Learning Local Appearances With Sparse Representation for Robust and Fast Visual Tracking

Tianxiang Bai, You-Fu Li, *Senior Member, IEEE*, and Xiaolong Zhou

***Abstract*—In this paper, we present a novel appearance model using sparse representation and online dictionary learning techniques for visual tracking. In our approach, the visual appearance is represented by sparse representation, and the online dictionary learning strategy is used to adapt the appearance variations during tracking. We unify the sparse representation and online dictionary learning by defining a sparsity consistency constraint that facilitates the generative and discriminative capabilities of the appearance model. An elastic-net constraint is enforced during the dictionary learning stage to capture the characteristics of the local appearances that are insensitive to partial occlusions. Hence, the target appearance is effectively recovered from the corruptions using the sparse coefficients with respect to the learned sparse bases containing local appearances. In the proposed method, the dictionary is undercomplete and can thus be efficiently implemented for tracking. Moreover, we employ a median absolute deviation based robust similarity metric to eliminate the outliers and evaluate the likelihood between the observations and the model. Finally, we integrate the proposed appearance model with the particle filter framework to form a robust visual tracking algorithm. Experiments on benchmark video sequences show that the proposed appearance model outperforms the other state-of-the-art approaches in tracking performance.**

***Index Terms*—Appearance model, dictionary learning, sparse representation, visual tracking.**

I. Introduction

In this paper, we address the visual tracking problem that establishes the correspondences of a general object between successive frames given its initial location in the first frame and no other information. In this case, the appearance of the target is the only available clue for tracking. This problem becomes challenging, especially considering the appearance of the target is nonstationary in the natural scenes, e.g., undergoing significant viewpoint, pose and illumination varying as well as partial occlusions.

For years, numerous contributions addressed the appearance representation and modeling problem for the aforementioned challenges from many and diverse points of view, such as manifold learning [1], subspace representation methods [2] discriminative method for classification [3], and kernel based filter [4], etc. We have no intention to provide a survey of this vast activity. Instead, we focus on one specific approach that is highly effective and promising for dynamic visual appearance representation in our prior work [6], [7]: the use of sparse representation-based appearance model.

Our previous work [6], [7] and related research [5], [10], [11] focus on the representation of the target appearance, which can be formulated as two perspectives. One is to represent the holistic target appearance as well as the occlusion via finding a sparse linear combination over a dictionary containing target and trivial templates [5]–[7], [37]. These methods, more or less, are extensions of recent study in face recognition via sparse representation [8] that attempt to recover the appearance from the corruptions with "cross-and-bouquet" error correction model [9]. The main drawback of these approaches is that their computational load is extensive, since the dictionaries they used are normally over complete. Our previous work [6], [7] can be classified into such holistic approach and attempted to speed up the tracking efficiency with structural sparsity imposed. Another approach is to represent the target appearance with the local appearance information that motivates this paper. These methods represent the target appearance, for example, by finding a sparse linear combination of sub-image samples in the tracked object region [10], by exploring the locality-constrained linear coding (LLC) framework with small image patches inside the target region [11], or by investigating the local image patches with a histogram-based method [36]. Unlike the holistic appearance representation methods that handle the occlusion as a sparse noise component with overcomplete dictionaries, these local appearance-based approaches exhibit comparative robustness against occlusions and tend to be more efficient, since the dictionaries they used are unnecessarily over complete.

Distinct from our previous work [6], [7], in this paper, we desire to address another fundamental problem: how to design or train dictionaries for the sparse representation-based model that can better model the dynamic visual appearance and facilitate the tracking performance. Most commonly the dictionaries are updated by a heuristic scheme that replaces the least important template [5], or randomly selected templates [10] with the current tracking result. However, the expressiveness of these template-based dictionaries is limited as the target appearance can only be represented by the subspace spanned by the raw templates directly cropped from the images, which makes it difficult to handle significant view

Manuscript received June 22, 2013; revised January 9, 2014, April 28, 2014, and May 30, 2014; accepted June 10, 2014. Date of publication July 10, 2014; date of current version March 13, 2015. This work was supported in part by the Research Grants Council of Hong Kong under Project CityU 118613 and in part by the National Natural Science Foundation of China under Grant 61273286. This paper was recommended by Associate Editor H. Qiao.

The authors are with the Department of Mechanical and Biomedical Engineering, City University of Hong Kong, Hong Kong (e-mail: tianxiangbai@gmail.com; meyfli@cityu.edu.hk; mexlz@hotmail.com).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCYB.2014.2332279

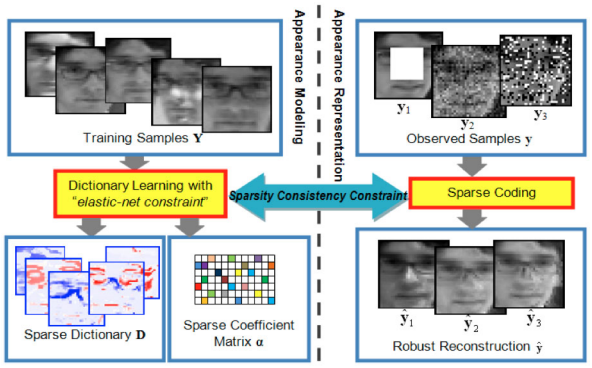


Fig. 1. Proposed appearance model. In the appearance modeling phase, a sparse dictionary is trained with observed samples online using the elastic-net constraint. During the appearance representation phase, new (corrupted) observations are sparsely approximated with the trained sparse dictionary under the sparsity consistency constraint.

or pose changes. Recently, the machine learning paradigm for dictionaries design is rapidly gaining interest for visual tracking. One direct benefit of exploiting the machine learning techniques for designing dictionaries is that a finer adaptation to the nonstationary appearance of the target becomes possible [13]. Integrating the incremental subspace learning scheme for dictionary design [6], [7], [34] were proposed for robust visual tracking, and obtained encouraging results on several bench mark data sets. However, they are still not efficient for online tracking tasks because of the use of overcomplete dictionary. In [11], a specific dictionary learning scheme, referred to as the k-selection, was developed for local appearance modeling. Moreover, a nonnegative dictionary learning-based algorithm [33], was proposed to solve the tracking problem. It updates the templates in the dictionary by capturing the distinctive aspects of the tracked object. More recently, a multitask sparse representation scheme is used to learn a dictionary with multiview features for tracking [29]. These dictionary learning-based methods show significant enhanced tracking accuracy and efficiency.

In this paper, we focus on the design of the appearance model. The philosophy behind the proposed appearance model is to represent and model the appearance using sparse representation and online dictionary learning. Fig. 1 shows an overview of the proposed appearance model. Our first contribution is the proposal of a novel appearance representation method based on sparse representation with the "sparsity consistency constraint" that can boost both generative and discriminative power of the model. The second contribution is the adoption of "elastic-net constraint" [14] with the online dictionary learning scheme that allows us to train a sparse undercomplete dictionary with local appearance information, which are more robust against the occlusions. The integration of the proposed robust similarity metric (RSM)-based observation model and particle filter framework for visual tracking is the third contribution. We present empirical results on publicly available benchmark video sequences, and show that the proposed appearance model can lead to more robust and efficient tracking than existing state-of-the-art algorithms in the literature.

II. Sparse Representation Over Learned Local Appearance Model

A. Sparse Representation-Based Holistic Appearance Model

In the previous work, sparse representation is usually used to obtain a sparse representation of the target appearance over an overcomplete holistic dictionary [5]. Mathematically, given a m-dimensional observed target appearance $\mathbf{y} \in \mathbb{R}^m$ (ordered lexicographically as a column vector) and a dictionary $\mathbf{D} \in \mathbb{R}^{m \times n}$ including n columns. The sparse representation problem can be formulated as the following ℓ_0-minimization problem:

$$\min_{\mathbf{a}} \|\mathbf{a}\|_0 \quad \text{subject to} \quad \mathbf{y} = \mathbf{D}\mathbf{a} \tag{1}$$

where $\|\cdot\|_0$ is the ℓ_0 norm, which counts the number of nonzero entries in the coefficient vector $\mathbf{a}$. In most of the existing work, the dictionary $\mathbf{D} = [\mathbf{T}\ \mathbf{E}]$ is constructed by holistic target template set $\mathbf{T} \in \mathbb{R}^{m \times d}$ and trivial template set $\mathbf{E} \in \mathbb{R}^{m \times m}$ [5]–[7], where d is the number of target templates. The target template set $\mathbf{T}$ is usually obtained or learned from the holistic tracking results from previous time intervals. The trivial template set $\mathbf{E}$ is an identity matrix that represents the occlusions.

