Exploiting Local Feature Fusion for Action Recognition

Jie Miao, Xiangmin Xu(✉), Xiaoyi Jia, Haoyu Huang, Bolun Cai, Chunmei Qing, and Xiaofen Xing

School of Electronic and Information Engineering,
South China University of Technology, Wushan RD., Tianhe District,
Guangzhou, People's Republic of China
xmxu@scut.edu.cn

Abstract. Densely sampled local features with bag-of-words models have been widely applied to action recognition. Conventional approaches assume that different kinds of local features are totally uncorrelated, and they are separately processed, encoded, and then fused at video-level representation. However, these local features are not totally uncorrelated in practice. To address this problem, multi-view local feature fusion is exploited for local descriptor fusion in action recognition. Specifically, tensor canonical correlation analysis (TCCA) is employed to obtain a fused local feature that carries the high-order correlation hidden among different types of local features. The high-order correlation local feature improves the conventional concatenation based fusion approach. Experimental results on three challenging action recognition datasets validate the effectiveness of the proposed approach.

Keywords: Action recognition · Multi-view · Local feature

1 Introduction

Vision-based action recognition has been an active research area in the recent decades, and it is the central part in many computer vision applications, such as intelligent video surveillance, human-computer interaction, video content analysis, and video retrieval. It is a changeling problem to recognize actions from unconstrained videos due to complex backgrounds, large intra-class variances, etc. To obtain a better feature for classification, videos are usually represented in different ways using multiple types of features. This kind of representation is termed as the multi-view features. In this paper, we exploit the multi-view local feature fusion for action recognition.

This work is supported in part by the National Natural Science Founding of China (61171142, 61401163), Science and Technology Planning Project of Guangdong Province, China (2011A010801005, 2014B010111003, 2014B010111006), Guangzhou Key Lab of Body Data Science (201605030011) and the Fundamental Research Funds for the Central Universities (2015ZZ032).

E. Chen et al. (Eds.): PCM 2016, Part II, LNCS 9917, pp. 221–230, 2016.
DOI: 10.1007/978-3-319-48896-7_22

Approaches based on densely sampled local features with bag-of-words models have been successfully applied to complex action recognition tasks. In these approaches, different types of local features are encoded independently by bag-of-words models, and then concatenated together as final representation for classification. This fusion strategy is regarded as the representation fusion. It is proposed by assuming that different types of local features are totally uncorrelated. In this way, they are processed separately without losing valuable information.

However, different types of local features are not totally uncorrelated in practice. There might be some correlation exist among these features that are useful for feature representation. A straightforward solution is the local feature fusion, which simply concatenate local features before feature encoding. This concatenation cannot capture the hidden connections among local features effectively. Furthermore, due to the increased input dimensionality, the conventional bag-of-words model in local feature fusion fails to encode features as good as that in representation fusion.

In this paper, we propose the high-order correlation local feature (HCF) to utilize the correlation among different types of local features. In our work, different types of local features are regarded as different views, tensor canonical correlation analysis [12] is employed to capture the high-order correlation among different views. Experimental results on three challenging action recognition datasets show that a significant improvement can be achieved by proposed approach.

2 Related Work

Recently, approaches based on densely sampled local features and bag-of-words models have been shown to be particularly successful. [25] first evaluated various local features and sampling strategies, and showed the effectiveness of densely sampled local features for action recognition. After that, more sampling strategies and local features were proposed. Dense trajectories [23] samples local patches in a dense grid from each frame and tracks them as trajectories using dense optical flow, it then extracts local features along trajectories and encodes them for video representation. Then the improved dense trajectories (IDT) was introduced by removing camera motion in videos [24]. A two-stream CNN was proposed by training two independent networks for appearance and motion representation, respectively [21]. All of these approaches use the representation-level fusion, which neglects the correlation be-tween different types of local features, and simply concatenating different features as final video representation. In our work, we propose a better fusion approach for different types of local features.

