Practical Action Recognition with Manifold Regularized Sparse Representations

Lining Zhang[1]
lining.zhang@port.ac.uk
Rinat Khusainov[1]
rinat.khusainov@port.ac.uk
John P. Chiverton[2]
john.chiverton@port.ac.uk

[1] School of Computing
University of Portsmouth
Portsmouth, UK
[2] School of Energy and Electronic Engineering
University of Portsmouth
Portsmouth, UK

Abstract

With the explosion of long term health conditions, monitoring human daily activities in home environment is one of the important issues in healthcare. Human action recognition in videos is one of the main topics in this context. Conventional representations are not very effective for encoding dense features extracted from videos. In this work, we propose a novel manifold regularized sparse representation (MRSR) method to encode dense features for human action recognition in assisted living. The new method can effectively incorporate a manifold regularization term to explore the geometric structure of the improved dense trajectories, which are very effective for learning action representations. By introducing a locality constraint, our method ensures each interest point is represented by its local closest words. Moreover, our method has an analytical solution and low computational complexity. Experimental results on different realistic databases show the effectiveness of the proposed algorithm for practical action recognition in assisted living.

1 Introduction

The U.K., like many other countries, is faced with an explosion of long term health conditions. In general, there are conditions that require condition management for many years, outside of the hospital setting. Recognizing human daily activities in the home environment is one of the important issues in healthcare. In computer vision, this problem can be classified as human action recognition, which is very important but also challenging [4, 12, 14, 20, 21, 26, 27]. Action recognition can also be applied to help solve many other real-world problems such as video surveillance, smart camera monitoring and human computer interaction.

In general, the challenges of human action recognition in videos come from difficulties, such as great intraclass variance, occlusion and clutter. A framework in such a field includes video representation and classification. Video representation learning is the procedure of acquiring features via interest point detection and representation. Action representation is obtained by encoding these features. Then, a classification model is learned for the final

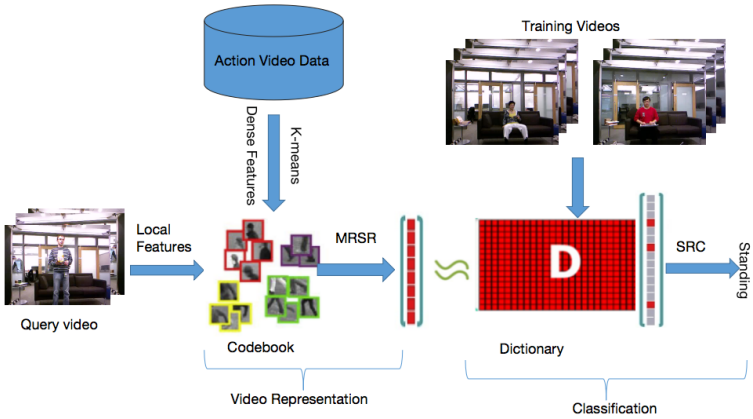


Figure 1: The framework of practical human action recognition for assisted living

action representation, which can be used to recognize the new action. Feature representations can be broadly divided into global representations [3, 5] and local representations [1, 7, 22]. Global representations first localize a person by background subtraction or tracking [28] and then represent the interest region as a whole. However, they are sensitive to noise, variations in viewpoint and partial occlusion. Local representations are based on the spatial temporal interest points and do not need to subtract the background or tracking. This means they are less sensitive to view-point changes, noise and partial occlusions.

In recent years, a variety of local features for data have been introduced [1, 7, 9, 10, 13, 22, 24], which have been widely applied to human action recognition [1, 7, 9, 10, 13, 22, 24]. A number of local spatial-temporal interest point detectors, e.g., Harris 3D detector [6], Cuboid detector [1], Hessian detector [22] and different descriptors, are all combined and then evaluated under the bag-of-features (BOF) recognition framework. Experimental results have shown that none of these local features can perform the best on all datasets. Among the tested descriptors, the combination of gradient and optical flow are the best choice [16] . To obtain the final action representation, The popular BOF model was applied, in which a class of codebook was first formed by utilizing the k-means algorithm in [1] . Each interest point was defined as that of its closest word and finally an action representation was given as a histogram of interest point information. However, conventional BOF representation cannot accurately describe an action since each interest point can only be represented by a single word, thus leading to a large reconstruction error. In BOF, the type of an interest point belongs to the type of the closest word. Thus, significantly different interest points may be assigned to the same type, which will decrease the performance of a human action recognition system.

