Robust Visual Tracking Using Flexible Structured Sparse Representation

Tianxiang Bai and Youfu Li, *Senior Member, IEEE*

***Abstract*—In this work, we propose a robust and flexible appearance model based on the structured sparse representation framework. In our method, we model the complex nonlinear appearance manifold and the occlusion as a sparse linear combination of structured union of subspaces in a basis library, which consists of multiple incremental learned target subspaces and a partitioned occlusion template set. In order to enhance the discriminative power of the model, a number of clustered background subspaces are also added into the basis library and updated during tracking. With the Block Orthogonal Matching Pursuit (BOMP) algorithm, we show that the new flexible structured sparse representation based appearance model facilitates the tracking performance compared with the prototype structured sparse representation model and other state of the art tracking algorithms.**

***Index Terms*—Appearance model, block orthogonal matching pursuit, sparse representation, visual tracking.**

I. Introduction

VISUAL tracking is one of the most well-known and fruitful research topics in computer vision communities as it can be widely applied in video surveillance [1], human computer interaction [2] and intelligent transportation system [3], etc. However, tracking objects undergoing significant viewpoint, pose and illumination variations has remained challenges. In this work, we concentrate on designing a robust and flexible sparse representation based appearance model that confronts the aforementioned challenges.

For years, various appearance models have found a rich expression in the computer vision literature, in particular, formulating with subspace representations [4]–[6] and sparse representations [7]–[11], [23], [25], [28], which motivated this work. Subspace representations are based on the core assumption that the appearance manifold can be linearly approximated by single or multiple low dimensional subspaces in a short time interval. These methods have been justified that they are effective approaches to model the appearance changes, such as pose and illumination variations. However, as mentioned in [5], they are sensitive to gross errors caused by significant occlusions. On the other hand, the sparse representation based appearance models exhibit promising performance against occlusions for robust visual tracking. These methods attempt to handle the occlusions as a sparse noise component and approximate the target appearance via seeking a sparse linear combination in an overcomplete basis library consisting of target and trivial templates [7]. Yet, these sparse representation based tracking algorithms are still far from practical applications as they suffer from the tremendous computational load. In addition, most of the basis libraries in these methods are composed of raw target templates that are directly sampled from the images, which limit their generative capabilities. Extensive improvements such as using feature extraction [8] and exploring the inherent structure of the data [10], [11] for computational complexity reduction are presented. In addition, there is also a growing interest in applying the machine learning techniques to enrich the descriptive abilities of the appearance model [10], [11], [23]. Moreover, a multitask learning based sparse representation [25], was proposed to solve the tracking problem. It generalized the prototype sparse representation based tracker [7] and showed significant enhanced tracking accuracy and efficiency by exploring the joint sparsity within the particle filter framework and the accelerated proximal gradient algorithm.

Recent works reported impressive improvements of tracking efficiency and robustness by using a structured sparse representation appearance model [10], [11], [29]. The object appearance is modeled as a sparse linear combination of structured union of subspaces instead of individual templates. It is verified that this structured model is suitable for practical visual tracking tasks by considering the predefined basis library structure and contiguous spatial distribution of occlusions. The computational load can also be significantly reduced by using the Block Orthogonal Matching Pursuit (BOMP) algorithm [12] to solve the structured sparse representation problem. However, visual drifts are still observed when there is extreme pose and illumination variation. This is because the above method models the object appearance with a single subspace and its descriptive abilities are limited for approximating the complex and nonlinear appearance manifold.

In this work, we propose a new appearance model that has richer expressiveness and stronger discriminative power based on the structured sparse representation framework. Different from the original structured sparse representation based tracking [10], [11], we model nonstationary object appearance manifold with a number of low dimensional linear subspaces rather than barely one subspace. On the other hand, a clustered

Manuscript received January 10, 2013; revised April 06, 2013; accepted June 16, 2013. Date of publication July 03, 2013; date of current version December 12, 2014. This work was supported by Research Grants Council of Hong Kong (Project No. CityU 118311), City University of Hong Kong (Project No. 7008176) and the National Natural Science Foundation of China (No. 51175087). Paper no. TII-13-0012.

T. Bai is with the Department of Research and Development, ASM Pacific Technology Limited, Kwai Chung, Hong Kong (e-mail: tianxiangbai@gmail.com).

Y. Li is with the Department of Mechanical and Biomedical Engineering, City University of Hong Kong, Kowloon, Hong Kong (e-mail: meyfli@cityu.edu.hk).

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/TII.2013.2272089

background subspace set is also added in to the basis library and is updated during tracking to enhance the discriminative power, which is ignored in the prototype structured sparse representation based method. Using the BOMP algorithm, the object appearance can be represented by a sparse union of subspaces in the target subspace set, while the continuous occlusion can be adaptively masked out by occlusion template sets. Finally, the proposed appearance model is integrated into a particle filter framework for tracking.

This paper is organized as follows. Section II briefs the principles of structured sparse representation based appearance model. The proposed new appearance model with multiple subspaces based target appearance learning and representation as well as the clustered background subspaces learning scheme are presented in Section III. Section IV describes the integration of the proposed appearance model and particle filter for visual tracking. Experiments proceed in Section V and this work is concluded in Section VI.

II. Structured Sparse Representation Based Appearance Model

The framework of structured sparse representation aims at approximating the observed object appearance $\mathbf{y} \in \mathbb{R}^L$ by a sparse linear combination over a basis library $\mathbf{A}$

$$\mathbf{y} = \mathbf{A}\boldsymbol{\omega} \tag{1}$$

where $\mathbf{A} = [\mathbf{A}_T \quad \mathbf{I}] \in \mathbb{R}^{L\times(d+L)}$ is a basis library consisting of target template set $\mathbf{A}_T \in \mathbb{R}^{L\times d}$ and the occlusion template set that is a identity matrix $\mathbf{I} \in \mathbb{R}^{L\times L}$. It is sensible to assume $\boldsymbol{\omega} = [\mathbf{x}\ \mathbf{e}] \in \mathbb{R}^{d+L}$ is a block sparse coefficient vector, if we consider the predefined basis library structure and continuous spatial distribution of occlusion in practical visual tracking tasks [10], [11]. The basis library $\mathbf{A}$ is a block-structured matrix, the basis vectors of which are sorted in blocks, thus enabling block-sparse representations for a variety of object appearances and occlusions. We can treat the basis library $\mathbf{A}$ as a concatenation of m column-blocks $\mathbf{A}[j]$ of size $L \times d_j$

$$\mathbf{A} = [\mathbf{A}[1]\ \mathbf{A}[2] \ldots \mathbf{A}[m]] \tag{2}$$

where d_j is the number of basis vectors that belong to the jth block. Accordingly, the sparse coefficient vector $\boldsymbol{\omega}$ can be denoted as a concatenation of m blocks $\boldsymbol{\omega}[j]$ of length d_j

$$\boldsymbol{\omega} = [\boldsymbol{\omega}[1]\ \boldsymbol{\omega}[2] \ldots \boldsymbol{\omega}[m]]^T. \tag{3}$$

The first block in $\mathbf{A}$ and $\boldsymbol{\omega}$ corresponds to the target template set $\mathbf{A}_T$ and target coefficient vector $\mathbf{x}$, respectively. Based on a local region analysis, the remaining blocks in $\mathbf{A}$ are used to represent the occlusions in the partitioned local regions [11].