Correlation is a powerful way to investigate and describe the relationship between two sets of data. There are many approaches proposed by utilizing correlation. [20] proposed a gradient-based subspace phase correlation for efficient image alignment estimations. Statistical methods was introduced to estimate vehicle count based on correlation estimation [18]. Canonical correlation analysis (CCA) is a straightforward method utilizing correlation for data fusion, it has

been widely used for multi-view applications, such as classification [6], regression [7] and clustering [1]. However, these approaches based on conventional CCA are limited by two views. Although the multi-view fusion can be achieved by enumerating each pair of views, the high-order correlation among multiple views will be ignored. Tensor canonical correlation analysis (TCCA) [12] was proposed as a multi-view dimension reduction method to handle the data of an arbitrary number of views by analyzing the covariance tensor of different views. Besides of correlation, lots of approaches for multi-view fusion based on different principles have been proposed. A non-linear multi-view dimension reduction method was proposed by utilizing the common structure among different views [27].

There are only few studies on the fusion of local features have been proposed. A local feature fusion approach using coupled multi-index frame was proposed for accurate image retrieval [29], but its performance is limited by nearest neighbor based bag-of-words models. Three types of simple concatenation based fusion strategy were investigated [17], and fused them as hybrid representation. Experimental results suggested that the representation-level fusion is the best fusion among three strategies, and more useful information can be obtained from different types of fusions. But this approach is inefficient due to the simple concatenating based fusion, and the improvement was mainly introduced by the large number of clusters used in bag-of-words models. A mixture model of probabilistic CCA was proposed to learn shared latent variables for utilizing the common part of different features in action recognition [2]. Similar to other approaches based on CCA, it is also limited by two views. Therefore, it fails to utilize rich correlation among all views.

3 Local Feature Fusion

In this section, we detail the proposed high-order correlation local feature (HCF) for action recognition. An illustration of our approach is shown in Fig. 1. First of all, we extract different types of local features using IDT [24]. Briefly, points are densely sampled in a grid from each frame and tracked as trajectories by dense optical flow, and then the features are extracted aligned with tracked trajectories.

There are four kinds of local features extracted by IDT. Histograms of oriented gradients (HOG) [4] encodes static edges and textures, which represent appearance information for action recognition. HOG is extracted directly from video frames. Histograms of oriented optical flow (HOF) [10] is obtained from horizontal and vertical optical flows. It captures both magnitudes and directions of motion. Horizontal and vertical motion boundary histograms (MBHx and MBHy) [5] are extracted from horizontal and vertical optical flow respectively. MBHx and MBHy utilize the gradient of optical flows, and they carry horizontal and vertical motion information respectively. These four kinds of local features represent different information of local video cubes. In conventional approach, they are assumed to be totally uncorrelated to each other. Therefore, they are processed and encoded independently, and then concatenated as the final feature for the classification.

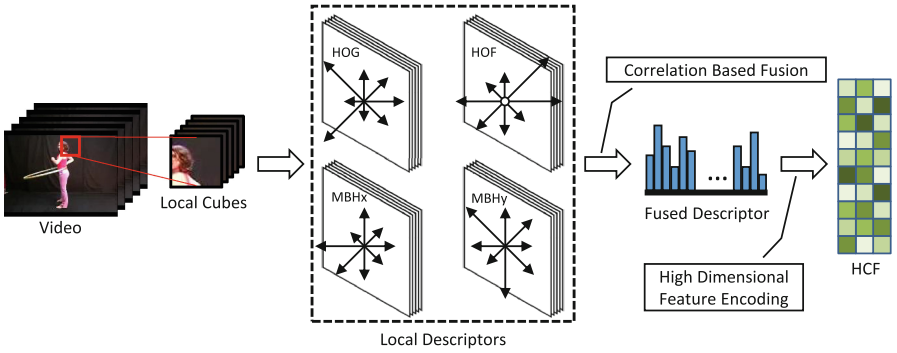

Fig. 1. The framework of the proposed approach.

We suggest that there are some important correlations hidden among them, which are neglected in conventional approaches. Regarding each type of features as a unique view, we perform a fusion by tensor canonical correlation analysis (TCCA) [12] to extract high-order correlation among different features. TCCA projects multiple views of features into a new subspace that maximizes correlation of multiple views. Different from other multi-view approaches that keep similar information among different views, TCCA is able to obtain new representation that carries high-order correlations among different views. These correlations encode the relationship of different types of features in each local video cubes. TCCA can be summarized as follows.

