Weighted Joint Sparse Representation Based Visual Tracking

Xiping Duan[1,2](✉), Jiafeng Liu[1], and Xianglong Tang[1]

[1] School of Computer Science and Technology,
Harbin Institute of Technology, Harbin, China
xpduan_1999@126.com, {jefferyliu,tangxl}@hit.edu.cn

[2] College of Computer Science and Information Engineering,
Harbin Normal University, Harbin, China

Abstract. Aiming at various tracking environments, a weighted joint sparse representation based tracker is proposed. Specifically, each object template is weighted according to its similarity to each candidate. Then all candidates are represented sparsely and jointly, and the sparse coefficients are used to compute the observation probabilities of candidates. The candidate with the maximum observation probability is determined as the object. The object function is solved by a modified accelerated proximal gradient (APG) algorithm. Experiments on several representative image sequences show that the proposed tracking method performs better than the other trackers in the scenarios of illumination variation, occlusion, pose change and rotation.

Keywords: Computer vision · Visual tracking · Kernel sparse representation

1 Introduction

Visual tracking is a hot topic of computer vision, and extensively applied into such fields as intelligent monitoring, car navigation, advanced human-computer interaction, and so forth. However, due to all kinds of internal and external factors, such as noises, occlusion, illumination and viewpoint variations, pose change, rotation, and so on, realizing the robust visual tracking is still a challenging task. The existing visual tracking methods can be categorized into the generative and the discriminative. The generative methods treat the visual tracking as a matching problem. For example, both Eigentracker [1] and meanshift tracker [2], which are proposed by Black and Comaniciu respectively, have desired real-time performance, but lack the necessary updating of target templates. IVT [3] proposed by Ross, adapts to the appearance change by incrementally learning a low dimensional domain. VTD [4] proposed by Kwon extends the traditional particle filter tracker with multiple motion models and multiple appearance models. The discriminative methods are regarded as a binary classification problem, and need to learn and update a classifier by sampling positive and negative samples. The representative methods are: Avidan et al. [5] proposed a SVM based tracker. Collins et al. [6] discriminate the target from background by online feature selection. Grabner et al. [7] proposed a online semi-supervised boosting method to avoid drifting away from the target and improve tracking performance. Babenko et al. [8] avoid bias and

S. Arik et al. (Eds.): ICONIP 2015, Part III, LNCS 9491, pp. 600–609, 2015.
DOI: 10.1007/978-3-319-26555-1_68

drift by multi-instance learning. Zhang et al. [9] formulate the tracking task as a binary classification problem in the compressed domain and obtain the real-time performance.

Recently, sparse representation based visual tracking methods are very popular. For instance, in [10–12], each candidate is represented sparsely as the linear combination of templates, with excellent tracking results. Due to the small sampling radium, candidates are similar to each other. [13] mines the similarity between candidates by joint sparse representation, improving the representation accuracy of candidates and the robustness of the whole tracking system. However, the aforementioned sparse representation based tracking methods assume that different templates have the same possibilities to represent each candidate. According to [14], the template which is similar to a candidate should have larger probability to represent the candidate.

In this paper, a weighted joint sparse representation based visual tracking method is proposed to robustly track the target in the complex environment. Specifically, to represent a candidate, all the templates are weighted according to their similarity to the candidate. Then, all the candidates are represented by the joint sparse representation to reflect the similarity among them. The sparse coefficient is used to compute the observation probability of the corresponding candidate. The candidate with the maximal observation probability is determined as the target. In this model, a template similar to a candidate has the larger possibility to represent the candidate. To adapt to the illumination change, a locally normalized feature is used for appearance representation, which is an illumination invariant feature. Experiments show that the proposed tracker can well adapt to the possible illumination variation, occlusion, pose change, rotation, and the influence of comprehensive conditions.