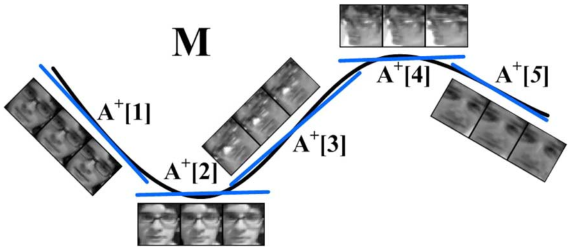


Fig. 1. Complex and nonlinear appearance manifold M can be approximated as a union of subspaces.

A block k-sparse vector $\boldsymbol{\omega}$ is defined as the nonzero values in the vector that are concentrated in k blocks only, as denoted by $\|\boldsymbol{\omega}\|_{2,0} \le k$, where

$$\|\boldsymbol{\omega}\|_{2,0} = \sum_{l=1}^{m} I(\|\boldsymbol{\omega}[l]\|_2 > 0). \tag{4}$$

The indicator function $I(\cdot)$ counts the number of blocks in $\boldsymbol{\omega}$ with nonzero Euclidean norm.

The coefficient vector $\boldsymbol{\omega}$ can then be estimated by approximating the observation $\mathbf{y}$ using the basis library $\mathbf{A}$ under a sparsity prior using the Block Orthogonal Matching Pursuit (BOMP) algorithm [12]. The structured sparse representation problem can be formulated as

$$\boldsymbol{\omega}^* = \arg\min_{\boldsymbol{\omega}} \|\mathbf{y} - \mathbf{A}\omega\|_2 \text{ subject to } \|\boldsymbol{\omega}\|_{2,0} \le T \tag{5}$$

where T is a parameter to impose the sparsity prior.

To capture the appearance variations, the raw target template set $\mathbf{A}_T$ is replaced by an Eigen template set $\mathbf{U}$ obtained by the SVD algorithm $\mathbf{A}_T = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T$ firstly. Then the Eigen templates $\mathbf{U}$ are updated with incremental PCA [13], [14] during tracking when new observed appearances are available. The learned Eigen templates provide a richer description than the raw templates because they span an optimal subspace that has the smallest reconstruction error with not only the current nut also the past appearance information.

III. Flexible Structured Sparse Representation Based Appearance Model

The prototype structured sparse representation framework can be used to represent the gradual appearance variations with a single learned Eigen template set based on the assumption that the appearance of the target can usually reside in a low dimensional submanifold in a short time interval [6]. However, in reality, the appearance manifold is complex and nonlinear [15]. The expressiveness of the single subspace model is limited and difficult to handle the nonlinear appearance manifold. Therefore a more flexible model that has richer descriptive capabilities for appearance representation is needed. In addition, the original model is essentially a generative method that ignores the background information so that it is prone to drift away from the target in the case of background cluttering.

Inspired by the previous problems, we propose a new structured sparse representation based appearance model that has richer descriptive capabilities and stronger discriminative power for tracking. As shown in Fig. 1, the new appearance model incrementally learns multiple low dimensional linear subspaces instead of single subspace representation to approximate the nonlinear appearance manifold. To improve the robustness of the model against background cluttering, we add a background subspace set into the basis library that provides additional discriminative power for the tracker. The new overcomplete basis library is constructed as

$$\mathbf{A} = [\mathbf{A}^+ \ \mathbf{A}^- \ \mathbf{I}] \tag{6}$$

where $\mathbf{A}^+ = [\mathbf{A}^+[1], \mathbf{A}^+[2] \dots \mathbf{A}^+[p]]$ is the new target subspace set that consists of p blocks that store the basis sets for the low dimensional subspaces learned from the target appearance, and $\mathbf{A}^- = [\mathbf{A}^-[1], \mathbf{A}^-[2] \dots \mathbf{A}^-[q]]$ is the background subspace set that includes q basis sets from the clustered background subspaces. Correspondingly, the block-sparse coefficient vector can be denoted as

$$\boldsymbol{\omega} = [\boldsymbol{\omega}^+ \boldsymbol{\omega}^- \mathbf{e}]^T \tag{7}$$

where $\boldsymbol{\omega}^+$ and $\boldsymbol{\omega}^-$ represent the decomposed coefficients that correspond to new target subspace and background subspace sets.

Although the learned multiple subspaces capture the nonlinear appearance manifold by allowing several possible subspace descriptions, the exact subspaces that represent the new observation is unknown a priori. In visual tracking, it is reasonable to assume that the new observed appearance can be well represented by only a few subspaces within the target subspace set because of the local linearity. We argue that the observation can be represented by, at most, one third of the subspaces in $\mathbf{A}^+$. Similar to [11], we also set the tolerance of occlusion in our model is one third of the sampling area. In this way, the structured sparse representation problem in (5) can be rewritten as

$$(\boldsymbol{\omega}^{+*}, \mathbf{e}^*) = \arg\min_{\boldsymbol{\omega}^+, \mathbf{e}} \left\| \mathbf{y} - [\mathbf{A}^+\mathbf{A}^-\mathbf{I}] \begin{bmatrix} \boldsymbol{\omega}^+ \\ \boldsymbol{\omega}^- \\ \mathbf{e} \end{bmatrix} \right\|_2$$
$$\text{subject to} \quad \|\boldsymbol{\omega}^+\|_{2,0} \le T_1 \text{ and } \|\mathbf{e}\|_{2,0} \le T_2, \tag{8}$$

where the parameters are set to $T_1 = \lfloor p/3 \rfloor$ and $T_2 = \lfloor (m - p - q)/3 \rfloor$, and $\lfloor \cdot \rfloor$ returns the nearest integer less than or equal to the value inside.

This new structured sparse representation problem can be also solved using BOMP by adding constraints on $\boldsymbol{\omega}^+$ and $\mathbf{e}$. As mentioned in [11], a valid observation could be better represented by the target templates rather than the background and occlusion templates. Additionally, a valid observation can achieve the highest correlation with the target basis sets in $\mathbf{A}^+$. Therefore, the matching stage of BOMP can act as a classifier that eliminates the outliers by judging whether the target subspace basis is picked in the first iteration. Given that the background subspace set is added into the basis library, more invalid observations are expected to be eliminated, such that the discriminative power and computational efficiency of the model is improved. From the second iteration, the background template library is removed from the basis library for computational load reduction.

A. Incremental Appearance Learning With Multiple Subspaces

In visual tracking, it is shown that the nonlinear appearance manifold can be approximated by a number of low dimensional linear subspaces [15]. In our approach, we employ a merging and insert strategy (see Algorithm 1) to perform an incremental learning procedure when the observed samples are given in a sequential manner [16]. Given that the maximum number of the subspaces is fixed, new subspaces are created with d dimension periodically by SVD during tracking and then inserted into the target subspace set. If the number of existing subspaces is equal to the predefined maximum, the two most similar subspaces are merged, and a space is vacated for the new created subspace.