Assuming that m views of features $\{X_l\}_{l=1}^{m}$ of N instances are given, where each view $X_l = [\mathbf{x}_{l1}, \mathbf{x}_{l2}, \cdots, \mathbf{x}_{lN}] \in \mathbb{R}^{d_l \times N}$ has been whitened, the covariance tensor among all views is represented as

$$\mathcal{C} = \frac{1}{N} \sum_{n=1}^{N} \mathbf{x}_{1n} \circ \mathbf{x}_{2n} \circ \cdots \circ \mathbf{x}_{mn}, \tag{1}$$

where the operate $\circ$ is the tensor outer product, and $\mathcal{C}$ is of dimension $d_1 \times d_2 \times \cdots \times d_m$.

According to [12], the major problem of TCCA is to maximize the correlation ρ among multiple types of features. It can be described as

$$\begin{aligned} \underset{\{\mathbf{h}_l\}}{\text{argmax}} \ \ & \rho = corr(\mathbf{z}_1, \mathbf{z}_2, ..., \mathbf{z}_m), \\ & s.t.\ \mathbf{z}_l^T \mathbf{z}_l = 1, l = 1, ..., m, \end{aligned} \tag{2}$$

where $corr(\mathbf{z}_1, \mathbf{z}_2, ..., \mathbf{z}_m) = (\mathbf{z}_1 \odot \mathbf{z}_2 \odot ... \odot \mathbf{z}_m)^T \mathbf{e}$ is the canonical correlation, the operator $\odot$ is the element-wise product, and $\mathbf{e} \in \mathbb{R}^N$is an all ones vector. Specifically, $\mathbf{z}_l = X_l^T \mathbf{h}_l, l = 1, ..., m$ are canonical variables, and $\mathbf{h}_l$ are called canonical vectors. Problem (2) is equivalent to solving the following formulation:

$$\begin{aligned} \underset{\{\mathbf{h}_l\}}{\text{argmax}} \ \ & \rho = \mathcal{C} \times_1 \mathbf{h}_1^T \times_2 \mathbf{h}_2^T \cdots \times_m \mathbf{h}_m^T, \\ & s.t.\ \mathbf{h}_l^T (C_l + \epsilon I) \mathbf{h}_l = 1, l = 1, ..., m, \end{aligned} \tag{3}$$

where I is an identity matrix, ϵ is a nonnegative trade-off parameter. The operator $\times_p$ is the p-mode production.

Specifically, C_l are self-covariance matrices represented as

$$C_l = \frac{1}{N}\sum_{n=1}^{N} \mathbf{x}_{ln}\mathbf{x}_{ln}^T,\ l = 1, ..., m.$$

The reformulated optimization problem is

$$\begin{aligned} \underset{\{\mathbf{u}_l\}}{\text{argmax}}\ \rho = \mathcal{M} \times_1 \mathbf{u}_1^T \times_2 \mathbf{u}_2^T \cdots \times_m \mathbf{u}_m^T, \\ s.t.\ \mathbf{u}_l^T\mathbf{u}_l = 1, l = 1, ..., m, \end{aligned} \tag{4}$$

where $\mathbf{u}_l = \tilde{C}_l^{1/2}\mathbf{h}_l$ are latent transform variables. And the definition of $\mathcal{M}$ is

$$\mathcal{M} = \mathcal{C} \times_1 \tilde{C}_1^{-1/2} \times_2 \tilde{C}_2^{-1/2} \cdots \times_m \tilde{C}_m^{-1/2}, \tag{5}$$

where $\tilde{C}_l = C_l + \epsilon I$.

The problem (4) is equivalent to

$$\underset{\{\mathbf{u}_l\}}{\text{argmax}} \left\| \mathcal{M} - \hat{\mathcal{M}} \right\|_F^2, \tag{6}$$

where $\hat{\mathcal{M}} = \rho\mathbf{u}_1 \circ \mathbf{u}_2 \circ \cdots \circ \mathbf{u}_m$. The problem (6) can be solved by Alternating Least Square (ALS) algorithm [8].

