Partial Occlusion Handling for Visual Tracking via Robust Part Matching

Tianzhu Zhang*, Kui Jia*, Changsheng Xu, Yi Ma, Narendra Ahuja
Institute of Automation, Chinese Academy of Sciences, P. R. China
Advanced Digital Sciences Center of Illinois, Singapore
School of Information Science and Technology, ShanghaiTech University
University of Illinois at Urbana-Champaign, Urbana, IL USA
*

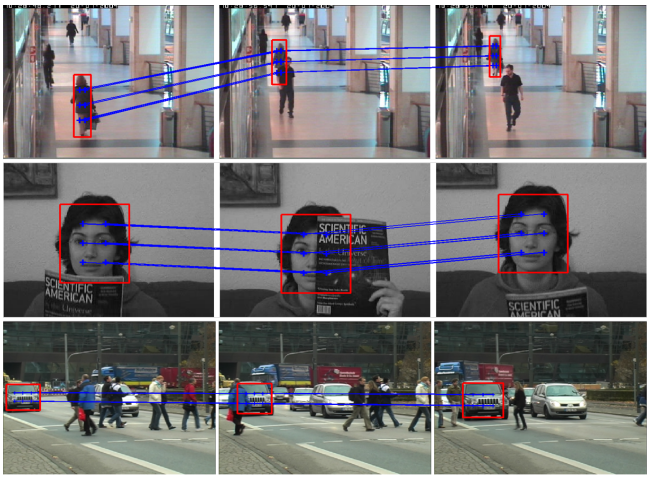
Abstract

Part-based visual tracking is advantageous due to its robustness against partial occlusion. However, how to effectively exploit the confidence scores of individual parts to construct a robust tracker is still a challenging problem. In this paper, we address this problem by simultaneously matching parts in each of multiple frames, which is realized by a locality-constrained low-rank sparse learning method that establishes multi-frame part correspondences through optimization of partial permutation matrices. The proposed part matching tracker (PMT) has a number of attractive properties. (1) It exploits the spatial-temporal locality-constrained property for robust part matching. (2) It matches local parts from multiple frames jointly by considering their low-rank and sparse structure information, which can effectively handle part appearance variations due to occlusion or noise. (3) The proposed PMT model has the inbuilt mechanism of leveraging multi-mode target templates, so that the dilemma of template updating when encountering occlusion in tracking can be better handled. This contrasts with existing methods that only do part matching between a pair of frames. We evaluate PMT and compare with 10 *popular state-of-the-art methods on challenging benchmarks. Experimental results show that PMT consistently outperform these existing trackers.*

1. Introduction

Visual tracking is one of the fundamental problems in computer vision. Its real-world applications range from video surveillance, autonomous vehicles, intelligent traffic control, human-computer interaction, etc. However, visual tracking is challenging due to significant object appearance variations caused by illumination change, occlusion, sensory noise, fast/abrupt object motion, and also cluttered background. Over the past years, tremendous efforts in visual tracking has been made to overcome these challenges, yielding a steady performance improvement. However, most of existing methods [4, 14, 26, 24, 20, 32] focus on modeling the holistic appearance of the target. As a result, the tracking is prone to fail especially in the presence of partial occlusion, as shown in Figure 1(a).

(a) Frames from three different video sequences with partial occlusion.



(b) Qualitative results via robust part matching of multiple frames.

Figure 1. (a) Frames from three different video sequences with partial occlusion. The ground truth track of each object is denoted in green. Clearly, occlusion renders the tracking problem very difficult. (b) The tracking results of our method are denoted with red bounding boxes. The blue cross marks denote the positions of parts, and the blue lines represent their correspondences. It is clear that the part based matching is robust to partial occlusion.

*indicates equal contribution

To design a robust tracking algorithm in spite of partial occlusion, researchers have developed sophisticated appearance models through statistical analysis [15], robust statistics [1, 10], model analysis [12], learning occlusion with likelihoods [21], and sparse representation [24, 35, 34, 5]. Among them, methods part-wisely modeling object appearance [1, 13, 9, 27, 18, 29, 25] become more popular partially because of their favorable property of robustness against partial occlusion. Indeed, when there exists partial occlusion, some parts of the object remain visible which provide reliable cues for tracking. Most of these methods can be viewed as tracking by part-based object matching over time in a video sequence. However, they have the following drawbacks. **(1)** Most of them track each part independently and ignore the collaborations among parts. Parts in one frame should be jointly matched to the corresponding parts in the consecutive frame. **(2)** Most of them establish part matchings between a pair of frames, while ignoring the very same object target appearing in other adjacent or history frames, which may provide additional constraints helpful for part matching. In addition, these methods usually propagate the part matching result in the present frame to subsequent frames, which may accumulate matching errors and are consequently prone to losing track. As a result, these existing part matching based tracking methods are still less reliable when more complicated factors of appearance variations appear in the video.

It is thus desirable that a globally consistent part matching among multiple frames can be established in visual tracking. To achieve this goal, we propose in this paper a new tracking algorithm based on the following observations: **(1)** In a short duration, if appearance of individual parts of object remains unchanged, their intensities in video frames should be similar. Representing appearance of an object part as a vector, the matrix formed by the vectors of the corresponding parts in multiple frames of the short duration should be low-rank, ideally rank-one. We are thus motivated to use the low-rank assumption as a criteria for part matching. **(2)** If there exist object appearance variations in images due to occlusion, object pose change, or illumination change, the low-rank assumption in (1) cannot be fully satisfied. To alleviate their negative effects, we may decompose out these appearance variations in images as sparse errors so that the low-rank assumption still applies. **(3)** Matching of individual parts from multiple adjacent frames should satisfy the locality-constrained property. In spatial domain, parts in one frame should be jointly matched to the parts in other frames. In temporal domain, the matched parts in adjacent frames should satisfy the constant-velocity motion constraint. **(4)** A dictionary of multi-mode target templates should be maintained and progressively updated to model the target appearance variations, which is critical to correct the track after occlusion. **(5)** Part matching across multiple frames is more robust than that between a pair of frames, as it can leverage additional constraints from other frames that contain the very same target.

Motivated by these observations, we propose in this paper a novel method, termed Part Matching Tracker (PMT), for robust visual tracking. PMT realizes part matching among multiple frames by optimizing a partial permutation matrix for each frame, using locality-constrained low-rank and sparsity of matched parts as criteria. Compared with existing part based visual tracking methods, our proposed PMT has three major contributions. **(1)** PMT has the spatial and temporal locality-constrained property, which enables our tracking of local parts to satisfy the constant-velocity motion constraint. **(2)** Part tracking using PMT is based on rank and sparsity optimization, which is potentially effective to model part appearance variations due to occlusion, illumination change, or target pose change over time. **(3)** Our tracker operates in a batch mode, in which multi-mode target templates and frames to be tracked are simultaneously taken into account to determine a global matching of corresponding parts. Even if occlusion happens, the error would not be propagated to the subsequent frames, and the track can be inferred from the observations before and after. Therefore, our tracker effectively cope with partial occlusion as shown in Figure 1(b). We intensively compare with 10 popular state-of-the-art methods on challenging benchmarks. Experimental results show that PMT consistently outperform these existing trackers.