Given that solving the minimum ℓ_0 norm is NP-hard, a common approximation method is to convert it into the following unconstrained, ℓ_1-regularized least square problem that imposes sparse solutions for $\mathbf{a}$ [23]:

$$\min_{\mathbf{a}} \|\mathbf{y} - \mathbf{D}\mathbf{a}\|_2^2 + \lambda \|\mathbf{a}\|_1 \tag{2}$$

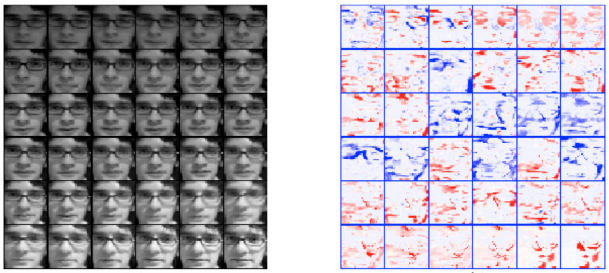
where λ is a regularization parameter that balances the trade-off between the reconstruction error and the sparsity, and $||\cdot||_1$ is the ℓ_1 norm that sums up the absolute value of the entries. This problem can be effectively and efficiently solved using convex optimization [15]. Although no direct analytic link exists between the value of the regularization parameter λ and the corresponding sparse level $||\mathbf{a}||_0$, the value of λ plays an important role in the proposed approach. The role of λ will be further explained in Section II-C.

B. From Holistic Sparsity to Local Sparsity

Distinct from the existing methods, we discard the holistic sparse representation model with trivial template set and online learn a local sparse dictionary instead. The sparse dictionary can capture the characteristics of the local target appearance and sparsely represent its appearance. In addition, we apply the online dictionary learning algorithm to train the dictionary rather than use the raw target templates [5] or Eigen templates [6], [7] in previous work. Given a collection of training images $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_t]$, the dictionary learning problem can be formulated as follows:

$$\min_{\mathbf{D},\mathbf{a}} \frac{1}{t} \sum_{i=1}^{t} \left(\|\mathbf{y}_i - \mathbf{D}\mathbf{a}_i\|_2^2 + \lambda \|\mathbf{a}_i\|_1 \right). \tag{3}$$

This combinatorial and nonconvex optimization problem can be solved using an iterative approach that consists of two (convex) steps, namely, the sparse representation step on a fixed $\mathbf{D}$ and the dictionary update step on a fixed $\mathbf{a}$. Several



(a) (b)

Fig. 2. Illustration of the dictionary learning results. (a) Training samples. (b) Learned sparse dictionary including the local face features (eyebrows, eyes, noses, mouths, facial contours, and glasses). Negative values are blue, positive values are red, and the zero values are represented in white. The parameter γ is set to 0.55.

methods have been proposed to solve the optimization problem, such as the MOD [17], K-SVD algorithm [18], and online dictionary learning [19]. Among these methods, the online dictionary learning algorithm [19] is significantly faster than other dictionary learning approaches. In addition, it is capable of handling dynamic data that can update the dictionary with new observed data and without storing the previous data.

Local feature extraction methods, such as sparse principal component analysis (SPCA), can better model and represent the appearance because they improve robustness against occlusions [22]. From the synthesis perspective, SPCA aims to construct a sparse basis, such that all the data have low reconstruction errors when decomposed [23]. This formulation is similar to the aforementioned online dictionary learning algorithm. An investigation shows that they are equivalent optimization problems by adding an elastic-net constraint to each columns of the dictionary $\mathbf{D}$ [19]. Thus, the previous dictionary learning problem becomes

$$\begin{aligned} &\min_{\mathbf{D},\mathbf{a}} \tfrac{1}{t}\textstyle\sum_{i=1}^{t}\left(\|\mathbf{y}_i-\mathbf{D}_i\|_2^2+\lambda\|\mathbf{a}_i\|_1\right) \\ &\textbf{subject to}\ \ \forall j=1,\ldots,n,\ \ \left\|\mathbf{d}_j\right\|_2^2+\gamma\left\|\mathbf{d}_j\right\|_1\le 1 \end{aligned} \tag{4}$$

where $\left\|\mathbf{d}_j\right\|_2^2+\gamma\left\|\mathbf{d}_j\right\|_1\le 1$ is the elastic-net constraint, and the parameter γ controls the levels of sparsity of each column $\mathbf{d}_j$ in dictionary $\mathbf{D}$. The elastic-net constraint is used because of its capability to preserve highly correlated entries in the dictionary that can better capture the local appearance of the target with the grouping effect. As shown in Fig. 2, the proposed method can learn the most discriminative, localized facial features on the training set, such as eyebrows, eyes, noses, mouths, facial contours, and glasses, while leaving the other parts zero. These features have already been verified to be very important for face recognition tasks by many face recognition algorithms [22]. This learned locally concentrated sparse dictionary is robust to occlusions because only a fraction of local features is corrupted.

The dictionary $\mathbf{D}\in\mathbb{R}^{m\times n}$ is typically overcomplete ($m<n$) for image processing and visual tracking applications because this redundant representation is generally suitable for representing a wider range of image variations such as image with occlusions. However, the computational cost for training an overcomplete dictionary is high, thus limiting its implementation for visual tracking. It has been shown that an object under pose variations, shape deformations, and illumination changes, individually or combined, sits on a low-dimensional subspace [2]. In visual tracking, using an overcomplete dictionary is unnecessary because the appearance of the target object has a much narrower dynamical range of expressiveness than general images. Our previous work shows that the undercomplete dictionary is effective to represent the target as well as the background visual appearances [12]. In addition, the trivial template set is discarded since the elastic-net constrained dictionary can handle the occlusion effectively. Thus, considering both efficiency and effectiveness, the undercomplete dictionary ($m>n$) is used for the visual tracking task. Moreover, we prefer that the training images and observed samples should be centered by subtracting a time varying mean in the sparse representation and dictionary learning. This is because considering the mean is able to enhance the generative capabilities of the model without sacrifice of the discriminative power. A detail experimental verification will be further presented in Section IV-B5. Given a data matrix $\mathbf{A}=[\mathbf{I}_1,\mathbf{I}_2,...,\mathbf{I}_a]\in\mathbb{R}^{m\times a}$ and a new observed data matrix $\mathbf{B}=[\mathbf{I}_{a+1},\mathbf{I}_{a+2},...,\mathbf{I}_{a+b}]\in\mathbb{R}^{m\times b}$. Denoting $\bar{\mathbf{I}}_A$ and $\bar{\mathbf{I}}_B$ are the sample mean of data matrix $\mathbf{A}$ and $\mathbf{B}$, respectively. As suggested in [2], the mean of the concatenation matrix $\mathbf{C}=[\mathbf{A}\ \mathbf{B}]\in\mathbb{R}^{m\times(a+b)}$ can be updated by

$$\bar{\mathbf{I}}_C=\frac{a}{a+b}\bar{\mathbf{I}}_A+\frac{b}{a+b}\bar{\mathbf{I}}_B. \tag{5}$$

C. Sparsity Consistency Constraint

Given a well-trained dictionary $\mathbf{D}$ that can represent a class of images (such as a variety of appearances regarding the object of interest) that have a certain sparsity level, the dictionary is assumed to be capable of representing the query image within the same class with an identical level of sparseness. The previous sparse representation and online dictionary learning scheme share the same ℓ_1-regularization parameter λ that bridges the appearance modeling and representation stages. In the proposed method, we use the LARS-Lasso algorithm [16] to solve both the sparse representation (2) and dictionary learning problems (4). Using an identical regularization parameter λ in sparse representation phase and dictionary learning phase can satisfy the sparsity consistency constraint. The proposed algorithm assumes the knowledge of sparsity, which is manifested by the regularization parameter λ.

Moreover, the bad observed samples from the background and occlusion are expected to facilitate a nonsparse representation with $\mathbf{D}$. This argument is sensible because similar claims were justified in recent studies on the role of sparse representation in image denoising [18], separation, and inpainting [21]. Imposing the sparsity consistency constraint incurs a higher reconstruction error with bad observations and boosts the discriminative power of the appearance model. This argument is corroborated by the quantitative experimental results presented in Section IV-B5.

Fig. 3 shows synthetic examples for recovering the corrupted sample from the experiments. Adding Gaussian noise,

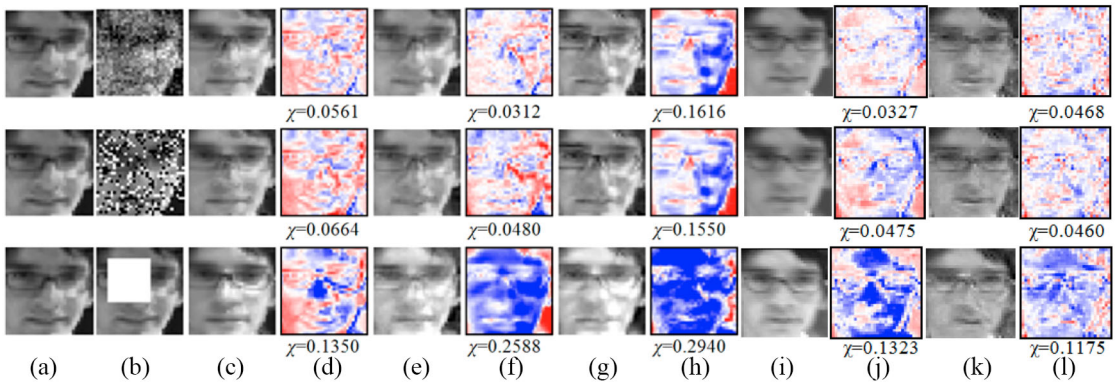


Fig. 3. Synthetic examples for reconstruction. (a) Actual test samples. (b) Corrupted test samples. (c) Reconstruction using the proposed method with sparsity consistency constraint and elastic-net constraint. (d) Reconstruction errors of the proposed method with sparsity consistency constraint and elastic-net constraint. (e) Reconstruction using sparse representation without the "elastic-net constraint." (f) Reconstruction errors using sparse representation without the "elastic-net constraint." (g) Reconstruction using PCA. (h) Reconstruction errors of PCA. (i) Reconstruction using SRTT. (j) Reconstruction errors using SRTT. (k) Reconstruction using SRSE. (l) Reconstruction errors using SRSE. Negative values are blue, positive values, are red, and the zero values are represented in white. χ is the root-mean-square error between the actual test samples and the reconstruction images.

