yuen, and R. Chellappa, "Joint sparse representation and robust feature-level fusion for multi-cue visual tracking," *IEEE Transactions on Image Processing*, vol. 24, no. 12, pp. 5826–5841, 2015. 1

[31] S. Zhang, H. Yao, and S. Liu, "Robust visual tracking using feature-based visual attention," *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 1150–1153, 2010. 1, 2

[32] B. Liu, J. Huang, L. Yang, and C. Kulikowski, "Robust tracking using local sparse appearance model and k-selection," *Proc. Int. Conference on Computer Vision and Pattern Recognition*, 2011. 2, 4

[33] Q. Wang, F. Chen, W. Xu, and M. Yang, "Online discriminative object tracking with local sparse representation," *Proceedings of IEEE Workshop on the Applications of Computer Vision*, pp. 425–432, 2012. 2, 4

[34] W. Zhong, H. Lu, and M. Yang, "Robust object tracking via sparsity-based collaborative model," *Proc. Int. Conference on Computer Vision and Pattern Recognition*, pp. 1838–1845, 2012. 2, 4

[35] X. Jia, H. Lu, and M. Yang, "Visual tracking via adaptive structural local sparse appearance model," *Proc. Int. Conference on Computer Vision and Pattern Recognition*, pp. 1822–1829, 2012. 2, 4

[36] R. Douglas, C. Koch, M. Mahowald, K. Martin, and H. Suarez, "Recurrent excitation in neocortical circuits," *Science*, vol. 269, no. 5226, pp. 981–985, 1995. 2

[37] M. Carandini and D. Heeger, "Summation and division by neurons in visual cortex," *Science*, vol. 264, pp. 1333–1336, 1994. 2

[38] N. Pinto, D. Cox, and J. DiCarlo, "Why is real-world visual object recognition hard," *PLoS Computational Biology*, vol. 4, no. 1, pp. 25–45, 2008. 2

[39] K. Jarrett, K. Kavukcuoglu, M. A. Ranzato, and Y. LeCun, "What is the best multi-stage architecture for object recognition?" *Proceedings of International Conference on Computer Vision*, pp. 2146–2153, 2009. 2

[40] A. Jepson, D. Fleet, and T. EI-Maraghi, "Robust online appearance models for visual tracking," *TPAMI*, vol. 25, no. 10, pp. 1296–1311, 2003. 2

[41] I. Matthews, T. Ishikawa, and S. Baker, "The template update problem," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 6, pp. 810–815, 2004. 2

[42] D. Ross, J. Lim, R. Lin, and M. Yang, "Incremental learning for robust visual tracking," *Int. Journal of Computer Vision*, vol. 77, pp. 125–141, 2008. 2, 4

[43] N. Wang, J. Wang, and D.-Y. Yeung, "Online robust non-negative dictionary learning for visual tracking," *Proc. Int. Conference on Computer Vision*, 2013. 2

[44] P. Pérez, C. Hue, J. Vermaak, and M. Gangnet, "Color-based probabilistic tracking," *Proceedings of European Conference on Computer Vision*, pp. 661–675, 2002. 2, 4

[45] A. Adam, E. Rivlin, and I. Shimshoni, "Robust fragments-based tracking using the integral histogram," *Proc. Int. Conference on Computer Vision and Pattern Recognition*, pp. 798–805, 2006. 2, 4

[46] A. Yilmaz, O. Javed, and M. Shah, "Object tracking: A survey," *ACM Computing Surveys*, vol. 38, no. 4, p. 13, 2006. 2

[47] H. Yang, L. Shao, F. Zheng, L. Wang, and Z. Song, "Recent advances and trends in visual tracking: A review," *Neurocomputing*, vol. 74, no. 18, pp. 3823–3831, 2011. 2

[48] S. Salti, A. Cavallaro, and L. D. Stefano, "Adaptive appearance modeling for video tracking: Survey and evaluation," *IEEE Transaction on Image Processing*, vol. 21, no. 10, pp. 4334–4348, 2012. 2

[49] X. Li, W. Hu, C. Shen, Z. Zhang, A. Dick, and A. van den Hengel, "A survey of appearance models in visual object tracking," *ACM Transactions on Intelligent Systems and Technology*, vol. 4, no. 4, p. 58, 2013. 2

