Low-rank Representation Based Action Recognition

Xiangrong Zhang[1], Yang Yang[2], Hanghua Jia[1], Huiyu Zhou[3], Licheng Jiao[1]

[1]Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education, Xidian University, Xi'an, China

[2]Institute of Automation, Chinese Academy of Sciences, Beijing, China

[3]Queen's University, Belfast, UK

xrzhang@mail.xidian.edu.cn, smileyoyon@gmail.com, H.Zhou@ecit.qub.ac.uk

***Abstract*—Human action recognition is an important problem in computer vision, which has been applied to many applications. However, how to learn an accurate and discriminative representation of videos based on the features extracted from videos still remains to be a challenging problem. In this paper, we propose a novel method named low-rank representation based action recognition to recognize human actions. Given a dictionary, low-rank representation aims at finding the lowest-rank representation of all data, which can capture the global data structures. According to its characteristics, low-rank representation is robust against noises. Experimental results demonstrate the effectiveness of the proposed approach on several publicly available datasets.**

Keywords—human action recognition; low-rank representation; video representation; sparse representation based classification

I. INTRODUCTION

Human action recognition is the process of recognizing the behavior in a real-world video, which has a wide range of applications, such as video summarization, human-machine interaction, and video surveillance. It is easy for human to recognize the behavior in a real-world video, but it is a challenging job for a computer. Although many impressive results have been reported on human action recognition, it still remains as a challenging problem [1] because of viewpoint changes, occlusions, illumination variations, and background clutters.

A common framework in human action recognition includes video representation and classification. Video representation is the process of acquiring features via interest point detection and feature representation and obtaining the behavior representation by encoding the features. In general, feature representations can be divided into two categories: global representations [2] and local representations [3]. Global representations allow a person to be localized by background subtraction or tracking, and then represent the region of interest as a whole. Local representations allow a video to be described as a collection of local descriptors or patches. In this paper, we use local representations to describe a video, which are less sensitive to view-point changes, noises, appearance and partial occlusions. When a video representation is available for an observed video sequence, human action recognition becomes a classification problem. In the stage of classification, many methods have been applied in the field of human action recognition, including the nearest neighbor (NN)/ k-Nearest Neighbor (k-NN) classifiers [4], Support Vector Machine (SVM) [5], and Sparse Representation based Classification (SRC) [6]. In our experiment, we use SRC to classify a query action.

In the previous work, the techniques of tracking or body pose estimation were used in human action recognition [7]. However it required accurate tracking or body pose estimation, which is difficult for realistic videos. In recent years, many approaches adopted an intermediate representation to describe a video, based on local spatio-temporal descriptors [3,8]. Although traditional Bag-of-Feature (BoF) [9] models with the local spatio-temporal descriptor could generate promising results [3], they could not accurately describe a behavior; because (1) each interest point is only represented by a single word, thus leading to a large reconstruction error, and (2) the type of an interest point completely depends on the type of the closest word, where different interest points may be assigned to the same type.

Sparse representation (SR) has been widely used and achieved promising results in pattern recognition. It is based on the idea that each data vector can be represented by a linear combination of a few atoms in the dictionary. Given a set of data vectors, SR allows the sparsest representation to be computed individually [10,11]. However, SR cannot capture the global information because there is no global constraint with its solution. Some modified SR based methods for action recognition are then proposed to improve the performance. In [12], a manifold-constrained term was incorporated into the objective function. This term can preserve the manifold-geometry of features. In [13], a Laplacian group sparse coding approach was proposed. This approach can encode a group of relevant features simultaneously, and allow as less atoms as possible to participate in the approximation. Meanwhile, by incorporating Laplacian regularization term, the similarity of the features can be preserved.

In this paper, we propose a novel method named low-rank representation (LRR) based action recognition to recognize human actions. To our best knowledge, it is the first time to apply LRR in the field of human action recognition. Different from robust PCA [14] aiming at matrix decomposition, LRR, which jointly obtains the representation of all data and seeks the lowest rank representation, is able to capture the global structures [15]. Experiments in [15] also demonstrated the effectiveness of LRR for robust subspace segmentation. We employ LRR to encode the interest points, because (1) some of

This work was supported in part by the National Natural Science Foundation of China (No. 61272282), the Program for New Century Excellent Talents in University (No. NCET-13-0948), and the Fundamental Research Funds for the Central Universities (No. K50511020011).

the interest points of one action are similar to each other and the corresponding coding coefficients are low-rank, thus LRR is feasible and (2) using this representation, we will be able to obtain a global representation, which will benefit the following classification. Given a set of action sequences, we employ Cuboid [3] to represent interest points (described in section 2)and then LRR is used to code those features (section 3). Our LRR will be demonstrated on the commonly used Weizmann, KTH datasets and UCF datasets (section 4).The experimental results show the promising performances of our method.