salt and pepper noise, and a white patch to the test sample performs the synthetic experiments. The images restored using sparse representation, with and without the elastic-net constraint, PCA, sparse representation with trivial templates (referred as SRTT) as in [5] and sparse representation with sparse error term (referred as SRSE) in [30] are displayed in Fig. 3(c), (e), (g), (i), and (k). The corresponding reconstruction errors are visualized and quantified using the root-mean-square error χ in Fig. 3(d), (f), (h), (j), and (l). The sparse representation-based methods (proposed method with and without the elastic-net constraint, SRTT, SRSE) demonstrate closer reconstruction with the actual image than the PCA-based least-squares restoration when random noise is added (the first and second rows of Fig. 3). This finding validates that sparse representation-based reconstruction is robust against noise and is consistent with recent investigations on image denoising using sparse representation [20]. The sparse representation without elastic-net constraint, SRTT, and SRSE are slightly better than the proposed method because the elastic-net constraint enforces the sparse property into the bases, such that less information is retained. However, the sparse representation without elastic-net constraint and the PCA-based reconstruction exhibits degenerative performance under gross corruption in the occluded case (the third row of Fig. 3) because it attempts to approximate the occluded patch with holistic features. The proposed method with the elastic-net constraint recovers the test image effectively because it relies on the local features that are insensitive to the gross errors attributed to occlusion. The SRTT and SRSE reconstruction can also reconstruct the original image well because both of these methods use the trivial or Laplacian error term to estimate the occlusion. However, these methods usually need more computational time to achieve the reconstruction and lack the consideration of discriminative capabilities.

D. Robust Similarity Metric

For visual tracking tasks, evaluating the similarities between the observed sample and the appearance model is necessary.

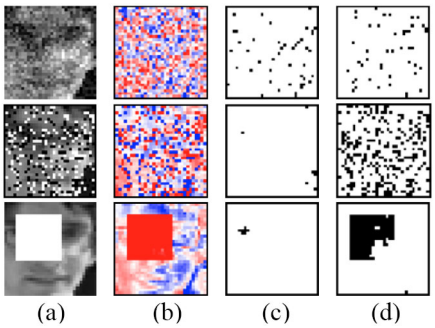


Fig. 4. Outlier identification results of using standard deviation and MAD. (a) Corrupted samples. (b) Reconstruction residuals (negative values are blue, positive values are red, and the zero values are represented in white). (c) Inferred outliers (indicated in black) with standard deviation. (d) Inferred outliers (indicated in black) with MAD.

The proposed appearance model can recover the image from corruptions and preserves the pixels in uncorrupted areas. The residuals of the uncorrupted pixels are small, whereas the corrupted pixels or outliers generate large positive or negative residuals [Fig. 4(b)]. To identify and eliminate the corrupted pixels, the residuals must be compared with the estimated standard deviation $\hat{\sigma}$ of the error scale [25]

$$\hat{\sigma} = 1.4826\ \mathrm{MAD}(\mathbf{r}) \tag{6}$$

where the $\mathrm{MAD}(\mathbf{r}) = \mathrm{median}_i(|r_i - \mathrm{median}_j(r_j)|)$ function returns the median absolute deviation (MAD) of the residual errors $\mathbf{r} = \|\mathbf{y}_t - \hat{\mathbf{y}}_t\|$. Estimating the standard deviation using MAD is more robust than its original calculation $\sigma = \sqrt{\mathrm{E}[(\mathbf{r} - \mathrm{E}(\mathbf{r}))^2]}$ that the distances from the mean are squared. Thus, the large deviations attributed to the outliers are weighted more heavily and can thus significantly affect the value of the standard deviation. On the contrary, the magnitude of the deviation in MAD is the absolute value, which is irrelevant to a few outliers. With the estimated $\hat{\sigma}$, the standardized residuals $|r_i/\hat{\sigma}|$ are computed and used to define a

weight vector $\mathbf{w}$ as follows:

$$w_i = \begin{cases} 1 & \text{if } \left|r_i/\hat{\sigma}\right| \leq \tau \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

where r_i is the reconstruction error that corresponds to the ith pixel. The residuals can be inferred as outliers with high probabilities in a Gaussian situation if they are larger than $3\hat{\sigma}$. Therefore, the threshold τ is typically set as 3 in the experiments. Fig. 4(c) and (d) shows the outlier identification results when using standard deviation and MAD. In the first case wherein Gaussian noise was added, the two methods yield similar results because the noise is governed by a Gaussian distribution. However, in cases in which gross corruptions (salt and pepper noise and occlusion) appear, the standard deviation is affected by large residuals and thus fails to reject the outliers. On the contrary, the proposed method can better identify the outliers because $\hat{\sigma}$ is estimated using MAD, which is robust against large residuals. Using the previously defined weight vector $\mathbf{w}$, the RSM is determined using the following root-mean-square error that ignores the outliers:

$$\text{RSM} = \sqrt{\frac{\sum w_i r_i^2}{\sum w_i}}. \tag{8}$$

III. Tracking Algorithm With Particle Filter

The proposed appearance model is embedded into a Bayesian inference framework to form a robust tracking algorithm. The model recursively updates the posterior distribution $p(\mathbf{x}_t|\mathbf{y}_{1:t})$ over the target state $\mathbf{x}_t$ given all observations $\mathbf{y}_{1:t} = \{\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_t\}$ until time t. By applying the Bayes' theorem, the Bayes filter can be written as follows:

$$p(\mathbf{x}_t|\mathbf{y}_{1:t}) \propto p(\mathbf{y}_t|\mathbf{x}_t) \int_{\mathbf{x}_{t-1}} p(\mathbf{x}_t|\mathbf{x}_{t-1}) p(\mathbf{x}_{t-1}|\mathbf{y}_{1:t-1}) d\mathbf{x}_{t-1} \tag{9}$$

where $p(\mathbf{y}_t|\mathbf{x}_t)$ is the observation model, and $p(\mathbf{x}_t|\mathbf{x}_{t-1})$ is the motion model. In the particle filter framework [26], the posterior distribution $p(\mathbf{x}_t|\mathbf{y}_{1:t})$ is recursively approximated by a set of weighted samples. The observation model indicates the likelihood between an observed target candidate and the appearance model. Using the defined RSM in (8), the observation model can be formulated as

$$p(\mathbf{y}_t|\mathbf{x}_t) = \exp^{-\zeta(\text{RSM})} \tag{10}$$

where ζ denotes the weighting parameter, which is set to m (dimension of the target appearance) in all experiments. The motion model predicts the current state given the previous state. In this paper, an affine image warping is used to model the target motion between two consecutive frames. The state vector $\mathbf{x}_t = (x_t, y_t, \eta_t, s_t, \beta_t, \phi_t)$ at time t is formulated using six parameters of affine transformation, where x_t, y_t denote the x and y translation, and $\eta_t, s_t, \beta_t, \phi_t$ represent the rotation angle, scale, aspect ratio, and skew direction at time t, respectively. Each parameter in $\mathbf{x}_t$ is governed by a Gaussian distribution around the previous state $\mathbf{x}_{t-1}$, and each parameter is assumed to be mutually independent as follows:

$$p(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathbb{N}(\mathbf{x}_t; \mathbf{x}_{t-1}, \Psi) \tag{11}$$

Algorithm 1 Proposed Tracking Algorithm

Input: The initial state of the target $\mathbf{x}_0 = (x_0, y_0, \eta_0, s_0, \beta_0, \phi_0)$.

1: **Initialization:** Construct the initial dictionary $\mathbf{D}_0 \in \mathbb{R}^{m \times n}$ with n labeled target samples.
2: For $t = n$ to N, where N is the total number of fames.
3: Generate P candidate samples $\mathbf{y}_i$ at state $\mathbf{x}_t^i$ based on the affine motion model (11).
4: For each $\mathbf{y}_i$, $i = 1{:}P$
5: Perform sparse representation (2) to approximate each sample.
6: Eliminate the outliers using (6) and (7), and calculate likelihood using (10) based on the RSM (8).
7: End for.
8: Obtain the current state $\hat{\mathbf{x}}_t$ using MAP, and store the tracking result $\mathbf{y}_t$.
9: Update the $\mathbf{D}_t$ with the tracking result using the online sparse dictionary learning scheme (5).
10: End for.

