Pattern Recognition 88 (2019) 75–89

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/patcog

Robust visual tracking via nonlocal regularized multi-view sparse representation

Bin Kang[a], Wei-Ping Zhu[b], Dong Liang[c,*], Mingkai Chen[d]

[a] College of Internet of Things, Nanjing University of Posts and Telecommunications, Nanjing 210003, China
[b] Department of Electrical and Computer Engineering, Concordia University, Montreal, Quebec H3G 1M8, Canada
[c] College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, China
[d] Key Lab of Broadband Wireless Communication and Sensor Network Technology, Nanjing University of Posts and Telecommunications, Nanjing 210003, China

ARTICLE INFO

Article history:
Received 20 March 2018
Revised 7 September 2018
Accepted 9 November 2018
Available online 10 November 2018

Keywords:
Sparse representation
Visual tracking
Multi-view learning
Dual group structure

ABSTRACT

The multi-view sparse representation based visual tracking has attracted increasing attention because the sparse representations of different object features can complement with each other. Since the robustness of different object features is actually not the same in challenging video sequences, it may contain unreliable features (the features with low robustness) in multi-view sparse representation. In this case, how to highlight the useful information of unreliable features for proper multi-feature fusion has become a tough work. To solve this problem, we propose a multi-view discriminant sparse representation method for robust visual tracking, in which we firstly divide the multi-view observations into different groups, and then estimate the sparse representations of multi-view group projections for calculating the observation likelihood. The advantages of the proposed sparse representation method are two-folds: 1) It can properly fuse the observation groups with reliable and unreliable features by using an online updated discriminant matrix to explore the group similarity in multi-feature space. 2) It introduces a nonlocal regularizer to enforce the spatial smoothness among the sparse representations of different group projections, which can enhance the robustness of multi-view sparse representation. Experimental results show that our method can achieve a better tracking performance than state-of-the-art tracking methods do.

1. Introduction

With the rapid development of multimedia and internet of things [1–3], there is a pressing demand for intelligent video technology such as visual tracking. A typical tracking algorithm includes a motion model and an observation model. The motion model aims to track the state of moving target, and the observation model evaluates the likelihood of each target observation to select the best one for the current frame. Designing the observation model is a piece of tough work in visual tracking because the target appearance often changes dramatically under occlusion, background clutter or illumination change etc. To overcome those challenges, lots of works have been done recently. According to different observation models, existing visual tracking algorithms can be categorized into discriminative trackers and generative trackers. The discriminative trackers cast the target tracking as a binary classification problem to distinguish the tracked target from the video background. The state-of-the-art methods on discriminative trackers include support vector machine based methods [4,5], online boosting [6–8], multiple instance learning based methods [9,10], compressed tracker[11] and correlation filter based methods [12,13] etc. The generative trackers typically search for an image region that best matches the object appearance. Recent efforts in this domain include subspace learning based tracking [14–16], matrix decomposition based tracking [17–19] and sparse representation based tracking [20] etc. Besides aforementioned observation models, the deep leaning based trackers [21–23] have attracted more attention due to the ability of nonlinear representation. The tracking performance of those methods often relies on a tedious off-line pre-training with tremendous amount of labeled training samples, thus the performance is sensitive to the choice of training samples and tends to be overfitting in the presence of label noise. In real world visual tracking, we may have a small number of labeled training samples or even only have non-labeled samples. In this case, how to achieve a robust visual tracking is worth giving the careful consideration.

Among existing generative trackers, sparse representation based visual tracking is the one that can use non-labeled samples to

* Corresponding author.
E-mail address: liangdong@nuaa.edu.cn (D. Liang).

https://doi.org/10.1016/j.patcog.2018.11.005
0031-3203/

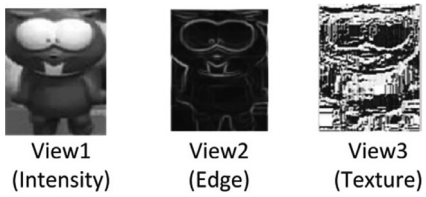

Fig. 1. An example of multi-views in visual tracking.

achieve visual tracking. Using sparse representation for visual tracking was first proposed by Mei [24], where the likelihood of target observation was evaluated through solving a series of regularized least square problems. Since this algorithm estimates the sparse representations of different particle observations separately, it ignores the particle relationships and makes the tracker prone to drift away. Although a lot of works [25–30] have been done to improve the performance of Mei's algorithm, those trackers may drift away from the target in long term video sequences because they only use pixel intensity to model the target appearance. The pixel intensity is robust to particle occlusion but sensitive to the shape deformation of moving target and illumination change. In computer vision, multi-view refers to different feature subsets used to represent particular characteristics of an object (see Fig. 1). Based on this concept, Hong et al. [31] proposed a multi-view based multi-task sparse representation method for visual tracking, in which different features can complement with each other to give better tracking performance as compared to single feature based tracking methods. The method in [31] was derived based on the assumption that all the features can work well in visual tracking. However, it may not be valid in the video sequences with severe occlusion because some feature observations, such as texture, are prone to be disturbed by occlusion or video noise. In fact, the robustness of a moving object feature can be varied by different kinds of appearance variations. For example, the histogram is robust to local distortion, but sensitive to background clutter. Those features with low robustness can be regarded as unreliable features due to the fact that they cannot be well represented by the corresponding feature dictionary. Fusing unreliable feature with high sparse representation error may degrade the tracking performance in challenging video sequences. Similar to Hong's algorithm, Hu et al. [32] also used multi-task multi-view sparse representation to model the target appearance. Since this algorithm could not discriminate the reliable and unreliable features during sparse representation, it may reduce the robustness of sparse representation results. To overcome the limitations in [31] and [32], Lan [33] proposed a multi-view based method to adaptively detect unreliable features and remove them during sparse representation. In fact, unreliable feature contains useful complementary information, and if used properly, it would enhance the tracking performance.

As aforementioned introduction, the key point in multi-view sparse representation based visual tracking is to properly fuse multi-view observations during sparse representation, which is a piece of tough work due to the following two challenges: (1) the unreliable views may disturb the fusing results, (2) it is clearly shown in Fig. 1 that there exist not only the potential similarity but also a large gap between different kinds of views. Exploiting the so called potential similarity can facilitate multi-view fusing. However, how to explore this similarity under multi-view gap is still an open problem.

Existing works such as Lan et al. [33] only focus on reducing the negative effect of unreliable views. As far as we know, there are few works that can simultaneously overcome two challenges in multi-view fusing. In this paper, we propose a multi-view discriminant learning based sparse representation method for robust visual tracking. Different from traditional multi-view sparse representation based tracking methods that directly use the sparse representations of multi-view observations to calculate the observation likelihood, our method firstly divides the multi-view observations into different groups, and then estimates the sparse representations of multi-view group projections for calculating the observation likelihood. Since the correlations between different observations of each view can be varied by the appearance variation, some observations may be very similar [34]. Dividing the multi-view observations into different groups and introducing group projections in sparse representation enable us to use multi-view learning to simultaneously exploit the group similarity in the same and different views, which can avoid the uncorrelated observation destroying the common sparsity and highlight the useful information in the unreliable observation groups (the observation groups with unreliable views).

The main contributions of this paper are summarized as follows:

1) We first propose a multi-view discriminant learning based sparse representation method to explore group similarity in the multi-feature space, which is then incorporated into a particle filter based framework to achieve robust visual tracking. The proposed method makes use of unreliable observation groups to achieve multi-view fusion and makes different observation groups more group discriminative.
2) In our sparse representation method, we propose a nonlocal regularizer to guarantee a robust tracking performance in severe object occlusion, pose variation etc. The nonlocal regularizer can simultaneously exploit both local and nonlocal relations among the sparse representations of group projections, enhancing the inherent consensus in different views.
3) We propose an adaptive alternating direction algorithm to solve the optimization problem involved in the proposed sparse representation method. The new reconstruction method can adaptively update the penalty parameter to achieve a fast convergence.

It is worth mentioning that in our previous work [35], the multi-view discriminant learning is introduced in the sparse representation model for the first time. The main differences between this paper and [35] are summarized as follows: Firstly, the sparse representation method in[35] only uses $l_{2,1}$ norm to constrain the sparse representation result, which may reduce the robustness of sparse representation because the reliable and unreliable view observations may not share the common sparse pattern when facing severe appearance variation. In this paper, we propose a nonlocal regularizer to enforce spatial smoothness among the sparse representations of different group projections, which can eliminate the negative effect caused by the sparse representations of the unreliable observations. Secondly, introducing the nonlocal regularizer in the multi-view sparse representation makes the optimization problem more complex. The reconstruction method in [35] cannot be directly used to solve this optimization problem. Here, we propose an adaptive alternating direction algorithm to solve this problem with fast convergence. Finally, we theoretically analyze the convergence of the proposed reconstruction method and increase the number of testing sequences for a thorough evaluation of the proposed tracking method.

This paper is organized as follows: in Section 2, we discuss the key problem in designing the sparse representation model. Section 3 illustrates our proposed sparse representation model in detail. Section 4 introduces how to use the proposed sparse representation model to achieve visual tracking. Experimental results and conclusions are presented in Sections 5 and 6, respectively.

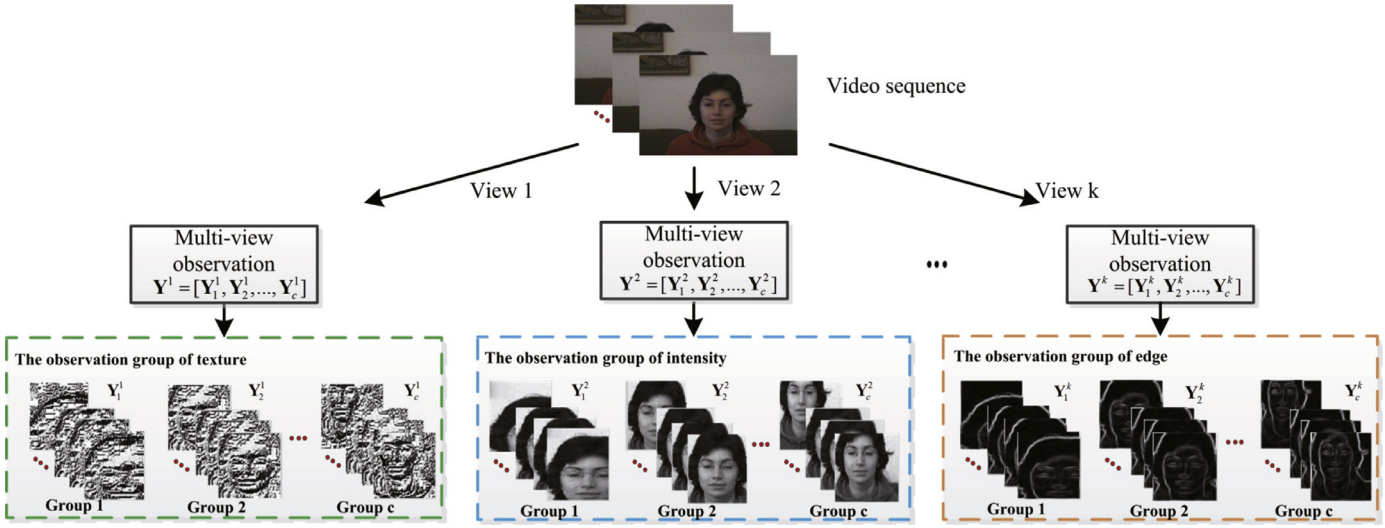

Fig. 2. The observation group similarity in different views.

2. Problem formulation

In this paper, multi-view refers to multiple features, e.g. color, shape and texture, that are used to represent a moving target. The views which do not work well during sparse representation are regarded as unreliable views. The multi-view sparse representation based visual tracking aims to use the sparse representation results to estimate the likelihood of multi-view observations. In this method, the tracking performance relies on the design of the sparse representation model. Here, suppose $\mathbf{Y}^k = [\mathbf{y}_1^k, \mathbf{y}_2^k, \ldots, \mathbf{y}_n^k]$ $(k = 1, 2, \ldots, m)$ denotes the observation matrix in the k-th view, where each column $\mathbf{y}_i^k$ means the i-th observation vector. Traditional multi-view based sparse representation model can be described as [32]

$$\min_{\mathbf{\Theta}} \frac{1}{2} \sum_{k=1}^{m} \|\mathbf{Y}^k - \mathbf{A}^k \mathbf{\Theta}^k\|_F^2 + \lambda \|\mathbf{\Theta}\|_{2,1}, \tag{1}$$

where $\mathbf{A}^k$ denotes the target template matrix in the k-th view. Problem (1) is aimed to seek the sparse representation matrix of $\mathbf{Y}^k$. The drawbacks of problem (1) are two-folds: (1) It cannot discriminate the contributions of multi-view observations because it uses the same weight for the sparse representation errors of different view observations. The unreliable views may give a high sparse representation error, thus causing a tracking drift in challenging video sequences. (2) It assumes that the columns in $\mathbf{Y}^k$ are highly correlated, hence the sparsity in $\mathbf{\Theta}$ is constrained by $l_{2,1}$ norm. This assumption may not be valid in challenging video sequences because the vectors in $\mathbf{Y}^k$ are easily disturbed by appearance variation. If some vectors are disturbed seriously, they are not highly correlated with adjacent vectors. In this case, only using $l_{2,1}$ norm to constrain $\mathbf{\Theta}$ may give a poor sparse representation result.