In this work, we propose a novel approach, Manifold Regularized Sparse Representation (MRSR), to encode features extracted by the state of the art improved dense trajectories [15]. Our MRSR incorporates a manifold regularization term, which can explore the manifold structure of improved dense trajectories and choose those local words that are on the same manifold with the interest points. By introducing a locality constraint to the MRSR, our algorithm can ensure that each interest point is represented by its local closest words. Moreover, compared with previous methods, MRSR has an analytical solution and is easy to calculate. Finally, Sparse Representation Classifier (SRC) is introduced to recognize the actions of interest. An illustration of the whole systems for human action recognition in

assisted living is given in Fig.1

The rest of the paper is organized as follows: in Section 2, we give a detailed description of our MRSR for human action recognition. Section 3 presents the experimental setup and a comparison of obtained results on different datasets. Section 4 concludes the paper with discussions.

2 The proposed approach

2.1 Improved Dense Trajectories

We first use the improved dense trajectories to extract local features of videos. Here, we briefly review the improved dense trajectories introduced in [15], which are extended from dense trajectories [17]. Firstly, the algorithm densely samples a set of points with a grid of 5 pixels over 8 spatial scales. The motion vectors are selected by thresholding the smallest eigenvalue of the autocorrelation matrix. These detected points are tracked by media filtering of the dense flow field [17]:

$$P_{t+1} = (x_{t+1}, y_{t+1}) = (x_t, y_t) + (\mathrm{M} * \omega_t)|_{(\bar{x}_t, \bar{y}_t)}, \tag{1}$$

where M is the median filter kernel, $*$ is convolution operation $\omega_t = (u_t, v_t)$ is the dense optical flow field of the tth frame, and $(\bar{x}_t, \bar{y}_t)$ is the rounded position of (x_t, y_t). To avoid the drift problem of tracking, the maximum length of a trajectory is set at 15 frames. Finally, those static trajectories are removed and other trajectories with sudden large displacements are also ignored [17]. For each trajectory, we compute several descriptors (trajectories, HOG, HOF and MBH) with same parameters in [17]. The final dimensions of the descriptors are 30 for trajectories, 96 for HOG, 108 for HOF and 192 for MBH.

2.2 Manifold Regularized Sparse Representation

After we have obtained the dense trajectories from the videos, we can encode the local features to obtain the action representation. Suppose we have obtained a set of d-dimensional dense trajectories with feature representation $X = [x_1, x_2, \ldots x_n] \in R^{d \times n}$ extracted from a video, where n is the number of local feature descriptors. Firstly, we use k-means algorithm to generate the codebook $B = [b_1, b_2, \ldots b_n] \in R^{d \times n}$ and each center is called a word.

Traditional BOF model has been widely used in computer vision and achieved comparatively good results. It solves the following problem:

$$\begin{aligned} &\hat{c}_i = \arg\min_{c_i} \|x_i - Bc_i\|_2^2, \\ &s.t. \|c_i\|_0 = 1, \|c_i\|_1 = 1, c_{i,j} \geq 0, j = 1, 2, \ldots, l, \end{aligned} \tag{2}$$

where $c_{i,j}$ is the jth element of the vector c_i. After encoding a human action from a video, BOF uses sum pooling method [8] with the formulation $z = \sum_{i=1}^{n} \hat{c}_i$ as the final action representation. Since local features from similar videos tend to lie on the same manifold, we propose a novel coding method called MRSR.

