Adaptive Multi-class Correlation Filters

Linlin Yang[1], Chen Chen[2], Hainan Wang[1], Baochang Zhang[1]*, and Jungong Han[3]

[1] School of Automation Science and Electrical Engineering, Beihang University
[2] Center for Research in Computer Vision, University of Central Florida
[3] Department of Computer Science and Digital Technologies, Northumbria University
bczhang@buaa.edu.cn

Abstract. Correlation filters have attracted growing attention due to their high efficiency, which have been well studied for binary classification. However, by setting the desired output to be a fixed Gaussian function, the conventional multi-class classification based on correlation filters becomes problematic due to the under-fitting in many real-world applications. In this paper, we propose an adaptive multi-class correlation filters (AMCF) method based on an alternating direction method of multipliers (ADMM) framework. Within this framework, we introduce an adaptive output to alleviate the under-fitting problem in the ADMM iterations. By doing so, a closed-form sub-solution is obtained and further used to constrain the optimization objective, simplifying the entire inference mechanism. The proposed approach is successfully combined with the Histograms of Oriented Gradients (HOG) features, multi-channel features and convolution features, and achieves superior performances over state-of-the-arts in two multi-class classification tasks including handwritten digits recognition and RGBD-based action recognition.

Keywords: Multi-class correlation filters, ADMM, adaptive output

1 Introduction

In the application of object detection and localization, correlation filters have shown to be competitive with far more complicated approaches, but using only a fraction of the computational power. Correlation filters take advantage of the fact that the convolution of two image patches is equivalent to an element-wise product in the fast Fourier transform (FFT) domain. Thus, by formulating the objective in the FFT domain, they can specify the desired output of a linear classifier for several translations or image shifts [7]. Bolme [1] proposed to learn

* This work was supported in part by the Natural Science Foundation of China under Contract 61272052 and Contract 61473086, in part by PAPD, in part by CICAEET, and in part by the National Basic Research Program of China under Grant 2015CB352501. The work of B. Zhang was supported by the Program for New Century Excellent Talents University within the Ministry of Education, China, and Beijing Municipal Science & Technology Commission Z161100001616005. Baochang Zhang is corresponding author.

a minimum output sum of squared error (MOSSE) filter for visual tracking on gray-scale images. Heriques *et al.* utilized correlation filters in a kernel space based on circulant structure (CSK) [7], which achieved very fast speed in tracking. By using HOG features, kernelized correlation filter (KCF) [7] was developed to improve the performance of CSK. Multi-channel correlation filters (MCCF) were developed to localize eye positions [5]. In [12], hierarchical convolutional features were combined with KCF to achieve robust tracking. Furthermore, maximum margin correlation filters (MMCF) [16], constraining the output at the target location, show better robustness to outliers. We can also find some works on correlation filters for multi-class tasks, such as the Distance Classifier Correlation Filter (DCCF) [13]. Although much success has been demonstrated, the existing works do not principally exploit adaptive output in the procedure of solving the optimized variable.

Problem and motivation: The multi-class output is a useful constraint and has been widely investigated in the state-of-the-art classifiers, i.e., support vector machine (SVM) [18] and Adaboost [17]. Many new applications have been developed [6]. However, to the best of our knowledge, the structured output constraint is neglected in correlation filters calculation. Since each class has its own correlation filter, each correlation filter finds the largest correlation response on its own sample set to discriminate among different classes. In traditional correlation filter methods, a fixed desired output setting could cause an underfitting problem since the learning process cannot converge when all samples are equally treated. As another intuition, the correlation filter of each class is iteratively computed, which can actually be considered as a prior to constrain the solution [26]. Different from existing works, this paper provides new insights into correlation filters from the following aspects:

• Based on an adaptive output, we propose an iterative procedure to calculate correlation filters and obtain a closed-form solution in each iteration.

• We use sub-filters derived from the previous steps to constrain the solution in the ADMM framework for an efficient correlation filter calculation.