2 Weighted Joint Sparse Representation Based Tracker

2.1 Weighted Joint Sparse Representation

Let $\mathrm{T} = \{\mathrm{T}_1, \mathrm{T}_2, \ldots, \mathrm{T}_M\}$ and $\{y_1, y_2, \ldots, y_N\}$ denote the target templates set and candidates of current frame respectively. To locate the target accurately, Zhang et al. proposed a multi-task learning based tracking method, which takes into account the similarity relation among candidates and represents all the candidates by solving the $\ell_{2,1}$ mixed norm regularized minimization problem [13].

$$\hat{c} = \arg\min_{c} \frac{1}{2} \sum_{i=1}^{N} \|\mathbf{y}_i - \mathbf{T}c_i\|_2^2 + \lambda \sum_{j=1}^{J} \left\|c^j\right\|_2 \tag{1}$$

where

$$c = [c_1, c_2, \ldots, c_N] = \left[c^1; c^2; \ldots, c^J\right] = \begin{pmatrix} c_1^1 & \cdots & c_N^1 \\ \vdots & \ddots & \vdots \\ c_1^J & \cdots & c_N^J \end{pmatrix}$$

is the sparse representation coefficient matrix. Specifically, the i th column c_i of c corresponds to the representation coefficient of the candidate $\mathbf{y}_i$, the j row c^j are the coefficients of all the candidates corresponding to the target template T_j, and c_i^j is the j th representation coefficient of the candidate $\mathbf{y}_i$ corresponding to the target template T_j. The regularization term of (2) is $\ell_{2,1}$ mixed norm, which can cause the joint sparse representation of candidates. By mining the similarity relation among candidates, [13] improves the representation accuracy and the tracking robustness compared with other sparse representation based trackers [10–12]. However, all the aforementioned trackers [10–13] assume that all the target templates have the same possibilities to represent a candidate. According to [14], the target template more similar to a candidate has larger possibility to be chosen to represent the candidate. So in this paper, before representing a candidate, all the target templates are weighted using their distance to the candidate. Then, considering the similarity relation among candidates, all the candidates are represented using joint sparse representation.

$$\hat{c} = \arg\min_{c} \frac{1}{2} \sum_{i=1}^{N} \|\mathbf{y}_i - \mathbf{T}c_i\|_2^2 + \lambda \sum_{j=1}^{J} \left\| w^j \odot c^j \right\|_2 \tag{2}$$

Compare with (1), the weight matrix

$$w = [w_1, w_2, \ldots, w_N] = \left[w^1; w^2; \ldots, w^J\right] = \begin{pmatrix} w_1^1 & \cdots & w_N^1 \\ \vdots & \ddots & \vdots \\ w_1^J & \cdots & w_N^J \end{pmatrix}$$

is introduced into (2), where w_i^j is the weight of the target template T_j representing the candidate $\mathbf{y}_i$, and calculated as the distance between them

$$w_i^j = \exp\left(\frac{\left\|y_i - \mathrm{T}_j\right\|_2}{\sigma}\right)$$

The internal $\left\|y_i - \mathrm{T}_j\right\|_2$ is the Euclidean distance between T_j and $\mathbf{y}_i$.

The sparse representation coefficients $\{c_i\}_{i=1}^N$ of candidates $\{\mathrm{y}_i\}_{i=1}^N$ are obtained by solving (2), and used to calculate the reconstruction errors $\{r_i\}_{i=1}^N$ of candidates.

$$r_i = \|\mathrm{y}_i - T \cdot c_i\|_2 \tag{3}$$

r_i is further used for calculating the observation probability of the candidate y_i.

$$P(\mathbf{y}_i|o) \propto \mu \cdot \exp(-r_i) \tag{4}$$

where μ is the normalization constant. Among a group of given candidates $\{y_1, y_2, \ldots, y_N\}$, the candidate y_i with minimum observation probability

$$i = \arg\min_{j} P\left(\mathbf{y}_j|o\right) \tag{5}$$

is determined as the target.

2.2 Solving of Objective Function

The key for the aforementioned weighted joint sparse representation is to solve (2), the corresponding objective function is denoted as

$$f(c) = \frac{1}{2}\sum_{i=1}^{N} \|\mathbf{y}_i - \mathbf{T}c_i\|_2^2 + \lambda\sum_{j=1}^{J} \left\|w^j \odot c^j\right\|_2 \tag{6}$$