We use all the words to represent an interest feature in order to reduce the reconstruction error. Motivated by the fact that locality is more essential than sparsity, we use the locality

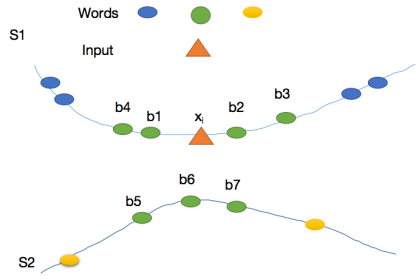


Figure 2: The input x_i and the words $b_1, b_2, b_3, b_4, b_5, b_6, b_7$. Our algorithm prefers to choose b_1, b_2, b_3, b_4, which span a low-dimensional manifold subspace around x_i rather than b_5, b_6, b_7.

as a constraint tern. Locality must lead to sparsity but not necessarily vice versa [19]. As discussed in [19], the locality constraint is smooth while the conventional sparse regularization term [25] is not. In addition, it can ensure that similar interest points have similar codes. In practice, we first find k-nearest neighbors to form a new codebook, in which the number of words should be more than 500 to accurately describe the action. Then, the coding coefficients are calculated as follows:

$$\hat{c}_i = \arg\min_{c_i} \|x_i - Bc_i\|_2^2 + \eta_0 ||c_i||_2^2 + \eta_1 \|d_i \odot c_i\|_2^2, \tag{3}$$

where $\odot$ denotes element-wise multiplication, $d_i = [||x_i - b_1||_2^2, ||x_i - b_2||_2^2, \ldots, ||x_i - b_l||_2^2]^T \in R^{l\times 1}, c_i \in R^{l\times 1}$.

To preserve the manifold geometry of local features, we introduced a manifold regularization term $||p_i c_i||_2^2$ as suggested in [2]:

$$\hat{c}_i = \arg\min_{c_i} \|x_i - Bc_i\|_2^2 + \eta_0 ||c_i||_2^2 + \eta_1 \|d_i \odot c_i\|_2^2 + \eta_2 ||p_i c_i||_2^2, \tag{4}$$

where $p_i = [p_{i1}, p_{i2}, \ldots, p_{il}] \in R^{d\times l}, p_{ij} = (x_i - b_j)/||x_i - b_j||^2$. The first term in Eq. 4 is the reconstruction error. Unlike BOF, we use multiple words, which are the neighbors of x_i to describe the interest features. The third term is a penalty function, ensuring that the similar patches will have similar codes. The fourth term will make the algorithm select words that lie in the same manifold as x_i [2] . An illustration of the proposed method is shown in Fig.2

Since Eq.4 is a strictly convex function, MRSR has an analytical solution:

$$\hat{c}_i = (B^T B + diag(\eta_0 1 + \eta_1 d_i \odot d_i) + \lambda_2 {p_i}^T p_i) \backslash B^T x_i, \tag{5}$$

where $1 \in R^{l\times 1}$ and $\backslash$ denote left matrix division. In experiments, we empirically set parameters in Eq.5 as follows: η_0 is $1e-4$, which is usually set as a small number; η_1 is $1e-3$; and η_2 is 1 thoughout the experiments. Further tuning the parameters can improve the performance of the algorithm.

Sum pooling [8] and max pooling [19, 25] schemes have been successfully used in pattern recognition. As in [29], we use a max pooling scheme to capture the global statistics

of an action in a video sequence and increase spatial and time translation invariance. Max pooling is defined as

$$z_i = \max(|\hat{c}_{i1}|, |\hat{c}_{i2}|, \ldots, |\hat{c}_{in}|), \tag{6}$$

These pooled features can then be normalized by sum normalization and l^2 normalization. In our work, we use the max pooling scheme combined with l^2 normalization as in [25].

2.3 Sparse Representation Classifier

To recognize human actions, we use the Sparse Representation Classifier [23] here since it can provide a comparable performance to the Support Vector Machine (SVM) classifier while there is only one parameter to be fixed. Suppose we have c classes of human actions and denote the training samples as $X = [X_1, \ldots, X_c]$ where X_i is the sub-set of training data from classes i. Let $\hat{y}$ be a testing data, then the main procedure of classification using SRC can be given as follows.