For easy of explanation, expressions are given for 1-dimensional (1-D) signals and can be extended to 2-D. ϕ is a kernel function and X^H is the Hermitian transpose of X. We define correlation operation as $*$, dot product as $\cdot$ and element-wise product as $\odot$. $\wedge$ denotes the Fourier transform and $\mathcal{F}^{-1}$ its inverse. Moreover we define $\mathbf{x} \otimes \mathbf{y} = max(\mathbf{x} * \mathbf{y})$ and $Gauss(x)$ denotes Gaussian distribution response with peak value x.

The rest of the paper is organized as follows. Section 2 describes the details of the proposed method, while the algorithm and discussion are stated in section 3. Experiments and results are presented in Section 4. Finally, Section 5 concludes the paper.

2 Proposed AMCF algorithm

We first revisit the one-vs-all (OVA) framework, based on which a new adaptive correlation filters scheme is proposed by considering a multi-class constraint in

the optimization process. A closed-form sub-solution is obtained in each iteration using the ADMM technique. The sub-solutions derived from ADMM iterations are used to constrain the problem in our framework.

Let $(\mathbf{x}_1, y_1), ..., (\mathbf{x}_m, y_m)$ be a set of training examples and assume that each example $\mathbf{x}_i$ is from a domain $X \subseteqq R_n$ and each label y_i is an integer from the set $Y = 1, ..., k$. A multi-class classifier can be seen as a mapping $H : \mathcal{X} \to \mathcal{Y}$. Based on the OVA framework for multi-class classification, we can compute the result according to

$$H(\mathbf{x}_i) = \underset{r=1}{\overset{k}{\operatorname{argmax}}}(W_r \otimes \phi(\mathbf{x}_i)), \tag{1}$$

where W is a matrix of size $k \times n$ over R and W_r is the t^{th} row of W indicating the correlation filter of the r^{th} class.

The objective is defined as the minimum squared-error between $Gauss(Y_{i,j})$ (the i^{th} training sample's gaussian desired output in j^{th} class) and the correlation response $W_j * \phi(\mathbf{x}_i)$. Y is a matrix consists of the maximum value of response. Moreover, the iterative process results in a subset of solutions, which can be used in a ADMM framework [26]. In the ADMM optimization, the variable could be considered from a subspace, which is also used here to constrain our problem. The optimization problem is then defined as follows:

$$\begin{aligned} \min_{W,\varepsilon} \quad & \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{k}\|Gauss(Y_{i,j}) - W_j * \phi(\mathbf{x}_i)\|^2 + \frac{\beta}{2}\sum_{i=1}^{m}\varepsilon_i^2 + \frac{\lambda}{2}\|W\|^2 \\ \textit{subject to } \forall\ i \quad & W_{y_i} \otimes \phi(\mathbf{x}_i) + \delta_{y_i,q_i} - W_{q_i} \otimes \phi(\mathbf{x}_i) = 1 - \varepsilon_i \\ & W \sim \mathcal{G} \end{aligned}, \tag{2}$$

where $q_i = \underset{j}{\operatorname{argmax}}\,(W_j \otimes \phi(\mathbf{x}_i))$ and $\delta_{i,j} = \begin{cases} 0 & i \neq j \\ 1 & i = j \end{cases}$, $\lambda > 0$ is a regularization constant and $\varepsilon_i \geq 0$ are soft constraints. $\mathcal{G}$ denotes a subspace, which is generated based on sub-solutions calculated in an iterative process. β is empirically given. To incorporate subspace prior $\mathcal{G}$ into our model, we propose to perform variable cloning $W \to N \sim \mathcal{G}$. This constraint is included in the ADMM framework to solve the optimization problem with high practicability [26]. The subspace constraint can be added to our approach to simplify the inference mechanism of our algorithm. Now the constrains $W \sim \mathcal{G}$ can be written as $W - N = 0$ and $N \in \mathcal{G}$.