II. FEATURE REPRESENTATION

To represent each video sequence, effective descriptors should be employed. By now, extensive methods have been published [16-19], which represent the interest points of human actions as local spatio-temporal features. Among those descriptors, Cuboid [3] is a popular approach and can generate a large number of features. We choose the popular Cuboid in our experiment. The Cuboid detector relies on separable linear filters for computing the response function of a video sequence. 2D Gaussian smoothing kernel and 1D Gabor filters [3] are applied along the spatial and temporal dimensions respectively instead of a 3D filter on the spatio-temporal domain. It can generate a rich set of interest points. A response function is calculated as follows:

$$\boldsymbol{R}=(\boldsymbol{I}*\mathrm{g}*\mathrm{h}_{\mathrm{ev}})^2+(\boldsymbol{I}*\mathrm{g}*\mathrm{h}_{\mathrm{ov}})^2 \tag{1}$$

where $g(x,y;\sigma)$ is the 2D Gaussian smoothing kernel applied in the spatial domain, * is the convolution operation, h_{ev} and h_{ov} are 1D Gabor filters applied temporally, and defined as:

$$\mathrm{h}_{ev}(t;\tau,w)=-\cos(2\pi tw)e^{-\mathrm{t}^2/\tau^2} \tag{2}$$

$$\mathrm{h}_{ov}(t;\tau,w)=-\sin(2\pi tw)e^{-\mathrm{t}^2/\tau^2} \tag{3}$$

We use $w=4/\tau$ as that used in [3], and there are essentially two free parameters σ and τ which correspond roughly to the spatial and temporal scales of the detector. Interest points corresponding to the local maximum of the response function and areas with spatially distinguishing features will induce a strong response. After interest points are found, we describe them using the Cuboid descriptors (we experimentally set the dimension of each descriptor to 100 in our paper). For more details, refer to [3].

III. LOW-RANK REPRESENTATION FOR ENCODING THE FEATURES

The linear representation of data has been widely and successfully employed in the area of signals processing recently. LRR is one of the successful cases, which can capture the global structure of data. In this section, we first review LRR and then apply it to encoding the features.

A. Review of LRR

Let $\boldsymbol{X}=[\boldsymbol{x}_1,\boldsymbol{x}_2,...,\boldsymbol{x}_n]\in\Re^{d\times n}$ be a matrix whose columns are n data samples drawn from independent spaces. Each column can be represented by a linear combination of a basis $\boldsymbol{A}=[\boldsymbol{a}_1,\boldsymbol{a}_2,...,\boldsymbol{a}_l]\in\Re^{d\times l}$:

$$\boldsymbol{X}=\boldsymbol{A}\boldsymbol{Z} \tag{4}$$

where $\boldsymbol{Z}=[\boldsymbol{z}_1,\boldsymbol{z}_2,...,\boldsymbol{z}_n]$ is the coefficient matrix with each $\boldsymbol{z}_i$ being the representation of $\boldsymbol{x}_i$. Thus, we can obtain infinitely many feasible solutions to Eq. (4).

We assume that the data is clean. Then, the following rank minimization problem is considered:

$$\min_{\boldsymbol{Z}} rank(\boldsymbol{Z}),\quad s.t.\ \boldsymbol{X}=\boldsymbol{A}\boldsymbol{Z} \tag{5}$$

In real applications, data is often noisy and even grossly corrupted, so we add a noise term $\boldsymbol{E}$ to Eq. (5). As a common practice in rank minimization problems, we replace the rank function with the nuclear norm. Now, we can obtain a low-rank recovery to $\boldsymbol{X}$ by solving the following convex optimization problem:

$$\min_{\boldsymbol{Z},\boldsymbol{E}} \|\boldsymbol{Z}\|_*+\lambda\|\boldsymbol{E}\|_{2,1},\quad s.t.\ \boldsymbol{X}=\boldsymbol{A}\boldsymbol{Z}+\boldsymbol{E} \tag{6}$$

where $\|\boldsymbol{Z}\|_*$ is the nuclear norm (i.e., the sum of the singular values) of $\boldsymbol{Z}$, which approximates the rank of $\boldsymbol{Z}$. Similar to [15], a relaxed constraint $\|\boldsymbol{E}\|_{2,1}=\sum_{j=1}^{n}\sqrt{\sum_{i=1}^{n}([\boldsymbol{E}]_{i,j})^2}$ is chosen. λ is a parameter that controls the effect of the noise matrix $\boldsymbol{E}$.