Given a testing sample $y \in R^m$ from the ith class, we first calculate its sparse coding coefficients by

$$\beta^* = \arg\min_{\beta} ||\hat{y} - X\beta||_F^2 + \gamma||\beta||_1, \tag{7}$$

Since y comes from the ith class, most nonzero coefficients are those associated with class i. We compute the residual as

$$e_i = ||y - D_i\hat{\beta}_i||_2, \tag{8}$$

where $\hat{\beta}_i$ is the coding coefficients associated with class i. Finally, we can do the classification via

$$label(\hat{y}) = \arg\min\{e_i\}. \tag{9}$$

3 Experiments

In experiments, we evaluate the effectiveness of our proposed algorithm on three public datasets: KTH dataset, UCF sports database and MSR Daily activity database. We compare different aspects of the algorithms and show the effectiveness of our scheme in encoding the features of videos. We compared our algorithm with BOF and Locality-constraint coding (LLC), which are two popular encoding methods for local features in computer vision. Here, we does not compare our algorithm with iDT with feature vector [15] since we want to show the effectiveness of the manifold structure in encoding dense features. We use the leave-one out cross validation to the evaluate the performance of our algorithm unless otherwise noted. Specifically it employs the actions from one person as the testing data and leave the remaining actions from other persons as the training data.

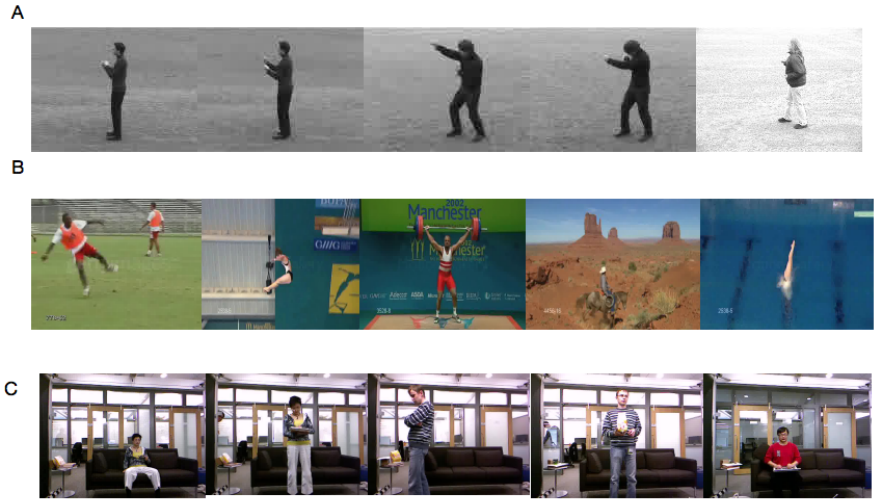


Figure 3: Examples of the three datasets: (a) KTH database, (b) UCF sports database, (c) MSR Daily Activity database

3.1 Datasets

KTH database is an important benchmark dataset that has been used to evaluate various human action recognition algorithm. It contains six actions: walking, jogging, running, boxing, hand waving and hand clapping. Twenty-five subjects in four different scenarios perform these actions. The scenarios include indoor, outdoor, changes in clothing and variations in scale. Overall it has 599 low-resolution video clips for one of the videos is missing.

UCF sports database is a set of 150 videos, which are collected from various broadcast sports channels such as BBC and ESPN. It contains 10 different actions: diving, golf swimming floor, walking. This dataset is challenging, with a wide range of scenarios and viewpoints.

MSR Daily Activity dataset is a daily living dataset captured by a Kinect device. There are 16 activity types: drink, eat, read book, call cellphone, write on paper, use laptop, use vacuum cleaner, cheer up, still, toss paper, play game, lay down on sofa, walk, play guitar, stand up, sit down. If possible, each subject performs an activity in two different positions: sitting on sofa and standing. There are totally 320 activity sequences.

Some of the example video frames from these three database are shown in Fig.3.

3.2 Comparison of BoF, LLC and MRSR

In this subsection, we aim to evaluate the performance of our method compared to previous popular encoding methods. Improved dense trajectories are firstly used to detect and describe the interest points. Subsequently, the k-means algorithm is employed to form words, which are set as 1,024. Finally, BOF, LLC and MRSR are utilized respectively to obtain the final action representation.