Output: The current state $\hat{\mathbf{x}}_t$ at each frame.

where $\Psi = (\psi^x, \psi^y, \psi^\eta, \psi^s, \psi^\beta, \psi^\phi)$ is the covariance matrix. The current state is then estimated by Maximum a Posteriori (MAP), which associates with the highest likelihood under the observation model. The proposed tracking algorithm is summarized in Algorithm 1.

IV. Experiments

A. Implementation

In this section, experiments are presented to demonstrate the efficiency and effectiveness of the proposed tracking algorithm. The proposed tracker is implemented using MATLAB on a 3 GHz machine with 2 GB RAM. For the online dictionary learning scheme, an undercomplete dictionary comprising 36 basis vectors is used. Each observed target sample is resized to a 32×32 patch. The parameters of the proposed method are fixed for all of the experiments, except for the covariance matrix Ψ of the motion model in (10). The variances ψ^x and ψ^y are set between 2 and 5 in anticipation of the x, y translation. $\psi^\eta, \psi^s, \psi^\beta$, and ψ^ϕ are normally set to a range of 0.001 to 0.05 to predict the variations of the rotation angle, scale, aspect ratio, and skew direction. The regularization parameter is set to $\lambda = 0.04$. The parameter γ is set to 0.55, such that each basis vector has approximately 25% nonzero entries. The choice of λ and γ will be discussed in Sections IV-B3 and IV-B4. The dictionary is updated every five frames to balance the computational efficiency and effectiveness of the appearance modeling.

A total of ten publicly available benchmark video sequences[1] are used to evaluate the performance of the proposed tracker. Random noises are manually added into the

[1]These video sequences are available at http://www.cs.toronto.edu/~dross/ivt/, http://cv.snu.ac.kr/research/~vtd/, http://vision.ucsd.edu/~bbabenko/project_miltrack.shtml and http://www.cvg.cs.rdg.ac.uk/PETS2001.

TABLE I
AVERAGED TRACKING LOCATION ERROR (PIXEL)

Video Clip	Frag	MIL Track	IVT	VTD	ℓ_1 Tracker	SSRT	MTT	LSST	DLSRVT	Proposed Tracker*	Proposed Tracker
PETS2001	77	62	49	25	88	32	38	42	68	24	**2**
Dudek	66	141	**7**	14	86	75	43	**7**	87	15	**7**
Occluded Face	8	27	15	10	9	27	25	7	**4**	25	8
Occluded Face2	13	20	9	9	35	11	13	6	6	10	**4**
David	107	23	**4**	27	57	5	56	6	10	5	5
David †	55	27	27	75	77	97	113	107	55	12	**10**
Trellis	81	60	61	54	65	12	29	63	95	12	**6**
Sylvester	17	11	71	**4**	44	20	11	48	14	15	13
Singer	39	15	6	5	**2**	8	**2**	**2**	**2**	11	**2**
Football	5	13	6	5	50	9	7	12	17	**3**	**3**
Football‡	11	13	9	53	27	12	13	48	29	7	**3**
Car	32	40	**2**	26	25	**2**	**2**	**2**	42	**2**	**2**
Overall	42	38	22	24	47	26	29	29	35	12	**5**

*Without the robust similarity measure
†Salt and pepper noise added
‡Gaussian noise added

TABLE II
TRACKING FAILURE RATE

Video Clip	Frag	MIL Track	IVT	VTD	ℓ_1 Tracker	SSRT	MTT	LSST	DLSRVT	Proposed Tracker*	Proposed Tracker
PETS2001	0.95	0.69	0.63	0.48	0.76	0.71	0.70	0.57	0.78	0.33	**0.03**
Dudek	0.69	0.60	0.03	0.08	0.51	0.56	0.62	0.05	0.77	0.16	**0.02**
Occluded Face	**0**	0.27	0.10	0.03	**0**	0.31	0.26	**0**	**0**	0.42	**0**
Occluded Face2	0.03	0.21	0.01	0.05	0.41	0.04	0.11	**0**	0.01	**0**	**0**
David	0.97	0.53	**0**	0.45	0.58	**0**	0.61	**0**	0.18	**0**	**0**
David †	0.79	0.63	0.34	0.78	0.73	0.72	0.66	0.69	0.73	0.09	**0.03**
Trellis	0.74	0.74	0.46	0.68	0.61	0.26	0.33	0.63	0.85	0.15	**0.05**
Sylvester	0.48	0.26	0.50	**0**	0.66	0.26	0.10	0.47	0.25	0.15	0.21
Singer	0.67	0.52	**0**	0.12	**0**	**0**	**0**	**0**	**0**	0.08	**0**
Football	0.13	0.47	0.09	0.03	0.54	0.19	0.22	0.10	0.07	**0.01**	**0.01**
Football‡	0.31	0.52	0.18	0.61	0.61	0.21	0.23	0.38	0.76	0.06	**0.02**
Car	0.89	0.86	**0**	0.45	0.37	**0**	**0**	**0**	0.91	0.04	**0**
Overall	0.50	0.47	0.23	0.23	0.48	0.28	0.30	0.25	0.38	0.15	**0.05**

*Without the robust similarity measure
†Salt and pepper noise added
‡Gaussian noise added

David and *Football* sequences to evaluate tracking performance against strong random disturbance. For comparison, we evaluate the proposed tracker using three of the latest sparse representation-based trackers, namely, the ℓ_1 tracker [5], the SSRT [6], and the MTT [29] as well as four other state-of-the-art trackers, namely, FragTrack [27], IVT [2], MILTrack [3], and VTD [28]. The source or binary codes of the trackers can be obtained from their respective project websites or authors. All the reference trackers are implemented using the parameter settings given in their respective papers or their default initialization. Given that SSRT, IVT, the ℓ_1 tracker, VTD, MTT, and the proposed tracker are Monte Carlo sampling-based methods, they all use 600 samples to track an object for fair comparison. The proposed algorithm runs at approximately 0.16 s per frame without code optimization; whereas the SSRT, ℓ_1 tracker and MTT with a 12×15 resized sampling patch run at approximately 1.6 s, 2 s, and 2∼3 s to process one frame, respectively. The proposed tracker is significantly more efficient than the other two sparse representation-based trackers.

B. Quantitative Evaluation

1) Comparison of Tracking Algorithms: The performances of the proposed tracker and reference trackers are

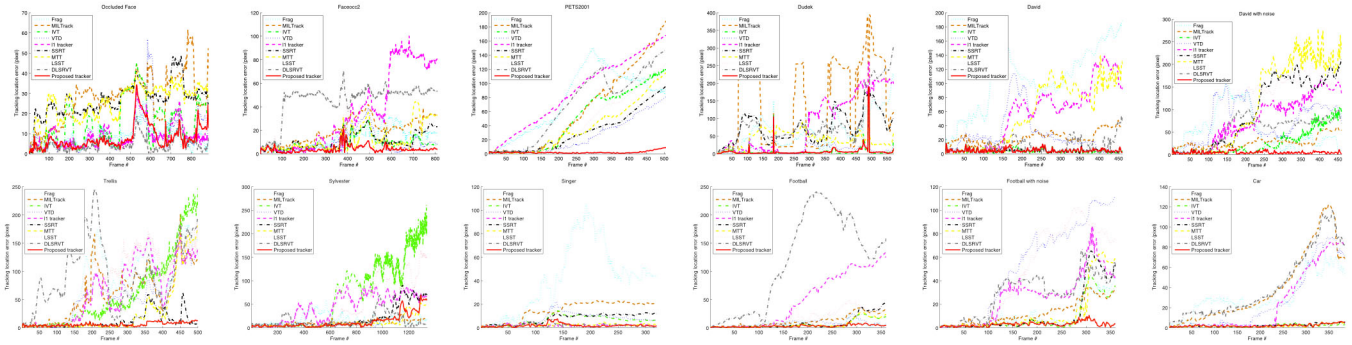


Fig. 5. Frame-by-frame comparison of the tracking location error. The tracking location error measures the Euclidean distance between the center of the tracking result and the ground truth in terms of pixel.