B. LRR for Encoding the Features

In this section, we present the method of encoding the features to obtain the behavior representation. Suppose we have obtained a set of d-dimensional local spatio-temporal features matrix $\boldsymbol{X}=[\boldsymbol{x}_1,\boldsymbol{x}_2,...,\boldsymbol{x}_n]\in\Re^{d\times n}$ extracted from a video. The codebook $\boldsymbol{A}=[\boldsymbol{a}_1,\boldsymbol{a}_2,...,\boldsymbol{a}_l]\in\Re^{d\times l}$ is generated from cluster centers by using the *k*-means algorithm reported in [3] to cluster all the local features. We obtain a low-rank recovery to $\boldsymbol{X}$ by solving Eq. (6). The optimization problem (6) is convex and can be solved by various methods. For efficiency, we adopt the Augmented Lagrange Multiplier (ALM) [20] method in this paper. We first transform (6) to the follow problem:

$$\min_{\boldsymbol{Z},\boldsymbol{E}} \|\boldsymbol{J}\|_*+\lambda\|\boldsymbol{E}\|_{2,1},\ s.t.\boldsymbol{X}=\boldsymbol{A}\boldsymbol{Z}+\boldsymbol{E},\ \boldsymbol{Z}=\boldsymbol{J} \tag{7}$$

This problem can be solved by the ALM method with a complexity of $O(n^3)$, which minimizes the following augmented Lagrangian function:

$$\begin{aligned} \boldsymbol{L} = \|\boldsymbol{J}\|_* + \lambda\|\boldsymbol{E}\|_{2,1} + <\boldsymbol{Y}_1, \boldsymbol{X} - \boldsymbol{A}\boldsymbol{Z} - \boldsymbol{E}> + \\ <\boldsymbol{Y}_2, \boldsymbol{Z} - \boldsymbol{J}> + \frac{\mu}{2}(\|\boldsymbol{X} - \boldsymbol{A}\boldsymbol{Z} - \boldsymbol{E}\|_F^2 + \|\boldsymbol{Z} - \boldsymbol{J}\|_F^2) \end{aligned} \quad (8)$$

The above problem is unconstrained. So it can be minimized with respect to $\boldsymbol{J}$, $\boldsymbol{Z}$ and $\boldsymbol{E}$, respectively, by fixing the other variables and then updating the Lagrange multipliers $\boldsymbol{Y}_1$ and $\boldsymbol{Y}_2$, where $\mu > 0$ is a penalty parameter. Based on the following lemma, its solution is outlined in Algorithm 1.Note that Step 1 and 3 of the algorithm are convex problems they both have closed-form solutions. Step 1 is solved via the Singular Value Thresholding (SVT) operator [21], while Step 3 is solved via the following lemma.

Lemma 1([22]): Let $\boldsymbol{Q}$ be a given matrix. If there is an optimal solution to

$$\boldsymbol{W}^* = \min_{\boldsymbol{W}} \alpha\|\boldsymbol{W}\|_{2,1} + \frac{1}{2}\|\boldsymbol{W} - \boldsymbol{Q}\|_F^2 \quad (9)$$

Then the *i*-th column of $\boldsymbol{W}^*$ is

$$[\boldsymbol{W}^*]_{:,i} = \begin{cases} \dfrac{\|\boldsymbol{Q}_{:,i}\|_2 - \alpha}{\|\boldsymbol{Q}_{:,i}\|_2}\boldsymbol{Q}_{:,i}, & \text{if } \|\boldsymbol{Q}_{:,i}\|_2 > \alpha; \\ 0, & \text{otherwise.} \end{cases} \quad (10)$$

The Sum pooling [9] or the max pooling [22]-[25] scheme has been successfully used in pattern recognition. As shown in [25], we use a max pooling scheme to capture the global statistics of an action in video sequences and increase spatial and time translation invariance. The max pooling is defined as:

$$\boldsymbol{y}_i = \max(|z_{i1}|, |z_{i2}|, \ldots, |z_{in}|), i = 1, 2, ..., l \quad (11)$$

Suppose one video has *n* local features, and the coding coefficient of these local features are $[\boldsymbol{z}_1, \boldsymbol{z}_2, ..., \boldsymbol{z}_n]$ and the size of the codebook is *l*. Then, after max pooling based on Eq. (11), we will obtain $\boldsymbol{Y} \in \Re^{l \times 1}$ to represent this video.

Algorithm 1: Solving Eq. (8) by ALM

Input:

n local features for one video $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_n] \in \Re^{d \times n}$, parameter λ.