2. Related Work

In general, visual tracking methods can be categorized as either generative or discriminative. Generative methods use appearance models to represent the target object and search for the most similar image regions to the generative model. Popular generative trackers include eigentracker [7], incremental tracker [26, 23], sparse trackers [24, 34, 5], visual tracking decomposition [22], and so on. A drawback of these methods is that they are not designed to distinguish between target and background patches, and are prone to drift. Discriminative methods formulate object tracking as a binary classification, which aims to find the target location that can best distinguish the target from the background. Popular discriminative methods include on-line boosting [14], ensemble tracking [3], online multiple instance learning [4], tracking-learning-detection [20], struck tracker [16], compressive tracker [31], etc. Most of these methods, however, delineate the entire tracked target by a single regular bounding box, which renders them sensitive to partial occlusion and damages tracking performance.

Part based visual tracking draws more recent attention. In [18, 27], multiple people tracking is achieved by part based model motivated by its successful application in human detection [11]. In [17], each part is tracked inde-

pendently, and the results are treated as multiple measurements [28]. Tracking is then achieved by identifying inconsistent measurements. The Frag tracker [1] models object appearance with histogram of local parts and combines votes of matching local patches. However, this template is not updated and therefore it is not expected to handle appearance changes. Nejhum et al. [25] model object shape in terms of a small number of rectangular blocks. The drawback is that it requires manual initialization of part locations carefully. Godec et al. [13] extend the hough forest to the online domain and integrate the voting method for tracking, regardless of the parts' spatial-temporal correlations.

Our formulation of leveraging low-rank sparse property for optimization of partial permutation matrices is similar to [19], which addresses feature matching across a set of images. However, [19] cannot address the visual tracking problem due to its ignorance of the fundamental spatial-temporal locality-constrained property. Furthermore, we formulate each corresponded part of the target object as a low-rank matrix, while all features to be matched are formulated into one low-rank matrix in [19]. Consequently, when using techniques in [19], matching of different parts may interfere with each other and may not be able to well address the partial occlusion problem in tracking.

3. Our Proposed Part Matching Tracker

In this section, we give details of our proposed PMT that is based on a locality-constrained low-rank sparse learning method to optimize partial permutation matrixes for the part correspondence problem among multiple frames.

3.1. Problem Setup

A typical setting of the tracking problem is that an object identified, either manually or automatically, in the first frame of a video sequence is tracked in the subsequent frames by estimating its bounding boxes as it moves. As discussed in previous sections, tracking methods that delineate the tracked object by a single regular bounding box will render them sensitive to partial occlusion and significantly impact their tracking performance. To address this problem, we attempt to adopt part based model to describe the target. The advantage of this model comes from the observation that under partial occlusion conditions, some parts of the object remain visible and distinguishable and can provide reliable cues for tracking. Therefore, if we can infer occlusion information from the confidence scores of individual parts, we can consequently utilize only the parts with high confidence to estimate the position of target over time.

To obtain the confidence scores of individual parts, we can adopt part matching methods. In the situation of visual tracking, a moving object appears in multiple frames of a video sequence. A straightforward approach is to locally build part correspondences between pairs of frames. However, pair-wise matching cannot leverage additional constraints from other frames that also contain the very same target. It may thus be less robust to noise and occlusion of parts. In addition, multi-mode target templates should be maintained to model the variations of target appearance over the history frames, which makes it possible to infer the tracker even after occlusion. Therefore, parts should be matched with a more *global and consistent property* across the sequence and in the target templates, in order to achieve robustness against occlusion. As discussed in Section 1, parts in a video volume have the *spatial-temporal locality-constrained property*, and appearance of the same local parts across frames have the *low-rank sparse property*. To exploit these properties, we propose a locality-constrained low-rank sparse learning method for robust part matching among multiple frames, which include both the multi-model target templates and frames to be tracked.

3.2. Problem Formulation

We sample K_1 target templates at and around the position of object in the first frame, as did in [4, 24, 14]. These target templates are of equal size. They will be progressively updated to incorporate variations of object appearance due to changes in illumination, viewpoint, etc. Target appearance remains the same only for a certain period of time. Eventually the templates are no longer accurate representations of the object appearance. A fixed target template is prone to the tracking drift problem, since it is insufficient to handle changes in appearance. Conversely, if the target templates are updated too often, irrelevant variations will be more possible to be introduced into the templates, causing tracking drift. In this work, we use the target template update scheme as in [24], where the tracking result is added to the template set if none of the templates are similar to the tracking result. For the template, their parts can be extracted by dividing each template into regular grids. Tracking of the object in the incoming frames is realized by matching its candidate parts to those in the target templates. Candidate parts in the incoming frames are simply sampled by particle filtering [2] at and around the parts of the previous tracking results by considering their recursive weights.

For the K_1 target templates, we extract n_k parts from each of them, $k = 1, \ldots, K_1$. These parts are all from the object target. We denote K_2 as the number of incoming frames to be tracked, and $K = K_1 + K_2$. For each of the incoming frames, we also sample n_k parts, $k = 1, \ldots, K_2$. The sampled parts from incoming frames are possibly background patches. For simplicity of notation, we use the same n_k to index parts of target templates and those of incoming frames, i.e., n_k for $k = 1, \ldots, K$. We denote the feature vectors associated with individual parts of any $k \in \{1, \ldots, K\}$ as $\mathbf{F}^k = [\mathbf{f}_1^k, \ldots, \mathbf{f}_{n_k}^k] \in R^{d \times n_k}$, and assume that these feature vectors in $\{\mathbf{F}^k\}_{k=1}^K$ are not corre-

sponded with respect to each other. Our interest is to find $n \leq n_k, \forall k \in \{1, \ldots, K\}$, intrinsic parts from each target template or incoming frame, and establish their correspondences. Because we model their correspondences based on multiple frames and target templates (K), our tracker has the globally consistent property. The correspondences of parts can be modeled by a partial permutation matrix $\mathbf{P^k} \in \mathcal{P}^k$ for each target template or incoming frame, where $\mathcal{P}^k$ is defined as follows:

$$\mathcal{P}^k = \{\mathbf{P}^k \big| \mathbf{P}^k \in \{0,1\}^{n_k \times n}, \mathbf{1}_{n_k}^\top \mathbf{P}^k = \mathbf{1}_n^\top, \mathbf{P}^k \mathbf{1}_n \leq \mathbf{1}_{n_k}\}, \tag{1}$$

where $\{0,1\}^{n_k \times n}$ denotes a $n_k \times n$ matrix whose elements are either 0 or 1 and $\mathbf{1}_n$ denotes a column vector of all 1 of length n. The term $\mathbf{1}_{n_k}^T \mathbf{P}^k = \mathbf{1}_n^T$ in Eq.(1) shows that the i^{th} part, $i = 1, \ldots, n$, is corresponded to only one of all the n_k sampled parts. The term $\mathbf{P}^k \mathbf{1}_n \leq \mathbf{1}_{n_k}$ constrains that each sampled part is corresponded to at most one of the n parts. Motion smoothness in visual tracking implies that the sampled parts of the i^{th} part in next frame must be associated to part i in current frame. So each of the rows of $\mathbf{P}^k$ that correspond to samples for part i in frame k must have exactly one element equal to 1. This constraint is written by defining a $n \times n_k$ matrix $\mathbf{A}^k$ whose i^{th} row flags the samples for part i in frame k, and requiring that $\mathbf{A}^k \mathbf{P}^k \mathbf{1}_n = \mathbf{1}_n$. Here, $\mathbf{A}^k \in \mathbb{R}^{n \times n_k}$ and its i^{th} row $\mathbf{A}_i^k$ is defined as follows: the elements from the $m_1^k + m_{i-1}^k + 1$ to $m_1^k + m_i^k$ are 1, and the others are zeros. The m_i^k is the number of sampled parts for the i^{th} part of the k^{th} image, and $n_k = m_1^k +, \ldots, + m_n^k$.

$$\begin{aligned}\mathcal{P}^k = \{\mathbf{P}^k \big| \mathbf{P}^k \in \{0,1\}^{n_k \times n}, \mathbf{1}_{n_k}^\top \mathbf{P}^k = \mathbf{1}_n^\top, \\ \mathbf{P}^k \mathbf{1}_n \leq \mathbf{1}_{n_k}, \mathbf{A}^k \mathbf{P}^k \mathbf{1}_n = \mathbf{1}_n\},\end{aligned} \tag{2}$$

As a result, Eq (2) can satisfy the spatial-temporal locality-constrained property of parts among multiple frames.

Features of the corresponding parts in different target templates or incoming frames should be linearly correlated. Let $\{\mathbf{P}^k\}_{k=1}^K$ in Eq (2) be the optimized partial permutation matrices such that parts are re-ordered and well corresponded, we thus have $\mathbf{D_i} = [\mathbf{F}^1 \mathbf{p_i}^1, \ldots, \mathbf{F}^K \mathbf{p_i}^K] \in \mathbb{R}^{d \times K}, i = 1, \ldots, n$, which stacks the features of the i^{th} part from target templates or incoming frames as a matrix, and $\mathbf{D_i}$ is rank deficient, ideally rank one. Here, $\mathbf{p_i}^k$ is the i^{th} column of $\mathbf{P}^k$ and $\mathbf{p_i}^k = \mathbf{P}^k \mathbf{e}_i$, where $\mathbf{e}_i$ denotes a unit column vector with all entries set to 0 except the k^{th} one, which is set to 1. Therefore, the problem of optimizing partial permutation matrices $\{\mathbf{P}^k\}_{k=1}^K$ can be formulated as the following rank minimization problem:

$$\begin{aligned}\min_{\{\mathbf{P}^k \in \mathcal{P}^k\}_{k=1}^K, \{\mathbf{L}_i\}_{i=1}^n} \sum_i \operatorname{rank}(\mathbf{L}_i) \\ \text{s.t.} \quad \mathbf{D}_i = \mathbf{L}_i, \quad i = 1, \ldots, n.\end{aligned} \tag{3}$$

In many visual tracking scenarios, target objects are often contaminated by noise, illumination change, object pose change, or partial occlusion. As a result, the parts characterizing the same local appearance information of object in different frames could vary. Thus the low-rank assumption used in (3) is likely to be violated. To improve the robustness, we introduce a sparse error term into (3) to model the noise of the data matrix $\mathbf{D}_i$, where we assume these errors are sparse and only appear in a small fraction of $\mathbf{D}_i$. Therefore, in the presence of noise or occlusion, the problem (3) can be refined as follows:

$$\begin{aligned}\min_{\{\mathbf{P}^k \in \mathcal{P}^k\}_{k=1}^K, \{\mathbf{L}_i, \mathbf{E}_i\}_{i=1}^n} \sum_i \operatorname{rank}(\mathbf{L}_i) + \lambda \|\mathbf{E}_i\|_0 \\ \text{s.t.} \quad \mathbf{D}_i = \mathbf{L}_i + \mathbf{E}_i, \quad i = 1, \ldots, n\end{aligned} \tag{4}$$

where $\|\cdot\|_0$ is ℓ_0-norm counting the number of nonzero entries, and $\lambda > 0$ is a parameter controlling the trade-off between rank property of $\mathbf{L}_i$ and sparsity of $\mathbf{E}_i$. As a result, the part matching problem in Eq (4) guarantees the low-rank sparse property.

4. Optimization

It is not tractable to solve the problem (4) due to the following aspects: (1) The two terms $\operatorname{rank}(\cdot)$ and $\|\cdot\|_0$ are non-convex, discrete-valued functions; (2) The entries of $\{\mathbf{P}^k\}_{k=1}^K$ are constrained to be binary. To make it tractable, we first make use of the convex surrogates $\|\cdot\|_*$ and $\|\cdot\|_1$ to replace $\operatorname{rank}(\cdot)$ and $\|\cdot\|_0$, respectively. Here, $\|\cdot\|_*$ denotes nuclear norm (sum of the singular values) and $\|\cdot\|_1$ is ℓ_1-norm. Applying the relaxation strategy to (4) yields

$$\begin{aligned}\min_{\{\mathbf{P}^k \in \mathcal{P}^k\}_{k=1}^K, \{\mathbf{L}_i, \mathbf{E}_i\}_{i=1}^n} \sum_i \|\mathbf{L}_i\|_* + \lambda \|\mathbf{E}_i\|_1 \\ \text{s.t.} \quad \mathbf{D}_i = \mathbf{L}_i + \mathbf{E}_i, \, i = 1, \ldots, n.\end{aligned} \tag{5}$$