Obtaining the solution $\mathbf{u}_l$, the canonical variable is $\mathbf{z}_l = X_l^T\tilde{C}_l^{-1/2}\mathbf{u}_l$. Define r as the dimensionality after the reduction ($r \leqslant min\{d_1, \cdots, d_m\}$), and let $U_l = [\mathbf{u}_l^{(1)}, \cdots, \mathbf{u}_l^{(r)}]$ and $\mathbf{z}_l^{(1)}, \cdots, \mathbf{z}_l^{(r)}$ be the column vectors of Z_l, the projected feature of the l-th view is

$$Z_l = X_l^T\tilde{C}_l^{-1/2}U_l. \tag{7}$$

Finally $\{Z_l\}_{l=1}^m$ are concatenated as $Z \in \mathbb{R}^{mr\times N}$ for subsequent processing. Regarding HOG, HOF, MBHx and MBHy as four different views, TCCA is performed to project them into a new subspace to obtain a fused feature-level representation that carries the high-order correlation among them.

4 Experiments

In this section, we report the experimental results of the proposed HCF. Firstly, we introduce the datasets used to evaluate the performance of HCF and detail the implementation parameters. Then we report the evaluation of different parameters used in our work. And we compare the performance of HCF with other feature fusion approaches for action recognition. Lastly, we compare our method with state-of-the-art approaches. We evaluate the proposed approach on three challenging action recognition datasets: HMDB51, UCF50 and Hollywood2. Some example frames of these datasets are shown in Fig. 2.

Fig. 2. Example frames from the action recognition datasets used in this paper. Rows from top to bottom are HMDB51, UCF50 and Hollywood2 respectively.

The HMDB51 dataset [9] consists of 51 action categories with 6,766 video sequences from different movies. We report the mean accuracy (mAcc.) over 3 train/test splits proposed in [9]. The UCF50 dataset [19] consists of 50 action categories divided into 25 groups. Following the standard evaluation protocol proposed in [19], we report the mean accuracy over 25 cross validation sets. The Hollywood2 dataset [13] consists of 3,669 movie clips collected from 69 different Hollywood movies. We follow the standard evaluation protocol proposed by [13], and report the mean average precision (mAP) over all classes.

4.1 Implementation Details

Here, we describe the implementation details of our experiments. All experiments were performed as described here, unless stated otherwise.

To extract local features, we follow the default parameters proposed in IDT [24]. Four kinds of local features were extracted, i.e. HOG, HOF, MBHx and MBHy. For preprocessing, principle component analysis (PCA) and whitening were performed to reduce the dimensionality by a factor of two. As the baseline approach, the representation-level fusion was performed. Here, for each type of local features, 256,000 features were randomly sampled from a dataset to train a GMM with 256 clusters, and Fisher vector was performed to get the encoded feature. After encoding, four kinds of feature were concatenated together as the final representation.

Instead of encoding fused features by conventional Fisher vector, we use sparse coding based Fisher vector [11] for encoding the fused features. The dimensionality r of each type of features after processed by TCCA was set to 45.

Thus the dimensionality of fused feature was 180. The dictionary size for sparse coding based Fisher vector was 256. And the number of nonzero coefficients for sparse coding was set to 15. Similar to conventional Fisher vector, normalization is required for better feature representation. In our work, we apply intra-normalization and L_2-normalization to the encoded feature. We concatenate HCF with conventional features for final classification. To better utilize both conventional single view features and the proposed HCF, an extra L_2-normalization is performed after the concatenation of all encode features. A one-against-all linear support vector machine was used for classification. The evaluation protocols for each dataset were then applied to produce the final results.

4.2 Experimental Results

First, we evaluate the impact of parameters in HCF on the HMDB51 dataset. Following the feature extraction process, we study the dimensionality of each view r kept by TCCA. As shown in Fig. 3, a proper r can capture more valuable connections among all views, but some interferences will be introduced while using a large r. In our experiments, the best performance is achieved at the point $r = 45$.

Second, we conduct experiments to evaluate the performance of HCF compared with other local feature fusion approaches. The results are shown in Table 1. Here the baseline approach use only representation-level fusion was proposed in [24]. The extra concatenation-based local feature fusion is able to slightly improve the recognition accuracy. As shown in Table 1, the proposed method outperforms MVSV and all kinds of fusion methods. Using both the correlation and the independence of multiple types of features, the proposed approach performs better than MVSV, which unable to make full use of the correlation information among four types of features. This indicates that HCF is more effective to capture the high-order correlation among different views simultaneously.