Initialize: $\boldsymbol{Z} = \boldsymbol{J} = 0$, $\boldsymbol{E} = 0$, $\boldsymbol{Y}_1 = \boldsymbol{Y}_2 = 0$,

$$\mu = 10^{-6}, \mu_{\max} = 10^6, \rho = 1.1, \text{and } \varepsilon = 10^{-8}.$$

While not converged do

Step 1: update $\boldsymbol{J}$, when fixing the other variables.

$$\boldsymbol{J} = \arg\min_{\boldsymbol{J}} \frac{1}{\mu}\|\boldsymbol{J}\|_* + \frac{1}{2}\|\boldsymbol{J} - (\boldsymbol{Z} + \boldsymbol{Y}_2/\mu)\|_F^2;$$

Step 2: update $\boldsymbol{Z}$, when fixing the other variables.

$$\boldsymbol{Z} = (\boldsymbol{I} + \boldsymbol{A}^T\boldsymbol{A})^{-1}(\boldsymbol{A}^T(\boldsymbol{X} - \boldsymbol{E}) + \boldsymbol{J} + (\boldsymbol{A}^T\boldsymbol{Y}_1 - \boldsymbol{Y}_2)/\mu);$$

Step 3: update $\boldsymbol{E}$, when fixing the other variables.

$$\boldsymbol{E} = \arg\min_{\boldsymbol{E}} \frac{\lambda}{\mu}\|\boldsymbol{E}\|_{2,1} + \frac{1}{2}\|\boldsymbol{E} - (\boldsymbol{X} - \boldsymbol{A}\boldsymbol{Z} + \boldsymbol{Y}_1/\mu)\|_F^2.$$

Step 4: update the multipliers

$$\boldsymbol{Y}_1 = \boldsymbol{Y}_1 + \mu(\boldsymbol{X} - \boldsymbol{A}\boldsymbol{Z} - \boldsymbol{E});$$
$$\boldsymbol{Y}_2 = \boldsymbol{Y}_2 + \mu(\boldsymbol{Z} - \boldsymbol{J});$$

Step 5: update the parameter μ by $\mu = \min(\rho\mu, \mu_{\max})$.

Step 6: check the convergence conditions:

$$\|\boldsymbol{X} - \boldsymbol{A}\boldsymbol{Z} - \boldsymbol{E}\|_\infty < \varepsilon, \|\boldsymbol{Z} - \boldsymbol{J}\|_\infty < \varepsilon.$$

End while

Output: $\boldsymbol{Z}$ and $\boldsymbol{E}$.

IV. EXPERIMENTAL RESULTS AND DISCUSSION

In this section, we will evaluate the performance of our proposed approach on the Weizmann [26], KTH [27] and UCF datasets [12, 13] and Weizmann robustness dataset. In the experiments, we use the leave-one-out cross validation (LOOCV) to evaluate the effectiveness of our algorithm unless otherwise stated. It employs actions from one person as the test samples, meanwhile leaving the remaining actions from other people as the training samples.

A. Action Datasets and Experimental Settings

We select three benchmark action datasets for performance evaluation. We also evaluate our approach with regards to its robustness. We first evaluate our proposed approach on the Weizmann dataset. This dataset contains 93 low-resolution video sequences from 9 different subjects, each of which performs 10 different actions including walking (walk), running (run), jumping (jump), galloping sideways (side), bending (bend), one-hand-waving (waveone), two-hands-waving (wavetwo), jumping in place (pjump), jumping jack (jack), and skipping (skip). One of the subjects performs walking, running and skipping twice. The dataset uses a fixed camera setting and a simple background. There is no occlusion or viewpoint change. Variations in spatial and temporal scale are also minimal. Examples of the Weizmann dataset can be seen in Fig. 1 (a).

We then evaluate our proposed approach on the KTH dataset. This dataset is relatively complex and can be considered as an important benchmark dataset to evaluate various human action recognition algorithms. KTH dataset contains 600 video clips in total. It consists of 6 types of human actions: walking (walk), jogging (jog), running (run), boxing (box), hand waving (hwav) and hand clapping (hclap). Each action is performed by 25 subjects under four different

environment settings: outdoors, outdoors with scale variation, outdoor with different clothes and indoors. Examples of the KTH dataset can be seen in Fig. 1 (b).