TABLE III
STATISTICAL TEST RESULTS FOR THE PROPOSED TRACKER AGAINST THE FRAGTRACK, MILTRACK, IVT, ℓ_1 TRACKER, SSRT, AND THE PROPOSED TRACKER WITHOUT THE ROBUST SIMILARITY METRIC

Test statistic $\bar{\Delta}/\sigma_{\bar{\Delta}}$		PETS 2001	Dudek	Occluded Face	Occluded Face2	David	David†	Trellis	Sylv	Singer	Football	Football‡	Car	Overall
Frag	LC	221.67	129.54	2.30	133.15	287.77	308.05	119.16	11.17	72.98	30.19	72.14	164.25	7.57
	FR	39.37	176.96	-9.43	N/A	N/A	203.37	40.89	9.91	N/A	33.07	N/A	37.06	9.67
MILTrack	LC	26.42	43.24	6.33	5.91	6.47	7.78	11.47	-3.38	20.13	17.92	47.03	23.24	6.56
	FR	22.89	9.12	3.81	2.52	6.57	6.97	36.42	1.82	13.33	8.69	44.06	18.64	14.98
IVT	LC	8.81	-0.16	9.70	23.44	-1.69	5.38	3.41	4.85	48.55	1.81	-8.42	4.17	5.64
	FR	18.08	1.96	7.22	1.12	N/A	9.00	4.52	7.61	N/A	1.74	N/A	5.11	8.13
VTD	LC	2.60	3.36	4.89	51.55	19.12	14.68	17.02	-12.80	2.61	15.70	5.23	2.68	6.15
	FR	12.12	1.19	3.10	25.86	8.41	23.24	39.72	-7.48	1.98	12.35	8.88	13.74	8.06
ℓ_1 Tracker	LC	2.74	7.97	2.67	13.07	5.07	11.36	5.70	5.09	0.67	3.01	7.70	5.61	9.76
	FR	8.37	15.30	-2.46	14.05	6.23	75.36	27.03	9.74	N/A	14.82	25.30	7.93	10.08
SSRT	LC	6.53	2.81	4.35	12.49	-0.40	16.38	4.61	3.94	9.85	3.82	0.18	9.21	4.83
	FR	28.41	3.30	2.30	3.57	3.57	39.40	5.83	1.06	1.12	13.99	1.12	21.66	7.94
MTT	LC	5.88	12.38	12.88	14.96	3.58	20.46	5.70	-1.14	4.02	4.53	-4.31	3.22	5.55
	FR	24.66	7.58	4.26	6.22	18.35	78.32	12.17	-2.32	NA	67.75	N/A	34.96	8.08
LSST	LC	36.76	1.49	-12.59	10.55	1.76	17.57	7.47	7.72	0.48	1.50	-0.97	1.69	5.89
	FR	23.29	8.93	-9.24	38.89	3.53	58.61	19.40	4.88	N/A	4.36	1.12	3.99	6.07
DLSRVT	LC	139.62	40.08	-18.09	1.28	15.74	3.68	14.17	2.62	-0.89	206.45	162.10	48.56	7.34
	FR	32.42	30.33	-9.43	1.15	12.32	80.62	20.51	1.61	N/A	99.53	161.58	32.09	9.74
Proposed Tracker*	LC	2.63	1.21	3.50	2.12	-0.72	1.38	1.56	1.06	32.81	-0.17	0.71	1.70	3.75
	FR	2.25	1.19	2.88	1.12	1.83	2.06	1.46	-2.22	15.55	0.62	1.12	1.98	4.97

LC: Location
FR: Failure rate
N/A: Not applicable
*Without the robust similarity measure
†Salt and pepper noise added
‡Gaussian noise added.

quantitatively evaluated in terms of the tracking location error and the tracking failure rate. The tracking location error measures the Euclidean distance between the center of the tracking result and the ground truth. The tracking location error is an effective metric for tracking accuracy evaluation when the candidate algorithms can track the target throughout the whole sequence. However, the tracking location error metric may yield an incorrect performance evaluation when the trackers completely lose the target. Thus, a failure rate metric, which indicates the percentage of frames in which the location error was less than 20% of the diagonal length of the rectangle enclosing the target, is also presented. A good tracker should achieve low values in both the tracking location error metric and the failure rate metric. The ground truth of the *David*, *Sylvester*, *Occluded Face*, *Occluded Face2*, and *Dudek* data sets are reported by [2] and [3]. For the other five video clips, the ground truth is manually labeled for quantitative comparison. For the probabilistic trackers, all the quantitative results are averaged over 25 runs. The quantitative results are summarized in Tables I and II and are shown in Fig. 5. The proposed tracker outperforms the other nine competitors in terms of the averaged tracking location error metric in all of the video sequences, except for the *Occluded Face*, *David* and *Sylvester* sequence. The proposed method also has the lowest failure rate in most of the test data sets. Considering the overall performance, the proposed tracker only has an averaged tracking location error of five pixels and a 5% failure rate, which are far lower than those of other trackers in all 10 video sequences that contain thousands of frames.

A standard, statistical, one-sided hypothesis test [6] is also conducted to evaluate the superior performance of the proposed tracker further. In this test, the null hypothesis H_0 indicates that the proposed tracker is not superior to the reference tracker. The alternative hypothesis H_1 indicates that the proposed tracker is significantly better than the others.

The sample performance differences at the jth repetition can be calculated as follows:

$$\Delta^j = C^j_{\text{REF}} - C^j_{\text{SRLT}} \tag{12}$$

where C^j_{REF} and C^j_{SRLT} denote the quantified performance of the reference trackers and the proposed sparse representation-based local appearance tracker, respectively. C^j represents the mean location error or the failure rate in run j. The hypothesis test is based on the sample mean of the above differences

$$\overline{\Delta} = \frac{1}{J}\sum_{j=1}^{J}\Delta^j \tag{13}$$

and its standard error

$$\delta_{\overline{\Delta}} = \sqrt{\frac{1}{J^2}\sum_{j=1}^{J}\left(\Delta^j - \overline{\Delta}\right)^2}. \tag{14}$$

The null hypothesis H_0 is rejected if the test statistic $\overline{\Delta}/\delta_{\overline{\Delta}}$ exceeds a threshold μ_α that represents a point on the standard Gaussian distribution, which corresponds to the upper-tail probability of α. The performance of the proposed tracker is significantly superior to the reference tracker if the test statistic is larger than $\mu_\alpha = 1.65$ ($\alpha = 0.05$). The results of the hypothesis testing on location error and failure rate with respect to different video sequences and all experiments are reported in Table III. The N/A marker indicates that the test is not applicable as the standard error $\delta_{\overline{\Delta}}$ becomes zero. Such cases usually stem from comparisons in which both the competitors produce 0% failure rates in all experiments (both Δ^j and $\overline{\Delta}$ are 0) or from comparisons that involve FragTrack, which yield consistent results against the proposed tracker and have 0% failure rates in all repetitions ($\Delta^j = \overline{\Delta}$). The alternative hypothesis H_1 is accepted by the majority of comparisons that use the reference tracker for different test video sequences and for overall performance.

2) Effect of the Robust Similarity Metric: The RSM is important for the improvement of tracking performance as shown in Tables I and II, in which the Proposed Tracker* denotes the proposed method without the robust similarity metric. The tracking location error and failure rate are reduced when the target is significantly affected by illuminations and occlusion (*Singer*, *Occluded Face*, and *Occluded Face2* sequences) when the RSM is used. Although the performance of the proposed method without the RSM yields comparative tracking performance in the *David* and *Football* sequences, the tracking failure rate increases significantly when strong random noise is added. These experimental results show that the use of the RSM helps because it can effectively eliminate the outliers and keep the informative pixels for tracking.

We also conduct experiments with the proposed method on replacing the MAD-based RSM by SRTT [5] and SRSE [30] inferring to further evaluate the effectiveness of the robust similarity metric. The comparison experiment results (Fig. 6) show that the MAD-based method outperforms the others. The SRTT and SRSE inferring have comparative performance in the *Faceocc2* and *David*† dataset, but fail to track the target

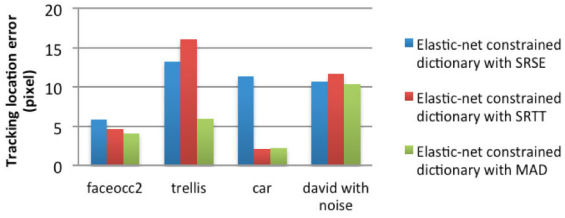


Fig. 6. Performance comparison of the proposed algorithm with SRTT, SRSE, and the proposed MAD-based robust similarity metric.

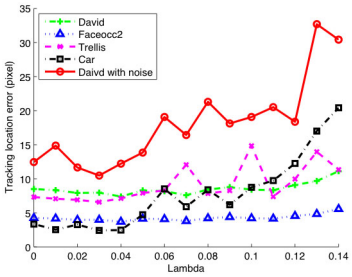
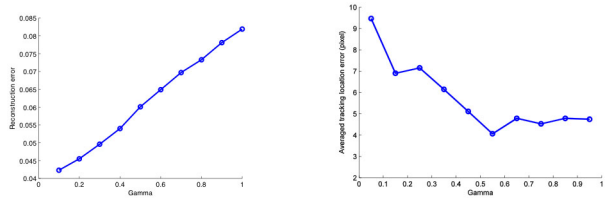


Fig. 7. Tracking location errors with varying parameter λ.

in the other two sequences. The different tracking results are probably due to the proposed sparsity consistency constraint. The sparsity consistency constraint unifies the dictionary learning and representation stages and enhances the generative and discriminative capabilities of the proposed method. The SRTT and SRSE, on the other hand, are not able to achieve the constraint because the trivial templates in SRTT and the Laplacian noise term affect the sparsity of the solutions.