Different from [32], we firstly divide n vectors in $\mathbf{Y}^k$ into c groups, i.e. $\mathbf{Y}^k = [\mathbf{y}_1^k, \mathbf{y}_2^k, \ldots, \mathbf{y}_n^k] = [\mathbf{Y}_1^k, \mathbf{Y}_2^k, \ldots, \mathbf{Y}_c^k]$, and then estimate the sparse representations of multi-view observation groups jointly. Since the correlation between different vectors in matrix $\mathbf{Y}^k$ can be varied by appearance variation, dividing $\mathbf{Y}^k$ into c groups enables us to explore the common sparsity according to the difference in vector correlation. The key to our sparse representation method is to exploit the group similarity during the sparse representation for properly fusing the reliable and unreliable observation groups. The group similarity is shown in Fig. 2. We can see that the observation groups not only have intra-view similarity, but also have inter-view similarity. The intra-view similarity means that the observations in the same group and the same view are highly correlated, while the inter-view similarity is that the same observation groups in different views contain inherent correlation. Due to the large gap between different views [55], directly exploiting the aforementioned group similarity is no longer applicable. Inspired by multi-view discriminant analysis [36], we use a discriminant matrix to project multi-view observation groups into a latent common space in which the between-group variations from both inter-view and intra-view are maximized, while the within-group variations from both inter-view and intra-view are minimized. In this case, the within-group similarity in the unreliable view can be enhanced, which would highlight the useful information in the unreliable observation groups. The multi-view group projections are denoted as $(\mathbf{P}^k)^T \mathbf{Y}_i^k$, where $\mathbf{P}^k$ is the learned discriminant matrix for the k-th view, $\mathbf{Y}_i^k$ $(\mathbf{Y}_i^k \subset \mathbf{Y}^k)$ denotes the i-th observation group in the k-th view. With this consideration, we form the following sparse representation based optimization problem,

$$\min_{\mathbf{\Theta}} \sum_{i=1}^{c} \sum_{k=1}^{m} \frac{1}{2} \|(\mathbf{P}^k)^T \mathbf{Y}_i^k - ((\mathbf{P}^k)^T \mathbf{A}^k) \mathbf{\Theta}_i^k\|_F^2 + \lambda_1 \|\mathbf{\Theta}\|_{2,1}, \tag{2}$$

where $\mathbf{\Theta} = [\mathbf{\Theta}_1^1, \ldots, \mathbf{\Theta}_i^k, \ldots, \mathbf{\Theta}_c^m]$, with $\mathbf{\Theta}_i^k = [[\mathbf{\Theta}_i^k]_1, [\mathbf{\Theta}_i^k]_2, \ldots, [\mathbf{\Theta}_i^k]_r]$ $(k = 1, 2, \ldots, m; i = 1, 2, \ldots, c)$ denoting the sparse representation result of the i-th group projection in the k-th view, and $[\mathbf{\Theta}_i^k]_j$ $(j = 1, 2, \ldots, r)$ being the j-th vector in matrix $\mathbf{\Theta}_i^k$. Problem (2) aims to estimate the sparse representation matrix of $(\mathbf{P}^k)^T \mathbf{Y}_i^k$. Compared with (1), problem (2) can obviously reduce the large sparse representation error caused by unreliable observations because $(\mathbf{P}^k)^T \mathbf{Y}_i^k$ can maximize the common information and minimize the disturbance in multi-view observation groups.

Inspired by [36], to learn the discriminant matrix $\mathbf{P}^k$ $(k = 1, 2, \ldots, m)$, the between-group variation from all views should be maximized while the within-group variation from all views should be minimized. This means that the trace of within-group scatter matrix $\mathbf{P}^T \mathbf{S} \mathbf{P}$ should be as small as possible. Meanwhile, the trace of between-group scatter matrix $\mathbf{P}^T \mathbf{D} \mathbf{P}$ should be as large as possible. Based on this observation, the discriminant matrix is learned by solving following problem:

$$\min_{\mathbf{P}} Tr(\mathbf{P}^T (\mathbf{S} - \mathbf{D}) \mathbf{P}), \tag{3}$$

where $\mathbf{P} = [(\mathbf{P}^1)^T, (\mathbf{P}^2)^T, \ldots, (\mathbf{P}^m)^T]^T$ with $\mathbf{P}^k$ denoting the discriminant matrix for the particle observations in the k-th view,

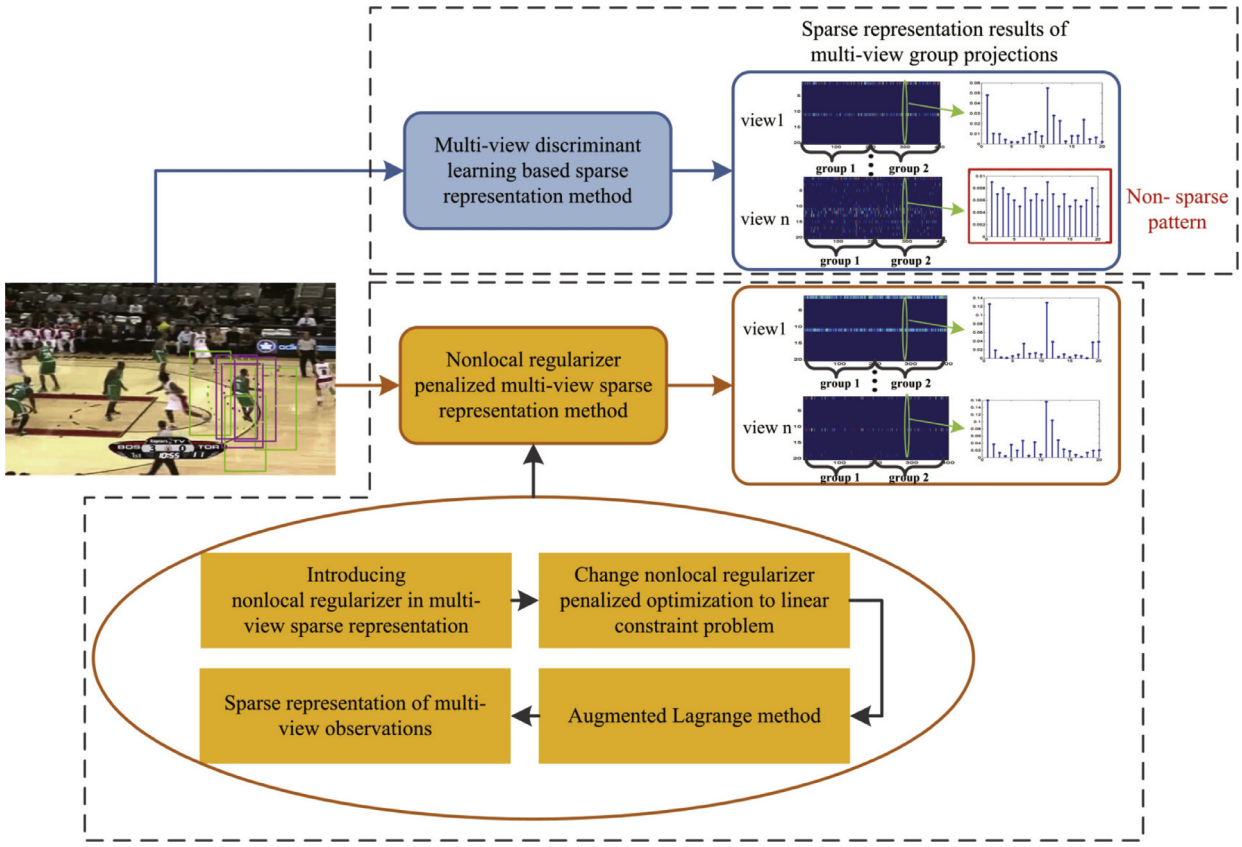

Fig. 3. Illustrate the difference between the multi-view discriminant learning based sparse representation method (Eq. (4)) and the nonlocal regularizer penalized multi-view sparse representation method (Eq. (5)). The sparse representation results of Eq. (4) may have non-sparse pattern, which will cause tracking drift in challenging video sequences. The detail of the ellipse is presented in Sections 3.1 and 3.2 . Main contributions in the proposed sparse representation method are highlighted with yellow boxes.

matrices $\mathbf{S}$ and $\mathbf{D}$ are two parameter matrices, which are used to calculate the within-group variation and the between-group variation, respectively. Here, we use the particle observations at the first frame as the training samples for calculating matrices $\mathbf{D}$ and $\mathbf{S}$ in a manner similar to that in [36].

Based on (2) and (3), the proposed multi-view discriminant learning based sparse representation method is formulated as

$$\min_{\mathbf{\Theta},\mathbf{P}} \sum_{i=1}^{c}\sum_{k=1}^{m}\frac{1}{2}\|(\mathbf{P}^k)^T\mathbf{Y}_i^k-((\mathbf{P}^k)^T\mathbf{A}^k)\mathbf{\Theta}_i^k\|_F^2+\lambda_1\|\mathbf{\Theta}\|_{2,1} +\lambda_2 Tr(\mathbf{P}^T(\mathbf{S}-\mathbf{D})\mathbf{P}). \quad (4)$$

Problem (4) integrates multi-view learning and sparse representation into a unified optimization model, in which, we can simultaneously achieve sparse representation and update the discriminant matrices. The matrix $\mathbf{P}^k$ is updated to explore the potential commonality between reliable and unreliable observation groups, making $(\mathbf{P}^k)^T\mathbf{Y}_i^k$ more group-discriminative in the latent common space. In this case we can properly fuse multi-view group projections when estimating the sparse representation of $(\mathbf{P}^k)^T\mathbf{Y}_i^k$. In (4), $(\mathbf{P}^k)^T\mathbf{A}^k$ highlights the potential commonality in multi-view template matrices.

3. The nonlocal regularizer penalized multi-view sparse representation

In (4), the sparse representations of multi-view group projections are arranged together to form $\mathbf{\Theta}$. As shown in Fig. 3, the multi-view discriminant learning based sparse representation method (Eq. (4)) may not guarantee the sparsity of multi-view sparse representation in challenging video sequences because it only uses $l_{2,1}$ norm to constrain the sparsity of $\mathbf{\Theta}$, making the sparse representations of unreliable observation groups may not share the same sparse pattern with that of reliable observation groups. To enforce the common sparsity in $\mathbf{\Theta}$, we propose to use a nonlocal regularizer in multi-view discriminant learning based sparse representation. The proposed regularizer can exploit the inherent similarity in the sparse representations of different group projections. Thus, we can enforce the spatial smoothness among the multi-view sparse representation results. The nonlocal regularizer penalized multi-view sparse representation is then formulated as

$$\min_{\mathbf{\Theta},\mathbf{P}} \sum_{i=1}^{c}\sum_{k=1}^{m}\frac{1}{2}\|(\mathbf{P}^k)^T\mathbf{Y}_i^k-((\mathbf{P}^k)^T\mathbf{A}^k)\mathbf{\Theta}_i^k\|_F^2+\lambda_1 g(\mathbf{\Theta}) +\lambda_2\|\mathbf{\Theta}\|_{2,1}+\lambda_3 Tr(\mathbf{P}^T(\mathbf{S}-\mathbf{D})\mathbf{P}), \quad (5)$$

where $g(\mathbf{\Theta})$ is the nonlocal regularizer, which is employed to enforce spatial smoothness among the sparse representations of different group projections. The concrete expression of $g(\mathbf{\Theta})$ will be derived in the next section.

3.1. The nonlocal regularizer

In (5), $\mathbf{\Theta}=[\mathbf{\Theta}_1^1,\ldots,\mathbf{\Theta}_i^k,\ldots,\mathbf{\Theta}_c^m]$, where $\mathbf{\Theta}_i^k=[[\mathbf{\Theta}_i^k]_1,[\mathbf{\Theta}_i^k]_2,\ldots,[\mathbf{\Theta}_i^k]_r]$ $(i=1,2,\ldots,c;\ k=1,2,\ldots,m)$. For notational simplicity, in this subsection, $\mathbf{\Theta}$ is rewritten as $\mathbf{\Theta}=[\boldsymbol{\theta}_1,\boldsymbol{\theta}_2,\ldots,\boldsymbol{\theta}_{mcr}]$, where $\boldsymbol{\theta}_i$ denotes the i-th vector in matrix $\mathbf{\Theta}$. The proposed regularizer $g(\mathbf{\Theta})$ is defined as

$$g(\mathbf{\Theta})=\sum_{\boldsymbol{\theta}_i\in\mathbf{\Theta}}\sum_{\boldsymbol{\theta}_j\in N_\theta}\phi(\|P(\boldsymbol{\theta}_i)-P(\boldsymbol{\theta}_j)\|_F), \quad (6)$$

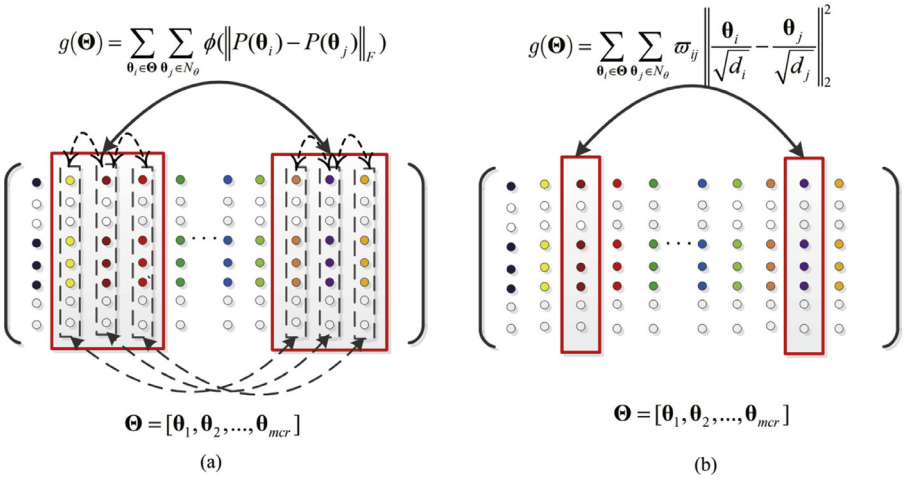

Fig. 4. The graph regularizer versus the proposed nonlocal regularizer. (a) The proposed nonlocal regularizer, (b) The graph regularizer.

where $\phi(\cdot)$ is the robust distance operator, which, for a scalar x, is defined as $\phi(x)=\sigma(1-e^{-\frac{x^2}{\sigma}})$, N_θ denotes the searching window, $P(\boldsymbol{\theta}_i)$ (or $P(\boldsymbol{\theta}_j)$) is an operator which is introduced to select adjacent elements centered at $\boldsymbol{\theta}_i$ i.e. $P(\boldsymbol{\theta}_i)=[\boldsymbol{\theta}_{i-u},\ldots,\boldsymbol{\theta}_{i-1},\boldsymbol{\theta}_i,\boldsymbol{\theta}_{i+1},\ldots,\boldsymbol{\theta}_{i+u}]$. Different from graph regularizer [26], here we introduce the operator $P(\boldsymbol{\theta}_i)$ in (6) for solving the MMV based inverse problem, which can make $g(\boldsymbol{\Theta})$ not only exploit the nonlocal similarity of $\boldsymbol{\Theta}$, but also consider the local interdependence in adjacent sparse representation results. The intuitive difference between the graph regularizer and our nonlocal regularizer is shown in Fig. 4. We can see that our regularizer adds the robustness of the sparse representation because it exploits the relationship between a group of adjacent vectors in $\boldsymbol{\theta}_i$ and the corresponding vectors in $\boldsymbol{\theta}_j$. On the other hand, the graph regularizer can only exploit the relationship between vector $\boldsymbol{\theta}_i$ and $\boldsymbol{\theta}_j$. Note that calculating $g(\boldsymbol{\Theta})$ in (6) is NP-hard due to its nonconvex nature. Inspired by [37], we use Majorize Minimize (MM) algorithm [38] to simplify (6). First, we have

$$\phi(\|P(\boldsymbol{\theta}_i)-P(\boldsymbol{\theta}_j)\|_F)\le s(i,j)\|P(\boldsymbol{\theta}_i)-P(\boldsymbol{\theta}_j)\|_F^2+b, \tag{7}$$

where

$$s(i,j)=\frac{\phi'(\|P(\boldsymbol{\theta}_i)-P(\boldsymbol{\theta}_j)\|_F)}{2\|P(\boldsymbol{\theta}_i)-P(\boldsymbol{\theta}_j)\|_F} \tag{8}$$

is a nonlinear function for measuring the similarity between $\boldsymbol{\theta}_i$ and $\boldsymbol{\theta}_j$. As parameter b in (7) is a constant, we can ignore it in the optimization process. Taking (7) into (6), we can obtain

$$g(\boldsymbol{\Theta})\le\sum_{\boldsymbol{\theta}_i\in\boldsymbol{\Theta}}\sum_{\boldsymbol{\theta}_j\in N_\theta}s(i,j)\|P(\boldsymbol{\theta}_i)-P(\boldsymbol{\theta}_j)\|_F^2. \tag{9}$$