Based on the Lagrangian method, we have:

$$\begin{aligned} L = & \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{k}\|Gauss(Y_{i,j}) - W_j * \phi(\mathbf{x}_i)\|^2 + \frac{\beta}{2}\sum_{i=1}^{m}\varepsilon_i^2 + \frac{\lambda}{2}\|W\|^2 \\ & + \sum_{i=1}^{m}\alpha_i(W_{y_i} \otimes \phi(\mathbf{x}_i) + \delta_{y_i,q_i} - W_{q_i} \otimes \phi(\mathbf{x}_i) - 1 + \varepsilon_i) \\ & + M \cdot (W - N) + \frac{\eta}{2}\|W - N\|^2 \end{aligned}, \tag{3}$$

where M is the Lagrangian multiplier. After transferring the constraint, we simplify the algorithm by means of ADMM [2]. Specifically, at each iterative step, we compute the p^{th} hyperplane of the t^{th} iteration using the result of the previous iteration (i.e., the $(t-1)^{th}$ iteration). We assume that there are s positive samples and $m-s$ negative samples in the p^{th} class. Similar to [7, 5], we pose this problem equivalently in the frequency domain based on Parseval's Theorem [14] to derive a closed-form and efficient solution:

$$\begin{aligned}
\hat{W}_p^t &= [(\lambda+\eta)I + 2\Phi(\widetilde{X})^H\Phi(\widetilde{X})]^{-1}[\Phi(\widetilde{X})^H\widetilde{Y}_p^t + \eta\hat{N}_p^{t-1} + \hat{M}_p^{t-1}] \\
\hat{N}_p^t &= mean(\hat{W}_p^{[1:t]}) \\
\hat{M}_p^t &= \hat{M}_p^{t-1} + \eta(\hat{W}_p^t - \hat{N}_p^t)
\end{aligned} , \tag{4}$$

and

$$\widetilde{X}_p^t = \begin{pmatrix} diag(\hat{\mathbf{x}}_1) \\ \cdots \\ diag(\hat{\mathbf{x}}_m) \end{pmatrix}; \widetilde{Y}_p^t = \begin{pmatrix} \hat{Gauss}(Y_{p,1}^t) \\ \cdots \\ \hat{Gauss}(Y_{p,m}^t) \end{pmatrix}; Y_p^t = Y_p^{t-1} + \beta \begin{pmatrix} \mathbf{a}_1 \\ \mathbf{a}_2 \end{pmatrix};$$

$$\mathbf{a}_1 = \begin{pmatrix} 1 + max(\mathcal{F}^{-1}(\hat{W}_{q_1}^{t-1} \odot \phi(\hat{\mathbf{x}}_1))) - \delta_{p,q_1} \\ \cdots \\ 1 + max(\mathcal{F}^{-1}(\hat{W}_{q_s}^{t-1} \odot \phi(\hat{\mathbf{x}}_s))) - \delta_{p,q_s} \end{pmatrix}$$

$$\mathbf{a}_2 = \begin{pmatrix} max(\mathcal{F}^{-1}(\hat{W}_{y_{s+1}}^{t-1} \odot \phi(\hat{\mathbf{x}}_{s+1}))) - 1 + \vartheta_{p,q_{s+1}} \\ \cdots \\ max(\mathcal{F}^{-1}(\hat{W}_{y_m}^{t-1} \odot \phi(\hat{\mathbf{x}}_m))) - 1 + \vartheta_{p,q_m} \end{pmatrix} .$$

$$q_i = \underset{j}{\operatorname{argmax}} \left(max(\mathcal{F}^{-1}(\hat{W}_j \odot \phi(\hat{\mathbf{x}}_i)))\right)$$

$$\vartheta_{p,q_i} = \begin{cases} 0 & q_i = p \\ 1 - max(\mathcal{F}^{-1}(\hat{W}_{y_i}^{t-1} \odot \phi(\hat{\mathbf{x}}_i))) & \\ \quad + max(\mathcal{F}^{-1}(\hat{W}_p^{t-1} \odot \phi(\hat{\mathbf{x}}_i))) & q_i \neq p \end{cases}$$

Here $mean(\cdot)$ denotes the mean of sub-filters and $diag(\cdot)$ is an operator that transforms a D dimensional vector into a $D \times D$ dimensional diagonal matrix.

3 Summary of the AMCF algorithm and discussion

The details of the proposed AMCF algorithm is shown in Algorithm 1. To demonstrate the advantages of AMCF, we focus on y (the maximum of response), which is the key component in the final classification. The value of y can be regarded as the weight of the training samples. When a sample is misclassified, $\|y\|$ increases, which means the weight decreases. It is known that under-fitting occurs in the setting of a fixed output since all the samples are treated equally. As evident from Fig. 1, both training and test errors are large without using adaptive y.