To simplify the subsequent notations, we change the variables and rewrite the formulation (5) as follows:

$$\begin{aligned}\min_{\{\theta^k\}_{k=1}^K, \{\mathbf{L}_i, \mathbf{E}_i\}_{i=1}^n} \sum_i \|\mathbf{L}_i\|_* + \lambda \|\mathbf{E}_i\|_1 \\ \text{s.t.} \quad \mathbf{D} = \mathbf{L} + \mathbf{E} \\ \theta^k \in \{0,1\}^{n_k n}, k \in \{1, \ldots, K\}, \\ \mathbf{Q}^k \theta^k = \mathbf{1}_n, \mathbf{H}^k \theta^k \leq \mathbf{1}_{n_k}, \mathbf{S}^k \theta^k = \mathbf{1}_n^\top,\end{aligned} \tag{6}$$

where $\theta^k = vec(\mathbf{P}^k)$, $vec(\mathbf{P}^k)$ is the vectorization of the matrix $\mathbf{P}^k$, $\mathbf{G}^k = \mathbf{I}_n \otimes \mathbf{F}^k \in \mathbb{R}^{dn \times nn_k}$, $\mathbf{D} = [(\mathbf{L}_1 + \mathbf{E}_1)^\top, \ldots, (\mathbf{L}_n + \mathbf{E}_n)^\top]^\top$, $\mathbf{D} = [\mathbf{G}^1 \theta^1, \ldots, \mathbf{G}^K \theta^K]$, $\mathbf{Q}^k = \mathbf{I}_n \otimes \mathbf{1}_{n_k}^\top \in \mathbb{R}^{n \times nn_k}$, $\mathbf{H}^k = \mathbf{1}_n^\top \otimes \mathbf{I}_{n_k} \in \mathbb{R}^{n_k \times nn_k}$, $\mathbf{S}^k = \mathbf{1}_n^\top \otimes \mathbf{A}^k \in \mathbb{R}^{dn \times nn_k}$, $\otimes$ is the Kronecker product, and $\mathbf{I}_n$ (or $\mathbf{I}_{n_k}$) is the identity matrix of size $n \times n$ (or $n_k \times n_k$). Here, we have used the fact $\operatorname{vec}(\mathbf{XYZ}) = (\mathbf{Z}^\top \otimes \mathbf{X})\operatorname{vec}(\mathbf{Y})$. The (6) involves jointly optimizing a set of K partial permutation matrices, exact solution of which is NP-hard. To get an approximate solution, we use the fast first-order Alternative Direction Method of Multiplier (ADMM) [8]. The general ADMM decomposes a global

problem into local subproblems that can be readily solved. For (6), ADMM decomposes optimization of $\mathbf{L}_i$, $\mathbf{E}_i$, and $\{\mathbf{P}^k\}_{k=1}^K$ into subproblems that update $\mathbf{L}_i$, $\mathbf{E}_i$, and each of $\{\mathbf{P}^k\}_{k=1}^K$, respectively. The augmented Lagrangian function of the optimization problem (6) can be written as:

$$\begin{aligned}
&\mathcal{L}\left(\{\mathbf{L}_i,\mathbf{E}_i\}_{i=1}^n,\{\theta^k\}_{k=1}^K,\mathbf{Y},u\right)\\
&=\sum_{i=1}^n\left(\|\mathbf{L}_i\|_*+\lambda\|\mathbf{E}_i\|_1\right)\\
&+\langle\mathbf{Y},\mathbf{D}-\mathbf{L}-\mathbf{E}\rangle+\frac{u}{2}\|\mathbf{D}-\mathbf{L}-\mathbf{E}\|_F^2 \qquad (7)\\
&\Rightarrow\min_{(\{\mathbf{L}_i,\mathbf{E}_i\}_{i=1}^n,\{\theta^k\}_{k=1}^K,\mathbf{Y},u)}\mathcal{L}\left(\{\mathbf{L}_i,\mathbf{E}_i\}_{i=1}^n,\{\theta^k\}_{k=1}^K,\mathbf{Y},u\right)
\end{aligned}$$

where $\mathbf{Y}\in\mathbb{R}^{dn\times K}$ is a matrix of Lagrange multipliers, u is a positive scalar, $\langle\cdot,\cdot\rangle$ denotes the matrix inner product, and $\|\cdot\|_F$ denotes the Frobenius norm. The ADMM algorithm iteratively updates one of the matrices $\mathbf{L}_i$, $\mathbf{E}_i$, $\{\mathbf{P}^k\}_{k=1}^K$, and the Lagrange multiplier $\mathbf{Y}$ by minimizing (7), while keeping the others fixed to their most recent values. Consequently, we obtain three update steps corresponding to the three sets of variables. The details are the following:

Step 1: Update $\mathbf{L}_i$ and $\mathbf{E}_i$, $\forall i\in\{1,\ldots,n\}$ (with others fixed): The minimization problem (7) w.r.t. $\{\mathbf{L}_i,\mathbf{E}_i\}_{i=1}^n$ can be decomposed into n independent subproblems (each of them is corresponding to one part.). The i^{th} subproblem to update $\mathbf{L}_i$ and $\mathbf{E}_i$ can be equivalently rewritten:

$$\begin{aligned}
\{\mathbf{L}_i,\mathbf{E}_i\}=&\arg\min_{\mathbf{L}_i,\mathbf{E}_i}\|\mathbf{L}_i\|_*+\lambda\|\mathbf{E}_i\|_1+\\
&\langle\mathbf{Y}_i,\mathbf{D}_i-\mathbf{L}_i-\mathbf{E}_i\rangle+\frac{u}{2}\|\mathbf{D}_i-\mathbf{L}_i-\mathbf{E}_i\|_F^2 \qquad (8)
\end{aligned}$$

Then, the solution of (8) can be obtained by solving the optimization problems in Eq (9) and Eq (10), respectively. Here, $\mathbf{X}=\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$ is the singular value decomposition of $\mathbf{X}$, $\mathcal{S}_\lambda(\mathbf{X}_{ij})=sign(\mathbf{X}_{ij})\max(0,|\mathbf{X}_{ij}|-\lambda)$ is the soft-thresholding operator, and $\mathcal{J}_\lambda(\mathbf{X})=\mathbf{U}\mathcal{S}_\lambda(\mathbf{\Sigma})\mathbf{V}^T$ is the singular value thresholding operator.