Note that Eq. (9) involves a weighted Frobenius norm for calculating patch differences. Suppose there is a matrix $\mathbf{F}=[\mathbf{f}_1,\mathbf{f}_2,\ldots,\mathbf{f}_n]$, $\|\mathbf{F}\|_F^2=\sum_{i=1}^n\|\mathbf{f}_i\|_2^2$. Based on the above definition, Eq. (9) can be rewritten as the weighted sum of vector differences, i.e.,

$$\begin{aligned}g(\boldsymbol{\Theta})&\le\sum_{\boldsymbol{\theta}_i\in\boldsymbol{\Theta}}\sum_{\boldsymbol{\theta}_j\in N_\theta}s(i,j)\sum_{a=-u}^{u}\|\boldsymbol{\theta}_{i+a}-\boldsymbol{\theta}_{j+a}\|_2^2\\&=\sum_{a=-u}^{u}\sum_{\boldsymbol{\theta}_{i+a}}\sum_{\boldsymbol{\theta}_{j+a}}s(i,j)\|\boldsymbol{\theta}_{i+a}-\boldsymbol{\theta}_{j+a}\|_2^2\\&\le\sum_{\boldsymbol{\theta}_i\in\boldsymbol{\Theta}}\sum_{\boldsymbol{\theta}_j\in N_\theta}\omega_{ij}\|\boldsymbol{\theta}_i-\boldsymbol{\theta}_j\|_2^2,\end{aligned} \tag{10}$$

with

$$\omega_{ij}=\sum_{a=-u}^{u}s(i-a,j-a), \tag{11}$$

For $\forall\boldsymbol{\theta}_i\in\boldsymbol{\Theta}$ and $\forall\boldsymbol{\theta}_j\in N_\theta$, we have $\boldsymbol{\theta}_{i+a}\in\boldsymbol{\Theta}$ and $\boldsymbol{\theta}_{j+a}\in N_\theta$. Through variable substitution, we can obtain the final result in (10), where ω_{ij} is calculated as the sum of similarity measure $s(i,j)$ between patch pairs in $P(\boldsymbol{\theta}_i)$ and $P(\boldsymbol{\theta}_j)$.

Based on (10) and [39], $g(\boldsymbol{\Theta})$ can be finally relaxed as $g(\boldsymbol{\Theta})\le Tr(\boldsymbol{\Theta}\mathbf{L}\boldsymbol{\Theta}^T)$, where $\mathbf{L}$ is the Laplacian matrix. The difference between our Laplacian matrix and the Laplacian matrix in [39] is that the weight ω_{ij} in our method is used to measure the similarity of the sparse representation results. Hence our Laplacian matrix can enforce the spatial smoothness among the sparse representations of different group projections. The Laplacian matrix in [39] is used to measure the similarity between different training data, which can highlight the difference between different classes.

Substituting $Tr(\boldsymbol{\Theta}\mathbf{L}\boldsymbol{\Theta}^T)$ for $g(\boldsymbol{\Theta})$ in Eq. (5), we can rewrite the nonlocal regularizer penalized multi-view sparse representation as

$$\begin{aligned}\min_{\boldsymbol{\Theta},\mathbf{P}}\ &\sum_{i=1}^{c}\sum_{k=1}^{m}\frac{1}{2}\|(\mathbf{P}^k)^T\mathbf{Y}_i^k-((\mathbf{P}^k)^T\mathbf{A}^k)\boldsymbol{\Theta}_i^k\|_F^2+\lambda_1Tr(\boldsymbol{\Theta}\mathbf{L}\boldsymbol{\Theta}^T)\\&+\lambda_2\|\boldsymbol{\Theta}\|_{2,1}+\lambda_3Tr(\mathbf{P}^T(\mathbf{S}-\mathbf{D})\mathbf{P}).\end{aligned} \tag{12}$$

Now problem (12) is a tractable problem which will be solved in the next section.

3.2. The reconstruction algorithm

Here, we will present the detailed reconstruction algorithm for the nonlocal regularizer penalized multi-view sparse representation in Fig. 3. Problem (12) is a non-constrained problem, and directly solving this problem using Accelerated Proximal Gradient (APG) algorithm [34] will slow down the convergence speed. Alternating Direction Method of Multipliers (ADMM) algorithm can give a faster convergence rate than APG algorithm [41], however, it always involves high computational complexity. To overcome the limitation of above algorithms, we propose an adaptive ADMM algorithm to solve problem (12), in which, we firstly use variable splitting method [40] to rewrite (12) as follows

$$\begin{aligned}\min_{\boldsymbol{\Theta},\mathbf{U},\mathbf{Z},\mathbf{P}}\ &\sum_{i=1}^{c}\sum_{k=1}^{m}\frac{1}{2}\|(\mathbf{P}^k)^T\mathbf{Y}_i^k-((\mathbf{P}^k)^T\mathbf{A}^k)\mathbf{Z}_i^k\|_F^2\\&+\lambda_1Tr(\mathbf{U}\mathbf{L}\mathbf{U}^T)+\lambda_2\|\boldsymbol{\Theta}\|_{2,1}+\lambda_3Tr(\mathbf{P}^T(\mathbf{S}-\mathbf{D})\mathbf{P}).\\ s.t.\ \ &\mathbf{Z}=\boldsymbol{\Theta},\mathbf{U}=\mathbf{Z}\end{aligned} \tag{13}$$

Reformulating (12) to (13) is aimed to change a difficult problem into a decomposable easy problem. After changing (12) into (13), we secondly merge the two constraints in (13) into a linear constraint and obtain

$$\begin{aligned}\min_{\mathbf{\Theta},\mathbf{U},\mathbf{Z},\mathbf{P}} \quad & \sum_{i=1}^{c}\sum_{k=1}^{m}\frac{1}{2}\|(\mathbf{P}^k)^T\mathbf{Y}_i^k - ((\mathbf{P}^k)^T\mathbf{A}^k)\mathbf{Z}_i^k\|_F^2 \\ & + \lambda_1 Tr(\mathbf{ULU}^T) + \lambda_2\|\mathbf{\Theta}\|_{2,1} + \lambda_3 Tr(\mathbf{P}^T(\mathbf{S}-\mathbf{D})\mathbf{P}), \\ s.t. \quad & \mathcal{B}(\mathbf{U}) + \mathcal{C}(\mathbf{Z}) = \mathcal{D}(\mathbf{\Theta}) \end{aligned} \tag{14}$$

where $\mathcal{B}$, $\mathcal{C}$ and $\mathcal{D}$: $R^{m\times n}\rightarrow R^{2m\times 2n}$ are linear operators which are defined as

$$\mathcal{B}(\mathbf{U}) = \begin{pmatrix}\mathbf{0} & \mathbf{0}\\ \mathbf{0} & \mathbf{U}\end{pmatrix}, \mathcal{C}(\mathbf{Z}) = \begin{pmatrix}\mathbf{Z} & \mathbf{0}\\ \mathbf{0} & -\mathbf{Z},\end{pmatrix},$$
$$\mathcal{D}(\mathbf{\Theta}) = \begin{pmatrix}\mathbf{\Theta} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix} \tag{15}$$

where element $\mathbf{0}$ in (15) is a zero matrix of the same size as $\mathbf{\Theta}$, $\mathbf{U}$ and $\mathbf{Z}$. Compared with (13), problem (14) can deal with all the constraints together to reduce the computational complexity.

Finally, we propose to use Augmented Lagrange method to solve problem (14). The flow chart of the Augmented Lagrange method is shown in Algorithm 1, and the detailed mathematical deduction for Algorithm 1 is shown in Appendix A. The advantage of Algorithm 1 is that we introduce the adjoint operators B^* and C^* in **Z-step** and **U-step**, respectively, to simplify the process of sparse coefficients estimation.

Algorithm 1 Augmented Lagrange Method for Solving Problem ([14]).

Input: $\mathbf{Y}_i^t$, $\mathbf{A}^t$, λ_1, λ_2 and λ_3
Output: $\mathbf{\Theta}$, $\mathbf{P}$
Initialize: $t=0$, $\mathbf{\Theta}^0=\mathbf{V}^0=\mathbf{Z}^0=\mathbf{U}^0=\mathbf{0}$, $\mathbf{P}=\mathbf{0}$
while $\|\mathbf{\Theta}^{t+1}-\mathbf{\Theta}^t\|_F^2 > 10^{-5}$ **do**
1. Using augmented Lagrange function to change (13) into a non-constraint problem.
2. ***P**-step*
Updating $\mathbf{P}^{t+1}$ by solving $\mathbf{D}^{-1}\mathbf{S}'\mathbf{P}=\lambda_3\mathbf{P}$.
3. ***Θ**-step*
Updating $\mathbf{\Theta}^{t+1}=\Gamma_{\frac{\lambda_2}{\beta^t}}(\mathbf{Z}^t+\frac{\lambda_2}{\beta^t}\mathbf{\Lambda}_{11}^t)$
4. *Z*-step
Updating $\mathbf{Z}^{t+1}=\tau(-\frac{1}{\beta^t}\mathbf{\Lambda}_{11}^t+\mathbf{\Theta}^{t+1}+\mathbf{U}^t+\frac{1}{\beta^t}\mathbf{\Lambda}_{22}^t+\frac{1}{\eta\beta^t}\mathbf{V}^t-\eta\bigtriangledown F(\mathbf{V}^t))$.
5. *U*-step
Updating $\mathbf{U}^{t+1}=(\mathbf{I}+\frac{\beta^t}{\lambda_1}\mathbf{L})^{-1}(\mathbf{Z}^{t+1}-\frac{1}{\beta^t}\mathbf{\Lambda}_{22}^t)$
6. $\mathbf{V}^{t+1}=\mathbf{Z}^{t+1}+\eta(\mathbf{Z}^{t+1}-\mathbf{Z}^t)$
7. $\mathbf{\Lambda}^{t+1}=\mathbf{\Lambda}^t+\beta^t(\mathcal{B}(\mathbf{U}^{t+1})+\mathcal{C}(\mathbf{Z}^{t+1})-\mathcal{D}(\mathbf{\Theta}^{t+1}))$
8. Updating Laplacian matrix $\mathbf{L}$
9. $\beta^{t+1}=min(\beta^{\max},\rho\beta^t)$
10. $t\longleftarrow t+1$
end while

3.3. Convergence and computational complexity analysis

Problem (14) is a convex but non-smooth problem. It is difficult to rigorously prove the convergence of the proposed Augmented Lagrange method. Convergence analysis of a general convex but non-smooth problem has been given in Tao and Yuan [43], where it is stated that if the Lagrange function is bounded, the Augment Lagrange Multiplier based reconstruction method can give a feasible solution. Based on [43], we have proved that the augmented Lagrange function of (14) is bounded in Appendix C, which can theoretically illustrate that Algorithm 1 is guaranteed to yield a feasible projection and sparse representation matrices. The computational complexity of each iteration in Algorithm 1 is mainly incurred by step 4, which is $\mathcal{O}(mcns^2)$, where n is the number of rows in matrix $\mathbf{P}$, s is the particle number of an observation group, m is the number of views and c is the number of observation groups in each view. In comparison, the computational complexity for solving the sparse representation method in [31] is $\mathcal{O}(2muv^2)$ (u is the original dimension of observation matrix, $u >= n$, v is the particle number of undivided observation matrix), and the complexity of multi-task tracker [26] is $\mathcal{O}(uvd)$ (d is the number of columns in template matrix). As a concrete example, if the number of views is 3, the group number is 8, the particle number without any division is 400, then the computational complexity of our method is in the order of 10^5, which is much lower than that required by [31] ($\mathcal{O}(2muv^2)\approx 10^7$). It is also lower than that reported in [26] where $d>u$, and $\mathcal{O}(uvd)\approx 10^6$.

3.4. Discussion

The proposed nonlocal regularizer penalized multi-view sparse representation method is closely related to the state-of-the-art tracking methods [31], [32] and [34]. Here, we will further discuss the difference between our method and those related works.

Difference from the work in [34]: In particle filter based visual tracking framework, the correlation between different particle observations are actually not the same, some observations may be very similar. Based on this observation, both [34] and our work divide particle observations into different groups for visual tracking. However, [34] explores group similarity in one view, while we proposed to explore the group similarity in the multi-feature space. Exploring the group similarity in multi-feature space is a challenging task because it not only requires to maximize the intra-group similarity in a certain view, but also requires to make sure that the same observation groups in different views can highlight their inherent commonality. For this purpose, our proposed sparse representation method (equation (12)) uses multi-view discriminant learning to simultaneously explore the intra-view and the inter-view similarity, which can guarantee that similar observation groups have similar sparse representation results.

Difference from the works in [31] and [32]: [31], [32] and our work are all to minimize the sum of the multi-view sparse representation errors to make different views complement with each other. In fact, [31] and [32] may not obtain the minimal sum of the multi-view sparse representation errors because they could not effectively resist the high sparse representation errors caused by unreliable views. Different from [31] and [32] that directly use multi-view observations to achieve sparse representation, we firstly divide multi-view observations into different groups, and then use the group projections to achieve sparse representation. The group projections are obtained by using the online updated projection matrices to project observation groups into a common subspace. Since the projection matrices are updated through exploring the multi-view group similarity, they can enforce the within-group similarity in the unreliable view. Based on this advantage, introducing group projections in multi-view sparse representation can highlight the useful complementary information of different observation groups. This means that the disturbance in unreliable observation groups can be reduced, which is good for minimizing the sparse representation errors of unreliable views. Moreover, the nonlocal regularizer in (12) can enforce the spatial smoothness among multi-view sparse representation results, which can further reduce the sparse representation errors of multi-view group projections.