As shown in Table 2, we compare the propose approach with state-of-the-art action recognition approaches. Results show that the proposed method can achieve competitive results on the three challenging action recognition datasets compared with other approaches. Here, IDT can be considered as the baseline

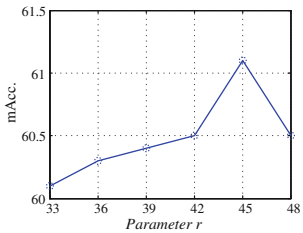

Fig. 3. Evaluation of the parameters on the HMDB51 dataset.

Table 1. The comparison of different fusion approaches.

	HMDB51 (mAcc.)	UCF50 (mAcc.)	Hollywood2 (mAP)
Baseline [24]	57.2%	91.2%	64.3%
Concatenation	58.6%	92.0%	66.7%
MVSV [2]	59.5%	92.0%	66.2%
Proposed	**61.1%**	**93.8%**	**67.3%**

Table 2. Comparison of different action recognition approaches on HMDB51, UCF50 and Hollywood2 datasets.

HMDB51	mAcc	UCF50	mAcc	Hollywood2	mAP
Causality [14]	58.7%	Movement [3]	90.0%	FV [15]	63.3%
SHVLAD [16]	59.8%	FV [15]	90.0%	IDT [24]	64.3%
Hybrid [17]	61.1%	IDT [24]	91.2%	GNMF [22]	56.8%
CNN [21]	59.4%	Causality [14]	92.5%	ICA [28]	54.1%
Pooling [26]	59.7%	Hybrid [17]	92.3%	Pooling [26]	67.5%
Proposed	**61.1%**	Proposed	**93.8%**	Proposed	**67.3%**

approach. Lots of efforts have been made for better feature encoding. Fisher vector was introduced to action recognition in [15], which compact feature set can be used. Granger causality is used to encode relationship of trajectory pairs [14]. SHVLAD employs high-order statistics and supervised learning to improve feature encoding [16]. A hybrid approach was proposed by combining many kinds of feature encoding approaches, and it gains lots of improvements on recognition accuracy. Using high-order correlation features, our approaches outperforms these approaches. Movement patterns histogram was proposed for action representation using motion information [3]. And the two-stream CNN was proposed by using CNN on both RGB video frames and stacked optical flows. Withe the motion information obtained from stacked optical flows, good results can be achieved by these approaches. The proposed approaches outperforms these action recognition approaches on HMDB51 and UCF50 datasets. And competitive results on Hollywood2 dataset can be achieved.

5 Conclusion

In this paper, we highlight the high-order correlation information among different types of local features in action recognition, which is neglected in conventional approaches. In particular, we proposed the high-order correlation local feature as an axillary feature to utilize these information. Experimental results show that the proposed approach able to improve conventional approach as an axillary feature. The proposed approach is by all means not limited to an action

recognition representation and could be applied in other applications that rely on densely sampled local features and bag-of-words models.

References

1. Blaschko, M.B., Lampert, C.H.: Correlational spectral clustering. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2008, pp. 1–8. IEEE (2008)
2. Cai, Z., Wang, L., Peng, X., Qiao, Y.: Multi-view super vector for action recognition. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 596–603. IEEE (2014)
3. Ciptadi, A., Goodwin, M.S., Rehg, J.M.: Movement pattern histogram for action recognition and retrieval. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8690, pp. 695–710. Springer, Heidelberg (2014). doi:10.1007/978-3-319-10605-2_45
4. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: Computer Vision and Pattern Recognition, vol. 1, pp. 886–893 (2005)
5. Dalal, N., Triggs, B., Schmid, C.: Human detection using oriented histograms of flow and appearance. In: European Conference on Computer Vision, pp. 428–441 (2006)
6. Farquhar, J., Hardoon, D., Meng, H., Shawe-taylor, J.S., Szedmak, S.: Two view learning: SVM-2K, theory and practice. In: Advances in Neural Information Processing Systems, pp. 355–362 (2005)
7. Kakade, S.M., Foster, D.P.: Multi-view regression via canonical correlation analysis. In: Bshouty, N.H., Gentile, C. (eds.) COLT 2007. LNCS, vol. 4539, pp. 82–96. Springer, Heidelberg (2007). doi:10.1007/978-3-540-72927-3_8
8. Kroonenberg, P.M., De Leeuw, J.: Principal component analysis of three-mode data by means of alternating least squares algorithms. Psychometrika **45**(1), 69–97 (1980)
9. Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., Serre, T.: HMDB: a large video database for human motion recognition. In: 2011 IEEE International Conference on Computer Vision (ICCV), pp. 2556–2563. IEEE (2011)
10. Laptev, I., Marszalek, M., Schmid, C., Rozenfeld, B.: Learning realistic human actions from movies. In: Computer Vision and Pattern Recognition, pp. 1–8 (2008)
11. Liu, L., Shen, C., Wang, L., van den Hengel, A., Wang, C.: Encoding high dimensional local features by sparse coding based fisher vectors. In: Advances in Neural Information Processing Systems, pp. 1143–1151 (2014)
12. Luo, Y., Tao, D., Wen, Y., Ramamohanarao, K., Xu, C.: Tensor canonical correlation analysis for multi-view dimension reduction. arXiv preprint arXiv:1502.02330 (2015)
13. Marszalek, M., Laptev, I., Schmid, C.: Actions in context. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2009, pp. 2929–2936. IEEE (2009)
14. Narayan, S., Ramakrishnan, K.R.: A cause and effect analysis of motion trajectories for modeling actions. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2633–2640. IEEE (2014)
15. Oneata, D., Verbeek, J., Schmid, C.: Action and event recognition with fisher vectors on a compact feature set. In: 2013 IEEE International Conference on Computer Vision (ICCV), pp. 1817–1824. IEEE (2013)