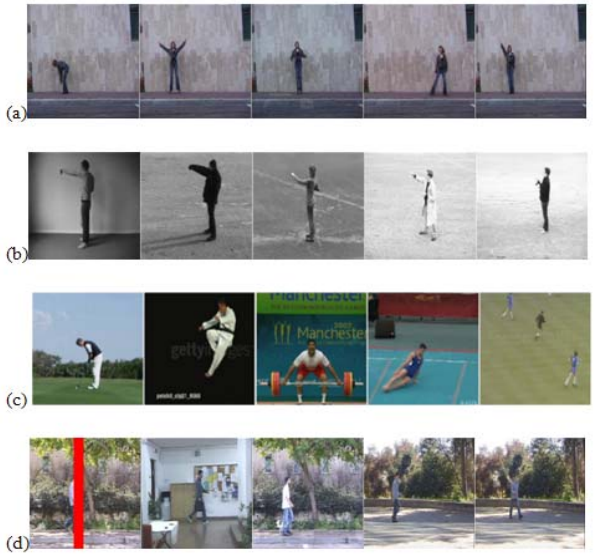


Fig. 1. Examples from the five public datasets: (a) Weizmann dataset; (b) KTH dataset; (c) UCF dataset; (d) Weizmann robustness dataset.

We also evaluate our proposed approach on the UCF dataset. This dataset contains 150 video sequences in total. It contains 10 different actions: diving, golf swing, horse riding, kicking, lifting, running, skating, swing bar, swing floor and walking. It is a challenging dataset with a wide range of scenarios and viewpoint variations. Examples of UCF dataset can be seen in Fig. 1(c).

Weizmann robustness: it includes 20 video sequences, ten of which are walking sequences under various difficult scenarios such as walking with partial occlusions, clothing changes and unusual walking styles. The others are walking sequences with ten different viewpoints (from 0^0 to 81^0 with the increasing speed of 9^0). Fig. 1 (d) shows some examples.

In all the experiments, we use Cuboid to extract and describe interest points. For Cuboid detector, we use standard spatial scale value 3 and temporal scale value 2. For the Cuboid descriptors, we use the optimal settings suggested in [3]. We set the dimension of each descriptor to 100 and normalize the feature matrix to the range of 0 to 1 in our experiments. Then k-means is employed to construct the dictionary, the number of the atoms in the dictionary is set 500. Finally, BoF, SR, LLC [24] and LRR are utilized respectively to obtain the final behavior representation.

B. Performance on Weizmann, KTH and UCF Dataset

We evaluate our algorithm on the Weizmann, KTH and UCF dataset. Four representation methods are compared under the same condition, which include BoF, SR, LLC and LRR. TABLE I shows the recognition results in the form of average recognition rate in comparison with different behavior representation schemes on three datasets. TABLE II shows the results of different classification schemes including CRC [28] and SRC on the Weizmann and KTH dataset. Confusion matrices in Fig. 2 shows in detail the average accuracy of the recognition of each action based on SRC classifier.

TABLE I. PERFORMANCE COMPARISON AMONG BOF, SR, LLC, AND LRR

Method	BoF	SR	LLC	LRR
Weizmann	94.7	94.7	95.6	96.7
KTH	87.7	93.2	93.5	93.2
UCF	73.3	75.3	79.3	87.6

TABLE II. PERFORMANCE COMPARISON ON WEIZMANN AND KTH DATASET

Method	CRC	SRC
Weizmann	96.9	96.7
KTH	90.8	93.2
UCF	82.7	87.6

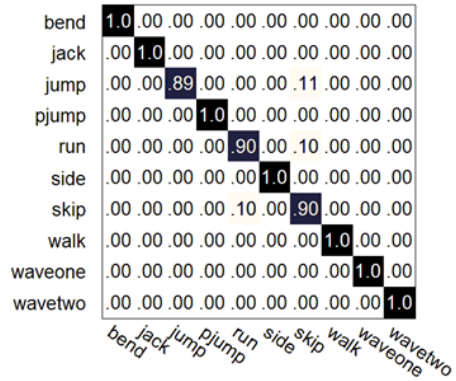
TABLE I shows the results on the Weizmann, KTH and UCF dataset. For the Weizmann dataset, the classification accuracy of LRR achieves 2% higher average accuracy than those of BoF and SR. Meanwhile, the classification accuracy of LRR achieves 1% higher accuracy than that of LLC, which considers a local structure. For the KTH dataset, the classification accuracy of LRR achieves higher average accuracy than that of BoF. In comparison with SR and LLC, the classification accuracy of LRR achieves competitive accuracy. For UCF dataset, LRR achieves the best average accuracy in comparison with other methods. On UCF dataset, we set the size of dictionary for BoF to 1000.We set the size of dictionary to 500 for the other representation methods.

TABLE II shows the recognition results on the three datasets based on different classification schemes. It can be seen that for Weizmann dataset, the classification accuracy of LRR combined with SRC achieves competitive accuracy. For KTH dataset, the classification accuracy of LRR combined with SRC achieves over 2% higher than from LRR combined with CRC. For UCF dataset, the classification accuracy of LRR combined with SRC achieves better result.