To solve (6), an auxiliary function is constructed

$$g(c) = \frac{1}{2}\sum_{i=1}^{N} \left\|\mathbf{y}_i - \mathbf{T} * (w_i)^{-1} \odot c_i\right\|_2^2 + \lambda\sum_{j=1}^{J} \left\|c^j\right\|_2 \tag{7}$$

where $(w_i)^{-1} = \left[\frac{1}{w_i^1}; \frac{1}{w_i^2}; \ldots; \frac{1}{w_i^{p+n}}\right]$, and $\odot$ is the element-wise multiplication operator, that is, the multiplication is conducted among corresponding elements of two matrices or two vectors. The same as (1), $g(c)$ is a multi-task joint sparse representation problem, and can be solved by a modified APG algorithm [15]. Suppose the optimal solutions of $g(c)$ and $f(c)$ are $\hat{c}' = \arg\min_{c} g(c)$ and $\hat{c} = \arg\min_{c} f(c)$ respectively, then similar to [13], there is the following correspondence

$$\hat{c} = w^{-1} \odot \hat{c}'$$

where

$$\begin{aligned} w^{-1} &= \left[(w_1)^{-1}, (w_2)^{-1}, \ldots, (w_N)^{-1}\right] \\ &= \left[\left(w^1\right)^{-1}; \left(w^2\right)^{-1}; \ldots; \left(w^{p+n}\right)^{-1}\right] \\ &= \begin{pmatrix} \frac{1}{w_1^1} & \cdots & \frac{1}{w_N^1} \\ \vdots & \ddots & \vdots \\ \frac{1}{w_1^{p+n}} & \cdots & \frac{1}{w_N^{p+n}} \end{pmatrix} \end{aligned} \tag{8}$$

2.3 Template Updating

A simple template updating strategy is adopted. (1) The target of the 1st frame is added into the target template set and not allowed be updated. (2) Determine whether the target appearance of the current frame greatly changes or not. If the amount of change is greater than a given threshold, update target templates set. Otherwise, not update. Specifically, if the similarity degree of the target y_i and the target template T_j is smaller than a given threshold τ, then the target y_i is used to update the target template T_k, where

$$\begin{aligned} i &= \arg\min_j P(\mathbf{y}_j|o), \\ j &= \arg\min_l \|\mathrm{T}_l - \mathrm{y}_i\|_2, \\ k &= \arg\max_l \|\mathrm{T}_l - \mathrm{y}_i\|_2. \end{aligned}$$

2.4 Tracking Algorithm

In this paper, the locally normalized feature is adopted for appearance representation, as shown in Fig. 1. In particular, the candidate or target template is firstly divided into several regions. Secondly the vector representation of each region is obtained by concatenating all the columns sequentially, and normalizing to the ℓ_2 unit length. Further the normalized vectors are concatenated into a longer vector as the feature vector of the candidate or target template.

The detailed algorithm of this paper is shown in Algorithm 1.

Algorithm 1 Weighted joint sparse representation based visual tracking algorithm

(1) Initialization: The target state x_0 of the 1st frame, initial target templates set $\mathrm{T} = \{\mathrm{T}_1, \mathrm{T}_2, \ldots, \mathrm{T}_M\}$;

(2) From the 2nd frame, execute the following steps continuously and alternately, till to the final frame.

a. Sample N candidates $\{x_1, x_2, \ldots, x_N\}$ within a circular region of radium r away from the target location of the previous frame, and obtain their feature vectors $\{y_1, y_2, \ldots, y_N\}$ as shown in Fig.1.

b. Obtain the sparse representation coefficient matrix $\hat{c}$ by solving (2), with the i th column c_i corresponding to the candidate y_i .

c. Compute the reconstruction error r_i and observation probability $P(\mathbf{y}_i|o)$ of the candidate x_i by (3) and (4) respectively.

d. Determine the candidate x_i satisfying $i = \arg\max_j P(y_j \mid o)$ as the target of current frame.

e. Adaptive template updating: Compute the similarity degree between the tracked target $\mathbf{y}_i$ and the most similar target template T_j ($j = \arg\min_l \|\mathrm{T}_l - \mathrm{y}_i\|_2$). If the similarity degree is less than the threshold τ , the target $\mathbf{y}_i$ is used to update target template set.