As far as computational complexity is considered, AMCF is very fast in the testing process given that there are only one element-wise product operation and one inverse transform operation in the FFT domain. When training K classes of D-dimensional feature vectors with T iterations, AMCF has a memory cost of $\mathcal{O}(KD)$ and a time cost of $\mathcal{O}(TKDlogD)$.

Algorithm 1 Adaptive multi-class correlation filters with subspace constraints

Set Y_p^t based on label, $\eta_p^0 = 0.25$, $\varepsilon_{p,best} = +\infty$, $\beta = 1$, iteration number $t = 1$
Initialize $\hat{\mathbf{W}}^{[0]}$,$\hat{\mathbf{N}}^{[0]}$ and $\hat{\mathbf{M}}^{[0]}$based on correlation filters
repeat
  for $i = 1 : label$ **do**
    $\mathbf{Y}_p^t = \mathbf{Y}_p^{t-1} + \beta \begin{pmatrix} \mathbf{a}_1 \\ \mathbf{a}_2 \end{pmatrix}$
    $\widetilde{\mathbf{Y}}_p^t = \begin{pmatrix} \hat{\mathbf{Gauss}}(\mathbf{Y^t_{p,1}}) \\ \cdots \\ \hat{\mathbf{Gauss}}(\mathbf{Y^t_{p,m}}) \end{pmatrix}$
    $\hat{\mathbf{W}}_p^t = [(\lambda + \eta_p^{t-1})\mathbf{I} + 2\Phi(\widetilde{\mathbf{X}})^H \Phi(\widetilde{\mathbf{X}})]^{-1}[\Phi(\widetilde{\mathbf{X}})^H \widetilde{\mathbf{Y}}_p^t + \eta_p^{t-1}\hat{\mathbf{N}}_p^{t-1} + \hat{\mathbf{M}}_p^{t-1}]$
    $\varepsilon_p = \|\hat{\mathbf{W}}^{[k+1]} - \hat{\mathbf{W}}^{[k]}\|_2$
    if $\varepsilon_p < \varepsilon_{p,best}$ **then**
      $\eta_p^t = \eta_p^{t-1}$
      $\varepsilon_{p,best} = \varepsilon_p$
    else
      $\eta_p^t = t\eta_p^{t-1}$
    end if
    $\hat{\mathbf{N}}_p^t = [(\lambda + \eta_p^t)\mathbf{I} + 2\Phi(\widetilde{\mathbf{X}})^H \Phi(\widetilde{\mathbf{X}})]^{-1}[\Phi(\widetilde{\mathbf{X}})^H \widetilde{\mathbf{Y}}_p^t + \eta_p^t \hat{\mathbf{W}}_p^t + \hat{\mathbf{M}}_p^{t-1}]$
    $\hat{\mathbf{M}}_p^t = \hat{\mathbf{M}}_p^{t-1} + \eta_p^t(\hat{\mathbf{M}}_p^t - \hat{\mathbf{N}}_p^t)$
    $t = t + 1$
  end for
until some stopping criterion

4 Experiments

To verify the effectiveness of the proposed AMCF algorithm in classification applications, we carry out experiments on a large-scale dataset, MNIST [10] dataset, for handwritten digits recognition and a challenging depth-based action datasets, MSRAction3D [11], for human action recognition.

For the MNIST dataset, we use features based on convolutional neural networks(CNNs) which exploit rich hierarchical features [12] in images. Hierarchical convolutional features are able to preserve the neighborhood relations and spatial locality of images in their latent higher-level feature representations.

For action recognition on the MSRAction3D, we use multi-channel HOG features computed from the Depth Motion Maps (DMMs) [24] due to their computational efficiency. According to [24], each depth action sequence generates

three DMMs corresponding to three projection views, i.e., front view (f), side view (s) and top view (t), denoted by DMM_f, DMM_s and DMM_t, respectively (see Fig. 2). We concatenate the HOG features extracted from the three DMMs to form the multi-channel HOG features as the final feature representation.