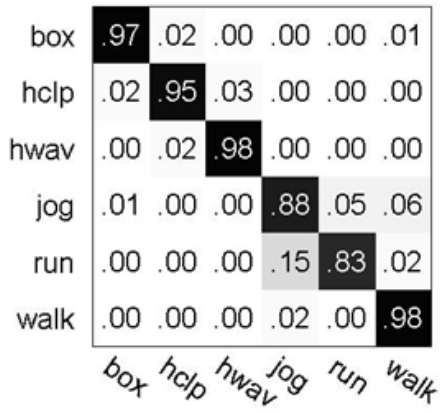
Though simple and effective, traditional BoF method would lead to relatively high reconstruction error by assigning each feature to its closest visual word. Although SR, LLC and LRR are linear coding methods, SR and LLC do not consider the relationship among the features, which encode the features individually. LRR encodes features as a whole. Thus, the relationship among the features is taken into consideration, which is the main difference from SR and LLC. Algorithm 1 shows the details.

Confusion matrices on the Weizmann, KTH and UCF datasets are shown in Fig. 2. In combination of the confusion matrix of the Weizmann dataset shown in Fig. 2(a), we find that some actions, e.g. jump and run, are wrongly classified as the action skip. In fact, those actions are similar in some key features which are considered to be included in intense response areas. From the confusion matrices of the KTH

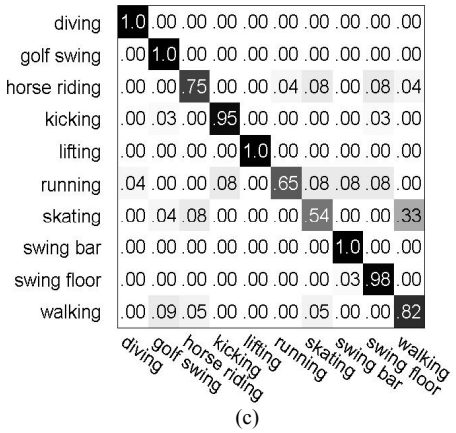
dataset shown in Fig. 2(b), we can see that there are several actions which are wrongly classified. Compared with the Weizmann dataset, the KTH dataset is more complex. We can also find that similar actions like "jog" and "run" are poorly recognized. However, those two types of actions "jog" and "run" are so difficult to classify that a human observer cannot accurately distinguish them.



(a)



(b)

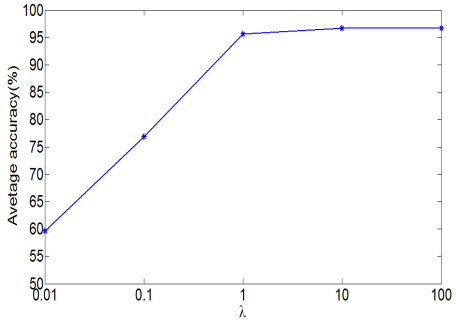


(c)

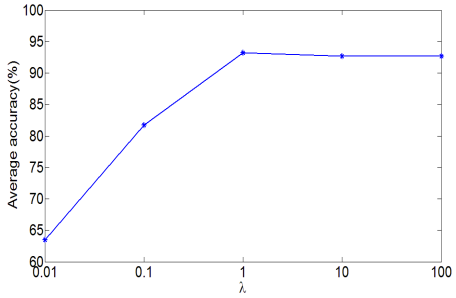
Fig. 2. Results based on LRR+SRC: (a) Weizmann; (b) KTH; (c) UCF datasets

C. *Parameter Analysis*

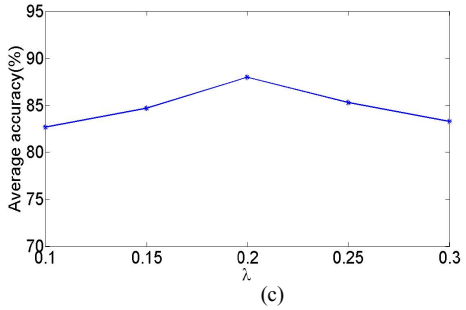
From the Eq. (7), we can see that there is a parameter λ in LRR. In this section, we mainly concern how to choose this parameter and how it affects the classification results. We conducted experiments on the Weizmann dataset and the KTH dataset to evaluate the sensitivity of the parameters λ. In the experiments, we choose LRR as the coding method and SRC as the classifier. The experimental results are shown in Fig. 3. We can see from Fig. 3(a) that when λ is less than 1, the recognition accuracy of LRR is not good. But when λ is larger than 1, the recognition accuracy of LRR is better and if λ is larger than 10, it achieves 96.7% accuracy. We set λ to 10 when we conduct experiments on Weizmann dataset. For KTH dataset, we can see from the Fig. 3(b), that when λ is larger than 1, the recognition accuracy of LRR is better and the best result is achieved when λ is around to 1. We set λ to 1when we conduct experiments on KTH dataset. For UCF dataset, we can see that from the Fig. 3(c), when λ is around 0.2, the accuracy of LRR is better. We set λ to 0.2 for this dataset and obtain the best result 87.6%.