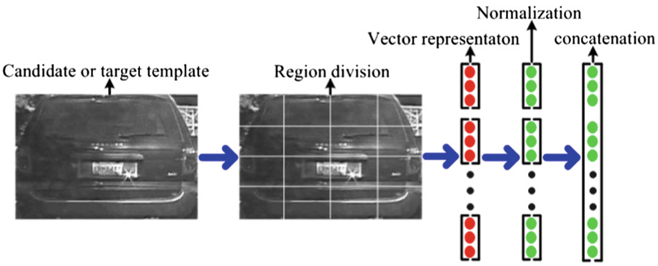


Fig. 1. Appearance representation

3 Experiments and Analyses

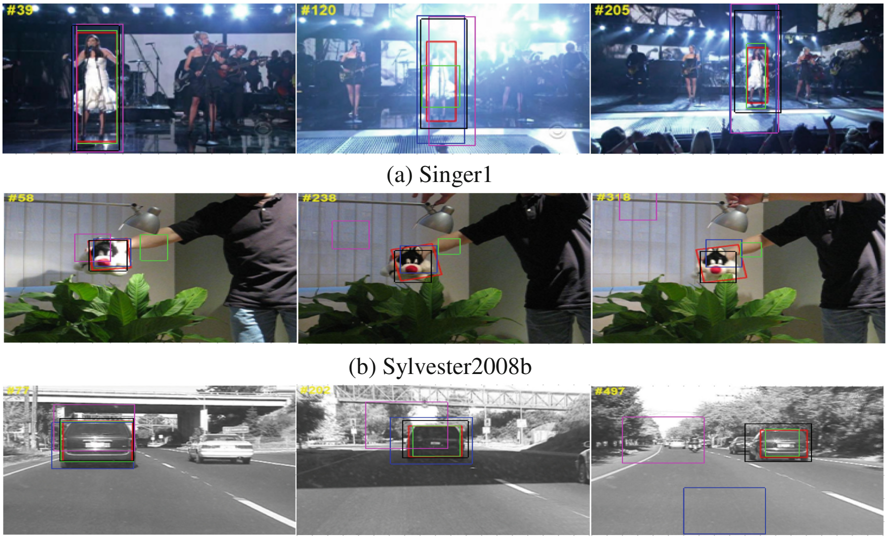
We implement the proposed tracker on the PC with AMD Sempron X2 198 CPU (2.5 GHz) and 3 GB memory, using MATLAB 2011b, and compare with other 4 trackers (MIL [8], CT [9], L1APG [10] and MTT [12]) in 3 different scenes to validate the performance. It should be mentioned that all the compared 4 trackers provide source codes. The related parameters are set to: the number M of target templates is 10, and the number of candidates N is 600. Each candidate or target template is scaled to 32*32 pixels and is divided into 4*4 regions to extract the locally normalized feature. The regularization parameter λ is set to 0.0002. The number of iterations in the optimization of the objective function is set to 5. And threshold τ of template updating is set to 0.4.

Scene 1: We select 3 videos (Singer1, Sylvester2008b and Car4) to validate the performance in the scene of illumination variation as shown in Fig. 2. It can be seen that the proposed tracker can track robustly under various illumination scenes. The reasons are two-folds. For one thing, the proposed method takes into account that the target template similar with the candidate has a greater possibility to represent the candidate, thus has higher reconstruction accuracy. For the other thing, the locally normalized feature, as an illumination invariant feature, can well adapt to the illumination variation.

Scene 2: In the videos Dudek and Sylvester2008b, the pose change and rotation of the tracked target take place as shown in Fig. 3. From the tracking result, the proposed tracker is more accurate than MTT, which is attributed to the weighting step before joint sparse representation of candidates.

Scene 3: In the videos Caviar1, Caviar2, Occlusion1 and Deduk, the tracked target may be occluded partially or severely as shown in Fig. 4. It can be concluded that the proposed tracker can reliably track in various occlusion scenes. The reason is that the proposed tracker makes the target template similar to the candidate has the greater possibility to represent the candidate, and obtains higher representation accuracy.

In the process of tracking, the tracked target suffers from influence of all kinds of comprehensive conditions actually. From the position error curve of Fig. 5, the pro-