4. Visual tracking framework

Here, we employ our proposed sparse representation method to achieve visual tracking. In this paper, the moving object is tracked under a particle filter framework, which mainly consists of two parts: the first part is to sample particles to generate multi-view observations using the particle filter method. The second part calculates the posterior probability of different particle samples using the sparse represent results from Algorithm 1. In the particle filter method, the state vector of a moving target at time t is denoted as $\mathbf{x}^t \in R^h$, and the observations of the state vector from time 1 to t are denoted as $\mathcal{Y}^t = \{\mathbf{y}^1, \mathbf{y}^2, \ldots, \mathbf{y}^t\}$. Using the Bayes rule, the posterior probability $p(\mathbf{x}^t|\mathcal{Y}^t)$ is calculated as $p(\mathbf{x}^t|\mathcal{Y}^t) \propto p(\mathbf{y}^t|\mathbf{x}^t) \int [p(\mathbf{x}^t|\mathbf{x}^{t-1}) p(\mathbf{x}^{t-1}|\mathcal{Y}^{t-1})] d\mathbf{x}^{t-1}$, where $p(\mathbf{y}^t|\mathbf{x}^t)$ is the observation likelihood and $p(\mathbf{x}^t|\mathbf{x}^{t-1})$ denotes the motion model. As it is very difficult to calculate $p(\mathbf{x}^t|\mathcal{Y}^t)$ directly using the aforementioned formula, the posterior probability is instead approximated by $p(\mathbf{x}^t|\mathcal{Y}^t) = \sum_{j=1}^{n} \omega_j^t \delta(\mathbf{x}^t - \mathbf{x}_j^t)$, where δ is the Dirac measure, $\mathbf{x}_j^t$ is the j-th sampled particle at time t, and ω_j^t is the particle importance weight, which is updated by $\omega_j^t = \omega_j^{t-1} p(\mathbf{y}^t|\mathbf{x}_j^t)$. Based on particle filter method, we use three features, namely intensity, texture and edge to represent $\mathcal{Y}^t$ for generating $\mathbf{Y}^k$ $(k = 1, 2, 3)$. To calculate $p(\mathbf{x}^t|\mathcal{Y}^t)$, the key is to compute $p(\mathbf{y}^t|\mathbf{x}_j^t)$.

At time t, suppose we have obtained the three-view observation matrices $\mathbf{Y}^1$, $\mathbf{Y}^2$ and $\mathbf{Y}^3$ using aforementioned particle filter method. Firstly, we divide each observation matrix into different sub-matrices (groups) by online k-means method [34]. Choosing online k-means because it can only use a newly arrived state vector to update the cluster centroid, we can avoid time-consuming re-clustering. To enhance the clustering performance, similar to [34], we use $\boldsymbol{v} = [u, v, \mathbf{q}^T]^T$ as state vector for observation clustering, which is robust to image noise and can make different observations more group-discriminative. In $\boldsymbol{v} = [u, v, \mathbf{q}^T]^T$, $[u, v]$ is the target coordinate and $\mathbf{q}^T$ is the target appearance of multi-views. In visual tracking, the cluster centroid is online updated by

$$\boldsymbol{\mu}_c^{new} = \boldsymbol{\mu}_c + \xi(\boldsymbol{v} - \boldsymbol{\mu}_c) \tag{16}$$

where $\boldsymbol{\mu_c}$ means the cluster centroid in the c-th group, $\boldsymbol{v}$ is the newly arrived state vector and ξ is the learning rate.

After the observation grouping process, secondly, we estimate $\boldsymbol{\Theta} = [\boldsymbol{\Theta}_1^1, \ldots, \boldsymbol{\Theta}_i^k, \ldots, \boldsymbol{\Theta}_c^3]$ $(k = 1, 2, 3; i = 1, 2, \ldots, c)$ by using Algorithm 1 to solve problem (14). When obtaining $\boldsymbol{\Theta}$, we then calculate the sparse representation errors of different observation groups. The i-th observation group error is calculated as

$$e(i) = \sum_{k=1}^{3} \|(\mathbf{P}^k)^T \mathbf{Y}_i^k - ((\mathbf{P}^k)^T \mathbf{A}^k) \boldsymbol{\Theta}_i^k\|_F^2, \ i = 1, 2, \ldots, c \tag{17}$$

Next, based on $e(i)$, we select an observation group with minimum sparse representation error to achieve observation likelihood estimation. Suppose the 1-th observation group has minimum sparse representation error, hence, the observation likelihood $p(\mathbf{y}^t|\mathbf{x}_j^t)$ is calculated by

$$p(\mathbf{y}^t|\mathbf{x}_j^t) = \frac{1}{\Gamma} exp(-\alpha \sum_{k=1}^{3} \|(\mathbf{P}^k)^T [\mathbf{Y}_1^k]_j - (\mathbf{P}^k)^T \mathbf{A}^k [\boldsymbol{\Theta}_1^k]_j\|_2^2), \tag{18}$$

where $[\mathbf{Y}_1^k]_j$ $(j = 1, 2, \ldots, r)$ means the j-th particle observation vector in the observation group $\mathbf{Y}_1^k$ and $[\boldsymbol{\Theta}_1^k]_j$ is the corresponding sparse representation result. After calculating $p(\mathbf{y}^t|\mathbf{x}_j^t)$, the final optimal tracking result for the t-th frame is calculated as $\bar{\mathbf{x}}^t = \frac{\sum_{j=1}^{r} \omega_j^t \mathbf{x}_j^t}{\sum_{j=1}^{r} \omega_j^t}$, where ω_j^t is the particle weight of the j-th particle observation. Since the proposed sparse representation method can use multi-view discriminant analysis to make $(\mathbf{P}^k)^T \mathbf{Y}_i^k$ group discriminative and highlight the useful information in unreliable observation groups, we can give an exact estimation for $e(i)$ and $p(\mathbf{y}^t|\mathbf{x}_j^t)$.

5. Experiments

In this section, we use the video sequences in CVPR2013 Visual Tracking Benchmark [44] to evaluate the performance of our proposed visual tracking algorithm. These video sequences are very challenging in the sense that they contain many adverse factors against visual tracking such as fast motion, large variation in pose and scale, occlusion and non-rigid object deformation etc. We compare the proposed tracking algorithm with 12 state-of-the-art methods: IVT[14], CT[11], l1-APG[25], MTT[26], LRT[17], STRUCK[45], CSK[46], TLD[47], Frag[48], KMS[49], OAB[50] and KCF[12]. Since our method and the existing ones like the l1-APG and MTT are all particle filter based sparse representation algorithms, the particle number is set equally as 400. To illustrate the effectiveness of the projection matrix $\mathbf{P}^k$ and the nonlocal regularizer $g(\boldsymbol{\Theta})$ in the proposed sparse representation method, we compare equations (1), (4) and (12) in our paper. Using Eq. (1) to achieve visual tracking is the multi-view sparse representation method without projection matrix and nonlocal regularizer. Eq. (4) is the multi-view discriminant learning based sparse representation method, which introduces projection matrix in the multi-view sparse representation. Finally, Eq. (12) is the nonlocal regularizer penalized multi-view sparse representation method to track moving object, which adds both projection matrix and the nonlocal regularizer into the sparse representation. For notational simplicity, we name the multi-view sparse representation method without projection matrix and nonlocal regularizer, the multi-view discriminant learning based sparse representation method [35] and the nonlocal regularizer penalized multi-view sparse representation method as **MVSR**, **MVDLSR**, and **NR-MVDLSR**, respectively.

Experimental setting: In our experiments, we use three complementary features to achieve visual tracking, which are intensity, local binary patterns (LBP)[51] and edges with canny operator. The target template matrices in three views have the same size, where $\mathbf{A}^i \in R^{256 \times 20} (i = 1, 2, 3)$. In these template matrices, the particle observation size is 16×16, and the number of target templates is 20 (10 for foreground templates and 10 for background templates). Currently, the demo code is available at the URL https://github.com/greatisgood123/MVDLSR.

5.1. Evaluation of cluster number

In this test, we choose a challenging video sequence called trellis to evaluate the relationship between the group number and the tracking performance. Choosing this sequence is because the target occupies a large space in video sequence, which can indicate the difference of tracking performance more clearly. In the experiment, we directly use online k-means [34] on multi-view observations to achieve group division without using any additional training process. During group division, we evaluate tracking performance with varying group number. From Fig. 5 we see that the proposed sparse representation method can give the best tracking performance when the particle observation is divided into 8 groups. If the group number is less than 8, some dissimilar particle samples may be involved in the particle observation groups and share a similar sparsity pattern with similar samples, thus degrading the tracking performance. If the group number is larger than 8, those similar particle samples cannot be grouped together, which would also degrade the tracking performance.

Table 1
FPS performance for different methods.

Tracker	NR-MVDLSR	IVT	CT	CSK	L1-APG	MTT	Frag	KMS	STRUCK	TLD	OAB	KCF
Compiler	matlab	matlab	matlab	matlab	C	matlab	C++	matlab	C++	matlab	matlab	matlab
FPS	1.1	10.2	13.3	85.1	13.2	0.3	2.2	4.6	0.12	12.6	16.0	30.2

Table 2
FPS performance for different spare representation based trackers.

Tracker	NR-MVDLSR	MVDLSR	LRT [17]	MTMVT [31]	DGSP [34]	JSRFFT [33]
Compiler	matlab	matlab	matlab	matlab	matlab	matlab
FPS	1.1	1.2	0.6	1.0	0.2	0.7

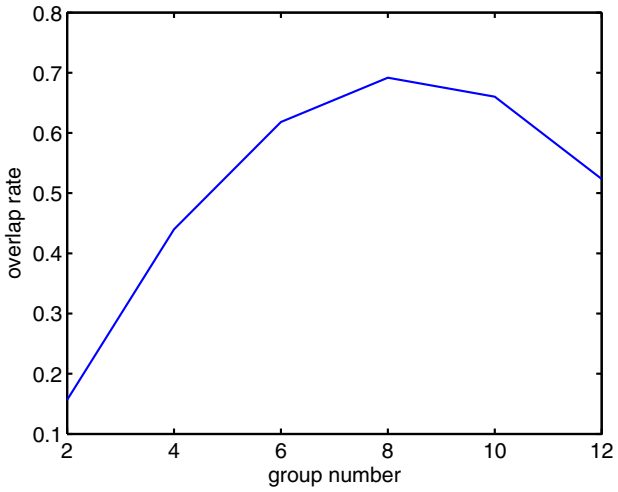

Fig. 5. Average overlap rate performance with varying candidate group numbers.

5.2. Runtime performance

To illustrate the computational complexity of the proposed NR-MVDLSR method, we test the average tracking speed (Frame num per Second, FPS) on a laptop with Inter(R) Core(TM) i3-2310M CPU @ 2.10 Hz (2GB RAM) (see Table 1), where different methods are all implemented on 30 video sequences.

To further illustrate the computational complexity of MVDLSR and NR-MVDLSR methods, we compare them with four well-known sparse representation methods. The testing result is shown in Table 2. From Tables 1 and 2, we can see that although NR-MVDLSR introduces multi-view learning and non-local regularizer in sparse representation to achieve visual tracking, its computational complexity is similar to traditional sparse representation based tracking methods.

5.3. Parameter analysis

There are four parameters η, λ_1, λ_2, λ_3 that require to be set in Algorithm 1. Inspired by [17], we randomly choose 10 challenging video sequences to select the optimal combination of four parameters according to the parameter sensitivity analysis. The detailed parameter analysis is discussed in the following.

Evaluation of η: The learning step parameter η controls the convergence rate of reconstruction algorithm. This parameter is not related to λ_1, λ_2 and λ_3. If the value of η is too small, the convergence speed would be slow. If the value of η is too large, it may cause vibration and no convergence. Inspired by parameter sensitivity analysis [17], to choose the value for η, we first fix λ_1, λ_2 and λ_3, and then test the running speed of reconstruction algorithm with different η values. The testing result is shown in Table 3, where *NaN* means the reconstruction method is not convergent, and FPS means the tracking speed (Frames per Second). From Table 3 we could see that with the increase of parameter η, the FPS is gradually increased. When the value of η is larger than 0.01, the reconstruction method will not get convergent, thus leading to an inoperative tracking result. Based on Table 3, we empirically set $\eta = 0.01$ for all the experiments.

Table 3
FPS performance with differnet η value.

η	0.001	0.005	0.01	0.05	0.1	0.5
FPS	0.4	0.6	1.1	*NaN*	*NaN*	*NaN*

Table 4
$\lambda_3 = 0.1$.

λ_1			
λ_2	0.1	0.15	0.2
0.1	0.67	0.67	0.68
0.5	0.72	0.73	0.60
1	0.75	0.68	0.71

Table 5
$\lambda_3 = 0.5$.

λ_1			
λ_2	0.1	0.15	0.2
0.1	0.71	0.69	0.60
0.5	0.67	0.69	0.63
1	0.70	0.68	0.71

Evaluation of λ_1, λ_2 and λ_3: In our proposed sparse representation model, λ_1 and λ_2 are two important parameters which control the smoothness and the sparsity of the sparse representation result, respectively. On one hand, if λ_1 and λ_2 are too large, it may cause over-smoothing and over-sparsity. On the other hand, if both are small, the sparse representation result will suffer undesired sparse pattern, resulting in the poor tracking performance. Besides λ_1 and λ_2, λ_3 is also critical for the proposed sparse representation model, which measures the contribution of multi-view discriminant learning. To find an optimal combination of three parameters, we firstly fix $\lambda_3 = 0.1$, and then calculate the average overlap rate over 10 video sequences with different combinations of λ_1 and λ_2. The value of λ_1 is selected from a predefined discrete set $\Lambda_1 = \{0.1, 0.5, 1\}$. The λ_2 is selected from $\Lambda_2 = \{0.1, 0.15, 0.2\}$. Thirdly, we fix $\lambda_3 = 0.5$ and $\lambda_3 = 1$, respectively, and re-calculate the average overlap rate with different combination of λ_1 and λ_2. The average overlap rate with different combinations of λ_1, λ_2 and λ_3 are shown in Tables 4, 5 and 6. From these tables, we can see that the proposed tracking method gives the highest average overlap rate when $\lambda_1 = 1$, $\lambda_2 = 0.1$ and $\lambda_3 = 0.1$. Hence, we

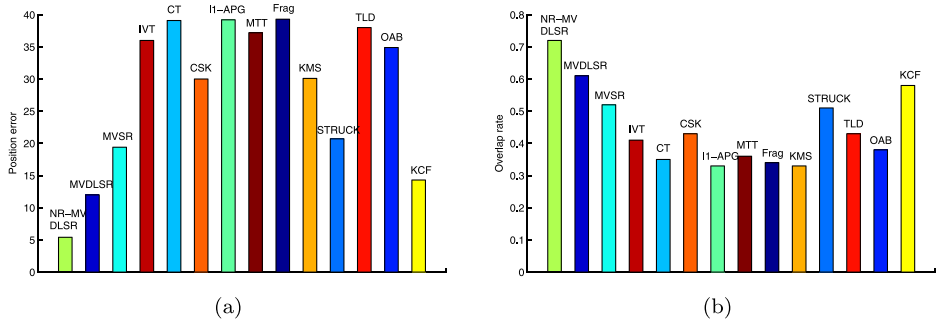

Fig. 6. Average tracking performance over 30 video sequences: (a) Mean value of position error, (b) Mean value of overlap rate.