3) Impact of the Sparsity Consistency Constraint: In this experiment, the proposed sparsity consistency constraint facilitates the tracking performance. The proposed algorithm is tested on the *David*, *Faceocc2*, *Trellis*, *Car*, and *David*† sequences that involves a wide range of challenges. Fig. 7 shows the tracking results with different values of the parameter λ. When the value of λ is moderate ($\lambda = 0.03 - 0.05$), the proposed model has better tracking accuracy than those without the sparsity consistency constraint ($\lambda = 0$). However, the averaged tracking location error increases significantly if the value of λ becomes large ($\lambda > 0.06$). Therefore, the parameter λ is set to 0.04 in all the experiments.

4) Effect of the Elastic-Net Constraint: In this experiment, we assess how the elastic-net constraint contributes to the robustness of the tracker against significant occlusion. We perform experiments on calculating the reconstruction error the tracking result of *David* sequence with regarding to varying value of gamma. The results in Fig. 8(a) shows that larger value of γ leads to larger reconstruction error because a larger γ yields more sparse loadings in each basis vector. We also assess how the elastic-net constraint contributes to the robustness of the tracker against significant occlusion. The tracking experiments are conducted on the *Faceocc2* sequence, with various values of the parameter γ. The averaged tracking location error curve for several values of γ is plotted in Fig. 8(b). In Fig. 8(b), the tracking performance improves when the value of γ increases. This finding validates the fact that adding the elastic-net constraint improves the robustness



(a) (b)

Fig. 8. Performance analysis with varying parameter γ. (a) Reconstruction error with regarding to varying γ. (b) Tracking location error with regarding to varying γ for the *Faceocc2* dataset.

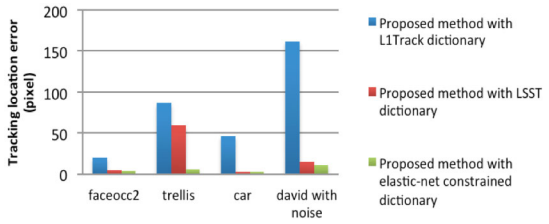
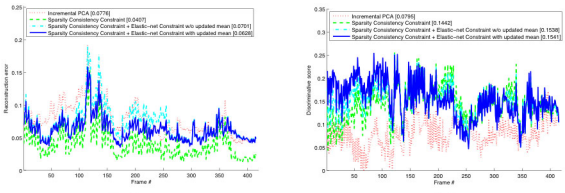


Fig. 9. Performance comparison of the proposed algorithm with ℓ_1 tracker dictionary, LSST dictionary, and the proposed elastic-net constrained dictionary.

of the tracker against occlusion. However, the tracking location error rises gradually if the value of γ becomes extremely large ($\gamma > 0.55$). The tracking location error gradually rises because sparser basis vectors are unlikely to be affected by occlusions. However, a larger proportion of zero entries in the basis vectors results in more information loss, thus degenerating the generative capabilities of the appearance model. In this paper, the parameter γ is set to 0.55 in all the experiments for a fair trade off.

In addition, we perform experiments to evaluate the elastic-net constrained dictionary by keeping the RSM and varying the dictionary. We use the dictionary with raw target templates trivial templates in [5] and dictionary learned with the sparse error term [30] to replace the proposed dictionary. Four representative data sets (*Trellis*, *Faceocc2*, *Car*, and *David*†) that covers challenges of pose and illumination variations, heavy occlusions, background cluttering, and high-level random noise are used in these experiments. As we can observe in Fig. 9, the proposed method with elastic-net constrained dictionary achieves the best results. The ℓ_1 tracker dictionary [5] provides unsatisfied results because it uses the raw templates that cannot represents the significant appearance variations and ambiguities caused by different poses and background cluttering (*Trellis*, *Car*, and *David*†). The LSST dictionary [30] yields better results than the ℓ_1 tracker dictionary because it is able to incrementally learn the target appearances. However, it fails in the Trellis dataset because of lacking of discriminative capability.

5) Generative and Discriminative Capabilities of the Proposed Appearance Model: To demonstrate the generative and discriminative capabilities of the proposed appearance model, the image patches from the target and background region are extracted in the *David* sequence. All the image patches are resized to 32 × 32 pixels. The conventional subspace analysis-based appearance model, incremental PCA [2],



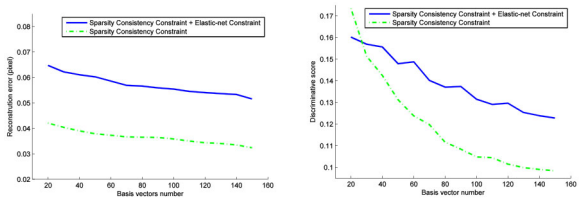
(a) (b)

Fig. 10. Performance analysis of the proposed appearance model with different constraints and updated means. (a) Generative power. (b) Discriminative power.

is used for comparison. For the proposed appearance model, an undercomplete dictionary comprising 36 basis vectors is used. Less than ten basis vectors are selected to represent the appearance if the regularization parameter λ is set to 0.04. The proposed appearance model is tested without the elastic-net constraint to investigate the impact of sparsity consistency constraint independently. In addition, we also test the proposed appearance model without updated mean. The incremental PCA algorithm retains the top ten eigenvectors for fair comparison. In these experiments, the first 36 image patches from the target region are used to initialize the proposed appearance model and the subspace-based appearance model. For the remaining frames and the tracker, the dictionary and PCA bases are updated every five frames. The reconstruction error and discriminative score in [11] are used to quantify the generative and discriminative capabilities of the appearance models, respectively. The discriminative score is defined as follows:

$$D(\mathbf{X}) = \left| E(\mathbf{X}^+) - E(\mathbf{X}^-) \right| \tag{15}$$

where $\mathbf{X}^+$ and $\mathbf{X}^-$ indicate the set of target and background image patches, and $E(\mathbf{X})$ is the reconstruction error. As shown in Fig. 10(a), the reconstruction error curves produced by the proposed appearance model with and without the elastic-net constraint are lower than that produced by the incremental PCA-based appearance model. Although the appearance model with the elastic-net constraint has a larger averaged reconstruction error (0.0628) than the appearance model without the constraint (0.0407), the former still has stronger generative power than the incremental PCA-based appearance model, with a mean error of 0.0776. As shown in Fig. 10(b), the discriminative score obtained from the incremental PCA-based appearance model is lower than the discriminative score from the proposed model. The proposed appearance model incurred averaged discriminative scores of 0.1541 and 0.1442 with and without the elastic-net constraint, respectively; whereas the incremental PCA-based appearance model yields a lower averaged discriminative score of 0.0795. In addition, the proposed appearance model without the updated mean leads to a larger reconstruction error 0.0701. Moreover, the discriminative scores of the proposed method with and without updated mean are similar. These findings verify that the proposed appearance model with the sparsity consistency constraint and updated mean improves generative power and discriminative capability. On the other hand, enforcing the elastic-net constraint in the model slightly increases the reconstruction error.



(a) (b)

Fig. 11. Performance analysis of the proposed appearance model with varying dictionary size. (a) Generative power. (b) Discriminative power.

However, the constraint improves discriminative capability and robustness against occlusions, making the model more suitable for tracking.

The generative and discriminative powers of the proposed appearance model are also evaluated with varying dictionary size, and the results are shown in Fig. 11. The reconstruction errors decrease slightly as the dictionary size grows, which verifies the argument that using an undercomplete dictionary is sufficient for visual tracking. Another reason for using an undercomplete dictionary is that the discriminative power of the model drops significantly if more basis vectors are included in the dictionary because adding more basis vectors contributes to the generative power of the target and the background representations. In this paper, the number of basis vectors is empirically determined and is set to 36 in all the experiments.

C. Qualitative Evaluation

1) Tracking Under Heavy Occlusions: The tracking performance against heavy occlusions is tested using the *Occluded Face*, *Occluded Face2*, *PETS2001*, and *Dudek* video sequences. The targets in the *Occluded Face* and *Occluded Face2* sequences undergo long durations of occlusion numerous times. The second clip has more challenging occlusions and large pose variations. The proposed tracker robustly tracked the face in the two sequences because the algorithm learns the local features for appearance representation, which is insensitive to occlusions. However, other methods can only roughly track the face and are not as accurate as the proposed tracker. In Fig. 12(a) and (b), MILTrack, MTT, and VTD are distracted by the book after a long duration of occlusion in frames 73, 217, and 614 of the Occluded Face sequence. The ℓ_1 tracker and DLSRVT are sensitive to pose variations and fails to locate the target from frame 594 and 171 of the *Occluded Face2* sequence, respectively. The *PETS2001* and *Dudek* sequences [Fig. 12(c) and (d)] are examples of how the proposed tracker outperforms the conventional tracking algorithms when the target is temporarily subjected to severe and full occlusions. Aside from occlusions, the two sequences also have the challenges such as nonrigid appearance variations, background cluttering, and pose changes. The proposed sparse representation-based local appearance model enables the proposed tracker to handle the occlusions and to distinguish the target from the cluttered background. FragTrack, MILTrack, and DLSRVT have unsatisfied performances in the two data sets, and the ℓ_1 tracker loses the target in the *Dudek* sequence when it encounters rapid motion. The IVT, LSST, and MTT tracker drift away from the target because of the occlusion caused by the bicyclist in the *PETS2001* sequence.