(a)



(b)



(c)

Fig. 3. The sensitivity of the parameters λ: (a) Weizmann; (b) KTH and (c) UCF dataset

D. Robustness Evaluation

The robustness of our proposed approach and the approach in [3] is evaluated on the Weizmann robustness dataset [12]. We divide this dataset into two parts, one is with viewpoint changes (s1), and the other is with occlusion (s2). When s1 is tested, we employ both the Weizmann dataset and the s2 as training samples; similarly, when s2 is tested, Weizmann dataset and the s1 are both used as training samples.

TABLE III and TABLE IV present the results under s1 and s2 respectively. We also compare our results with the best results using random sample reconstruction (RSR) in [10].

TABLE III. RESULTS ON THE WEIZMANN ROBUSTNESS DATASET: PERFORMANCE UNDER S1

Test samples (walking in n°)	**BoF+NN[3]**	**Guha[10]**	**Ours**
n=0	walk	walk	walk
n=9	walk	walk	walk
n=18	walk	walk	walk
n=27	walk	walk	walk
n=36	walk	walk	walk
n=45	bend	walk	walk
n=54	walk	walk	walk
n=63	bend	walk	walk
n=72	walk	walk	skip
n=81	walk	skip	skip

TABLE IV. RESULTS ON THE WEIZMANN ROBUSTNESS DATASET: PERFORMANCE UNDER S2

Test samples	**BoF+NN[3]**	**Guha[10]**	**Ours**
walk with a bag	bend	walk	walk
walk with a briefcase	side	walk	walk
walk with a dog	waveone	walk	walk
knees up	bend	walk	walk
limp	walk	walk	walk
moonwalk	walk	walk	walk
no feet	waveone	walk	walk
norm walk	walk	walk	walk
occluded by a pole	walk	walk	walk
walk in a skirt	waveone	walk	walk

TABLE III and TABLE IV show that our approach exhibits robustness against viewpoint changes and small occlusion. In fact, each of the four actions performed by each people has viewpoint changes in the KTH dataset. We use the same descriptor as [3] while local motion pattern (LMP) descriptor is used in [10].

From TABLE III, we can see that our method is tolerant up to 63°, it is better than the method in [3]. When the viewpoint is more than 63°, our approach becomes invalid. We can see from TABLE IV that under the method in [3], many test samples are mistakenly labeled. Consequently, BoF representation is inaccurate and does not provide semantic information. Thus, LRR model is robust against occlusion and viewpoint changes to an extent.

V. CONCLUSION

In this paper, we presented a low-rank representation scheme to encode local spatio-temporal features. Given a set of local features, low-rank representation aims at finding the lowest-rank representation jointly. Thus, it can capture the global structure of local features from one action of each person. Specifically, a codebook is first created by utilizing the k-means algorithm. Then, local spatio-temporal features from one action of each person are represented by the codebook under the low-rank constraint. Finally, sparse representation based classification is used to recognize the actions. Experimental results demonstrate the effectiveness of the proposed approach on the Weizmann, KTH, and UCF datasets. Finally, we evaluate the robustness of our proposed approach on the Weizmann robustness dataset and the results show that our approach can work properly with certain occlusion and disturbance.

REFERENCES

[1] R. Poppe, "A survey on vision-based human action recognition," Image and Vision Computing, vol.28, no. 6, pp. 976-990, 2010.

[2] Bobick, A. F. and J. W. Davis, "The recognition of human movement using temporal templates." IEEE Tran. Pattern Analysis and Machine Intelligence, vol. 23, no. 3, pp: 257-267, 2011.

[3] P. Dollár, V. Rabaud, G. Cottrell, S. Belongie, "Behavior recognition via sparse spatio-temporal features," In: VS□PETS, pp. 65-72, 2005.

[4] I. Laptev, T. Lindeberg, "Velocity adaptation of Space–time interest points," Pattern Recognition, vol. 1, pp. 52-56, 2004.

[5] H. Wang, M.M. Ullah, A. Klaser, I. Laptev, C. Schmid, "Evaluation of local spatio-temporal features for action recognition," in Proceedings of British Machine Vision Conference, 2009.