Algorithm 1: Incremental appearance learning with multiple subspaces

```
Input: A sequence of samples ${\mathbf{y}_1, \dots, \mathbf{y}_N}$, the dimension
of the target subspace $d$, and the maximum number of the
target subspaces $p$.

Form a new $d$-dimensional subspace
  $\Phi^{new} = (\mathbf{U}^{new}, \mathbf{\Lambda}^{new}, n^{new})$ every $d$ frames:
  $\Phi^{new} \leftarrow \{\mathbf{y}_i, \dots, \mathbf{y}_{i+d}\}$
If an empty block $\mathbf{A}^+[l], l \in [2, p]$ exists, then
  $\mathbf{A}^+[l] \leftarrow \mathbf{U}^{new}$.
Else
  $(j,k)^* = \arg\max Sim(\mathbf{A}^+[j], \mathbf{A}^+[k]), j,k \in [1, \dots, p], j \neq k$,
  $\mathbf{A}^+[j] = \mathbf{A}^+[j] \cup \mathbf{A}^+[k]$ and $\mathbf{A}^+[k] \leftarrow \mathbf{U}^{new}$.
End if.

Output: the target subspace set $\mathbf{A}^+$.
```

Mathematically, let $\Phi = \{\Phi_1, \dots, \Phi_p\}$ represent the object appearance manifold and $\Phi_i, i \in [1, \dots, p]$ denote the local submanifold. Let $\Phi_i = (\mathbf{U}_i, \mathbf{\Lambda}_i, n_i)$ denotes an eigenspace model that approximates the ith local submanifold, where $\mathbf{U}_i, \mathbf{\Lambda}_i, n_i$ represent the eigenvectors, eigenvalues, and the total numbers of samples that form the subspace, respectively. Without loss of generality, zero mean is assumed by removing the sample mean. In the proposed algorithm, the similarity between two subspaces is determined by the canonical angles [17]. Given two subspaces $\Phi_1 = (\mathbf{U}_1, \mathbf{\Lambda}_1, n_1)$ and $\Phi_2 = (\mathbf{U}_2, \mathbf{\Lambda}_2, n_2)$, such that $a = dim(\Phi_1) \ge dim(\Phi_2) = b \ge 1$, then there are b canonical angles between the two subspaces. The similarity of two subspaces can be represented as

$$Sim(\Phi_1, \Phi_2) = \prod_{k=b-d+1}^{b} \sigma_k \left(\mathbf{U}_1^T \mathbf{U}_2\right) \tag{9}$$

where σ_k is the kth sorted eigenvalue computed using SVD. This similarity metric is an approximate measurement using the d largest principal angles.

For the subspace merging procedure, we use the method proposed in [18] to update the model incrementally. To keep this paper self-contained, we briefly introduce the algorithm here. The first step is to construct an orthonormal basis set $\mathbf{U}' = [\mathbf{U}_1 \ \mathbf{v}]$ that spans the subspaces Φ_1 and Φ_2, where $\mathbf{v} = orth(\mathbf{H})$

and $orth(\cdot)$denotes the orthogonalization process. $\mathbf{H}$ contains the residues of each of the eigenvectors in $\mathbf{U}_2$ with respect to the eigenspace of Φ_1

$$\mathbf{G} = \mathbf{U}_1^T \mathbf{U}_2, \tag{10}$$

$$\mathbf{H} = \mathbf{U}_2 - \mathbf{U}_2 \mathbf{G}. \tag{11}$$

By denoting the covariance matrices of subspaces Φ_1 and Φ_2 as $\mathbf{C}$ and $\mathbf{D}$, the combined covariance matrix can easily be obtained by

$$\mathbf{E} = \frac{n_1}{n_1 + n_2}\mathbf{C} + \frac{n_2}{n_1 + n_2}\mathbf{D}. \tag{12}$$

The second step is to solve a new eigenproblem with respect to the combined covariance matrix $\mathbf{E}$:

$$\mathbf{U}'^T \mathbf{E} \mathbf{U}' = \frac{n_1}{n_1 + n_2}\begin{bmatrix} \mathbf{\Lambda}_1 & 0 \\ 0 & 0 \end{bmatrix} + \frac{n_2}{n_1 + n_2}\begin{bmatrix} \mathbf{G}\mathbf{\Lambda}_2\mathbf{G}^T & \mathbf{G}\mathbf{\Lambda}_2\mathbf{\Gamma}^T \\ \mathbf{\Gamma}\mathbf{\Lambda}_2\mathbf{G}^T & \mathbf{\Gamma}\mathbf{\Lambda}_2\mathbf{\Gamma}^T \end{bmatrix} = \mathbf{R}\mathbf{\Lambda}\mathbf{R}^T, \tag{13}$$

where $\mathbf{\Gamma} = \mathbf{v}^T \mathbf{U}_2$, and $\mathbf{\Lambda}$ is the eigenvalue of the merged subspace. The eigenvector is then obtained by a linear transform that rotates the basis set $\mathbf{U} : \mathbf{U} = \mathbf{U}'\mathbf{R}$.

B. Incremental Background Learning With Clustered Subspaces

The background or negative samples are readily obtained compared with the target appearance samples. Let $L(\mathbf{z})$and L_t^* denote the location of the observation $\mathbf{z}$ and the object location, respectively. We randomly sample the background image patches from an annular region $\mathbf{Z} = \{\mathbf{z} : \alpha < \| L(\mathbf{z}) - L_t^* \| < \beta\}$, where α and β are thresholds that define the annular area. In this work, we set the value of α to the diagonal length of the box that encloses the target, while β is assigned to twice of α. These settings allow us to extract the negative data not too close or too far from the target such that the background subspaces are able to eliminate the invalid observation. The negative samples exhibit more diverse properties because the background is usually cluttered. Therefore, we use a different strategy to learn the background information with multiple subspaces (see Algorithm 2).

Algorithm 2: Incremental background Learning with Clustered Subspaces

```
Input: A collection of background samples $\{\mathbf{z}_1, \dots, \mathbf{z}_r\}$,
the dimension of the background subspace $d$, and the
maximum number of the background subspaces $q$.
1. Group the background samples into $q$ clusters using the $k$-means algorithm.
2. Form $q$ new $d$-dimensional subspaces $\Phi_l^{new} = (\mathbf{U}_l^{new}, \mathbf{\Lambda}_l^{new}, n_l^{new}), l = 1, \dots, q$ with the clustered samples.
3. If $\mathbf{A}^-$ is empty then
4.     $\mathbf{A}^-[l] \leftarrow \Phi_l^{new}, l = 1, \dots, q.$
5. Else
6.     For $l = 1$ to $q$.
7.         $j^* = \arg\max Sim(\mathbf{A}^-[j], \Phi_l^{new}), j \in [1, \dots, q].$
8.         $\mathbf{A}^-[j] \leftarrow \mathbf{A}^-[j] \cup \Phi_l^{new}.$
9.     End for.
10. End if.

Output: the target subspace library $\mathbf{A}^-$.
```

To initialize the background subspace set $\mathbf{A}^-$, we apply the K-means algorithm to group the sampled image patches into q clusters. Then we create q low dimensional subspaces for each of the cluster and add them into the background subspace set $\mathbf{A}^-$. Once the clustered background subspaces are initialized, they are updated incrementally. During the tracking process, the online sampled background image patches are firstly clustered and form q subspaces. Then we find the most similar existing background subspace with the new created subspace and then perform the merging process as mentioned previously.

IV. Tracking Algorithm With the Particle Filter

We embed the proposed appearance model 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 the observation $\mathbf{y}_{1:t} = \{\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_t\}$ up to and including time t. By applying the Bayes' theorem, the Bayes filter can be written as

$$p(\mathbf{x}_t|\mathbf{y}_{1:t}) \propto p(\mathbf{y}_t|\mathbf{x}_t) \int_{\mathbf{x}_{t-1}} p(\mathbf{x}_t|\mathbf{x}_{t-1}) p(\mathbf{x}_{t-1}|\mathbf{y}_{1:t-1}) d\mathbf{x}_{t-1}, \tag{14}$$

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 [19], 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 similarity between an observed target candidate and the recovered image as determined by

$$p\left(\mathbf{y}_t|\mathbf{x}_t^i\right) = \begin{cases} 0 & \text{if } \mathbf{y}_t^i \text{ is outlier} \\ \exp^{-\lambda \mathbf{r}_t} & \text{else} \end{cases} \tag{15}$$

where λ denotes the weighting parameter, and $\mathbf{r}_t = \|\mathbf{y}_t - \hat{\mathbf{y}}_t\|_2$ is the residual between the observed target sample $\mathbf{y}_t$and the recovered image $\hat{\mathbf{y}}_t = \mathbf{A}_t \boldsymbol{\omega}_t$. We set $\lambda = 5$ in all our experiments similar to the prior work [11], [22].