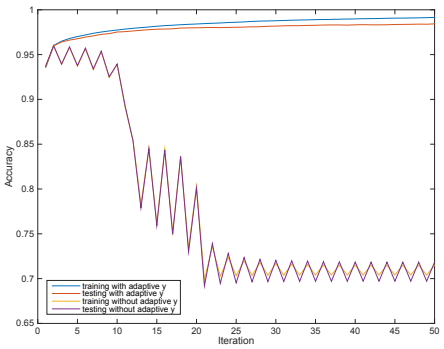

Fig. 1. Convergence comparison between the correlation filters without adaptive y and the proposed AMCF method on the MNIST dataset using hierarchical features.

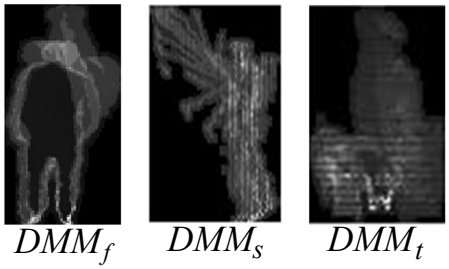

Fig. 2. Three DMMs of a depth action sequence "high throw".

4.1 MNIST dataset

We perform handwritten digits recognition on the widely used MNIST [10] dataset. It has a training set of 60,000 examples, and a test set of 10,000 examples. We use the convolutional features extracted by LeNet [9] to encode the original MNIST digits. The same experimental setup in [8] is followed. The proposed method achieves the best performance as shown in Table 1.

4.2 MSRAction3D dataset

The MSRAction3D dataset consists of depth sequences captured by a Kinect device. It includes 20 actions performed by 10 subjects. Each subject performed each action 2 or 3 times. The size of each depth image is 240 $\times$ 320 pixels. The same experimental setup in [20, 4, 25] is adopted. The same parameters reported in [3] were used here for the sizes of DMMs and block. A total of 20 actions are employed and one half of the subjects (1, 3, 5, 7 and 9) are used for training and the remaining subjects are used for testing. The recognition rates of our method and existing approaches are listed in Table 2. It is clear that our method achieves better performance than other competing methods.

Table 1. Recognition results comparison on the MNIST dataset.

Method	Accuracy(%)
LeNet [9]	98.9
MCCF [5]	93.5
SVM poly 4 [9]	98.9
K-NN Euclidean [9]	97.6
PCA + quadratic [9]	96.7
Ours	**98.9**

Table 2. Recognition results comparison on the MSRAction3D dataset.

Method	Accuracy(%)
DMM-HOG [24]	85.5
Random Occupancy [21]	86.5
HON4D [15]	88.9
Actionlet Ensemble [22]	88.2
Depth Cuboid [23]	89.3
Vemulapalli et al. [19]	89.5
Ours	**92.3**

5 Conclusions

In this paper, we propose an efficient framework for the multi-class correlation filters classification with subspace constraints in the ADMM framework. The new insight focuses on an adaptive output and also subspace constraints on the solution. The experimental results on handwritten digits recognition and MSR action recognition show that the proposed algorithm performs favorably against the state-of-the-art methods. Future work will focus on improving the kernel method and accelerating the convergence speed.