Table 6
$\lambda_3 = 1$.

λ_1			
λ_2	0.1	0.15	0.2
0.1	0.65	0.65	0.68
0.5	0.66	0.65	*NaN*
1	0.64	0.64	*NaN*

empirically set four parameters in Algorithm 1 as $\eta = 0.01$, $\lambda_1 = 1$, $\lambda_2 = 0.1$ and $\lambda_3 = 0.1$.

5.4. Quantitative tracking performance

In this section, we will give the quantitative evaluation over 30 video sequences. The quantitative visual tracking performance is evaluated by four kinds of objective measures [44]: the position error, the overlap rate, the precision plot and the success plot. The position error is defined as the Euclidean distance between the central location of the tracked bounding box and the manually labeled ground truth. The overlap rate is defined as $\frac{area(B_T \cap B_G)}{area(B_T \cup B_G)}$, where B_T and B_G are the tracked bounding box of each frame and the corresponding ground truth, respectively. The precision plot indicates accumulated position errors under different location error thresholds. The success plot reflects the accumulated successful rates versus different overlap thresholds, where the successful rate counts the number of video frames where the overlap rate is larger than 0.5. The position error and the overlap rate are the objective measures to evaluate the tracking performance for each video frame, while the precision and success plots can illustrate the overall tracking performance.

Firstly, we test the average tracking performance over 30 video sequences. The average position error and the average overlap rate of one video frame are denoted as ave_p and ave_o, respectively. The mean values of ave_p and ave_o over 30 video sequences are shown in Fig. 6. It is seen from Fig. 6(a) that the smaller the position error, the higher the tracking accuracy, and the position error of our proposed NR-MVDLSR is 5.4, which is obviously smaller than other methods. This means that our method can still track the moving target in all selected video sequences. In Fig. 6(b), the large value of overlap rate means the tracker can use a bounding box with an appropriate scale to track the target. This figure shows that our method can give better overlap rate performance than other methods.

These 30 selected sequences contain five adverse factors against visual tracking such as: occlusion, motion blur, scale variation, illumination change and pose variation. Hence, in the next experiment, we divide the test video sequence into 5 groups. The detailed information about the video group is shown in Table 7.

Based on Table 7, we test the mean value of ave_p and ave_o over different video groups to illustrate our tracking performance in different scenes (see Tables 8 and 9).

From Tables 8 and 9 we can clearly see that the proposed NR-MVDLSR method ranks top two among all trackers. This means that our proposed sparse representation model can give the good tracking performance when facing different adverse factors against visual tracking. The NR-MVDLSR, MVDLSR and MVSR use multi-feature to achieve visual tracking, hence they can give obviously lower position error than the single-feature based sparse representation trackers such as l1-APG and MTT. MVDSL gives a better tracking performance than MVSR method because it introduces multi-view discriminant learning into the sparse representation. Since MVDSL only uses $l_{2,1}$ norm to constrain the sparse representations of multi-view observation projections, it may not guarantee the low position error when facing severe motion blur. Compared with MVDSL, the proposed NR-MVDLSR method adds a non-local regularizer into the multi-view discriminant learning based sparse representation model to smooth the sparse representations of multi-view group projections, which can obviously reduce the position error and enhance the overlap rate in challenging video sequences. In visual tracking, KCF is a well-known tracking method. Through the comparison with KCF, we can clearly see the advantage of NR-MVDLSR. In above experiment, the position error performance for motion blur test is not better than that for other adverse factors because the motion blur will destroy the inherent correlation between different pixels. Hence it is a tough work to overcome this adverse factor. Here, we use Tables 10 and 11 to show the detailed tracking performance of different methods in motion blur video groups to further illustrate tracking accuracy of our method. From Tables 10 and 11 we could see that although our method could not give the best tracking performance in crossing and duderk sequences, the tracking accuracy of NR-MVDLSR is similar to that of KCF.

Since the precision and success plots are two well-known objective measures for testing the overall tracking performance, we now adopt these two measures to test our tracking performance over 30 video sequences (see Fig. 7). In Fig. 7(a) and (b), the area under curve (AUC) of each precision and success plot indicates the rank of different tracking algorithms. Based on this observation, we can clearly see that the NR-MVDLSR method ranks first on the success and precision plots.

5.5. Qualitative tracking performance

In this section, we select ten challenging sequences to show the qualitative tracking performance (see Fig. 8). The video sequence selecting strategy is that: we randomly select two video sequences from each video group. This test can give a direct impression of the tracking performance when the target facing different adverse factors.

Table 7
The detail information about the video groups for experiments.

Adverse factors	Video sequence
Occlusion	Faceocc1, Faceocc2, Football, Coke, Subway, Jogging, Lemming
Motion blur	Crossing, Singer2, Jumping, Dudek, Mountainbike, Deer
Scale variation	Car4, Singer1, Walking2, Carscale, Fleetface, Freeman4
Illumination change	Trellis, Skating1, Car11, David Indoor, Fish
Pose variation	Basketball, Shaking, Bolt, Mhyang, Boy, Sylvester

Table 8
Mean value of position error over different video groups. The best two results are denoted as bold and italic.

Seq. / Meth.	NR-MVDLSR	MVDLSR	MVSR	IVT	CT	CSK	l1-APG	MTT	Frag	KMS	STRUCK	TLD	OAB	KCF
Occlusion	**5.3**	*14.8*	24.5	34.8	20.1	26.5	41.5	36.9	24.4	33.4	12.6	25.1	20.8	16.6
Motion blur	**7.2**	16.4	24.9	67.2	66.8	45.0	46.0	62.6	47.0	27.8	23.9	73.2	35.5	*8.7*
Scale variation	**5.4**	*11.6*	12.4	14.8	36.5	18.1	17.6	25.0	33.8	41.1	22.5	21.8	29.2	24.1
Illumination change	*5.0*	**4.9**	13.0	18.0	30.4	18.4	23.3	18.8	38.9	20.4	19.0	25.8	23.3	7.7
Pose variation	**4.2**	*10.5*	20.5	42.7	43.7	40.9	64.8	39.8	55.0	25.5	28.0	44.4	66.0	12.7

Table 9
Mean value of overlap rate over different video groups. The best two results are denoted as bold and italic.

Seq. / Meth.	NR-MVDLSR	MVDLSR	MVSR	IVT	CT	CSK	l1-APG	MTT	Frag	KMS	STRUCK	TLD	OAB	KCF
Occlusion	**0.72**	*0.58*	0.49	0.41	0.45	0.47	0.31	0.38	0.47	0.30	0.59	0.50	0.42	0.56
Motion blur	**0.67**	0.58	0.45	0.34	0.31	0.35	0.36	0.31	0.30	0.41	0.53	0.43	0.41	*0.62*
Scale variation	**0.74**	*0.60*	0.56	0.54	0.29	0.39	0.43	0.40	0.31	0.22	0.41	0.43	0.29	0.41
Illumination change	**0.77**	*0.70*	0.63	0.52	0.35	0.45	0.34	0.44	0.24	0.33	0.55	0.46	0.41	0.66
Pose variation	**0.72**	0.62	0.52	0.28	0.35	0.46	0.24	0.28	0.37	0.41	0.47	0.34	0.35	*0.65*

Table 10
Detailed position error performance over motion blur video group. The best two average results are denoted as bold and italic.

Seq. / Meth.	NR-MVDLSR	MVDLSR	MVSR	IVT	CT	CSK	l1-APG	MTT	Frag	KMS	STRUCK	TLD	OAB	KCF
Crossing	4.4	5.8	23.8	18.5	*3.2*	6.7	42.3	30.9	21.3	5.7	3.3	13.8	4.2	**2.9**
Singer2	12.1	56.5	71.6	175.9	101.0	104.1	135.4	140.8	58.7	20.9	101.1	253.5	105.8	**7.5**
Jumping	**4.1**	5.6	11.5	38.2	62.6	15.8	24.4	41.2	4.3	47.8	5.9	*4.6*	58.8	12.8
Duderk	11.7	12.7	25.7	**9.8**	16.5	13.7	23.4	14.3	44.6	45.3	17.9	18.7	25.3	*10.2*
Mountainbike	**5.9**	6.5	7.5	8.1	94.2	*6.1*	13.2	10.3	102.3	30.7	8.7	106.9	9.3	6.2
Deer	**5.1**	11.2	9.1	152.6	123.4	123.4	37.4	138.0	51.0	16.1	*6.7*	41.9	9.7	12.8

Table 11
Detailed overlap rate performance over motion blur video group. The best two average results are denoted as bold and italic.

Seq. / Meth.	NR-MVDLSR	MVDLSR	MVSR	IVT	CT	CSK	l1-APG	MTT	Frag	KMS	STRUCK	TLD	OAB	KCF
Crossing	0.63	0.60	0.24	0.31	*0.66*	0.49	0.17	0.22	0.29	0.56	0.61	0.43	0.65	**0.70**
Singer2	*0.55*	0.35	0.18	0.04	0.29	0.03	0.04	0.04	0.16	0.28	0.03	0.02	0.02	**0.67**
Jumping	**0.66**	0.58	0.42	0.21	0.04	0.17	0.36	0.20	0.61	0.10	0.58	*0.65*	0.08	0.28
Dudek	0.68	0.65	0.49	*0.72*	0.63	0.69	0.52	0.66	0.51	0.51	0.61	0.61	0.49	**0.72**
Mountainbike	**0.75**	*0.72*	0.71	0.70	0.18	0.69	0.64	0.67	0.12	0.48	0.66	0.26	0.62	0.71
Deer	**0.72**	0.59	0.65	0.03	0.03	0.04	0.41	0.04	0.08	0.52	*0.66*	0.58	0.62	0.63

1) Occlusion: In Fig. 8(a), the faceocc1 sequence is used to test the tracking performance under occlusion. In this sequence, a woman's face undergoes the partial and severe occlusion by a book. From the tracking performance of different methods we can see that OAB method is not robust to face occlusion. Our method can still give an exact tracking result in the entire video sequence. Besides faceocc1 sequence, Fig. 8(b) gives a test on jogging video sequence. This sequence is challenging because the runner is totally occluded by a lamppost. From the tracking results we can see that most tracking methods begin to drift at 181-th frame because the runner suffers a total occlusion after this frame. Clearly, NR-MVDLSR, OAB, TLD and Frag methods are robust to this kind of occlusion. Since MVDLSR method only uses $l_{2,1}$ norm to regularize the sparse representations of multi-view observation projections, it could not give an exact tracking performance in jogging sequence.
2) Motion blur: Motion blur means the target region is blurred due to the motion of target or camera. Jumping and deer video sequences are all suffered from severe motion blur. Fig. 8(c) gives the tracking performance of jumping sequence. From this test we can see that CT, KMS, KCF, MTT and OAB give a poor tracking performance in this sequence while NR-MVDLSR, TLD and MVDLSR can still track the motion of the boy's face. Fig. 8(d) is the tracking performance of deer sequence. From this test we can see that NR-MVDLSR and STRUCK methods give a better tracking performance than other 12 methods do.

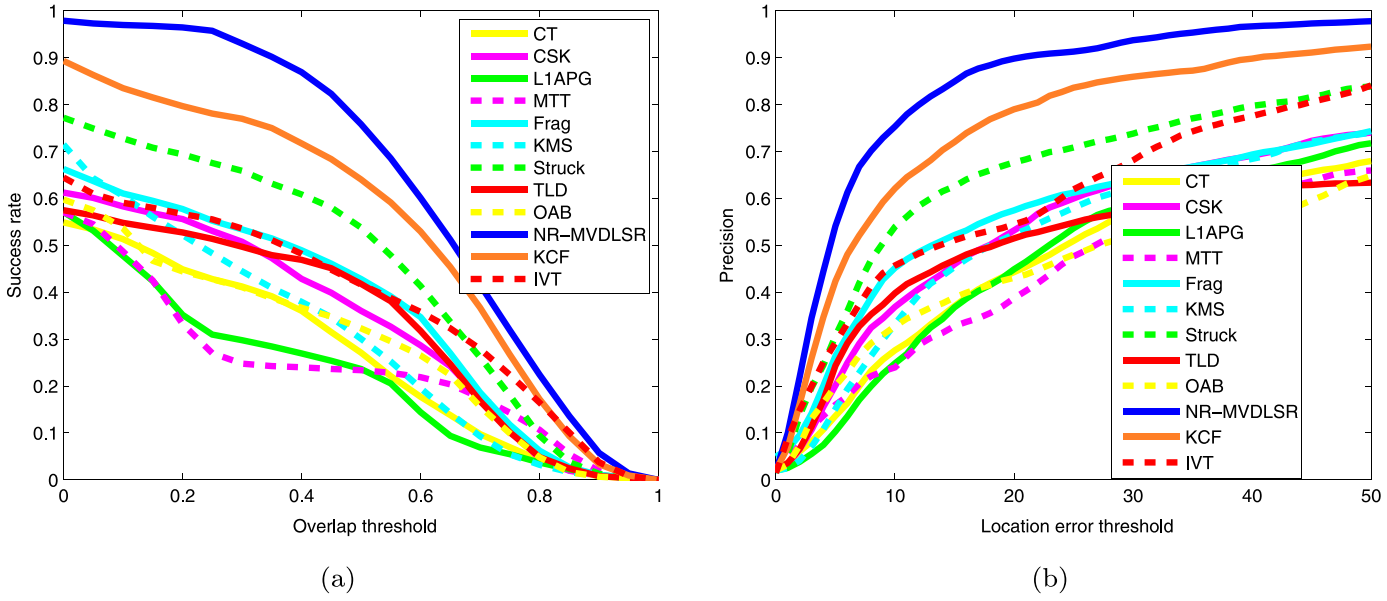

Fig. 7. Success and precision plots over 30 video sequences: (a) Success plot, (b) Precision plot.

Fig. 8. Qualitative tracking results on the randomly selected frames with some challenging factors: (a)-(b) occlusion, (c)-(d) motion blur, (e)-(f) scale variation, (g)-(h) illumination change, (i)-(j) pose variation.