The motion model predicts the current state based on the previous state. In this paper, an affine image warping is used to model the target motion between two consecutive frames. We formulate the state vector $\mathbf{x}_t = (x_t, y_t, \eta_t, s_t, \beta_t, \phi_t)$ at time t with six parameters of affine transformation wherex_t, y_t denote the x, 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 their previous state $\mathbf{x}_{t-1}$ and are assumed to be mutually independent:

$$p(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathbb{N}(\mathbf{x}_t; \mathbf{x}_{t-1}, \boldsymbol{\psi}) \tag{16}$$

where $\boldsymbol{\psi}$ is a diagonal covariance matrix. The current state is then estimated by maximum a posterior (MAP) that associates with the highest likelihood under the observation model. The tracking algorithm is summarized in Algorithm 3.

Algorithm 3: Proposed tracking algorithm

Initialization: Construct the first target subspace $\mathbf{A}^{+}[1]$ by manually labeling the first 5 frames in a test video sequence. Sample the background image patches from an annular region $\mathbf{Z} = \{\mathbf{z} : \alpha < \|L(\mathbf{z}) - L_t^*\| < \beta)$ and group them into q clusters. Then create background subspace set $\mathbf{A}^{-}$.

```
For $t = 6$ to $N$, where $N$ is the total number of frames.
    Generate $P$ candidate samples $\mathbf{y}_i$ at state $\mathbf{x}_t^i$ according to the affine motion model (16).
    For each $\mathbf{y}_t^i, i = 1 : P$
        Perform BOMP to solve (8). Break the BOMP loops and return $\mathbf{y}_t^i$ as an outlier if the target subspaces are not picked in the first iteration.
        Calculate likelihood with (15) according to the observation model.
    End for.
    Obtain the current state $\hat{\mathbf{x}}_t$ using MAP and store the tracking result $\mathbf{y}_t$.
    Update the basis library $\mathbf{A}_t$ every 5 frames with algorithm 1 and 2.
End for.
```

V. Experiments

The proposed tracking algorithm is implemented using MATLAB on a 3 GHz machine with 2 GB RAM. The maximum numbers of the target subspaces and background subspaces are set to $p = 12$ and $q = 4$, respectively. The dimensions of target and background subspaces are both 5. In other words, the length of the blocks in the target and background subspace set is 5. Each observed sample is resized to a 12×15 patch and is partitioned into six 6×5 subimages for local region analysis. Correspondingly, the length of the remaining blocks in the occlusion template set is set to 30. The first five frames are manually labeled to initialize the tracker and the model is updated every five frames.

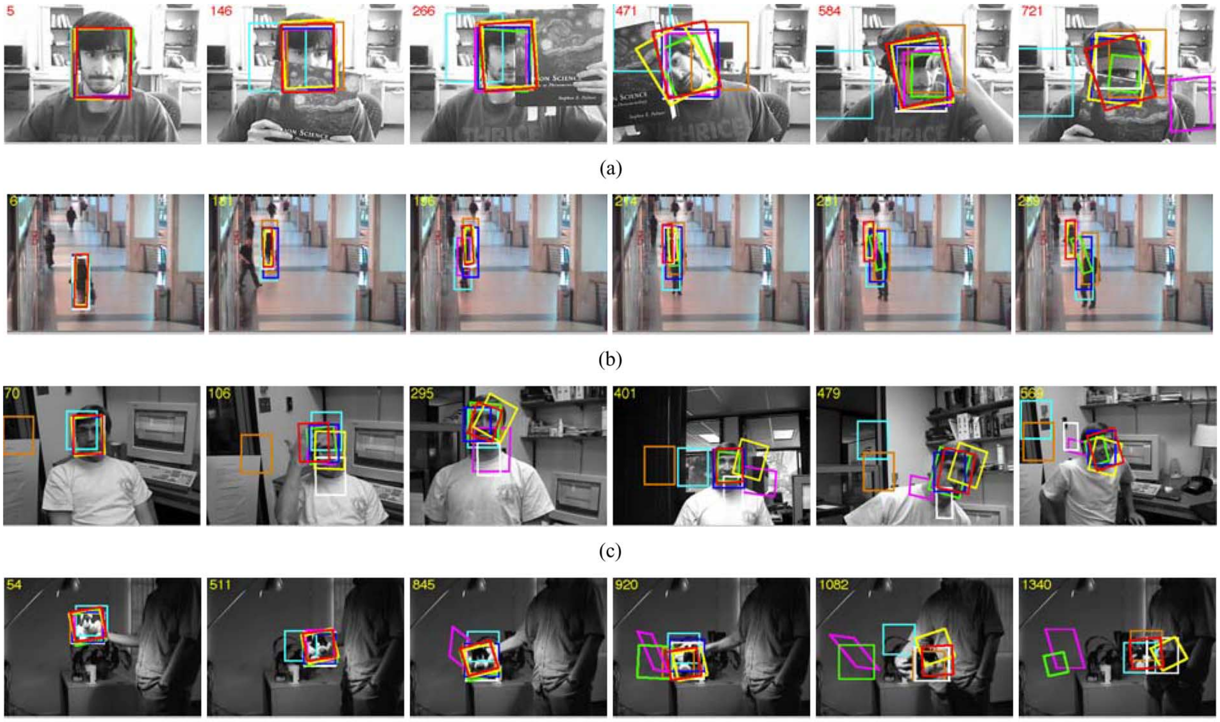
For comparison, we evaluate the proposed tracker against two latest sparse representation based trackers, namely, the ℓ_1 tracker [9], multitask sparse learning based tracker (MTT) [25] and the structured sparse representation based tracker (SSRT) [11], as well as four other state-of-the-art trackers: mean shift (MS) [20], incremental visual tracker (IVT) [5], multiple instance learning based tracker (MILTrack) [21], and visual tracking decomposition (VTD) [22]. The MS algorithm is implemented with the function in OpenCV. The source or binary codes of other six trackers can be obtained in the corresponding project website or the authors. All the reference trackers are implemented with the parameter settings given in their respective papers or use their default initialization. However, as SSRT, IVT, ℓ_1 tracker, MTT, VTD and the proposed tracker are Monte Carlo sampling based methods, we used the same number of samples, 600, to track an object for fair comparison. For the particle filter based trackers (i.e., ℓ_1 tracker, SSRT, MTT, IVT, and the proposed method), we set the parameters of both the motion and observation model as consistent as possible. All the trackers start with the same initial position of the videos. For the probabilistic trackers, the quantitative results are obtained by calculating the mean over five runs. Our tracker runs at about 1 frame per second, which is faster than the prototype structured sparse representation based tracker (SSRT) [11] (1.5–1.8 seconds per frame), the ℓ_1 tracker [7] (2 seconds per frame), and the sparse representation based tracker [23] (9 seconds per frame). The reason for yielding more efficient tracking performance compared to the other methods is the use of background learning scheme that is able to identify the invalid observations in the early stage of the algorithm so that the computational can be effectively reduced. Eight publicly available benchmark video sequences from the prior works [5], [21], [22] are used to evaluate the performance of our tracker under the challenges of significant occlusion, pose, scale and illumination variations as well as background cluttering. The details of the selected datasets are listed in Table I, where the start and the end frames, the initial position for tracking (indicated by the image coordinates) and the main challenges are included. It is also worth to notice that the maximum number of tracking frames for VTD tracker is 1000 with the released source code. The representative tracking results are shown in Figs. 2 and 3.