2) Tracking Under Significant Pose and Illumination Variations: The *David*, *Trellis*, *Sylvester*, and *Singer* sequences are used to evaluate the performance of the proposed tracker under severe pose and illumination variations. In the *David* data set [Fig. 12(e)], the proposed tracker successfully tracks the face of the person during the whole sequence. The ℓ_1 tracker, MTT, and VTD lose the target from the frame 113 to 171. However, the VTD algorithm resumed tracking the target after the person turned his head to face the front (frame 296). MILTrack and DLSRVT can roughly locate the target, but tracking results are less accurate.

The *Trellis* data set [Fig. 12(f)] provides a challenging scenario in which the person undergoes a combination of significant illumination and poses variations as well as background cluttering. The experimental results show that only the proposed tracker can track the target throughout the whole video. The proposed method is robust because the appearance model has a rich generative power to represent appearance variations and a strong discriminative capability to prevent visual drifts caused by the cluttered background. However, VTD, LSST, DLSRVT, and MILTrack drift from the target at an early stage. The IVT fails to track the target because of drastic pose and illumination changes in frame 327.

The *Sylvester* sequence [Fig. 12(g)] has challenging lighting, scale, and pose changes. The VTD, SSRT, MTT, DLSRVT, and the proposed tracker perform well in this clip, but the IVT, LSST, and ℓ_1 tracker fail in frame 579, 694, and 940, respectively. FragTrack and MILTrack are also capable of tracking the animal doll, but yield large tracking errors. Fig. 12(h) shows the tracking results from the *Singer* sequence. The proposed tracker can stably track the singer even with dramatic lighting variations onstage. The IVT, VTD, and SSRT algorithms are vulnerable to failure after the illumination changes frames 130, 160, and 273.

3) Tracking Under Background Cluttering: In the *Football* and *Car* sequences, the goal is to track the football player and the car that are moving against similar backgrounds. In Fig. 12(i), VTD and the proposed tracker are capable of tracking the player correctly. However, the VTD algorithm has a slightly larger location error in the tracking experiments (frames 163, 256, and 292). The proposed tracker works well in such a challenging background cluttering scenario because it uses an appearance model that has strong generative and discriminative powers. On the other hand, the IVT, MTT, and SSRT are distracted by the similar helmets of other players and the ambiguous background after frame 292. In the *Car* sequence [Fig. 12(j)], MILTrack and VTD are gradually distracted by the background, and fail to track the car after frame 81 and 179, respectively. The DLSRVT and ℓ_1 tracker also start to drift from the car in the 237th and 295th frame, respectively. The proposed tracker and MTT successfully and stably track the car in the whole sequence.

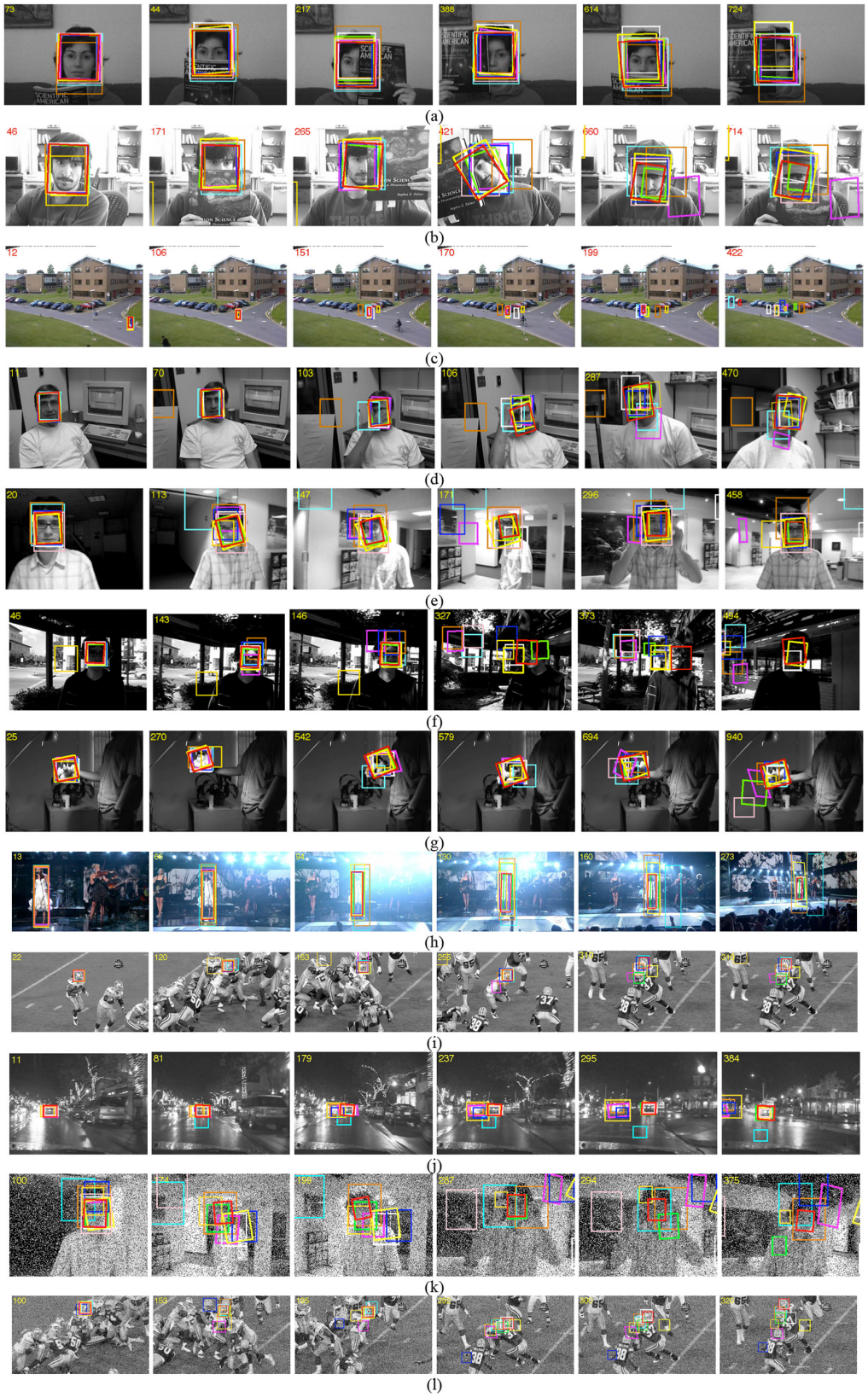


Fig. 12. Screen shots of the comparison of tracking results. (a) Occluded Face. (b) Occluded Face2. (c) PETS2001. (d) Dudek. (e) David. (f) Trellis. (g) Sylvester. (h) Singer. (i) Football. (j) Car. (k) David †. and (l) Football‡. The results of the proposed tracker, FragTrack, MILTrack, IVT, VTD, ℓ_1 Tracker, SSRT, MTT, LSST, and DLSRVT are indicated by the red, cyan, orange, green, blue, magenta, yellow, white, pink, and gold boxes, respectively.

4) Tracking Under Severe Random Noise: The tracking results are shown in Fig. 12(k) and (l), which include severe random noise (salt and pepper and Gaussian noises). The proposed method tracked the face accurately and robustly in the *David*† sequence, although disturbances from the combined random noise and pose and illumination variations

occurred. Notably, the other algorithms (i.e., IVT and SSRT), which can successfully handle the same sequence without noise, failed in this corrupted case. The proposed tracker can track the helmet throughout the *football*‡ sequence successfully. VTD and LSST drifted into the background when random noise was present in the sequence.

V. Conclusion

An appearance model that uses sparse representation with an online sparse dictionary learning scheme has been presented. Instead of commonly used subspace representations, the sparse representation scheme with a sparsity constraint, which has richer descriptive capability and stronger discriminative power, is used for appearance representation. The target appearance is modeled using an online dictionary learning approach with an elastic-net constraint that induces sparsity in the dictionary. The online learned sparse dictionary is robust to the occlusions because it models the target appearance with local features. Furthermore, using an undercomplete dictionary is sufficient for visual tracking tasks, thereby facilitating a more efficient implementation compared with other sparse representation-based algorithms. Moreover, an RSM has been presented to evaluate the similarities between the observed sample and the learned appearance model. An affine particle filter is integrated with the proposed appearance model to form a robust visual tracking algorithm. The proposed tracker is compared with seven state-of-the-art trackers using ten challenging benchmark video sequences to validate the robustness. The qualitative and quantitative results indicate that the proposed tracker is more accurate than the reference trackers.

References

[1] H. Qiao, P. Zhang, B. Zhang, and S. Zheng, "Learning and intrinsic-variable preserving manifold for dynamic visual tracking," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 40, no. 3, pp. 868–880, Jun. 2010.

[2] D. A. Ross, J. Lim, R. S. Lin, and M.-H. Yang, "Incremental learning for robust visual tracking," *Int. J. Comput. Vis*., vol. 77, nos. 1–3, pp. 125–141, 2008.

[3] B. Babenko, S. Belongie, and M. H. Yang, "Visual tracking with online multiple instance learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Miami, FL, USA, Jun. 2009, pp. 983–990.