References

1. Bolme, D.S., Beveridge, J.R., Draper, B.A., Lui, Y.M.: Visual object tracking using adaptive correlation filters. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2544–2550. IEEE (2010)
2. Boyd, S., Parikh, N., Chu, E., Peleato, B., Eckstein, J.: Distributed optimization and statistical learning via the alternating direction method of multipliers. Foundations and Trends® in Machine Learning 3(1), 1–122 (2011)
3. Chen, C., Jafari, R., Kehtarnavaz, N.: Action recognition from depth sequences using depth motion maps-based local binary patterns. In: Proceedings of the IEEE Winter Conference on Applications of Computer Vision. pp. 1092–1099. IEEE (2015)
4. Chen, C., Liu, M., Zhang, B., Han, J., Jiang, J., Liu, H.: 3d action recognition using multi-temporal depth motion maps and fisher vector. In: Proceedings of International Joint Conference on Artificial Intelligence. pp. 3331–3337 (2016)
5. Galoogahi, H., Sim, T., Lucey, S.: Multi-channel correlation filters. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 3072–3079 (2013)
6. Han, J., Zhou, P., Zhang, D., Cheng, G., Guo, L., Liu, Z., Bu, S., Wu, J.: Efficient, simultaneous detection of multi-class geospatial targets based on visual saliency modeling and discriminative learning of sparse coding. ISPRS Journal of Photogrammetry and Remote Sensing 89, 37–48 (2014)
7. Henriques, J.F., Caseiro, R., Martins, P., Batista, J.: High-speed tracking with kernelized correlation filters. IEEE Transactions on Pattern Analysis and Machine Intelligence 37(3), 583–596 (2015)
8. Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T.: Caffe: Convolutional architecture for fast feature embedding. In: Proceedings of the ACM International Conference on Multimedia. pp. 675–678. ACM (2014)

9. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proceedings of the IEEE 86(11), 2278–2324 (1998)
10. LeCun, Y., Cortes, C., Burges, C.J.: The mnist database of handwritten digits (1998)
11. Li, W., Zhang, Z., Liu, Z.: Action recognition based on a bag of 3d points. In: Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops. pp. 9–14 (2010)
12. Ma, C., Huang, J.B., Yang, X., Yang, M.H.: Hierarchical convolutional features for visual tracking. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 3074–3082 (2015)
13. Mahalanobis, A., Kumar, B.V., Sims, S.: Distance-classifier correlation filters for multiclass target recognition. Applied Optics 35(17), 3127–3133 (1996)
14. Oppenheim, A.V., Willsky, A.S., Nawab, S.H.: Signals and Systems. Pearson (2014)
15. Oreifej, O., Liu, Z.: Hon4d: Histogram of oriented 4d normals for activity recognition from depth sequences. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 716–723 (2013)
16. Rodriguez, A., Boddeti, V.N., Kumar, B.V., Mahalanobis, A.: Maximum margin correlation filter: A new approach for localization and classification. IEEE Transactions on Image Processing 22(2), 631–643 (2013)
17. Shen, C., Lin, G., van den Hengel, A.: Structboost: Boosting methods for predicting structured output variables. IEEE Transactions on Pattern Analysis and Machine Intelligence 36(10), 2089–2103 (2014)
18. Tsochantaridis, I., Joachims, T., Hofmann, T., Altun, Y.: Large margin methods for structured and interdependent output variables. In: Journal of Machine Learning Research. pp. 1453–1484 (2005)
19. Vemulapalli, R., Arrate, F., Chellappa, R.: Human action recognition by representing 3d skeletons as points in a lie group. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 588–595 (2014)
20. Vieira, A.W., Nascimento, E.R., Oliveira, G.L., Liu, Z., Campos, M.F.: On the improvement of human action recognition from depth map sequences using space–time occupancy patterns. Pattern Recognition Letters 36, 221–227 (2014)
21. Wang, J., Liu, Z., Chorowski, J., Chen, Z., Wu, Y.: Robust 3d action recognition with random occupancy patterns. In: Proceedings of the European Conference on Computer Vision, pp. 872–885. Springer (2012)
22. Wang, J., Liu, Z., Wu, Y., Yuan, J.: Mining actionlet ensemble for action recognition with depth cameras. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1290–1297. IEEE (2012)
23. Xia, L., Aggarwal, J.: Spatio-temporal depth cuboid similarity feature for activity recognition using depth camera. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2834–2841 (2013)
24. Yang, X., Zhang, C., Tian, Y.: Recognizing actions using depth motion maps-based histograms of oriented gradients. In: Proceedings of the ACM International Conference on Multimedia. pp. 1057–1060. ACM (2012)
25. Yang, Y., Zhang, B., Yang, L., Chen, C., Yang, W.: Action recognition using completed local binary patterns and multiple-class boosting classifier. In: Proceedings of Asian Conference on Pattern Recognition. pp. 336–340 (2015)
26. Zhang, B., Perina, A., Murino, V., Del Bue, A.: Sparse representation classification with manifold constraints transfer. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4557–4565 (2015)