[6] Wright J, Yang A Y, Ganesh A, Sastry S. S and Ma Y "Robust face recognition via sparse representation," IEEE Trans. Pattern Analysis and Machine Intelligence, 2009, 31(2): 210-227.

[7] H. Zhou, Y. Yuan, C. Shi, "Object tracking using SIFT features and mean shift," Computer Vision and Image Understanding, vol. 113, no. 3, pp. 345-352, 2009.

[8] J. Niebles, H. Wang, L. Fei-Fei, "Unsupervised learning of human action categories using spatial-temporal words," International Journal of Computer Vision, vol. 79, no. 3, pp.299-318, 2008.

[9] S. Lazebnik, C. Schmid, J. Ponce, "Beyond bags of features: spatial pyramid matching for recognizing natural scene categories," in Proceedings of IEEE Conference on Computer Vision and Pattern Recongnition, 2006, pp. 2169-2178.

[10] T. Guha, R. Ward, "Learning sparse representations for human action recognition," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 34, no. 8, pp. 1576-1588, 2012.

[11] Q. Wei, X. Zhang, Y. Kong, W. Hu, "Compact visual codebook for action recognition, "in Proceedings of the 17th IEEE International Conference on Image Processing, 2010,pp. 3805-3808.

[12] X. Zhang, Y. Yang, L.C. Jiao, F. Dong: "Manifold-constrained coding and sparse representation for human action recognition," Pattern Recognition, vol.46, no.7, pp.1819–1831, 2013.

[13] X. Zhang, H. Yang, L. Jiao, Y. Yang, F. Dong, "Laplacian group sparse modeling of human actions," Pattern Recognition, DOI:10.1016/j.patcog.2014.02.007, 2014.

[14] D. L. T. Fernando and J. B. Michael, "Robust Principal Component Analysis for Computer Vision," in Proceedings of the IEEE International Conference on Computer Vision, 2001, pp. 362-369.

[15] G. Liu, Z. Lin, and Y. Yu, "Robust subspace segmentation by low-rank representation," in Proceedings of the 27th International Conference on Machine Learning, 2010, pp. 663-670.

[16] I. Laptev, M. Marszalek, C. Schmid, "Learning realistic human actions from movies," in Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2008, pp.1-8.

[17] A. Kläser, M. Marszalek, C. Schmid, "A spatio-temporal descriptor based on 3D-gradients,"In the British Machine Vision Conference, 2008.

[18] P. Scovanner, S. Ali, M. Shah, "A 3-dimensional SIFT descriptor and its application to action recognition" In: ACM Multimedia, 2007.

[19] G. Willems, T. Tuytelaars, L. Van Gool, "An efficient dense and scale-invariant spatio-temporal interest point detector," In European Conference on Computer Vision, 2008. pp. 650-663.

[20] Z. Lin, M. Chen, L. Wu, and Y. Ma, "The augmented Lagrangemultiplier method for exact recovery of corrupted low-rank matrices," UIUC Technical Report UILU-ENG-09-2215, Tech. Rep., 2009

[21] J. Cai, E. Candes, and Z. Shen, "A Singular Value Thresholding Algorithm for Matrix Completion," SIAM Journal on Optimization, vol. 20, no. 4, pp. 1956-1982, 2010.

[22] J. Yang, W. Yin, Y. Zhang, Y. Wang. "A fast algorithm for edge preserving variational multichannel image restoration," SIAM Journal on Imaging Sciences, vol. 2, no. 2, pp. 569–592, 2009.

[23] Y. Zhu, X. Zhao, Y. Fu, "Sparse coding on local spatial-temporal volumes for human action recognition," in Proceedings of the10th Asian Conference on Computer Vision, New Zealand, 2010, pp. 660-671.

[24] J. Wang, J. Yang, K. Yu, "Locality-constrained linear coding for image classification," in Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2010, pp. 3360-3367.

[25] J. Yang, K. Yu, Y. Gong, T. Huang, "Linear spatial pyramid matching using sparse coding for image classification," in Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 1794-1801.

[26] M. Blank, L. Gorelick, E. Shechtman, M. Irani, R. Basri, "Action as space–time shapes," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 29, no. 12, pp. 2247-2253, 2005.

[27] C. Schuldt, I. Laptev, B. Caputo, "Recognizing human actions: a local SVM approach," in Proceedings of the 17th International Conference on Pattern Recognition, 2004, pp. 32–36.

[28] L. Zhang, M. Yang, X. Feng, "Sparse representation or collaborative representation: Which helps face recognition?," in Proceedings of the IEEE International Conference on Computer Vision, 2011, pp. 471-478.