TABLE I
Details of the Video Sequences

Video Clip	Frames	Initial Position	Challenges
Dudek	1~573	188,192	Large pose variation and short time full occlusion
Faceocc2	1~814	158,107	Significant and long duration occlusion as well as large pose variation
OLSR2cor	1~260	121,153	Non-rigid deformation and significant occlusion
David	1~462	160,106	Significant illumination and pose changes
Trellis	1~502	200,100	Difficult illumination conditions and large pose changes
Sylvester	1~1344	145,77	Uneven illumination and pose variations
Singer	1~351	68,136	Drastic illumination and scale variations
Football	1~362	225,85	Background cluttering and occlusion

A. Qualitative Analysis

In the *Faceocc2* and *OneLeaveShopReeter2cor* sequences [Fig. 2(a) and (b)], the objectives are to track the face and pedestrian with significant occlusions. It is worth notice that there are also challenges regarding to pose variation and nonrigid deformation in the *Faceocc2* and *OneLeaveShopReeter2cor* sequences, respectively. The VTD, SSRT, MTT, and the proposed method perform well in the first sequence. The proposed method has a tracking performance similar to that of the prototype structured sparse representation based tracker because they use the same strategy (partitioned occlusion template set) to handle the occlusions. The ℓ_1 tracker is robust to occlusion, but fails to handle the pose variations. The other methods can also track the target, but are not as accurate as the proposed method. As shown in Fig. 2(b), SSRT, ℓ_1 tracker, MTT and



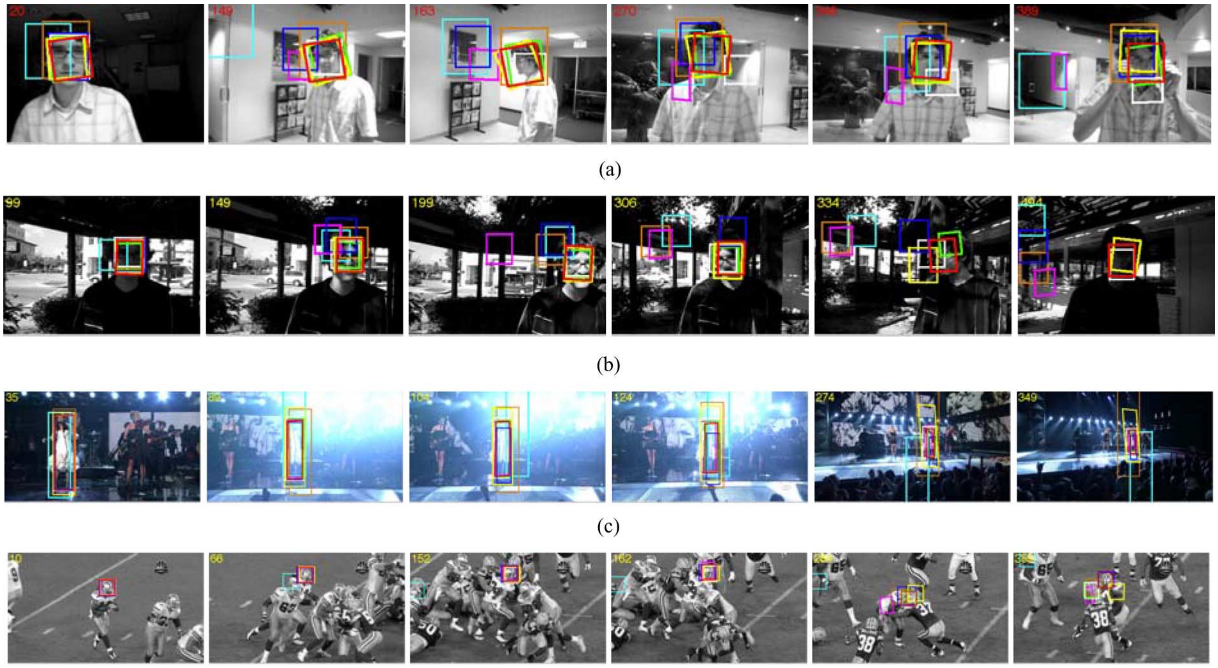
(a)

(b)

(c)

(d)

Fig. 2. Screen shots of comparison tracking results. The results of the proposed tracker, MS, MILTrack, IVT, VTD, ℓ_1 Tracker, MTT, and SSRT are indicated by red, cyan, orange, green, blue, magenta, white, and yellow boxes. (a) Faceocce2. (b) OneLeaveShopReeter2cor. (c) Dudek. (d) Sylvester.



(a)

(b)

(c)

(d)

Fig. 3. Screen shots of comparison tracking results. The results of the proposed tracker, MS, MILTrack, IVT, VTD, ℓ_1 Tracker, MTT, and SSRT are indicated by red, cyan, orange, green, blue, magenta, white, and yellow boxes. (a) David. (b) Trellis. (c) Singer. (d) Football.

our tracker can successfully track the woman in the whole sequence, but visual drifts of the ℓ_1 tracker and SSRT were observed at frame #196 and #214. The other four trackers fail to track the target when the occlusion presents.

The *Dudek*, *Sylvester*, and *David* sequences [Fig. 2(c) and (d) and 3(a)] are used to test the flexibilities of the proposed appearance model for tracking objects undergoing large pose variations. Given that the proposed method adapts the nonlinear appearance manifold with multiple subspaces, it always outperforms the SSRT, which uses only one single subspace for appearance modeling and representation in these sequences. The proposed tracker can track the targets successfully in all these sequences. However, the ℓ_1 tracker fails to track the target undergoing significant pose variations such as at frame #845 in

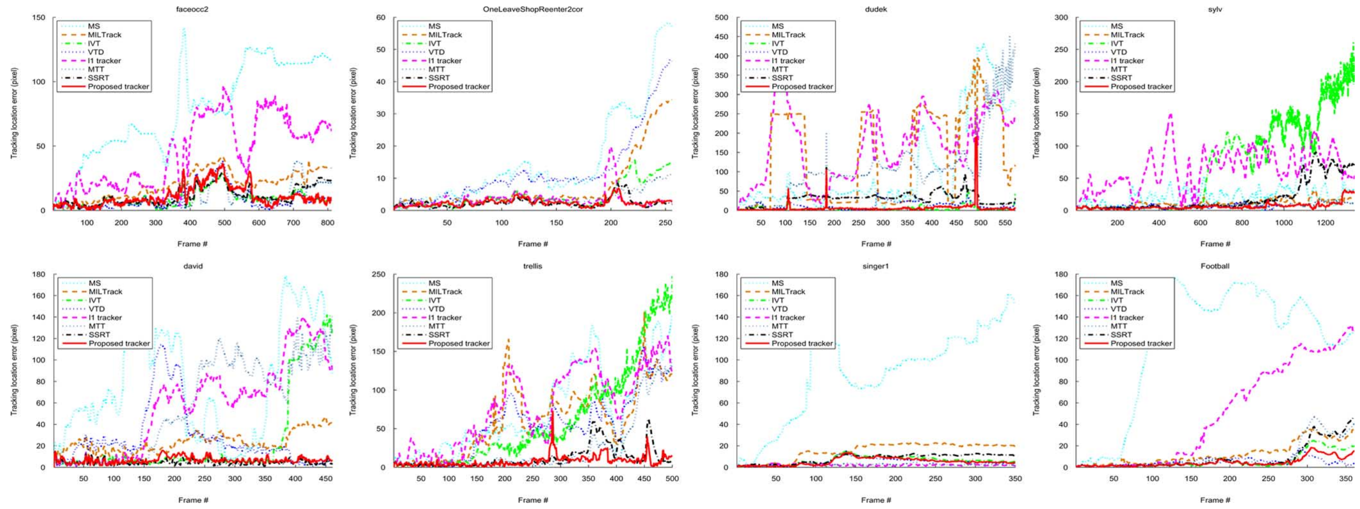


Fig. 4. Tracking location error plots.

the *Sylvester* sequence and at frame #150 in the *David* sequence. The IVT and SSRT lose the target as a result of a combination of drastic pose and illumination changes in the *Sylvester* sequence. The VTD fails to track the face at frame #150 in the *David* sequence probably because it requires more samples (e.g., 800) to produce better accuracy. We also observe that MTT may drift away from the target when it undergoes a combination of large pose and illumination variations in the *Dudek* and *David* sequences. The MILTrack is capable of tracking the target in the *Sylvester* sequences, but drifts to the background in the *Dudek* and *Trellis* video clip.