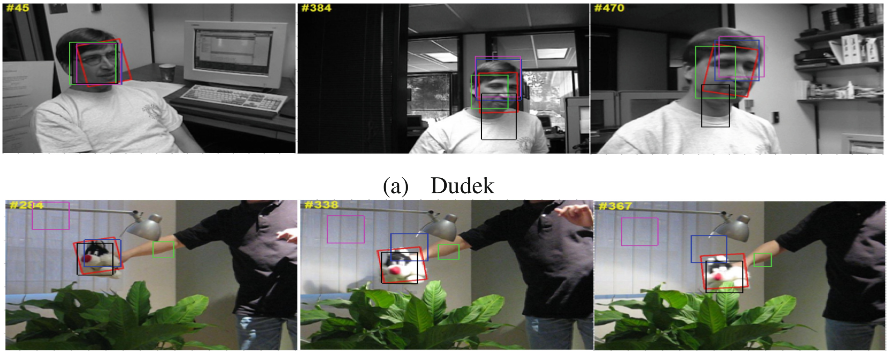


(a) Singer1

(b) Sylvester2008b

(c) Car4

Fig. 2. The tracking result in illumination variation

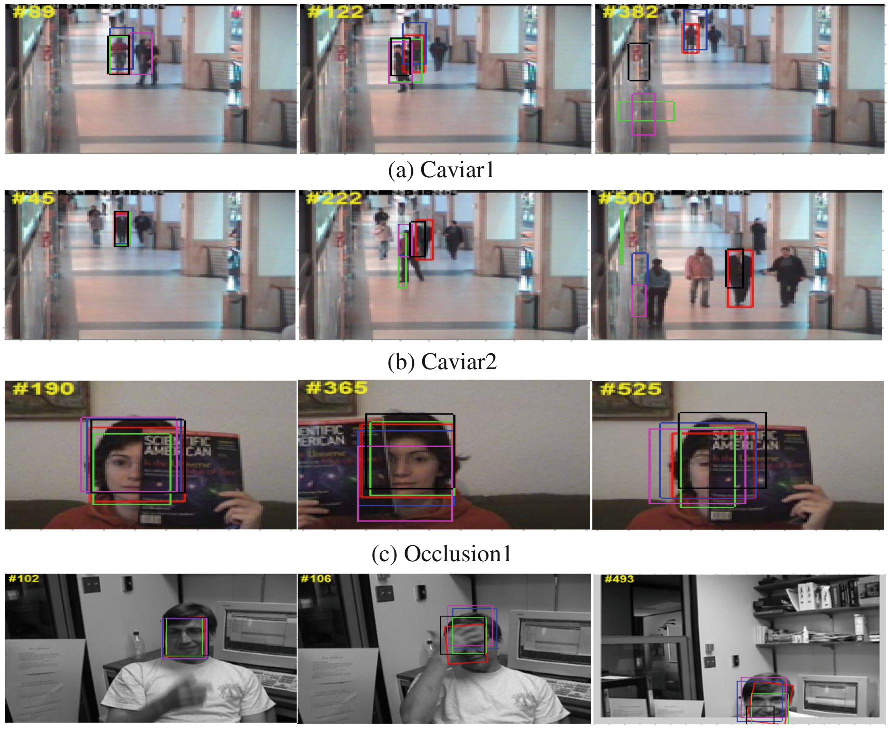


(a) Dudek

(b) Sylvester2008b

Fig. 3. The tracking result in the scene of the pose change and rotation

posed tracker can reliably track the target under various tracking conditions. Compared with other 4 trackers, the average tracking error of the proposed tracker is smallest, as shown in Table 1.



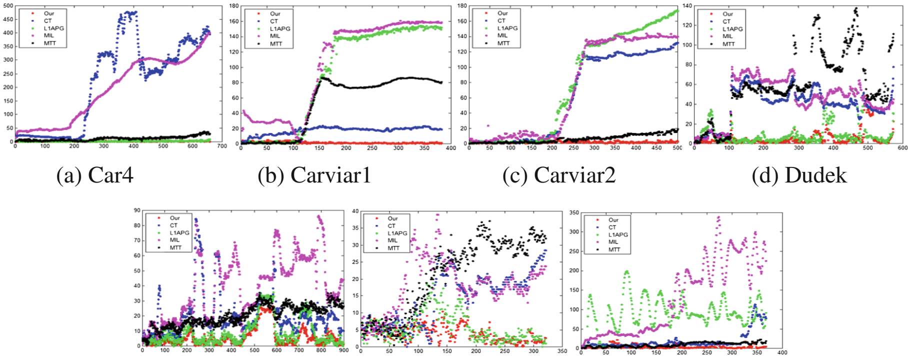
(a) Caviar1

(b) Caviar2

(c) Occlusion1

(d) Deduk

Fig. 4. The tracking result in the scene of occlusion



(a) Car4

(b) Carviar1

(c) Carviar2

(d) Dudek

(e) Occlusion1

(f) Singer1

(g) Sylvester2008b

Fig. 5. Position error curve

Table 1. Average tracking error (center point error, unit: pixel)