[50] C. Liu and H. Wechsler, "Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition," *IEEE Transactions on Image Processing*, vol. 11, no. 4, pp. 467–476, 2002. 3

[51] Y. N. Wu, Z. Si, H. Gong, and S. C. Zhu, "Learning active basis model for object detection and recognition," *International Journal of Computer Vision*, vol. 90, no. 2, pp. 198–235, 2010. 3

[52] A. Doucet, N. de Freitas, and N. Gordon, "Sequential monte carlo methods in practice," *Springer-Verlag*, 2001. 4

[53] M. Isard and A. Blake, "Condensation—conditional density propagation for visual tracking," *International Journal of Computer Vision*, vol. 29, no. 1, pp. 5–28, 1998. 4

[54] B. Babenko, M. Yang, and S. Belongie, "Visual tracking with online multiple instance learning," *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 983–990, 2009. 4

[55] H. Grabner and H. Bischof, "On-line boosting and vision," *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, vol. 1, pp. 260–267, 2006. 4

[56] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (VOC) challenge," *International Journal of Computer Vision*, vol. 88, no. 2, pp. 303–338, 2010. 5

Shengping Zhang (M'13) received the Ph.D. degree in computer science from the Harbin Institute of Technology, Harbin, China, in 2013. He is currently an Associate Professor with the School of Computer Science and Technology, Harbin Institute of Technology at Weihai. He had been a Post-Doctoral Research Associate with Brown University, Providence, RI, USA, and a Visiting Student Researcher with University of California at Berkeley, Berkeley, CA, USA. He has authored or co-authored over 30 research publications in refereed journals and conferences. His research interests include sparse coding and its applications in computer vision.

Dr. Zhang is also an Associate Editor of Signal Image and Video Processing.

Xiangyuan Lan (S'14) received the B.Eng. degree in computer science and technology from the South China University of Technology, China, in 2012. He is currently pursuing the Ph.D. degree with the Department of Computer Science, Hong Kong Baptist University, Hong Kong. In 2015, he was a Visiting Scholar with the Computer Vision Laboratory, University of Maryland Institute for Advanced Computer Studies, University of Maryland, College Park, USA.

Mr. Lan' current research interests include computer vision and pattern recognition, particularly, feature fusion and sparse representation for visual tracking.

Yuankai Qi received the B.S. and M.S. degrees from Harbin Institute of Technology, China, in 2011 and 2013, respectively, and is currently working toward the Ph.D. degree in computer science and technology at Harbin Institute of Technology, China. His research interests include object tracking, sparse coding, and machine learning.

Pong C. Yuen (M'93–SM'11) received his B.Sc. degree in Electronic Engineering with First Class Honours in 1989 from City Polytechnic of Hong Kong, and his Ph.D. degree in Electrical and Electronic Engineering in 1993 from The University of Hong Kong. He joined the Hong Kong Baptist University in 1993 and, currently is a Professor and Head of the Department of Computer Science.

Dr. Yuen was a recipient of the University Fellowship to visit The University of Sydney in 1996. In 1998, Dr. Yuen spent a 6-month sabbatical leave in The University of Maryland Institute for Advanced Computer Studies (UMIACS), University of Maryland at college park. From June 2005 to January 2006, he was a visiting professor in GRAVIR laboratory (GRAphics, VIsion and Robotics) of INRIA Rhone Alpes, France. Dr. Yuen was the director of Croucher Advanced Study Institute (ASI) on biometric authentication in 2004 and the director of Croucher ASI on Biometric Security and Privacy in 2007.

Dr. Yuen has been actively involved in many international conferences as an organizing committee and/or technical program committee member. He was the track co-chair of International Conference on Pattern Recognition (ICPR) 2006 and the program co-chair of IEEE Fifth International Conference on Biometrics: Theory, Applications and Systems (BTAS) 2012. Currently, Dr. Yuen is an Editorial Board Member of Pattern Recognition and Associate Editor of IEEE Transactions on Information Forensics and Security, and SPIE Journal of Electronic Imaging. He is also serving as a Hong Kong Research Grant Council Engineering Panel Member.

Dr. Yuen's current research interests include video surveillance, human face recognition, biometric security and privacy.