Fig. 3(b) and(c) shows the tracking results in the *Trellis* and *Singer* sequences that involves challenges of severe illumination changes. The proposed appearance model shows stronger capabilities to adapt the illumination and pose variations, and yields more accurate tracking result than SSRT in the *Trellis* sequence. There is a significant visual drift at frame #334 in the *Trellis* sequence produced by MTT, but MTT can resume tracking the target after a few frames. The other five trackers lose the target successively as shown in Fig. 3(b). In the *Singer* sequence, the proposed tracker, VTD, MTT and the ℓ_1 tracker can stably track the singer even when it under dramatic lighting variations at the stage. The MS, IVT, and SSRT algorithms are vulnerable to drift after the illumination and scale changes.

Fig. 3(d) presents the tracking results under background cluttering environment using the *Football* sequence. The proposed tracker can robustly track the target, but SSRT has difficult at frame #285. It is proven that the introduction of clustered background subspaces learning enhance the discriminative capabilities of tracker. In the sequence, VTD also produces accurate tracking results; however, the other trackers are hijacked by other objects looking similar to the target.

B. Quantitative Analysis

1) Performance of the Tracking Algorithms: We use the averaged tracking location error to quantify the performance of the proposed tracker and the reference trackers in our experiments. In our work, the averaged location errors measure the Euclidean distance between the tracking window center and the ground truth. The location error with respect to the frame number and the averaged location error are summarized in Fig. 4 and Table III, respectively. In most sequences, the proposed tracker has the lower averaged location errors than the original SSRT and has better tracking performance than other reference trackers.

In addition, we used the PASCAL score based detection rate and f-score similar to the work [26] to further evaluate the tracking performance. The detection rate is interpreted by two indices: precision and recall. The precision is defined as the number of true positives divided by the number of retrieved instances, while recall is the number of true positives divided by the total number of frames that contains target of interest (i.e., the sum of retrieved instances and negative falses). Given the score based on the PASCAL challenge [27]

$$score = \frac{area(ROI_T \cap ROI_{GT})}{area(ROI_T \cup ROI_{GT})} \tag{17}$$

where ROI_T is the tracked bounding box and ROI_{GT} is the ground truth bounding box. A frame is indicated as true positive when its PASCAL score is exceeds 0.5, and is identified as retrieved instance when its PASCAL score is larger than 0. False negatives are counted if the PASCAL score is zero but the target is still visible. Precision typically indicates the extent of visual drift if there is still overlap area between the tracked result and ground truth. On the other hand, recall usually interprets the tracking failure level since the cases of no overlap area are also counted. The f-score measures the tracking accuracy by considering both the precision and recall

$$f{-}score = 2 \cdot \frac{precision \cdot recall}{precision + recall}. \tag{18}$$

Table II summarized the detection rates and f-scores on eight different dataset. It was shown that the proposed tracker achieves the higher detection rate and f-score than the original SSRT and has the best or second best performance than the other reference trackers in most of the sequences.

2) Performance of Flexible Structured Sparse Representation: Although the experimental results show that the flexible structured sparse representation strategy facilitates robustness of visual tracking, the contributions of multiple subspaces

TABLE II
DETECTION RATES. RED BOLD FONT INDICATES THE BEST PERFORMANCE, GREEN ITALICS FONT INDICATES THE SECOND BEST

Video Clip	Detection Rates	MS	MIL	IVT	VTD	ℓ_1T	MTT	SSRT	Our method
Dudek	Recall	0.32	0.48	**0.97**	*0.84*	0.49	0.19	0.40	**0.97**
	Precision	0.30	0.79	**0.97**	*0.84*	0.66	0.29	0.41	**0.97**
	f-score	0.25	0.52	**0.97**	*0.84*	0.56	0.22	0.40	**0.97**
Faceocc2	Recall	0.08	0.47	*0.80*	0.71	0.52	0.77	*0.80*	**0.83**
	Precision	0.13	0.47	*0.80*	0.71	0.60	0.77	*0.80*	**0.83**
	f-score	0.10	0.47	*0.80*	0.71	0.55	0.77	*0.80*	**0.83**
OLSR2cor	Recall	0.70	0.79	0.88	0.79	*0.97*	0.93	0.92	**0.99**
	Precision	0.70	0.82	0.88	0.79	*0.97*	0.95	0.92	**0.99**
	f-score	0.70	0.80	0.88	0.79	*0.97*	0.94	0.92	**0.99**
David	Recall	0.12	0.20	0.80	0.43	0.39	0.32	**0.88**	*0.86*
	Precision	0.23	0.20	**0.90**	0.50	0.78	0.45	*0.88*	0.86
	f-score	0.15	0.20	0.84	0.46	0.50	0.36	**0.88**	*0.86*
Trellis	Recall	0.16	0.20	0.53	0.29	0.38	0.63	*0.70*	**0.79**
	Precision	0.28	0.41	**0.86**	0.66	0.71	0.71	0.75	*0.79*
	f-score	0.20	0.27	0.65	0.40	0.49	0.67	*0.72*	**0.79**
Sylvester	Recall	0.29	0.69	0.50	**1**	0.3	0.93	0.73	*0.94*
	Precision	0.30	0.69	*0.95*	**1**	0.51	0.93	0.84	*0.95*
	f-score	0.30	0.69	0.65	**1**	0.37	0.93	0.78	*0.94*
Singer	Recall	0.33	0.22	0.66	0.68	**1**	*0.93*	0.34	0.89
	Precision	0.37	0.22	0.66	0.68	**1**	*0.93*	0.34	0.89
	f-score	0.35	0.22	0.66	0.68	**1**	*0.93*	0.34	0.90
Football	Recall	0.16	0.63	0.87	**0.93**	0.45	0.78	0.64	*0.92*
	Precision	0.86	0.64	0.91	**0.93**	0.80	0.82	0.69	*0.92*
	f-score	0.27	0.64	0.89	**0.93**	0.56	0.80	0.66	*0.92*

TABLE III
TRACKING LOCATION ERROR (PIXELS). RED BOLD FONT INDICATES THE BEST PERFORMANCE, GREEN ITALICS FONT INDICATES THE SECOND BEST

Video Clip	MS	MIL	IVT	VTD	ℓ_1T	SSRT	MT T	Our method
Dudek	122	141	*7*	14	168	75	114	**6**
Face2	78	20	13	**9**	34	*11*	*11*	*11*
Olsr2cor	15	7	4	12	*3*	**2**	4	**2**
David	80	23	24	27	57	**5**	57	*7*
Trellis	71	60	124	54	65	*12*	24	**8**
sylv	22	9	29	**6**	36	21	*7*	**6**
Singer	87	15	6	*5*	**2**	8	**2**	*5*
Football	117	13	*6*	**5**	50	9	10	**5**