[4] C. Shen, J. Kim, and H. Wang, "Generalized kernel-based visual tracking," *IEEE Trans. Circuits Syst. Video Technol*., vol. 20, no. 1, pp. 119–130, Jan. 2010.

[5] X. Mei and H. Ling, "Robust visual tracking using ℓ_1 minimization," in *Proc. IEEE Conf. Comput. Vis.*, Kyoto, Japan, Sep./Oct. 2009, pp. 1436–1443.

[6] T. Bai and Y. F. Li, "Robust visual tracking with structured sparse representation appearance model," *Pattern Recognit.*, vol. 45, no. 6, pp. 2390–2404, Jun. 2012.

[7] T. Bai and Y. F. Li, "Robust visual tracking using flexible structured sparse representation," *IEEE Trans. Ind. Inf.*, vol. 10, no. 1, pp. 538–547, Feb. 2014.

[8] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Trans. Pattern Anal. Mach. Intell*., vol. 31, no. 2, pp. 210–227, Feb. 2009.

[9] J. Wright and Y. Ma, "Dense error correction via ℓ_1-minimization," *IEEE Trans. Inf. Theory*, vol. 56, no. 7, pp. 3540–3560, Jun. 2010.

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

[11] B. Liu, J. Huang, L. Yang, and C. Kulikowsk, "Robust tracking using local sparse appearance model and K-selection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit*., Providence, RI, USA, Jun. 2011, pp. 1313–1320.

[12] T. Bai, Y. F. Li, and X. Zhou, "Monocular human motion tracking with discriminative sparse representation," *Adv. Robot.*, vol. 28, no. 6, pp. 403–414, 2014.

[13] R. Rubinstein, A. M. Bruckstein, and M. Elad, "Dictionaries for sparse representation modeling," *Proc. IEEE*, vol. 98, no. 6, pp. 1045–1057, Jun. 2010.

[14] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *J. Roy. Stat. Soc. Ser. B-Stat. Methodol*., vol. 67, no. 2, pp. 301–320, 2005.

[15] D. L. Donoho and Y. Tsaig, "Fast solution of ℓ_1-norm minimization problems when the solution may be sparse," *IEEE Trans. Inf. Theory*, vol. 54, no. 11, pp. 4789–4812, Nov. 2008.

[16] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, "Least angle regression," *Ann. Stat.*, vol. 32, no. 2, pp. 407–499, 2004.

[17] K. Engan, S. O. Aase, and J. Hakon Husoy, "Method of optimal directions for frame design," in *Proc. IEEE Conf. Acoust., Speech Signal Process.*, vol. 5. Phoenix, AZ, USA, Mar. 1999. pp. 2443–2446.

[18] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Trans. Signal Process.*, vol. 54, no. 11, pp. 4311–4322, Nov. 2006.

[19] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online learning for matrix factorization and sparse coding," *J. Mach. Learn. Res*., vol. 11, pp. 19–60, Jan. 2010.

[20] M. Elad, M. A. T. Figueiredo, and Y. Ma, "On the role of sparse and redundant representations in image processing," *Proc. IEEE*, vol. 98, no. 6, pp. 972–982, Jun. 2010.

[21] M. Elad, J. Starck, D. Donoho, and P. Querre, "Simultaneous cartoon and texture image inpainting using morphological component analysis (MCA)," *J. Appl. Comput. Harmon. Anal*., vol. 19, no. 3, pp. 340–358, 2005.

[22] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld, "Face recognition: A literature survey," *ACM Comput. Surv*., vol. 35, no. 4, pp. 399–458, Dec. 2003.

[23] H. Zou, T. Hastie, and R. Tibshirani, "Sparse principal component analysis," *J. Comput. Graph. Stat.*, vol. 15, no. 2, pp. 265–286, 2006.

[24] T. Zhou, D. Tao, and X. Wu, "Manifold elastic net: A unified framework for sparse dimension reduction," *Data Mining Knowl. Discov.*, vol. 22, no. 3, pp. 340–371, 2011.

[25] L. Sachs, *Applied Statistics: A Handbook of Techniques*, New York, NY, USA: Springer-Verlag, 1984.

[26] M. S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp, "A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking," *IEEE Trans. Signal Process.*, vol. 50, no. 2, pp. 174–188, Feb. 2002.

[27] A. Adam, E. Rivlin, and I. Shimshoni, "Robust fragments-based tracking using the integral histogram," in *Proc. IEEE Conf. Comput. Vis.*, New York, NY, USA, Jun. 2006.

[28] J. Kwon and K. M. Lee, "Visual tracking decomposition," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, San Francisco, CA, USA, Jun. 2010, pp. 1269–1276.

[29] T. Zhang, B. Ghanem, S. Liu, and N. Ahuja, "Robust visual tracking via structured multi-task sparse learning," *Int. J. Comput. Vis*., vol. 101, no. 2, pp. 367–383, Jan. 2013.

[30] D. Wang, H. Lu, and M. H. Yang, "Least soft-threshold squares tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Portland, OR, USA, Jun. 2013, pp. 2371–2378.

[31] Q. Wang, F. Chen, W. Xu, and M. H. Yang, "Online discriminative object tracking with local sparse representation," in *Proc. IEEE Conf. Applicat. Comput. Vis.*, Breckenridge, CO, USA, Jan. 2012, pp. 425–432.

[32] K. Zhang, L. Zhang, and M. H. Yang, "Real-time compressive tracking," in *Proc. 12th Eur. Conf. Comput. Vis.*, Florence, Italy, Oct. 2012, pp. 864–877.

[33] N. Wang, J. Wang, and D. Y. Yeung, "Online robust non-negative dictionary learning for visual tracking," in *Proc. IEEE Conf. Comput. Vis.*, Sydney, NSW, Australia, Dec. 2013, pp. 657–664.

[34] D. Wang, H. Lu, and M. H. Yang, "Online object tracking with sparse prototypes," *IEEE Trans. Image Process*., vol. 22, no. 1, pp. 314–325, Jan. 2013.

[35] Z. Hong, X. Mei, D. Prokhorov, and D. Tao, "Tracking via robust multi-task multi-view joint sparse representation," in *Proc. IEEE Conf. Comput. Vis.*, Sydney, NSW, Australia, Dec. 2013, pp. 649–656.

[36] W. Zhong, H. Lu, and M. H. Yang, "Robust object tracking via sparsity-based collaborative model," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Providence, RI, USA, Jun. 2012, pp. 1838–1845.

[37] Z. Hong, X. Mei, and D. Tao, "Dual-force metric learning for robust distracter-resistant tracker," in *Proc. 12th Eur. Conf. Comput. Vis.*, Florence, Italy, Oct. 2012, pp. 513–527.

Tianxiang Bai received the B.S. and M.S. degrees in mechanical engineering from Guangzhou University, Guangzhou, China, and from Guangdong University of Technology, Guangzhou, in 2006 and 2009, respectively, and the Ph.D. degree in robot vision from the Department of Mechanical and Biomedical Engineering, City University of Hong Kong, Hong Kong, in 2012.

From 2008 to 2009, he was a Visiting Researcher at the Department of Mechanical Engineering, Korea Advanced Institute of Science and Technology, Daejeon, Korea. He joined ASM Pacific Technology Ltd., Hong Kong, in 2012, and is currently a Senior Computer Vision Engineer with the Research and Development Department. His current research interests include robot vision and machine learning, especially for visual tracking and defect inspection in semiconductor manufacturing.

Dr. Bai was a recipient of the T. J. Tarn Best Paper in Robotics at the IEEE International Conference on Robotics and Biomimetics in 2012, the Third Prize in IEEE Region-10 Postgraduate Student Paper Contest in 2012, and the First Prize in the IEEE Hong Kong Section (PG) Student Paper Contest in 2011.

You-Fu Li (SM'01) received the B.S. and M.S. degrees in electrical engineering from the Harbin Institute of Technology, Harbin, China, and the Ph.D. degree in robotics from the Department of Engineering Science, University of Oxford, Oxford, U.K., in 1993.

From 1993 to 1995, he was a Research Staff at the Department of Computer Science, the University of Wales, Wales, U.K. He joined City University of Hong Kong, Hong Kong, in 1995, and is currently a Professor with the Department of Mechanical and Biomedical Engineering. His current research interests include robot sensing, robot vision, 3-D vision, and visual tracking.

Dr. Li has served as an Associate Editor for the IEEE TRANSACTIONS ON AUTOMATION SCIENCE AND ENGINEERING, the IEEE ROBOTICS AND AUTOMATION MAGAZINE, and an Editor for CEB, IEEE International Conference on Robotics and Automation.

Xiaolong Zhou received the B.S. and M.S. degrees in mechanical engineering from Fuzhou University, Fuzhou, China, in 2007 and 2010, respectively, and the Ph.D. degree in mechanical and biomedical engineering from City University of Hong Kong, Hong Kong, in 2013.

He is currently an Assistant Professor at the College of Computer Science and Technology, Zhejiang University of Technology, Zhejiang, China. His current research interests include visual tracking, pattern recognition, and 3-D reconstruction.