These three action representation methods are further compared under the same condition and the Nearest Neighbor (NN) classifier is selected for the classification stage. Table 1 shows the recognition results in the form of average recognition accuracy we find that

Table 1: Performance comparison accuracy (%) of BoF, LLC and MRSR on KTH, UCF Sports and MSR Daily activity databases

Method	BoF	LLC	MRSR
KTH	86.5	88.1	**90.5**
UCF Sports	80.5	82.6	**85.6**
MSR Daily	87.3	88.7	**90.4**

Table 2: Performance comparison accuracy (%) among the different classification schemes including NN, SVM and MRSR on the KTH database.

Method	NN	SVM	SRC
Accuracy	90.2	93.2	**94.8**

the proposed MRSR achieves the highest average recognition accuracy, compared to BOF and LLC. We can find that the proposed MRSR representation achieves the highest average recognition rate on the KTH database, UCF Sports database and MSR Daily database while the BOF model performs worst. Compared with LLC, MRSR which considers the intrinsic manifold structure of the words, is more beneficial for recognition.

3.3 Evaluation on the KTH database

We also evaluate our algorithm on the KTH dataset with different classification schemes. Table 2 compares different classification schemes including NN, SVM and MRSR on the KTH dataset. As shown in Table 2, SRC achieves best result compared with the other two classification methods.

Although there are fewer types of actions in the KTH dataset, the KTH dataset is more challenging due to different scenarios and scale variations. As there are better sufficient training samples for each action, we can notice that the SRC is significiantly better than NN in terms of accuracy.

We also show the confusion matrices in Fig.4. We find that NN does not recognize the actions "jog" and "run". While SVM and SRC show much better performance. SRC performs slightly better than SRC in recognizing "walk", "jog", "hand wave" and "hand clapping", which may be because the manifold constraint in MRSR has better discriminative ability.

3.4 UCF Sports database

In this subsection, we evaluate our proposed approach on the UCF sports action database. Compared with the KTH database, it is a more challenging database for action recognition. The confusion matrices of LLC and MRSR methods with SRC in Fig.5. From the experimental results, we can notice that MRSR with SRC can perform better on "glolf swimming", "skating" and "swing floor" action classes because our method considers intrinsic manifold structure of the dense features from action videos while LLC can only incorporate the locality information of dense features. It can be seen that our method effectively encodes dense

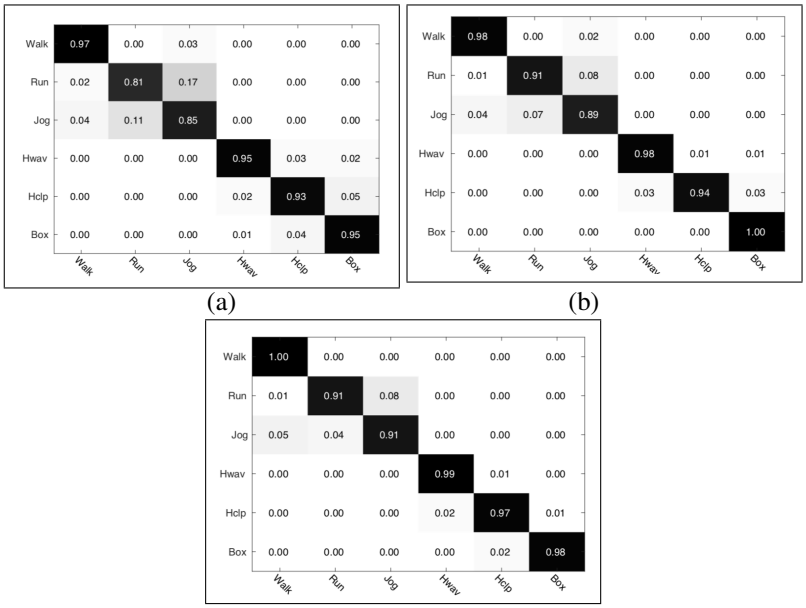


Figure 4: Confusion matrices for KTH dataset for different classification methods.The rows are the actual action label and the columns are predicted ones. (a) NN used for classification, (b) SVM used for classification and (c) SRC used for classification