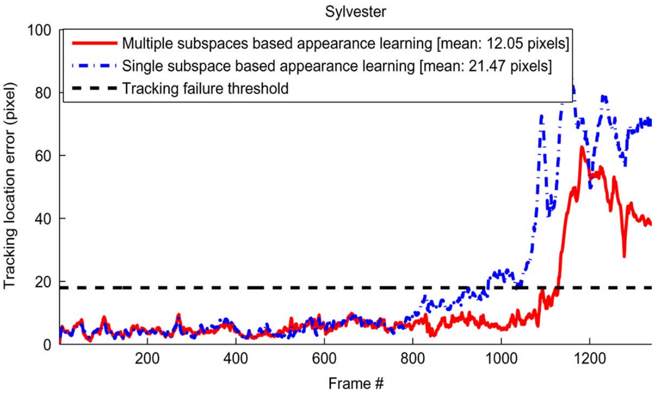


Fig. 5. Tracking location error plots for comparison of single subspace based appearance learning and multiple subspaces based appearance learning. The tracking failure threshold is defined as the tracking location error that is 20% of the diagonal length of the rectangle enclosing the target in the first frame.

based appearance learning and clustered background subspaces learning are still unclear. To demonstrate the effectiveness of the proposed appearance model compared to the prototype structured sparse representation model, we conduct tracking experiments on the *Sylvester* sequence without cluster background subspaces learning scheme. We use the *Sylvester* video sequence because the target undergoes a combination of significant pose and illumination changes, which is a representative example of tracking nonlinear appearance variations. This setting allows us to eliminate the influence of the background subspaces learning and make sure that the tracking performance only depends on multiple subspaces learned from the target appearance. As shown in Fig. 5, the location error curve for the multiple subspaces based appearance learning scheme with the mean error 12.05 pixels is lower than SSRT that uses only a single subspace to represent the target appearance with the mean error 21.47 pixels. Fig. 6 describes how the proposed algorithm actively selects appropriate subspaces to represent the target appearance in *Sylvester* sequence. Fig. 6(d) shows that only one or two subspaces are selected to represent the target appearance. From frame #500 to #504 and from frame #541 to #544, the BOMP algorithm picks the 11st and 3rd subspaces to represent the frontal and side appearances of the target in Fig. 6(a) and (b), respectively. It is shown that the first principle components of these two selected target subspaces appear similar frontal and side appearance. At frame #571 [Fig. 6(c)], our method selected two subspaces from the target subspace set that indicates the union of these subspaces can also be used to represent the target appearance. With help of such a flexible multiple subspaces representation, the proposed appearance model leads to more accurate tracking in spite of severe pose and illumination changes.

We also assess the contribution of the clustered background subspaces learning scheme and how this procedure facilitates the discriminative power or the model. The *Football* sequence

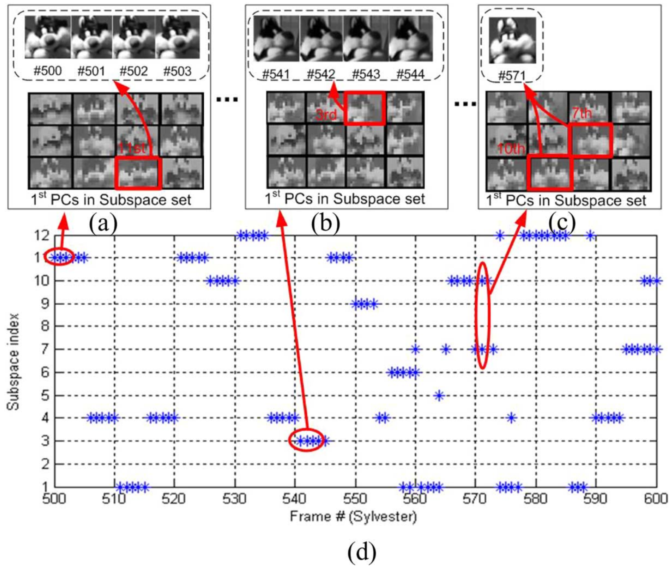


Fig. 6. (a)–(c) Flexible appearance representation of the proposed model. (d) Selected subspaces (indicated by *) with respect to frame number.

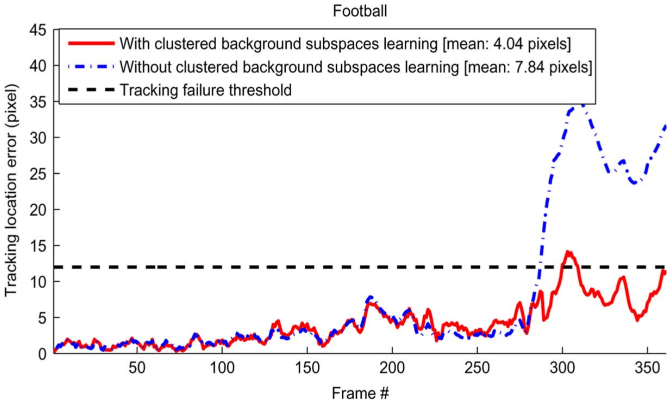


Fig. 7. Tracking location error plots for comparison of with and without clustered background subspaces learning. The tracking failure threshold is defined as the tracking location error that is 20% of the diagonal length of the rectangle enclosing the target in the first frame.

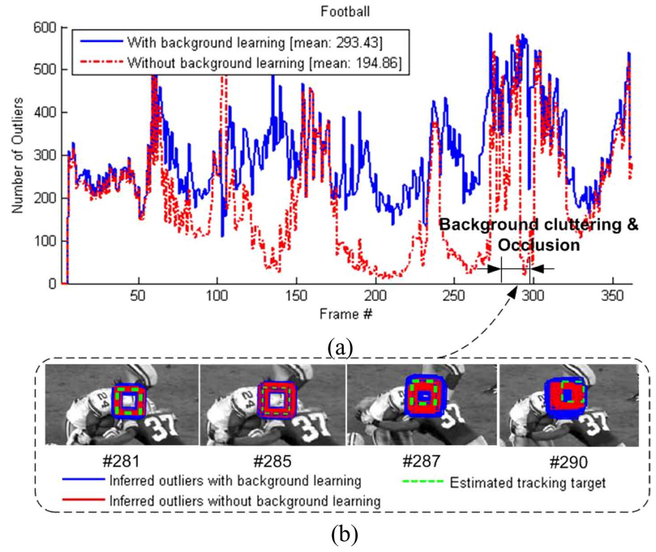


Fig. 8. Comparison of the outlier elimination scheme with and without background learning.

provides another representative example of tracking a target in the situation of background cluttering. We rerun the tracker without adding the background subspace set into the basis library. In this case, a larger averaged tracking location error (7.84 pixels) than that with clustered background subspaces learning strategy (4.04 pixels) is observed. The location error curves are plotted in Fig. 7. Although the two curves are similar with each other in the first 255 frame, a significant drift of the approach without using the background information is observed due to the background cluttering and occlusion, and this leads to a dramatic increase of the tracking location errors in the last 100 frames. The mechanisms of yielding more accurate tracking results with clustered background subspaces learning scheme are shown in Fig. 8. As shown in Fig. 8(a), the proposed method with background learning scheme inferred averaged 293.43 particles as outliers, which are far more than that without learning the background information. Fig. 8(b) provides the intermediary results and illustrates that the proposed model with background learning infers more outliers that covers larger area around the target than that without using the background information. The reason for contributing a more accurate tracking performance is that, the more outliers are identified, the larger effective sampling size is obtained to reduce the effects of degeneracy in the particle filtering procedure [11], [24].