$$\begin{aligned}
\mathbf{L}_i&=\arg\min_{\mathbf{L}_i}\frac{1}{u}\|\mathbf{L}_i\|_*+\frac{1}{2}\left\|\mathbf{L}_i-\mathbf{D}_i+\mathbf{E}_i-\frac{\mathbf{Y}_i}{u}\right\|_F^2\\
&=\mathcal{J}_{\frac{1}{u}}\left(\mathbf{D}_i-\mathbf{E}_i+\frac{\mathbf{Y}_i}{u}\right) \qquad (9)\\
\mathbf{E}_i&=\arg\min_{\mathbf{E}_i}\frac{\lambda}{u}\|\mathbf{E}_i\|_1+\frac{1}{2}\left\|\mathbf{E}_i-\mathbf{D}_i+\mathbf{L}_i-\frac{\mathbf{Y}_i}{u}\right\|_F^2\\
&=\mathcal{S}_{\frac{\lambda}{u}}\left(\mathbf{D}_i-\mathbf{L}_i+\frac{\mathbf{Y}_i}{u}\right) \qquad (10)
\end{aligned}$$

Step 2: Update θ^k, $\forall k\in\{1,\ldots,K\}$ (with others fixed): The minimization problem (7) with respect to $\{\theta^1,\ldots,\theta^K\}$ can be decoupled into K independent subproblems, each of which corresponds to $\{\theta^k\}$ and can be equivalently formulated as the following integer constrained convex quadratic programming (QP) problem:

$$\begin{aligned}
\theta^k=&\arg\min_{\theta^k}\frac{1}{2}\theta^{k\top}\mathbf{G}^{k\top}\mathbf{G}^k\theta^k+\mathbf{e}_k^\top(\frac{Y}{u}-\mathbf{D})^\top\mathbf{G}^k\theta^k\\
\text{s.t.}&\theta^k\in\{0,1\}^{n_kn},\mathbf{Q}^k\theta^k=\mathbf{1}_n,\mathbf{H}^k\theta^k\leq\mathbf{1}_{n_k},\mathbf{S}^k\theta^k=\mathbf{1}_n^\top \qquad (11)
\end{aligned}$$

where $\mathbf{D}=\mathbf{L}+\mathbf{E}$. This is a NP-hard problem. However, as proved for a very similar problem in [19], $\theta^{k\top}\mathbf{G}^{k\top}\mathbf{G}^k\theta^k=\left\|\mathbf{G}^k\theta^k\right\|_2^2=\sum_{i=1}^n\left\|\mathbf{F}^k\mathbf{p}_i^k\right\|_2^2$ is a constant value if the features are normalized. Therefore, the quadratic term of problem (11) can be removed to get the linear programming problem (12). By relaxing the binary constraint to a real value between 0 and 1, the problem (12) can be exactly solved by a standard LP solver.

$$\begin{aligned}
\theta^k=&\arg\min_{\theta^k}\mathbf{e}_k^\top(\frac{Y}{u}-\mathbf{D})^\top\mathbf{G}^k\theta^k \qquad (12)\\
\text{s.t.}&\mathbf{0}_{nn_k}\leq\theta_k\leq\mathbf{1}_{nn_k},\mathbf{Q}^k\theta^k=\mathbf{1}_n,\mathbf{H}^k\theta^k\leq\mathbf{1}_{n_k},\mathbf{S}^k\theta^k=\mathbf{1}_n^\top
\end{aligned}$$

Step 3: Update Multiplier $\mathbf{Y}$: We update the Lagrange multipliers in Eq (13), where $\rho>1$.

$$\mathbf{Y}=\mathbf{Y}+u(\mathbf{D}-\mathbf{L}-\mathbf{E});\quad u=\rho u \qquad (13)$$

5. Experimental Results

Datasets: We evaluate tracking performance on 16 publicly available video sequences, which are captured in different scenarios and contain challenging appearance variations due to occlusion, object pose and scale changes, illumination change, and abrupt motion.

Implementation Details: In all experiments, the number of target templates is set to $K_1=5$ as in most of existing trackers. We set $K_2=3$, and $\lambda=1$ (in Eq (4)). We adopt Geometric Blur [6] as the feature to characterize each part. As a trade-off between effectiveness and speed, n and n_k are set to 6, 100, respectively. Here, we employ a simple heuristic to determine the number of parts (n) within the tracking object as in [30] - we divide the object into six parts of either 3×2 or 2×3 depending on its aspect ratio.

Baselines: Our PMT tracker is analyzed and compared with 10 state-of-the-art tracking methods, FRAGT (fragment-based tracker [1]), VTD (visual tracking decomposition [22]), L_1T (ℓ_1 minimization tracker [24]), IVT (incremental subspace visual tracker [26]), MIL (multiple instance learning tracker [4]), LRST (low-rank sparse tracker [33]), TLD (tracking-learning-detection [20]), CT (compressive tracking [31]), Struck (structured output tracker [16]), and OAB (online AdaBoost [14]). We implement these trackers using publicly available source codes or binaries provided by the authors. The parameters of these trackers are adjusted to show the best tracking performance. For fair comparisons, the same initializations are set to all methods. The supplementary material contains result videos.

Evaluation Metrics: For quantitative comparison, two popular evaluation metrics are used. The first metric is the

Table 1. The average center location errors of 11 trackers on 16 sequences. For each sequence, the smallest and second smallest distances are denoted in red and blue, respectively.

Video	PMT	CT	IVT	MIL	OAB	Frag	VTD	Struck	L_1T	LRST	TLD
tud	**8.5**	55.1	25.9	51.2	26.2	10.8	43.1	17.8	**10.3**	30.2	16.7
trellis	**14.1**	42.4	54.0	37.3	41.5	55.7	47.8	28.3	31.1	**26.5**	50.9
sylv	**4.6**	13.5	39.4	15.3	10.4	6.8	7.4	4.7	14.5	**4.5**	5.9
soccer	**15.4**	79.6	97.8	46.3	65.3	41.4	**10.5**	41.0	58.5	16.3	29.8
skating	**4.3**	84.9	74.9	49.2	39.3	63.3	**5.0**	51.9	20.1	**5.0**	99.3
singer	**3.7**	5.9	9.8	11.1	63.0	26.9	**4.4**	4.5	5.3	5.6	44.1
girl	**3.5**	17.4	**3.2**	12.4	11.0	7.4	11.4	18.6	5.0	4.0	8.3
face1	**6.2**	19.0	9.1	34.3	17.2	7.9	8.7	8.4	7.0	9.6	14.8
face2	**7.8**	24.0	8.3	10.2	20.8	48.2	11.8	**6.5**	15.2	8.1	13.3
david	**10.8**	32.4	15.6	30.3	26.4	73.0	64.9	46.7	16.2	16.0	**14.1**
coke11	**7.0**	11.1	58.5	13.7	11.3	71.0	62.7	**4.0**	12.1	9.6	11.6
car4	**3.4**	86.3	6.4	53.8	88.1	127.3	27.0	**4.3**	8.5	5.8	6.9
biker	**18.4**	**16.0**	76.8	29.6	22.0	104.4	17.3	48.0	29.4	47.7	86.9
osow	**1.8**	15.2	3.0	11.6	4.6	5.6	3.3	4.7	**2.0**	6.8	11.1
olsr2	**3.8**	56.8	24.0	23.8	12.5	57.6	44.3	14.3	4.7	38.1	49.5
olsr1	**2.7**	8.3	**2.9**	9.8	68.3	4.0	3.4	5.0	3.6	5.0	10.9

Table 2. The average overlap scores of 11 trackers on 16 sequences. For each sequence, the best and the second best scores are denoted in red and blue, respectively.