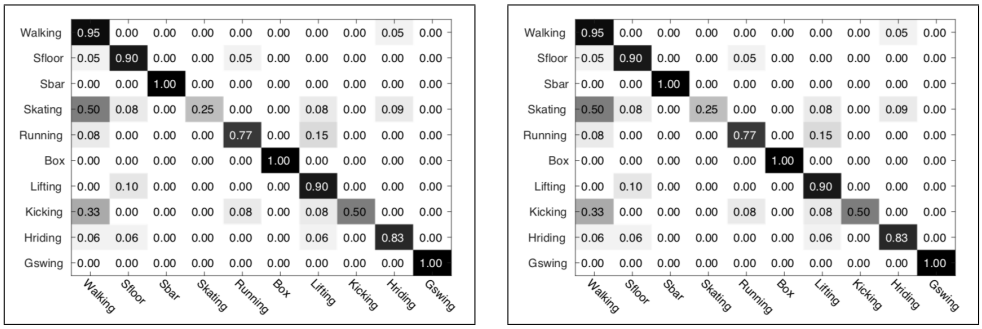
features on the same manifold and is more useful for learning action representations. Thus, we can conclude that MRSR is more effective to encode the features than LLC.

3.5 Evaluation on the MSR Daily activity database

The MSR Daily activity database is designed to cover daily activities in a home living environment. In this work, all experiments are conducted on the RGB channel of the database. We apply the cross-subject setting to evaluate the proposed algorithm on this dataset. Half of the subjects are used as training samples, while the other half are used as testing samples. We compared the MRSR with Dynamic Temporal Warping [11] and Random Occupancy Pattern method [18]. In Table 3, the experimental results of our proposed algorithm are compared with two popular algorithms on the MSR Daily activity database. Compared with Dynamic Temporal Warping [11] and Random Occupancy Pattern [18], our method can effectively encode the dense features on the same manifold, which is more effective for recognizing the human daily activity.

4 Conclusion and future work

In this work, we have proposed the Manifold Regularized Sparse Representation (MRSR) method to encode the dense features for human action recognition in assisted living. The new algorithm can incorporate the manifold regularization term to explore the manifold



(a) (b)

Figure 5: The confusion matrices for UCF sports data: (a) LLC+ SRC (b) MRSR+SRC.

Table 3: Recognition accuracy (%) comparison for MSR Daily activity dataset.

Method	Accuracy
Dynamic Temporal Warping [11]	0.423
Random Occupancy Pattern [18]	0.747
MRSR	**0.885**

structure of the improved dense trajectories, which are very effective for learning action representations. By introducing a locality constraint, MRSR ensures that each interest point is represented by its closest words. Moreover, compared with previous methods, MRSR has an analytical solution and is easy to calculate. Experimental results on different realistic databases have shown the effectiveness of the proposed algorithm to represent the human action recognition in assisted living. For future work, we are planing to apply our method to more realistic larger databases.

References

[1] Piotr Dollár, Vincent Rabaud, Garrison Cottrell, and Serge Belongie. Behavior recognition via sparse spatio-temporal features. In *Visual Surveillance and Performance Evaluation of Tracking and Surveillance, 2005. 2nd Joint IEEE International Workshop on*, pages 65–72. IEEE, 2005.

[2] Ehsan Elhamifar and René Vidal. Sparse manifold clustering and embedding. In *Advances in neural information processing systems*, pages 55–63, 2011.

[3] Hao Jiang and David R Martin. Finding actions using shape flows. In *European Conference on Computer Vision*, pages 278–292. Springer, 2008.

[4] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1725–1732, 2014.

[5] Vili Kellokumpu, Guoying Zhao, and Matti Pietikäinen. Human activity recognition using a dynamic texture based method. In *BMVC*, volume 1, page 2, 2008.

[6] Ivan Laptev. On space-time interest points. *International journal of computer vision*, 64(2-3):107–123, 2005.

[7] Ivan Laptev, Marcin Marszalek, Cordelia Schmid, and Benjamin Rozenfeld. Learning realistic human actions from movies. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008.

[8] Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *null*, pages 2169–2178. IEEE, 2006.

[9] Changhong Liu, Yang Yang, and Yong Chen. Constructing visual vocabularies using sparse coding for action recognition. In *Information Engineering and Computer Science, 2009. ICIECS 2009. International Conference on*, pages 1–4. IEEE, 2009.

[10] Jingen Liu and Mubarak Shah. Learning human actions via information maximization. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008.