VI. Conclusion

In this paper, we have presented a novel robust and flexible appearance model based on the structured sparse representation framework. The basis library is constructed with target subspace set, background subspace set and partitioned occlusion template set. The target appearance manifold is represented by a sparse union of low dimensional subspaces in the target subspace set. An insert and merging strategy is proposed to learn the multiple target subspaces incrementally. The background information is also introduced into the background subspace set by learning the clustered negative image patches to improve the discriminative capabilities. We use the BOMP algorithm to solve the new structured sparse representation problem. The proposed appearance model and an affine particle filter are integrated to form a robust visual tracking algorithm. Both the qualitative and quantitative results show that our tracker is more accurate than the original structured sparse representation based tracker and the other state of the art methods.

References

[1] X. Zhou, Y. F. Li, and B. He, "Game-theoretical occlusion handling for multitarget visual tracking," *Pattern Recognit.*, vol. 46, no. 10, pp. 2670–2684, Oct. 2013.

[2] C. Tran and M. M. Trivedi, "3-D posture and gesture recognition for interactivity in smart space," *IEEE Trans. Ind. Inf.*, vol. 8, no. 1, pp. 178–187, Feb. 2012.

[3] M. Kafai and B. Bhanu, "Dynamic Bayesian networks for vehicle classification in video," *IEEE Trans. Ind. Inf.*, vol. 8, no. 1, pp. 100–109, Feb. 2012.

[4] M. J. Black and A. D. Jepson, "EigenTracking: Robust matching and tracking of articulated objects using a view-based representation," *Int. J. Comput. Vis.*, vol. 26, no. 1, pp. 63–84, 1998.

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

[6] M. Yang, Z. Fan, J. Fan, and Y. Wu, “Tracking nonstationary visual appearances by data-driven adaptation,” *IEEE Trans. Image Process.*, vol. 18, no. 7, pp. 1633–1644, July 2009.

[7] X. Mei and H. Ling, “Robust visual tracking and vehicle classification via sparse representation,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 11, pp. 2259–2272, Nov. 2011.

[8] B. Liu, L. Yang, J. Huang, P. Meer, L. Gong, and C. Kulikowski, “Robust and fast collaborative tracking with two stage sparse optimization,” in *Proc. 11th Eur. Conf. Comput. Vis.*, Hersonissos, Greece, Sep. 5–11, 2010, pp. 624–637.

[9] S. Kwak, W. Nam, B. Han, and J. Han, “Learning occlusion with likelihoods for visual tracking,” in *Proc. IEEE Conf. Comput. Vis.*, Barcelona, Spain, Nov. 6–13, 2011.

[10] T. Bai, Y. F. Li, and Y. Tang, “Structured sparse representation appearance model for robust visual tracking,” in *Proc. IEEE Conf. Robot. Automat.*, Shanghai, China, May 9–13, 2011, pp. 4399–4404.

[11] 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.

[12] Y. C. Eldar, P. Kuppinger, and H. Bolcskei, “Block-sparse signals: Uncertainty relations and efficient recovery,” *IEEE Trans. Sig. Proc.*, vol. 58, no. 6, pp. 3042–3054, 2010.

[13] M. Brand, “Incremental singular value decomposition of uncertain data with missing values,” in *Proc. 7th Eur. Conf. Comput. Vis.*, Copenhagen, Demark, May 27–June 2, 2002, pp. 707–720.

[14] A. Levy and M. Lindenbaum, “Sequential Karhunen-Loeve basis extraction and its application to images,” *IEEE Trans. Image Process.*, vol. 9, no. 8, pp. 1371–1374, 2000.

[15] K. C. Lee, J. Ho, M. H. Yang, and D. Kriegman, “Visual tracking and recognition using probabilistic appearance manifolds,” *Comput. Vis. Image Understanding*, vol. 99, no. 3, pp. 303–331, Sep. 2005.

[16] Q. Yu, T. B. Dinh, and G. Medioni, “Online tracking and reacquisition using co-trained generative and discriminative trackers,” in *Proc. 10th Eur. Conf. Comput. Vis.*, Marseille, France, Oct. 12–18, 2008, pp. 678–691.

[17] A. Bjoerck and G. G. Golub, “Numerical methods for computing angles between linear subspaces,” *Math. Comp.*, vol. 27, no. 123, pp. 579–594, Jul. 1973.

[18] P. Hall, D. Marshall, and R. Martin, “Merging and splitting eigenspace models,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 9, pp. 1042–1049, Sep. 2000.

[19] 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, 2002.

[20] D. Comaniciu, V. Ramesh, and P. Meer, “Kernel-based object tracking,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 5, pp. 564–577, May 2003.

[21] B. Babenko, M.-H. Yang, and S. Belongie, “Robust object tracking with online mulitiple instance learning,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 8, pp. 1619–1632, Aug. 2011.

[22] J. Kwon and K. M. Lee, “Visual tracking decomposition,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, San Francisco, CA, USA, June 13–18, 2010, pp. 1269–1276.

[23] F. Chen, Q. Wang, S. Wang, W. Zhang, and W. Xu, “Object tracking via appearance modeling and sparse representation,” *Image Vis. Comput.*, vol. 29, no. 11, pp. 787–796, Oct. 2011.

[24] H. Chen and Y. F. Li, “Enhanced particles with pseudolikelihoods for three-dimensional tracking,” *IEEE Trans. Ind. Electron.*, vol. 56, no. 8, pp. 2992–2997, Aug. 2009.

[25] T. Zhang, B. Ghanem, S. Liu, and N. Ahuja, “Robust visual tracking via multitask sparse learning,” in *Proc. IEEE Conf. Comput. Vis.*, Providence, RI, USA, Nov. 6–13, 2012, pp. 2042–2049.

[26] S. Stalder, H. Grabner, and L. Van Gool, “Cascaded confidence filtering for improved tracking-by-detection,” in *Proc. 11th Eur. Conf. Computer Vision*, Hersonissos, Greece, Sep. 5–11, 2010, pp. 624–637.

[27] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, “The pascal visual object classes (VOC) challenge,” *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–308, 2009.

[28] 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.

[29] T. Bai, Y. F. Li, and J. Liu, “Structured compressive sensing for robust and fast visual tracking,” in *Proc. 2012 IEEE Sensors*, Taipei, Taiwan, Oct. 28–31, 2012, pp. 1–4.

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

From 2008 to 2009, he was a visiting Researcher in the Department of Mechanical Engineering at Korea Advanced Institute of Science and Technology, Republic of Korea. After receiving the Ph.D. degree, he joined ASM Pacific Technology Limited as a Computer Vision Engineer in the Department of Research and Development. His research interests are robot vision and machine learning, especially for visual tracking and visual inspection for semiconductor manufacturing.

Youfu Li (S’91–M’92–SM’01) received the B.S. and M.S. degrees in electrical engineering from the Harbin Institute of Technology China, in 1982 and 1986, respectively, 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 Postdoctoral Research Staff in the Department of Computer Science at the University of Wales, Aberystwyth, U.K. He joined City University of Hong Kong in 1995. His research interests include robot sensing, sensor guided manipulation, robot vision, 3Dvision, visual tracking.

Dr. Li has served as an Associate Editor of the IEEE TRANSACTIONS ON AUTOMATION SCIENCE AND ENGINEERING (T-ASE) and is currently serving as Associate Editor of the *IEEE Robotics and Automation Magazine (RAM)*.