Video	PMT	CT	IVT	MIL	OAB	Frag	VTD	Struck	L_1T	LRST	TLD
tud	**0.85**	0.32	0.56	0.38	0.56	0.68	0.40	0.61	**0.81**	0.51	0.71
trellis	**0.52**	0.22	0.39	0.35	0.46	0.29	0.31	**0.50**	0.38	0.48	0.21
sylv	**0.78**	0.59	0.47	0.58	0.67	0.74	0.73	**0.76**	0.58	**0.78**	0.70
soccer	**0.28**	0.15	0.14	0.12	0.10	0.19	**0.35**	0.13	0.14	0.26	0.17
skating	**0.63**	0.01	0.07	0.23	0.37	0.19	**0.61**	0.29	0.47	0.59	0.07
singer	**0.78**	0.45	0.48	0.41	0.18	0.26	**0.66**	0.46	0.63	0.65	0.40
girl	**0.67**	0.32	**0.68**	0.45	0.53	0.60	0.55	0.41	0.64	0.65	0.59
face1	**0.89**	0.73	0.84	0.58	0.77	0.87	0.82	**0.85**	0.84	0.82	0.57
face2	**0.75**	0.54	0.79	0.72	0.59	0.38	0.70	**0.77**	0.67	0.74	0.57
david	**0.73**	0.41	0.36	0.42	0.43	0.23	0.26	0.38	0.50	0.50	**0.60**
coke11	**0.71**	0.47	0.10	0.43	0.41	0.06	0.06	**0.74**	0.46	0.72	0.45
car4	**0.82**	0.24	0.74	0.27	0.22	0.23	0.47	0.49	0.62	**0.80**	0.57
biker	**0.45**	**0.45**	0.31	0.43	0.44	0.27	**0.47**	0.38	0.39	0.42	0.30
osow	**0.94**	0.56	0.83	0.56	0.71	0.77	0.88	0.81	**0.91**	0.74	0.65
olsr2	**0.82**	0.29	0.44	0.35	0.47	0.27	0.34	0.50	**0.75**	0.31	0.28
olsr1	**0.88**	0.71	**0.86**	0.67	0.17	0.78	0.81	0.77	0.81	0.77	0.68

center location error which is the Euclidean distance between the central locations of the tracked targets and the manually labeled ground truth. The other is the Pascal VOC overlap score. Given the tracked bounding box ROI_T and the ground truth bounding box ROI_{GT}, the overlap score is computed as $score = \frac{area(ROI_T \cap ROI_{GT})}{area(ROI_T \cup ROI_{GT})}$.

5.1. Quantitative and Qualitative Evaluation

Table 1 and Table 2 report the average center location error and Pascal VOC overlap score of the 11 trackers on each of the 16 video sequences. Figure 2 plots the frame-by-frame center location errors (highlighted in different colors) obtained by the 11 trackers for the 4 of the 16 video sequences. Figure 2, Table 1, and Table 2 tell that our proposed PMT achieves the best tracking performance on most video sequences. In particular, PMT obtains more robust tracking results in the presence of complicated appearance changes caused by occlusion, drastic pose variation, background clutter, illumination change, and abrupt motion, etc.

Figure 3 shows qualitative tracking results of the 11 trackers over several representative frames of the 16 video sequences. For an example of occlusion in the "olsr2" sequence, tracking of the woman is lost by all other trackers at frame 200 as she is partially occluded by a man. The other trackers lock onto the man, so their errors increase for the rest of the sequence, as shown in Figure 3. Another example is the "tud" sequence, where the target vehicle is occluded by crossing pedestrians. The MIL, VTD, OAB, and CT methods drift away from the target object when occlusion occurs. On the other hand, the L_1T, TLD, and our PMT methods perform well. In the other sequences with occlusion, such as, "osow", "faceocc", "coke11", "faceocc2", the proposed PMT performs at least the second best. The "car4", "car11", and "sylv" video sequences contain illumination changes. Take "car4" as an example, the OAB, Frag, and VTD methods start to drift from the target at frame 185 when the vehicle goes through the overpass. The MIL and CT algorithms start drift away from the target object at frame 210. The L_1 and TLD approaches are able to track the target although with some errors. On the other hand, the target object is successfully tracked by our PMT and Struck algorithms throughout the entire sequence despite large illumination changes. The "david", "singer", and "trellis" contains significant illumination changes and pose variations. On the "trellis" sequence, Frag, and VTD begin to drift away from the target after frame 172 because of the changing lighting conditions. Due to the combination of lighting and head pose changes, IVT, Frag, and C-T fail to track the target after the 367^{th} frame. Both our tracker and Struck successfully track the target across the whole video sequence, although our tracker locates the head more accurately. The sequences "girl" and "skating" contain abrupt motion, pose change, and partial occlusion. On the "girl" sequence, the proposed tracker are capable of tracking the target for the entire sequence. Other trackers experience drift at different time. The "soccer" sequence contains abrupt motion and background clutter. Compared with other trackers, PMT achieves the better results and can track the target object despite scale and pose changes as well as occlusion by confetti at most of the frames. In contrast, other methods (IVT, L_1 , OAB, MIL, and Frag) fail to track the target reliably. The "biker" sequence contains scenes with abrupt motion and large pose change. Nevertheless, our PMT performs well throughout the entire sequence with more stable tracking results.