3) Scale variation: In the car4 video sequence (see Fig. 8(e)), there is a drastic change of scale and illumination when the car goes underneath the overpass. NR-MVDLSR, MVDLSR and MVSR methods can perform well in the whole sequence while CT, Frag, CSK, OAB and KCF methods cannot adaptively suit the change of the target appearance, hence they give a poor tracking performance. In the walking2 video sequence (see Fig. 8(f)), the scale of the women's appearance would become more and more smaller when the target is far away from the camera. From this test we can clearly see that the proposed NR-MVDLSR method is robust to the scale variation in walking2 sequence.
4) Illumination change: Trellis and skating1 sequences are suffered from severe illumination change. From Fig. 8(g) we can see that when the illumination of target's face changes dramati-

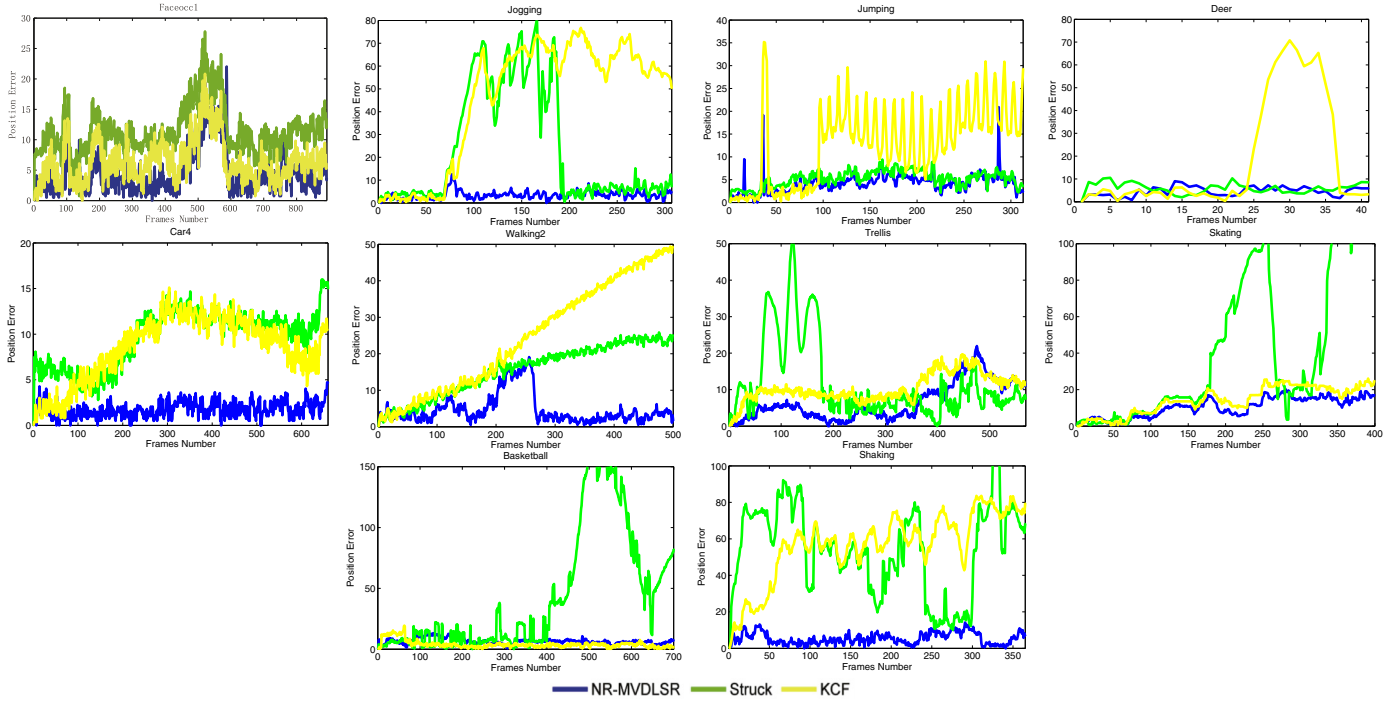

Fig. 9. Each frame position error over 10 video sequences.

cally, such as the 272-th frame, OAB, FRAG, CT and CSK methods begin to drift. The proposed NR-MVDLSR method can still give a better tracking performance in the whole sequence because it is robust against the severe illumination change. In Fig. 8(h), the illumination in the skating arena would be frequently changed. Moreover, the player would also be suffered from occlusion and pose variation. From this test we can see that NR-MVDLSR and MVDLSR methods can give a better tracking performance than other 12 methods do.

5) Pose variation: The tests in Fig. 8(i) and (j) are very difficult because there is severe pose variation in these two video sequences. MVDLSR method fails to track the target in the shaking video sequence whereas NR-MVDLSR can still accurately track the moving target in two video sequences.

Fig. 8 only uses 2 random selected frames to illustrate the qualitative tracking performance of different tracking methods. To illustrate the performance of our proposed method more clearly, we also give each frame position error (see Fig. 9) for these 10 selected video sequences in qualitative evaluation. For a clear display, we only choose two methods, which have good tracking performance in Fig. 8, as comparison to carry out this test. From Fig. 9, we can clearly see that our method still maintains small position errors over 10 very challenging video sequences.

5.6. The failure case

Although the proposed sparse representation method can give a good tracking performance in aforementioned experiments. It could not guarantee a good tracking performance in motorrolling video sequence (see Fig. 10). Motorrolling sequence is very challenging because it contains very large scale changes and fast rotation. The possible reason for the tracking failure in motorrolling sequence is that the template updating strategy cannot timely capture the appearance changes, and thus the target cannot be well represented by multi-view dictionaries. Online multi-view dictionary learning technology may solve this problem, however it is out of our scope in this paper.

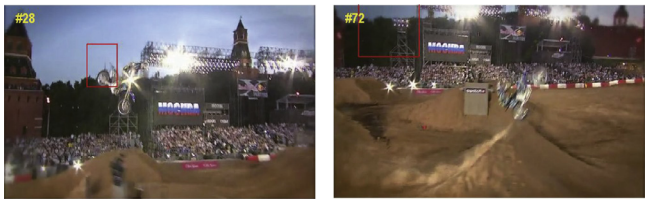

Fig. 10. Randomly selecting two frames as example to show the failure cases in motorrolling video sequence.

6. Conclusion and future work

In this paper, we have proposed a nonlocal regularizer penalized multi-view discriminant sparse representation method for visual tracking. By exploiting the group similarity using multi-view discriminant learning and adopting a nonlocal regularizer to enforce the spatial smoothness among the sparse representations of different group projections, the proposed method can properly fuse reliable and unreliable observation groups to enhance the robustness of visual tracking in severe occlusion, illumination change or pose variation. Experimental results illustrated that the proposed method can give a superior performance in challenging video sequences as compared to a number of known methods in literature. In this paper, the multi-views for visual tracking have the same dimension. To extend our sparse representation to other computer vision applications, our future work is to build a more general multi-view sparse representation model with flexible size of feature subsets.

Acknowledgments

This work is partly supported by the National Key R&D Program of China under Grant 2017YFB0802300, the National Natural Science Foundation of China (Grants Nos. 61801242, 61601223, 61571240 and 61871235), the Natural Science Foundation of Jiangsu Province (Grants Nos. BK20170915 and BK20150756), and the

Fundamental Research Funds for the Central Universities (No. NS2016091).

Appendix A

Here, we discuss the detailed mathematical deduction of Algorithm 1. To solve problem (14), we first adopt augmented Lagrange function to rewrite (14) as

$$\begin{aligned}\mathcal{L}(\mathbf{\Theta},\mathbf{U},\mathbf{Z},\mathbf{P},\mathbf{\Lambda},\beta) = &\frac{1}{2}\sum_{i=1}^{c}\sum_{k=1}^{m}\|(\mathbf{P}^k)^T\mathbf{Y}_i^k - ((\mathbf{P}^k)^T\mathbf{A}^k)\mathbf{Z}_i^k\|_F^2\\ &+\lambda_1 Tr(\mathbf{U}\mathbf{W}\mathbf{U}^T)+\lambda_2\|\mathbf{\Theta}\|_{2,1}\\ &+\lambda_3 Tr(\mathbf{P}^T(\mathbf{S}-\mathbf{D})\mathbf{P})\\ &+ <\mathbf{\Lambda},\mathcal{B}(\mathbf{U})+\mathcal{C}(\mathbf{Z})-\mathcal{D}(\mathbf{\Theta})>\\ &+\frac{\beta}{2}\|\mathcal{B}(\mathbf{U})+\mathcal{C}(\mathbf{Z})-\mathcal{D}(\mathbf{\Theta})\|_F^2,\end{aligned} \tag{A.1}$$

where $\mathbf{\Lambda}=\begin{pmatrix}\mathbf{\Lambda}_{11} & \mathbf{\Lambda}_{12}\\ \mathbf{\Lambda}_{21} & \mathbf{\Lambda}_{22}\end{pmatrix}$ is the Lagrange multiplier matrix, with $\mathbf{\Lambda}_{\mathbf{ij}}\in R^{m\times n}$ $(i=1,2;j=1,2)$ being its submatrices, and $\beta>0$ is the penalty parameter. Since it is very difficult to choose an optimal value for β in advance, we adopt a simple and efficient rule to adaptively update it to further accelerate the convergence rate (see the 9 step in Algorithm 1). Problem (A1) becomes a non-constrained problem, which can be solved by iteratively minimizing the augmented Lagrange function and updating the Lagrange multiplier as follows,

$$(\mathbf{\Theta}^{t+1},\mathbf{Z}^{t+1},\mathbf{U}^{t+1},\mathbf{P}^{t+1}) = \min_{\mathbf{\Theta},\mathbf{U},\mathbf{Z},\mathbf{P}}\mathcal{L}(\mathbf{\Theta},\mathbf{U},\mathbf{Z},\mathbf{P},\mathbf{\Lambda},\beta), \tag{A.2}$$

$$\mathbf{\Lambda}^{t+1} = \mathbf{\Lambda}^t+\beta(\mathcal{B}(\mathbf{U}^{t+1})+\mathcal{C}(\mathbf{Z}^{t+1})-\mathcal{D}(\mathbf{\Theta}^{t+1})). \tag{A.3}$$

Note that it is difficult to solve (A2) directly because it requires to simultaneously minimize four variables. Next, we propose to use an alternating strategy to divide (A2) into four sub-problems, referred to as **P**-step, **Θ**-step, **Z**-step and **U**-step.

P-step is to update projection matrix **P**. Here, we fix **Θ**, **Z** and **U**, and update **P** by solving the following problem

$$\begin{aligned}\min_{\mathbf{P}}\ &\sum_{i=1}^{c}\sum_{k=1}^{m}\frac{1}{2}\|(\mathbf{P}^k)^T\mathbf{Y}_i^k-((\mathbf{P}^k)^T\mathbf{A}^k)\mathbf{Z}_i^k\|_F^2\\ &+\lambda_3 Tr(\mathbf{P}^T(\mathbf{S}-\mathbf{D})\mathbf{P}).\end{aligned} \tag{A.4}$$

Using the relationship between matrix trace and Frobenius norm, we can simplify problem (A4) as

$$\min_{\mathbf{P}} Tr(\mathbf{P}^T\mathbf{M}\mathbf{P})+\lambda_3 Tr(\mathbf{P}^T(\mathbf{S}-\mathbf{D})\mathbf{P}), \tag{A.5}$$

where $\mathbf{Q}^k=\sum_{i=1}^{c}(\mathbf{Y}_i^k-\mathbf{A}^k\mathbf{Z}_i^k)(\mathbf{Y}_i^k-\mathbf{A}^k\mathbf{Z}_i^k)^T$, $\mathbf{M}=diag(\mathbf{Q}^1,\mathbf{Q}^2,\ldots,\mathbf{Q}^m)$. Let $\mathbf{S}'=\lambda_3\mathbf{S}+\mathbf{M}$, then problem (A5) becomes

$$\min_{\mathbf{P}}\lambda_3 Tr(\mathbf{P}^T(\mathbf{S}'-\lambda_3\mathbf{D})\mathbf{P}), \tag{A.6}$$

which can be solved directly by setting its first derivative to zero, giving

$$\mathbf{D}^{-1}\mathbf{S}'\mathbf{P}=\lambda_3\mathbf{P}. \tag{A.7}$$

The eigenvector matrix $\mathbf{P}^*$ with respect to $\mathbf{D}^{-1}\mathbf{S}'$ becomes the solution to problem (A7).

By fixing **P**, **Z** and **U**, the **Θ**-step aims to update matrix $\mathbf{\Theta}^{t+1}$ by solving the following problem

$$\min_{\mathbf{\Theta}}\lambda_2\|\mathbf{\Theta}\|_{2,1}+\frac{\beta^t}{2}\|\mathcal{B}(\mathbf{U}^t)+\mathcal{C}(\mathbf{Z}^t)-\mathcal{D}(\mathbf{\Theta})+\frac{1}{\beta^t}\mathbf{\Lambda}^t\|_F^2. \tag{A.8}$$

Ignoring the constant elements in (A8), we can obtain

$$\mathbf{\Theta}^{t+1}=\Gamma_{\frac{\lambda_2}{\beta^t}}(\mathbf{Z}^t+\frac{\lambda_2}{\beta^t}\mathbf{\Lambda}_{11}^t), \tag{A.9}$$

where $\Gamma_\alpha(\cdot)$ is a matrix operator [42]. Suppose there is a matrix **X**, such that $\Gamma_\alpha(\mathbf{X})$ outputs a matrix in which the i-th row of $\Gamma_\alpha(\mathbf{X})$ is updated as

$$[\Gamma_\alpha(\mathbf{X})](i,:)=\begin{cases}\left(\frac{\|\mathbf{X}(i,:)\|_2-\alpha}{\|\mathbf{X}(i,:)\|_2}\right)\mathbf{X}(i,:) & \|\mathbf{X}(i,:)\|_2>\alpha\\ \mathbf{0} & otherwise\end{cases} \tag{A.10}$$

where $[\Gamma_\alpha(\mathbf{X})](i,:)$ means the i-th row in $\Gamma_\alpha(\mathbf{X})$, $\mathbf{X}(i,:)$ means the i-th row in **X**, **0** is a zero vector which has the same size as $\mathbf{X}(i,:)$, and α is a soft thresholding.