Sequences	CT	L1APG	MIL	MTT	Our
Car4	217.6553	2.7661	191.7701	12.0800	2.0373
Caviar1	17.2037	89.9965	103.6075	53.1033	1.4344
Caviar2	61.9205	77.0259	72.5072	6.1791	1.8378
Dudek	39.4552	8.2842	45.4436	58.8086	5.0877
Occlusions1	20.1084	8.8239	39.3074	20.0080	5.7169
Singer1	14.4684	5.7062	18.1904	21.2293	3.9776
Sylvester2008b	18.2245	89.8053	128.9541	9.5268	4.0926

4 Conclusions

Aiming at the complex tracking environment such as illumination changes, occlusion, pose changes, rotation, and so on, a weighted joint sparse representation based tracking method is proposed with 4 characteristics: (1) The similarity relation among candidates is reflected by joint sparse representation. (2) By weighting target template to reflect the different similarity degree between a target template and a candidate. (3) The objective function is solved by modifying the APG algorithm. (4) A simple template updating strategy is used to adapt to the target appearance change.

Acknowledgment. This work is supported by Scientific Research Fund of Heilongjiang Provincial Education Department (NO: 12541238), Dr. Scientific Research Foundation of Harbin Normal University (KGB201216), the National Science Foundation of China (61173087), Heilongjiang Provincial University Engineering R&D Center of Machine Vision and Intelligent Detection, and Heilongjiang Provincial Education Department Key Laboratory of Intelligent Education and Information Engineering.

References

1. Black, M.J., Jepson, A.D.: Eigentracking: robust matching and tracking of articulated objects using a view-based representation. Int. J. Comput. Vis. **26**(1), 63–84 (1998)
2. Comaniciu, D., Ramesh, V., Meer, P.: Kernel-based object tracking. IEEE Trans. Pattern Anal. Mach. Intell. **25**(5), 564–577 (2003)
3. Ross, D.A., Lim, J., Lin, R.S., Yang, M.H.: Incremental learning for robust visual tracking. Int. J. Comput. Vis. **77**(1–3), 125–141 (2008)
4. Kwon, J., Lee, K.M.: Visual tracking decomposition. In: 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1269–1276. IEEE (2010)
5. Avidan, S.: Support vector tracking. IEEE Trans. Pattern Anal. Mach. Intell. **26**(8), 1064–1072 (2004)
6. Collins, R.T., Liu, Y., Leordeanu, M.: Online selection of discriminative tracking features. IEEE Trans. Pattern Anal. Mach. Intell. **27**(10), 1631–1643 (2005)
7. Grabner, H., Leistner, C., Bischof, H.: Semi-supervised on-line boosting for robust tracking. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part I. LNCS, vol. 5302, pp. 234–247. Springer, Heidelberg (2008)

8. Babenko, B., Yang, M.H., Belongie, S.: Robust object tracking with online multiple instance learning. IEEE Trans. Pattern Anal. Mach. Intell. **33**(8), 1619–1632 (2011)
9. Zhang, K., Zhang, L., Yang, M.H.: Real-time compressive tracking. In: European Conference on Computer Vision, pp. 864–877 (2012)
10. Mei, X., Ling, H.: Robust visual tracking using ℓ 1 minimization. In: 2009 IEEE 12th International Conference on Computer Vision, pp. 1436–1443. IEEE (2009)
11. Mei, X., Ling, H., Wu, Y., et al.: Minimum error bounded efficient ℓ 1 tracker with occlusion detection. In: 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1257–1264. IEEE (2011)
12. Zhang, S., Yao, H., Zhou, H., et al.: Robust visual tracking based on online learning sparse representation. Neurocomputing **100**, 31–40 (2013)
13. Zhang, T., Ghanem, B., Liu, S., et al.: Robust visual tracking via structured multi-task sparse learning. Int. J. Comput. Vis. **101**(2), 367–383 (2013)
14. Tang, X., Feng, G., Cai, J.: Weighted group sparse representation for undersampled face recognition. Neurocomputing **145**, 402–415 (2014)
15. Yuan, X.T., Liu, X., Yan, S.: Visual classification with multitask joint sparse representation. IEEE Trans. Image Process. **21**(10), 4349–4360 (2012)