16. Peng, X., Wang, L., Qiao, Y., Peng, Q.: Boosting VLAD with supervised dictionary learning and high-order statistics. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8691, pp. 660–674. Springer, Heidelberg (2014). doi:10.1007/978-3-319-10578-9_43
17. Peng, X., Wang, L., Wang, X., Qiao, Y.: Bag of visual words and fusion methods for action recognition: comprehensive study and good practice. Eprint Arxiv (2014)
18. Peng, Z., Yi, S., Bei, H.: Statistical methods to estimate vehicle count using traffic cameras. Multidimension. Syst. Signal Process. **20**(2), 121–133 (2009)
19. Reddy, K.K., Shah, M.: Recognizing 50 human action categories of web videos. Mach. Vis. Appl. **24**(5), 971–981 (2013)
20. Ren, J., Vlachos, T., Zhang, Y., Zheng, J., Jiang, J.: Gradient-based subspace phase correlation for fast and effective image alignment. J. Vis. Commun. Image Represent. **25**(7), 1558–1565 (2014)
21. Simonyan, K., Zisserman, A.: Two-stream convolutional networks for action recognition in videos. In: Advances in Neural Information Processing Systems, pp. 568–576 (2014)
22. Wang, H., Yuan, C., Hu, W., Ling, H., Yang, W., Sun, C.: Action recognition using nonnegative action component representation and sparse basis selection. IEEE Trans. Image Process. **23**(2), 570–581 (2014)
23. Wang, H., Kläser, A., Schmid, C., Liu, C.L.: Dense trajectories and motion boundary descriptors for action recognition. Int. J. Comput. Vision **103**(1), 60–79 (2013)
24. Wang, H., Schmid, C.: Action recognition with improved trajectories. In: IEEE International Conference on Computer Vision, ICCV 2013, pp. 3551–3558, December 2013
25. Wang, H., Ullah, M.M., Klser, A., Laptev, I., Schmid, C.: Evaluation of local spatio-temporal features for action recognition. In: British Machine Vision Conference (2009)
26. Wang, P., Cao, Y., Shen, C., Liu, L., Shen, H.T.: Temporal pyramid pooling based convolutional neural networks for action recognition. arXiv preprint arXiv:1503.01224 (2015)
27. Xia, T., Tao, D., Mei, T., Zhang, Y.: Multiview spectral embedding. IEEE Trans. Syst. Man Cybern. B Cybern. **40**(6), 1438–1446 (2010)
28. Zhang, S., Yao, H., Sun, X., Wang, K., Zhang, J., Lu, X., Zhang, Y.: Action recognition based on overcomplete independent components analysis. Inf. Sci. **281**, 635–647 (2014)
29. Zheng, L., Wang, S., Liu, Z., Tian, Q.: Packing and padding: coupled multi-index for accurate image retrieval. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1947–1954 (2014)