In **Z**-step, $\mathbf{Z}^{t+1}$ is updated by solving the following problem

$$\begin{aligned}\min_{\mathbf{Z}}\ &\frac{1}{2}\sum_{i=1}^{c}\sum_{k=1}^{m}\|(\mathbf{P}^k)^T\mathbf{Y}_i^k-((\mathbf{P}^k)^T\mathbf{A}^k)\mathbf{Z}_i^k\|_F^2\\ &+\frac{\beta^t}{2}\|\mathcal{B}(\mathbf{U}^t)+\mathcal{C}(\mathbf{Z})-\mathcal{D}(\mathbf{\Theta}^{t+1})+\frac{1}{\beta^t}\mathbf{\Lambda}^t\|_F^2.\end{aligned} \tag{A.11}$$

In (A11), let $F(\mathbf{Z})=\sum_{i=1}^{c}\sum_{k=1}^{m}\|(\mathbf{P}^k)^T\mathbf{Y}_i^k-((\mathbf{P}^k)^T\mathbf{A}^k)\mathbf{Z}_i^k\|_F^2$, $\Omega(\mathbf{Z})=\|\mathcal{B}(\mathbf{U}^t)+\mathcal{C}(\mathbf{Z})-\mathcal{D}(\mathbf{\Theta}^{t+1})+\frac{1}{\beta^t}\mathbf{\Lambda}^t\|_F^2$, where $F(\cdot)$ and $\Omega(\cdot)$ are differentiable functions. Applying composite gradient mapping [54] to problem (A11), we can obtain

$$\mathbf{Z}^{t+1}=\min_{\mathbf{Z}}F(\mathbf{V}^t)+ <\triangledown F(\mathbf{V}^t),\mathbf{Z}>+\frac{1}{2\eta}\|\mathbf{Z}-\mathbf{V}^t\|_F^2+\frac{\beta^t}{2}\Omega(\mathbf{Z}), \tag{A.12}$$

where η is a step-size parameter. Problem (A12) can be solved by setting its partial derivative with respect to **Z** to zero, leading to

$$\frac{1}{\eta}(\mathbf{Z}-\mathbf{V}^t+\eta\triangledown F(\mathbf{V}^t))+\beta^t\mathcal{C}^*\left(\mathcal{B}(\mathbf{U}^t)+\mathcal{C}(\mathbf{Z})-\mathcal{D}(\mathbf{\Theta}^{t+1})+\frac{1}{\beta^t}\mathbf{\Lambda}^t\right)=0, \tag{A.13}$$

where

$$\begin{aligned}[\triangledown F(\mathbf{V}^t)]_i^k = &-((\mathbf{P}^k)^T\mathbf{Y}_i^k)^T((\mathbf{P}^k)^T\mathbf{Y}_i^k-((\mathbf{P}^k)^T\mathbf{A}^k)\mathbf{V}_i^k)\\ &i=1,2,\ldots,c\quad k=1,2,\ldots,m.\end{aligned} \tag{A.14}$$

In (A13), $\mathcal{C}^*(\cdot):R^{2m\times 2n}\to R^{m\times n}$ is the adjoint operator. The property of this operator is shown in Appendix B. Rearranging (A13), we can obtain

$$\begin{aligned}\mathcal{C}^*(\mathcal{C}(\mathbf{Z})) = &-\frac{1}{\eta\beta^t}(\mathbf{Z}-\mathbf{V}^t+\eta\triangledown F(\mathbf{V}^t))\\ &-\mathcal{C}^*(\mathcal{B}(\mathbf{U}^t)-\mathcal{D}(\mathbf{\Theta}^{t+1})+\frac{1}{\beta^t}\mathbf{\Lambda}^t),\end{aligned} \tag{A.15}$$

where

$$\begin{aligned}&\mathcal{C}^*(\mathcal{B}(\mathbf{U}^t)-\mathcal{D}(\mathbf{\Theta}^{t+1})+\frac{1}{\beta^t}\mathbf{\Lambda}^t)\\ &=\mathcal{C}^*\begin{pmatrix}\frac{1}{\beta^t}\mathbf{\Lambda}_{11}^t-\mathbf{\Theta}^{t+1} & \frac{1}{\beta^t}\mathbf{\Lambda}_{12}^t\\ \frac{1}{\beta^t}\mathbf{\Lambda}_{21}^t & \mathbf{U}^t+\frac{1}{\beta^t}\mathbf{\Lambda}_{22}^t\end{pmatrix}\end{aligned} \tag{A.16}$$

Based on the property of operator $\mathcal{C}^*$, equation (A16) can be simplified as

$$\begin{aligned}&\mathcal{C}^*\begin{pmatrix}\frac{1}{\beta^t}\mathbf{\Lambda}_{11}^t-\mathbf{\Theta}^{t+1} & \frac{1}{\beta^t}\mathbf{\Lambda}_{12}^t\\ \frac{1}{\beta^t}\mathbf{\Lambda}_{21}^t & \mathbf{U}^t+\frac{1}{\beta^t}\mathbf{\Lambda}_{22}^t\end{pmatrix}\\ &=\frac{1}{\beta^t}\mathbf{\Lambda}_{11}^t-\mathbf{\Theta}^{t+1}-\mathbf{U}^t-\frac{1}{\beta^t}\mathbf{\Lambda}_{22}^t.\end{aligned} \tag{A.17}$$

Substituting (A17) into (A15), we can obtain

$$\begin{aligned}\mathcal{C}^*(\mathcal{C}(\mathbf{Z})) = &-\frac{1}{\beta^t}\mathbf{\Lambda}_{11}^t+\mathbf{\Theta}^{t+1}+\mathbf{U}^t+\frac{1}{\beta^t}\mathbf{\Lambda}_{22}^t\\ &-\frac{1}{\eta\beta^t}(\mathbf{Z}-\mathbf{V}^t+\eta\triangledown F(\mathbf{V}^t)),\end{aligned} \tag{A.18}$$

with

$$\mathcal{C}^*(\mathcal{C}(\mathbf{Z})) = \mathcal{C}^*\begin{pmatrix}\mathbf{Z} & \mathbf{0}\\ \mathbf{0} & -\mathbf{Z}\end{pmatrix} = 2\mathbf{Z}. \tag{A.19}$$

Similar to (A17), (A19) is also obtained by using the property of operator $\mathcal{C}^*$. Based on (A18) and (A19), we can finally obtain

$$\mathbf{Z}^{t+1} = \tau\Big(-\frac{1}{\beta^t}\mathbf{\Lambda}_{11}^t + \mathbf{\Theta}^{t+1} + \mathbf{U}^t + \frac{1}{\beta^t}\mathbf{\Lambda}_{22}^t + \frac{1}{\eta\beta^t}\mathbf{V}^t - \eta \bigtriangledown F(\mathbf{V}^t)\Big), \tag{A.20}$$

where $\tau = \frac{\eta\beta^t}{1+2\eta\beta^t}$.

In **U**-step, $\mathbf{U}^{t+1}$ is updated by solving the following problem

$$\min_{\mathbf{U}} \lambda_1 Tr(\mathbf{ULU}^T) + \frac{\beta^t}{2}\|\mathcal{B}(\mathbf{U}) + \mathcal{C}(\mathbf{Z}^{t+1}) - \mathcal{D}(\mathbf{\Theta}^{t+1}) + \frac{1}{\beta^t}\mathbf{\Lambda}^t\|_F^2. \tag{A.21}$$

Problem (A21) is differentiable and can be solved by setting its first order derivative to zero, obtaining

$$\lambda_1\mathbf{UL} + \beta^t\mathcal{B}^*(\mathcal{B}(\mathbf{U}) + \mathcal{C}(\mathbf{Z}^{t+1}) - \mathcal{D}(\mathbf{\Theta}^{t+1}) + \frac{1}{\beta^t}\mathbf{\Lambda}^t) = 0, \tag{A.22}$$

where $\mathcal{B}^*(\cdot)$ is another adjoint operator. The property of this operator is also shown in Appendix B. Similar to (A15), we rearrange (A22) as

$$\mathcal{B}^*(\mathcal{B}(\mathbf{U})) = -\frac{\beta^t}{\lambda_1}\mathbf{UL} - \mathcal{B}^*(\mathcal{C}(\mathbf{Z}^{t+1}) - \mathcal{D}(\mathbf{\Theta}^{t+1}) + \frac{1}{\beta^t}\mathbf{\Lambda}^t), \tag{A.23}$$

where

$$\begin{aligned}&\mathcal{B}^*(\mathcal{C}(\mathbf{Z}^{t+1}) - \mathcal{D}(\mathbf{\Theta}^{t+1}) + \frac{1}{\beta^t}\mathbf{\Lambda}^t)\\ &= \mathcal{B}^*\begin{pmatrix}\frac{1}{\beta^t}\mathbf{\Lambda}_{11}^t + \mathbf{Z}^{t+1} - \mathbf{\Theta}^{t+1} & \frac{1}{\beta^t}\mathbf{\Lambda}_{12}^t\\ \frac{1}{\beta^t}\mathbf{\Lambda}_{21}^t & \frac{1}{\beta^t}\mathbf{\Lambda}_{22}^t - \mathbf{Z}^{t+1}\end{pmatrix}.\end{aligned} \tag{A.24}$$

Based on the property of operator $\mathcal{B}^*$, equation (A24) can be simplified as

$$\begin{aligned}&\mathcal{B}^*\begin{pmatrix}\frac{1}{\beta^t}\mathbf{\Lambda}_{11}^t + \mathbf{Z}^{t+1} - \mathbf{\Theta}^{t+1} & \frac{1}{\beta^t}\mathbf{\Lambda}_{12}^t\\ \frac{1}{\beta^t}\mathbf{\Lambda}_{21}^t & \frac{1}{\beta^t}\mathbf{\Lambda}_{22}^t - \mathbf{Z}^{t+1}\end{pmatrix}\\ &= \frac{1}{\beta^t}\mathbf{\Lambda}_{22}^t - \mathbf{Z}^{t+1}.\end{aligned} \tag{A.25}$$

Substituting (A25) into (A23), we can obtain

$$\mathcal{B}^*(\mathcal{B}(\mathbf{U})) = -\frac{\beta^t}{\lambda_1}\mathbf{UL} - \frac{1}{\beta^t}\mathbf{\Lambda}_{22}^t + \mathbf{Z}^{t+1}, \tag{A.26}$$

with

$$\mathcal{B}^*(\mathcal{B}(\mathbf{U}) = \mathbf{U}. \tag{A.27}$$

where (A27) is obtained by using the property of operator $\mathcal{B}^*$. Based on (A26) and (A27), we finally obtain

$$\mathbf{U}^{t+1} = \Big(\mathbf{I} + \frac{\beta^t}{\lambda_1}\mathbf{L}\Big)^{-1}(\mathbf{Z}^{t+1} - \frac{1}{\beta^t}\mathbf{\Lambda}_{22}^t). \tag{A.28}$$

Appendix B

Here, we discuss the property of adjoint operators B^* and C^*.

Let $\mathcal{C}^*(\cdot)$ and $\mathcal{B}^*(\cdot)$ be the adjoint operators of $\mathcal{C}(\cdot)$ and $\mathcal{B}(\cdot)$, respectively. Inspired by [41], we have the following property

$$< \mathcal{C}(\mathbf{Z}), \mathbf{\Lambda} >=< \mathbf{Z}, \mathcal{C}^*(\mathbf{\Lambda}) > . \tag{B.1}$$

$$< \mathcal{B}(\mathbf{U}), \mathbf{\Lambda} >=< \mathbf{U}, \mathcal{B}^*(\mathbf{\Lambda}) > . \tag{B.2}$$

Through the definition of operator $\mathcal{C}(\cdot)$ and $\mathcal{B}(\cdot)$ in equation (15), we can obtain

$$\begin{aligned}< \mathcal{C}(\mathbf{Z}), \mathbf{\Lambda} > &= Tr\left(\begin{pmatrix}\mathbf{Z} & \mathbf{0}\\ \mathbf{0} & -\mathbf{Z}\end{pmatrix}\begin{pmatrix}\mathbf{\Lambda}_{11} & \mathbf{\Lambda}_{12}\\ \mathbf{\Lambda}_{21} & \mathbf{\Lambda}_{22}\end{pmatrix}^T\right)\\ &= Tr(\mathbf{Z}\mathbf{\Lambda}_{11}^T - \mathbf{Z}\mathbf{\Lambda}_{22}^T)\\ &=< \mathbf{Z}, \mathbf{\Lambda}_{11} - \mathbf{\Lambda}_{22} > .\end{aligned} \tag{B.3}$$

$$\begin{aligned}< \mathcal{B}(\mathbf{U}), \mathbf{\Lambda} > &= Tr\left(\begin{pmatrix}\mathbf{0} & \mathbf{0}\\ \mathbf{0} & \mathbf{U}\end{pmatrix}\begin{pmatrix}\mathbf{\Lambda}_{11} & \mathbf{\Lambda}_{12}\\ \mathbf{\Lambda}_{21} & \mathbf{\Lambda}_{22}\end{pmatrix}^T\right)\\ &= Tr(\mathbf{U}\mathbf{\Lambda}_{22}^T)\\ &=< \mathbf{U}, \mathbf{\Lambda}_{22} > .\end{aligned} \tag{B.4}$$

Based on (A1) and (A3), the adjoint operator $\mathcal{C}^*(\cdot)$ can be calculated as

$$\mathcal{C}^*(\mathbf{\Lambda}) = \mathbf{\Lambda}_{11} - \mathbf{\Lambda}_{22}. \tag{B.5}$$

Based on (A2) and (A4), $\mathcal{B}^*(\cdot)$ can be calculated as

$$\mathcal{B}^*(\mathbf{\Lambda}) = \mathbf{\Lambda}_{22}. \tag{B.6}$$

Appendix C

Let $\mathcal{L}^{t+1} = \mathcal{L}(\mathbf{\Theta}^{t+1}, \mathbf{U}^{t+1}, \mathbf{Z}^{t+1}, \mathbf{P}^{t+1}, \mathbf{\Lambda}^t, \beta^t)$ and $e^t = \|\mathcal{B}(\mathbf{U}^t) + \mathcal{C}(\mathbf{Z}^t) - \mathcal{D}(\mathbf{\Theta}^t)\|_F^2$. We want to prove that the augmented Lagrange function in Algotithm 1 is bounded, which means that

$$\mathcal{L}^{t+1} - \mathcal{L}^t \le \frac{\beta^t + \beta^{t-1}}{2}e^t \quad t = 0, 1, \ldots, n.$$

Proof. Given

$$\mathcal{L}^{t+1} = \mathcal{L}(\mathbf{\Theta}^{t+1}, \mathbf{U}^{t+1}, \mathbf{Z}^{t+1}, \mathbf{P}^{t+1}, \mathbf{\Lambda}^t, \beta^t), \tag{C.1}$$

we can obtain

$$\begin{aligned}\mathcal{L}^{t+1} &\le \mathcal{L}(\mathbf{\Theta}^{t+1}, \mathbf{U}^{t+1}, \mathbf{Z}^{t+1}, \mathbf{P}^t, \mathbf{\Lambda}^t, \beta^t)\\ &\le \mathcal{L}(\mathbf{\Theta}^{t+1}, \mathbf{U}^{t+1}, \mathbf{Z}^t, \mathbf{P}^t, \mathbf{\Lambda}^t, \beta^t)\\ &\le \mathcal{L}(\mathbf{\Theta}^{t+1}, \mathbf{U}^t, \mathbf{Z}^t, \mathbf{P}^t, \mathbf{\Lambda}^t, \beta^t)\\ &\le \mathcal{L}(\mathbf{\Theta}^t, \mathbf{U}^t, \mathbf{Z}^t, \mathbf{P}^t, \mathbf{\Lambda}^t, \beta^t)\\ &= \mathcal{L}^t + < \mathbf{\Lambda}^t - \mathbf{\Lambda}^{t-1}, \mathcal{B}(\mathbf{U}^t) + \mathcal{C}(\mathbf{Z}^t) - \mathcal{D}(\mathbf{\Theta}^t) >\\ &\quad + \frac{\beta^t - \beta^{t-1}}{2}\|\mathcal{B}(\mathbf{U}^t) + \mathcal{C}(\mathbf{Z}^t) - \mathcal{D}(\mathbf{\Theta}^t)\|_F^2\\ &= \mathcal{L}^t + \beta^{t-1}\|\mathcal{B}(\mathbf{U}^t) + \mathcal{C}(\mathbf{Z}^t) - \mathcal{D}(\mathbf{\Theta}^t)\|_F^2\\ &\quad + \frac{\beta^t - \beta^{t-1}}{2}\|\mathcal{B}(\mathbf{U}^t) + \mathcal{C}(\mathbf{Z}^t) - \mathcal{D}(\mathbf{\Theta}^t)\|_F^2.\end{aligned} \tag{C.2}$$

Therefore

$$\mathcal{L}^{t+1} - \mathcal{L}^t \le \frac{\beta^t + \beta^{t-1}}{2}e^t \quad t = 0, 1, \ldots, n. \tag{C.3}$$

To prove e^t is bounded, we should prove Λ^t is bounded. This proof is similar to Lemma 1 in [52]. Based on this observation, we use Theorem 4 of [53] to prove Λ^t is bounded. Hence $e^t = (\frac{\Lambda^t - \Lambda^{t-1}}{\beta^{t-1}})^2$ is bounded.