5.2. Discussion and Analysis

We present more illustrative tracking examples in this section to demonstrate the effectiveness of PMT for robust visual tracking. In particular, Figure 4 demonstrates a process of partial occlusion, where for each of the three sets of incoming frames to be tracked (around frames 248, 266, 295), we only show one of the target templates on the left and one of the tracked frames on the right, due to space limit. When the partial occlusion starts at the incoming frame 248, PMT still matches its parts to those of the target template shown on the top-left image. However, due to partial occlusion, the three parts on the left of the face in the top-right image rank higher in terms of matching confidence. Appearance of the partially occluded face is very different from that in the target templates, PMT thus updates the target template as shown in the middle-left image. After updat-

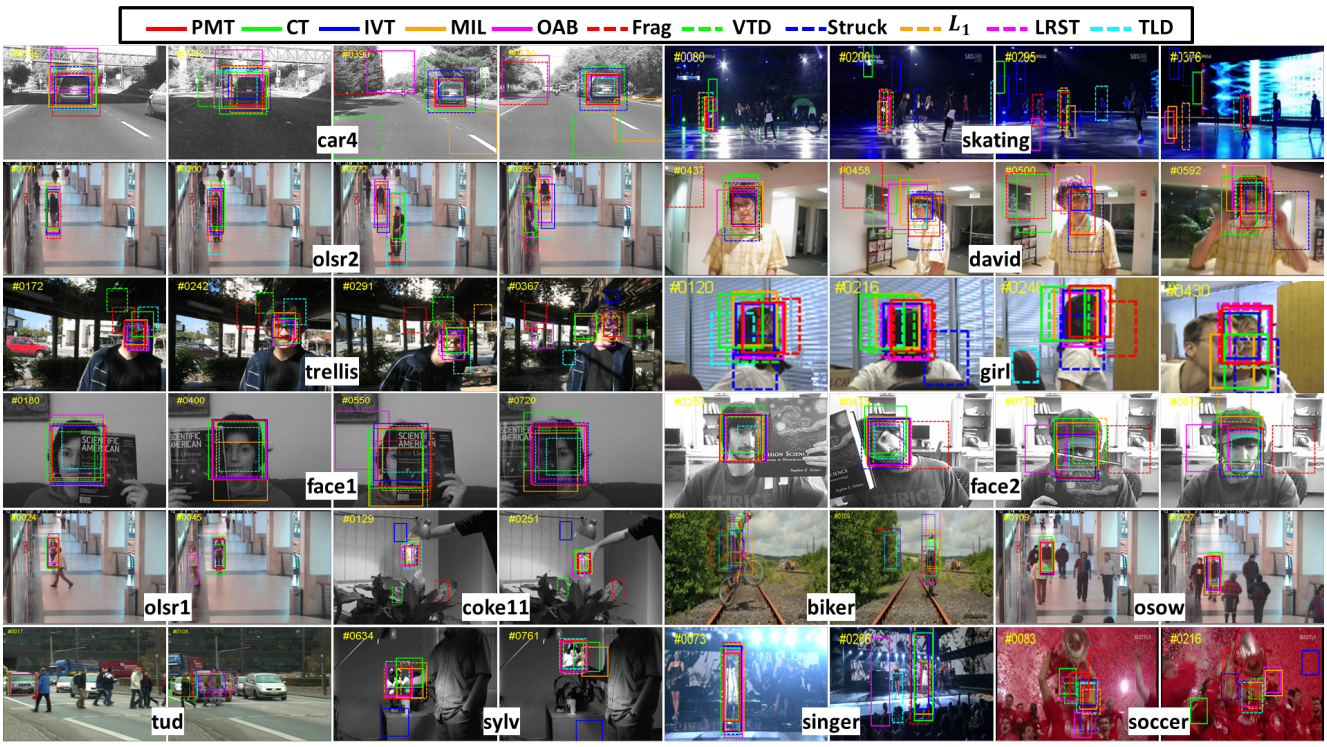


Figure 3. Tracking results of 11 trackers (denoted in different colors) on 16 video sequences. Frame numbers are overlayed in yellow. See text for details. Results best viewed on high-resolution displays.

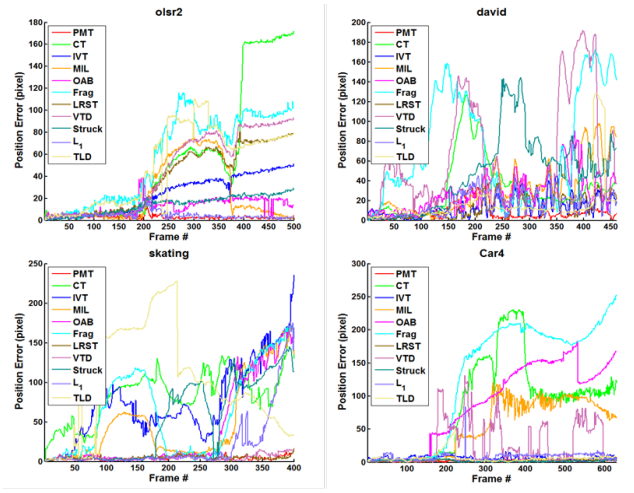


Figure 2. Quantitative comparison of the 11 trackers with the center location error on the 4 video sequences.

ing, most of the six parts at frame 266 match well with the updated target template, as shown by the confidence scores in the middle-right image. This update of target template is important when partial occlusion remains for a longer duration of time, otherwise matching may fail and tracking drifts. When the face re-appears at frame 295, PMT matches its parts to those of earlier target templates that contain no occlusion, as shown in the bottom image of Figure 4. The tracking process continues successfully, which shows

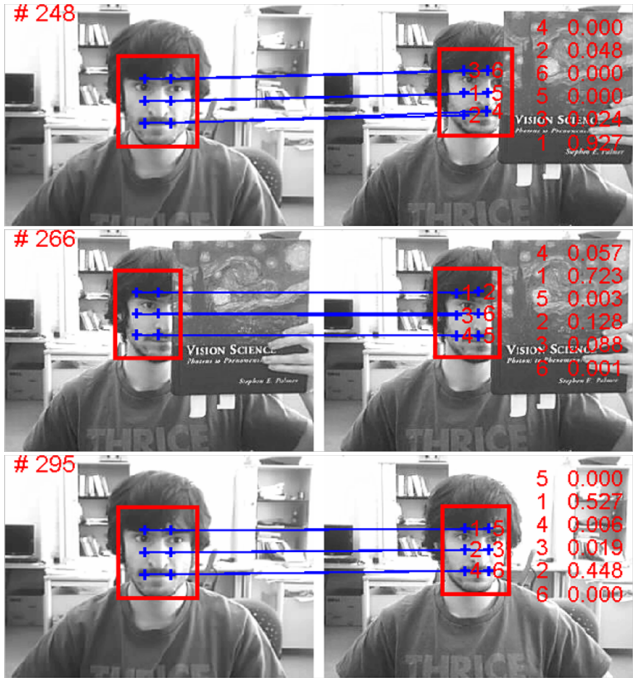


Figure 4. Illustration of PMT's robustness against partial occlusion. The numbers of "1" to "6" index different parts of the face. "1" ranks highest and "6" ranks lowest in terms of confidence score of part matching. *More explanations are in Section 5.2.*

the effectiveness of our proposed PMT.

Figure 5. Our PMT can track well even when there are errors of part matching due to occlusion, illumination changes, etc.