[11] Meinard Müller and Tido Röder. Motion templates for automatic classification and retrieval of motion capture data. In *Proceedings of the 2006 ACM SIGGRAPH/Eurographics symposium on Computer animation*, pages 137–146. Eurographics Association, 2006.

[12] Xiaojiang Peng, Limin Wang, Xingxing Wang, and Yu Qiao. Bag of visual words and fusion methods for action recognition: Comprehensive study and good practice. *Computer Vision and Image Understanding*, 150:109 – 125, 2016. ISSN 1077-3142.

[13] Qiang Qiu, Zhuolin Jiang, and Rama Chellappa. Sparse dictionary-based representation and recognition of action attributes. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 707–714. IEEE, 2011.

[14] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *Advances in neural information processing systems*, pages 568–576, 2014.

[15] Heng Wang and Cordelia Schmid. Action recognition with improved trajectories. In *Proceedings of the IEEE international conference on computer vision*, pages 3551–3558, 2013.

[16] Heng Wang, Muhammad Muneeb Ullah, Alexander Klaser, Ivan Laptev, and Cordelia Schmid. Evaluation of local spatio-temporal features for action recognition. In *BMVC 2009-British Machine Vision Conference*, pages 124–1. BMVA Press, 2009.

[17] Heng Wang, Alexander Kläser, Cordelia Schmid, and Cheng-Lin Liu. Dense trajectories and motion boundary descriptors for action recognition. *International journal of computer vision*, 103(1):60–79, 2013.

[18] Jiang Wang, Junsong Yuan, Zhuoyuan Chen, and Ying Wu. Spatial locality-aware sparse coding and dictionary learning. In *Asian Conference on Machine Learning*, pages 491–505, 2012.

[19] Jinjun Wang, Jianchao Yang, Kai Yu, Fengjun Lv, Thomas Huang, and Yihong Gong. Locality-constrained linear coding for image classification. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 3360–3367. IEEE, 2010.

[20] Limin Wang, Yuanjun Xiong, Dahua Lin, and Luc Van Gool. Untrimmednets for weakly supervised action recognition and detection. In *CVPR*, pages 4325–4334, 2017.

[21] Yifan Wang, Jie Song, Limin Wang, Luc Van Gool, and Otmar Hilliges. Two-stream sr-cnns for action recognition in videos. In *BMVC*, 2016.

[22] Geert Willems, Tinne Tuytelaars, and Luc Van Gool. An efficient dense and scale-invariant spatio-temporal interest point detector. In *European conference on computer vision*, pages 650–663. Springer, 2008.

[23] John Wright, Allen Y Yang, Arvind Ganesh, S Shankar Sastry, and Yi Ma. Robust face recognition via sparse representation. *IEEE transactions on pattern analysis and machine intelligence*, 31(2):210–227, 2009.

[24] Xunshi Yan and Yupin Luo. Making full use of spatial-temporal interest points: an adaboost approach for action recognition. In *Image Processing (ICIP), 2010 17th IEEE International Conference on*, pages 4677–4680. IEEE, 2010.

[25] Jianchao Yang, Kai Yu, Yihong Gong, and Thomas Huang. Linear spatial pyramid matching using sparse coding for image classification. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1794–1801. IEEE, 2009.

[26] Bowen Zhang, Limin Wang, Zhe Wang, Yu Qiao, and Hanli Wang. Real-time action recognition with deeply-transferred motion vector cnns. *IEEE Transactions on Image Processing*, 27(5):2326–2339, 2018.

[27] Yue Zhao, Yuanjun Xiong, Limin Wang, Zhirong Wu, Xiaoou Tang, and Dahua Lin. Temporal action detection with structured segment networks. *ICCV, Oct*, 2, 2017.

[28] Huiyu Zhou, Yuan Yuan, and Chunmei Shi. Object tracking using sift features and mean shift. *Computer vision and image understanding*, 113(3):345–352, 2009.

[29] Yan Zhu, Xu Zhao, Yun Fu, and Yuncai Liu. Sparse coding on local spatial-temporal volumes for human action recognition. In *Asian Conference on Computer Vision*, pages 660–671. Springer, 2010.