This proof implies the upperbound of augmented Lagrange function. Based on [43], if $\sum_{k=1}^{\infty}(\beta^t)^{-2}\beta^{k+1} < +\infty$, the upperbound of augmented Lagrange function can imply that any accumulation points of $\mathbf{U}^t$, $\mathbf{Z}^t$, $\mathbf{P}^t$ and $\mathbf{\Theta}^t$ can approach a feasible solution. □

References

[1] L. Zhou, On data-driven delay estimation for media cloud, IEEE Trans. Multimed. 18 (5) (2016) 905–915.
[2] L. Zhou, Qoe-driven delay announcement for cloud mobile media, IEEE Trans. Circuits Syst. Video Technol. 27 (1) (2017) 84–94.
[3] L. Zhou, Mobile device-to-device video distribution: theory and application, ACM Trans. Multimed. Comput., Commun. Appl. 12 (3) (2015) 1253–1271.

[4] S. Avidan, Support vector tracking, IEEE Trans. Pattern Anal. Mach. Intell. 26 (8) (2004) 1064–1072.
[5] S. Zhang, X. Yu, Y. Sui, S. Zhao, L. Zhang, Object tracking with multi-view support vector machines, IEEE Trans. Multimed. 17 (3) (2015) 265–278.
[6] H. Grabner, H. Bischof, On-line boosting and vision, in: IEEE Conference on Computer Vision and Pattern Recognition, 2006.
[7] H. Grabner, C. Leistner, H. Bischof, Semi-supervised on-line boosting for robust tracking, in: Proc. ECCV, 2008.
[8] F. Yang, H. Lu, M. Yang, Robust visual tracking via multiple kernel boosting with affinity constraints, IEEE Trans. Circuits Syst. Video Technol. 24 (2) (2014) 242–254.
[9] B. Babenko, Y. Ming-Hsuan, S. Belongie, Visual tracking with online multiple instance learning, in: IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 983–990.
[10] K. Zhang, H. Song, Real-time visual tracking via online weighted multiple instance learning, Pattern Recognit. 46 (1) (2013) 397–411.
[11] K. Zhang, L. Zhang, M.H. Yang, Fast compressive tracking, IEEE Trans. Pattern Anal. Mach Intell 36 (10) (2014) 2002–2015.
[12] J.F. Henriques, R. Caseiro, P. Martins, J. Batista, High-speed tracking with kernelized correlation filters, IEEE Trans. Pattern Anal. Mach. Intell. 37 (3) (2015) 583–596.
[13] M. Mueller, N. Smith, B. Ghanem, Context-aware correlation filter tracking, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.
[14] D.A. Ross, J. Lim, R.S. Lin, M.H. Yang, Incremental learning for robust visual tracking, Int. J. Comput. Vis. 77 (1) (2008) 125–141.
[15] Y. Xie, W. Zhang, Y. Qu, Y. Zhang, Discriminative subspace learning with sparse representation view-based model for robust visual tracking, Pattern Recognit. 47 (3) (2014) 1383–1394.
[16] Y. Sui, S. Zhang, L. Zhang, Robust visual tracking via sparsity-induced subspace learning, IEEE Trans. Image Process. 24 (12) (2015) 4686–4700.
[17] T. Zhang, S. Liu, A. Narendra, M.H. Yang, G. Bernard, Roubust visual tracking vis consistent low-rank sparse learning, Int. J. Comput. Vis. 111 (2) (2014) 171–190.
[18] Y. Wu, B. Shen, H. Ling, Visual tracking via online nonnegative matrix factorization, IEEE Trans. Circuits Syst. Video Technol. 24 (3) (2014) 374–383.
[19] H. Zhang, S. Hu, X. Zhang, L. Luo, Visual tracking via constrained incremental non-negative matrix factorization, IEEE Signal Process. Lett. 22 (9) (2015) 1350–1353.
[20] S. Zhang, H. Yao, X. Sun, X. Lu, Sparse coding base visual tracking: review and experimental comparison, Pattern Recognit. 46 (7) (2013) 1772–1788.
[21] Z. Chi, H. li, H. Lu, M. Yang, Dual deep network for visual tracking, IEEE Trans. Image Process. 26 (4) (2017) 2005–2015.
[22] C. Ma, J. Huang, X. Yang, M.H. Yang, Hierarchical convolutional features for visual tracking, in: IEEE International Conference on Computer Vision (ICCV), 2015.
[23] L. Bertinetto, J. Valmadre, J.F. Henriques, A. Vedaldi, P. Torr, Fully-convolutional siamese networks for object tracking, in: European Conference on Computer Vision (ECCV), 2016.
[24] X. Mei, H. Ling, Robust visual tracking and vehicle classification via sparse representation, IEEE Trans. Pattern Anal. Mach. Intell. 33 (11) (2011) 2259–2272.
[25] C. Bao, Y. Wu, H. Ling, H. Ji, Real time robust l1 tracker using accelerated proximal gradient approach, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2012, pp. 1830–1837.
[26] T. Zhang, G. Bernard, S. Liu, A. Narendra, Robust visual tracking via structured multi-task sparse learning, Int. J. Comput. Vis. 101 (2) (2013) 367–383.
[27] B. Ma, J. Shen, Y. Liu, H. Hu, L. Shao, X. Li, Visual tracking using strong classifier and structural local sparse descriptors, IEEE Trans. Multimed. 17 (10) (2015) 1818–1828.
[28] S. Zhang, H. Zhou, F. Jiang, X. Li, Robust visual tracking using structurally random projection and weighted least squares, IEEE Trans. Circuits Syst. Video Technol. 25 (11) (2015) 1749–1760.
[29] T. Zhang, A. Bibi, B. Ghanem, In defense of sparse tracking: circulant sparse tracker, IEEE Computer Vision and Patter Recognition, 2016.
[30] Y. Zhou, J. Han, X. Yuan, Z. Wei, R. Hong, Inverse sparse group lasso model for robust object tracking, IEEE Trans. Multimed. 19 (8) (2017) 1798–1810.
[31] X. Mei, Z. Hong, D. Prokhorov, D. Tao, Robust multitask multiview tracking in videos, IEEE Trans. Neural Netw. Learn. Syst. 26 (11) (2015) 2874–2890.
[32] W. Hu, W. Li, X. Zhang, S.J. Maybank, Single and multiple object tracking using a multi-feature joint sparse representation, IEEE Trans. Pattern Anal. Mach. Intell. 37 (4) (2015) 816–833.
[33] X. Lan, A.J. Ma, P.C. Yuan, R. Chellappa, Joint sparse representation and robust feature-level fusion for multi-cue visual tracking, IEEE Trans. Image Process. 24 (12) (2015) 5826–5841.
[34] F. Li, H. Lu, D. Wang, Y. Wu, K. Zhang, Dual group structured tracking, IEEE Trans. Circuits Syst. Video Technol. 26 (9) (2016) 1697–1708.
[35] B. Kang, D. Liang, S. Zhang, Robust visual tracking via multi-view dicriminant based sparse representation, in: International Conference on Image Processing (ICIP), 2017.
[36] M. Kan, S. Shan, H. Zhang, S. Lao, X. Chen, Multi-view discriminant analysis, IEEE Trans. Pattern Anal. Mach. Intell. 38 (1) (2016) 188–194.
[37] Z. Yang, M. Jacob, Nonlocal regularization of inverse problems: a unified variational framework, IEEE Trans. Image Process. 22 (8) (2013) 3192–3203.
[38] M.A.T. Figueiredo, J.M. Bioucas-Dias, R.D. Nowak, Majorization-minimization algorithms for wavelet based image restoration, IEEE Trans. Image Process. 16 (12) (2007) 2980–2991.
[39] S. Gao, W.H. Tsang, L.T. Chia, Laplacian sparse coding, hypergraph laplacian sparse coding, and applications, IEEE Trans. Pattern Anal. Mach. Intell. 35 (1) (2013) 92–104.
[40] M. Afonso, J. Bioucas-Dias, M. Figueiredo, An augmented lagrangian approach to the constrained optimization formulation of imaging inverse problem, IEEE Trans. Image Process. 20 (3) (2011) 681–695.
[41] Y. Hu, D. Zhang, J. Ye, X. Li, X. He, Fast and accurate matrix completion via truncated nuclear norm regularization, IEEE Trans. Pattern Anal. Mach. Intell. 35 (9) (2013) 2117–2130.
[42] G. Liu, Z. Lin, Y. Yu, Robust subspace segmentation by low-rank representation, in: International Conference on Machine Learning, 2010.
[43] M. Tao, X. Yuan, Recovering low-rank and sparse components of matrices from incomplete and noisy observations, SIAM J. Optim. 21 (1) (2011) 57–81.
[44] Y. Wu, J. Lim, M.H. Yang, Online object tracking: a benchmark, in: Proc. CVPR, 2013.
[45] S. Hare, A. Saffari, P.H.S. Torr, Struck: structured output tracking with kernels, in: IEEE International Conference on Computer Vision (ICCV), 2011.
[46] J. Henriques, R. Caseiro, P. Martins, J. Batista, Exploiting the circulant structure of tracking-by-detection with kernel, in: European Conference on Computer Vision (ECCV), 2012.
[47] Z. Kalal, K. Mikolajczyk, J. Matas, Tracking-learning-detection, IEEE Trans. Pattern Anal. Mach. Intell. 34 (7) (2012) 1409–1422.
[48] A. Adam, E. Rivlin, I. Shimshoni, Robust fragments-based tracking using the integral histogram, in: IEEE Conference on Computer Vision and Pattern Recognition, 2006.
[49] D. Comaniciu, V. Ramesh, P. Meer, Kernel-based object tracking, IEEE Trans. Pattern Anal. Mach. Intell. 25 (2003) 564–577.
[50] H. Grabner, M. Grabner, H. Bischof, Real-time tracking via on-line boosting, in: British Machine Vision Conference, 2006.
[51] T. Ojala, M. Pietikainen, T. Maepaa, Multiresolution gray-scale and rotation invariant texture classification with local binary patterns, IEEE Trans. Pattern Anal. Mach. Intell. 24 (7) (2002) 971–987.
[52] O. Oreifej, X. Li, M. Shah, Simultaneous video stabilization and moving object detection in turbulence, IEEE Trans. Pattern Anal. Mach. Intell. 35 (2) (2013) 450–462.
[53] Z. Lin, M. Chen, L. Wu, Y. Ma, The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices, 2009, In UIUC Technical Report.
[54] X. Yuan, X. Liu, S. Yan, Visual classification with multitask joint sparse representation, IEEE Trans. Image Process. 21 (10) (2012) 4349–4360.
[55] C. Xu, D. Tao, C. Xu, Large-margin multi-view information bottleneck, IEEE Trans. Pattern Anal. Mach. Intell. 36 (8) (2014) 1559–1572.

Bin Kang: received the M.S. degree in Circuits and Systems, and the Ph.D. degree in Electrical Engineering from Lanzhou University and Nanjing University of Posts and Telecommunications, in 2011 and 2016, respectively. He is currently a lecturer at College of Internet of Things, Nanjing University of Posts and Telecommunications. His research interests include computer vision and pattern recognition.

Wei-Ping Zhu: received the B.E. and M.E. degrees from Nanjing University of Posts and Telecommunications, and the Ph.D. degree from Southeast University, Nanjing, China in 1982, 1985 and 1991, respectively, all in Electrical Engineering. He was a post-doctoral fellow from 1991 to 1992 and a research associate from 1996 to 1998 in the Department of Electrical and Computer Engineering, Concordia University, Montreal, Canada. During 1993–1996, he was an Associate Professor in the Department of Information Engineering, Nanjing University of Posts and Telecommunications. From 1998 to 2001, he worked in telecommunication industries in Ottawa, Canada, including Nortel Networks and SR Telecom Inc. Since 2001, he has been a full-time Faculty Member with the Department of Electrical and Computer Engineering, Concordia University, where he is currently a Full Professor. Since 2008, he has been also an Adjunct Professor with the Nanjing University of Posts and Telecommunications. His research interests include digital signal processing fundamentals and image processing.

Dong Liang: received the B.S. degree in Telecommunication Engineering and the M.S. degree in Circuits and Systems from Lanzhou University, China, in 2008 and 2011, respectively. In 2015, he received Ph.D. at Graduate School of IST, Hokkaido University, Japan. He is now an Assistant Professor in Pattern Recognition and Neural Computing Lab., College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics (NUAA). His research interests include computer vision and pattern recognition.

Mingkai Chen: received the M.S. degree in Electrical Engineering from Fuzhou University. He is studying for the Ph.D. course at Department of Communication and Information Engineering, Nanjing University of Posts and Telecommunications, China.