Our PMT seems less prone to noise, part matching error, or partial occlusion. This is further demonstrated in Figure 5, where when some of the parts cannot be matched to those in the target templates, other parts of the object are less influenced and their matchings still make tracking successful. This is consistent with confidence scores of the matching of different parts shown in the right of Figure 5.

6. Conclusion

In this paper, we proposed a locality-constrained low-rank sparse learning method to effectively optimize optimal partial permutation matrices for the part correspondence among multiple frames for visual tracking. By using the three properties (locality-constrained property, low-rank sparse property, and globally consistent property), our tracker is robust for partial occlusion. We extensively analyze the performance of our tracker on challenging real-world video sequences and show it outperforms 10 state-of-the-art tracking methods.

Acknowledgment

This study is supported by the research grant for the Human Sixth Sense Programme at the Advanced Digital Sciences Center from Singapore's Agency for Science, Technology and Research (A*STAR). C. Xu is supported by 973 Program Project No. 2012CB316304 and NSFC 61225009.

References

[1] A. Adam, E. Rivlin, and I. Shimshoni. Robust fragments-based tracking using the integral histogram. In *CVPR*, pages 798–805, 2006.

[2] M. S. Arulampalam, S. Maskell, N. J. Gordon, and T. Clapp. A tutorial on particle filters for online nonlinear/non-gaussian bayesian tracking. *IEEE Transactions on Signal Processing*, 50(2):174–188, 2002.

[3] S. Avidan. Ensemble tracking. In *CVPR*, pages 494–501, 2005.

[4] B. Babenko, M.-H. Yang, and S. Belongie. Visual tracking with online multiple instance learning. In *CVPR*, 2009.

[5] C. Bao, Y. Wu, H. Ling, and H. Ji. Real time robust l_1 tracker using accelerated proximal gradient approach. In *CVPR*, 2012.

[6] A. C. Berg and J. Malik. Geometric blur for template matching. In *CVPR*, 2001.

[7] M. J. Black and A. D. Jepson. Eigentracking: Robust matching and tracking of articulated objects using a view-based representation. *IJCV*, pages 63–84, 1998.

[8] S. Boyd, N. Parikh, E. Chu, and J. Peleato, B. andEcjstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. In *Foundations and Trends in Machine Learning*, 2010.

[9] L. Cehovin, M. Kristan, and A. Leonardis. An adaptive coupled-layer visual model for robust visual tracking. In *ICCV*, 2011.

[10] P. Chockalingam, N. Pradeep, and S. Birchfield. Adaptive fragments-based tracking of non-rigid objects using level sets. In *ICCV*, 2009.

[11] P. F. Felzenszwalb, R. B. Girshick, D. A. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE Trans. Pattern Anal. Mach. Intell.*, 32(9):1627–1645, 2010.

[12] V. Gay-Bellile, A. Bartoli, and P. Sayd. Direct estimation of nonrigid registrations with image-based self-occlusion reasoning. *IEEE Trans. PAMI*, 32(1):87–104, 2010.

[13] M. Godec, P. M. Roth, and H. Bischof. Hough-based tracking of non-rigid objects. In *ICCV*, 2011.

[14] H. Grabner, M. Grabner, and H. Bischof. Real-Time Tracking via On-line Boosting. In *BMVC*, 2006.

[15] B. Han and L. Davis. On-line density-based appearance modeling for object tracking. In *ICCV*, 2005.

[16] S. Hare, A. Saffari, and P. H. Torr. Struck: Structured output tracking with kernels. In *ICCV*, 2011.

[17] G. Hua and Y. Wu. Measurement integration under inconsistency for robust tracking. In *CVPR*, 2006.

[18] H. Izadinia, I. Saleemi, W. Li, and M. Shah. $(mp)^2t$: Multiple people multiple parts tracker. In *ECCV*, 2012.

[19] K. Jia, T.-H. Chan, Z. Zeng, G. Wang, T. Zhang, and Y. Ma. ROML: A Robust Feature Correspondence Approach for Matching Objects in A Set of Images. *submitted to International Journal of Computer Vision*.

[20] Z. Kalal, K. Mikolajczyk, and J. Matas. Tracking-learning-detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(7):1409–1422, 2012.

[21] S. Kwak, W. Nam, B. Han, and J. H. Han. Learning occlusion with likelihoods for visual tracking. In *ICCV*, 2011.

[22] J. Kwon and K. M. Lee. Visual tracking decomposition. In *CVPR*, 2010.

[23] X. Li, W. Hu, Z. Zhang, X. Zhang, M. Zhu, and J. Cheng. Visual tracking via incremental log-euclidean riemannian subspace learning. In *CVPR*, 2008.

[24] X. Mei and H. Ling. Robust Visual Tracking and Vehicle Classification via Sparse Representation. *TPAMI*, 33(11):2259–2272, 2011.

[25] S. M. S. Nejhum, J. Ho, and M.-H. Yang. Visual tracking with histograms and articulating blocks. In *CVPR*, 2008.

[26] D. Ross, J. Lim, R.-S. Lin, and M.-H. Yang. Incremental Learning for Robust Visual Tracking. *International Journal of Computer Vision*, 77(1):125–141, 2008.

[27] G. Shu, A. Dehghan, O. Oreifej, E. Hand, and M. Shah. Part-based multiple-person tracking with partial occlusion handling. In *CVPR*, 2012.

[28] G. Wang, D. Forsyth, and D. Hoiem. Improved object categorization and detection using comparative object similarity. *TPAMI*, 35(10):2442–2453, 2013.

[29] B. Yang and R. Nevatia. Online learned discriminative part-based appearance models for multi-human tracking. In *ECCV*, 2012.

[30] R. Yao, Q. Shi, C. Shen, Y. Zhang, and A. van den Hengel. Part-based visual tracking with online latent structural learning. In *CVPR*, 2013.

[31] K. Zhang, L. Zhang, and M.-H. Yang. Real-time compressive tracking. In *ECCV*, 2012.

[32] T. Zhang, B. Ghanem, and N. Ahuja. Robust multi-object tracking via cross-domain contextual information for sports video analysis. In *ICASSP*, 2012.

[33] T. Zhang, B. Ghanem, S. Liu, and N. Ahuja. Low-rank sparse learning for robust visual tracking. In *ECCV*, 2012.

[34] T. Zhang, B. Ghanem, S. Liu, and N. Ahuja. Robust visual tracking via multi-task sparse learning. In *CVPR*, 2012.

[35] T. Zhang, B. Ghanem, S. Liu, and N. Ahuja. Robust visual tracking via structured multi-task sparse learning. *International Journal of Computer Vision*, 101(2):367–383, 2013.