IEICE TRANS. INF. & SYST., VOL.E101–D, NO.3 MARCH 2018

LETTER

Action Recognition Using Low-Rank Sparse Representation

Shilei CHENG[†a)], ***Nonmember***, **Song GU**[††], ***Member***, **Maoquan YE**[†], ***and*** **Mei XIE**[†], ***Nonmembers***

SUMMARY Human action recognition in videos draws huge research interests in computer vision. The Bag-of-Word model is quite commonly used to obtain the video level representations, however, BoW model roughly assigns each feature vector to its nearest visual word and the collection of unordered words ignores the interest points' spatial information, inevitably causing nontrivial quantization errors and impairing improvements on classification rates. To address these drawbacks, we propose an approach for action recognition by encoding spatio-temporal log Euclidean covariance matrix (ST-LECM) features within the low-rank and sparse representation framework. Motivated by low rank matrix recovery, local descriptors in a spatial temporal neighborhood have similar representation and should be approximately low rank. The learned coefficients can not only capture the global data structures, but also preserve consistent. Experimental results showed that the proposed approach yields excellent recognition performance on synthetic video datasets and are robust to action variability, view variations and partial occlusion.

key words: *human action recognition, low-rank sparse representation, bag of word model, sparse coding representation, low-rank representation*

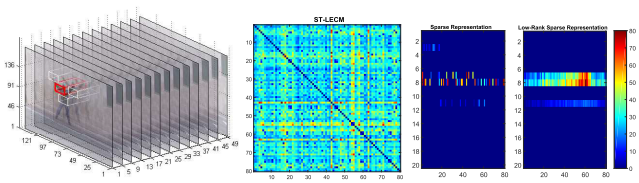
1. Introduction

Human action recognition has been an important topic in the field of computer vision. Although many significant results have been reported on human action recognition, it still remains a challenging problem due to occlusions, view variations, and background cluster [1]. Human action representation is the process of computing the time evolution of human silhouettes, the action cylinders, the space-time shapes, and the local $3D$ patch descriptors and it is usually considered as a key issue in action recognition. They can be classified into two categories: global representations [2] and local representations [3]. Sparse representation (SR) has been widely used and achieved promising results in pattern recognition [4]. It mainly develop within the framework of particle filter, where it is used to measure the similarity between the particle and the dictionary with the reconstruction error. According to Liu *et al.* [4], the local structure of information can be captured by the sparse representation of data. However, sparse representation is sensitive to noise which is frequently happened in occlusions and background clutter. Thus we need fuse global structure of data to enhance the robust of action representation.

In order to capture the global structure of the data, Liu *et al.* [5] propose the low-rank representation (LRR) with the hope that if two samples are close in the intrinsic geometry over the data distribution, they will have a large similarity coefficient. Unlike the sparse representation, the aim of low-rank representation is to find the lowest rank representation of all the test sample jointly [6]. Zhang *et al.* [7] has firstly proposed LRR in human action recognition and the results demonstrate LRR approach can capture the global structure of local features. However, LRR approach is not only computationally expensive for large sets of features but also it only captures the global Euclidean structure, while the local manifold structure, which is often important for the view variations in action recognition, has been ignored.

From Fig. 1 (c), we see that the local features have inconsistent representation, i.e. their features are similar but their codes and the supports of their codes are not. This is because SR represents each feature independently. However, the representation learnt by LRS are joint sparse, i.e. a few but the same visual words are used to represent all the local features together, which renders the representation consistent and more robust to noise.

In recent years, there has been a growing interest in deep learning based action recognition approaches, that can learn multiple layers of feature hierarchies and automatically build high-level representations of the raw video. The typical methods include 3D ConvNets [8], Deep ConvNets [9]. However, these deep learning based methods fail to outperform conventional hand-crafted features. One problem of deep learning methods is that they require a large number of labeled videos for training, while most available datasets are relatively small. Moreover, most of current deep learning based action recognition methods largely ignore the intrinsic difference between temporal domain and



(a) (b) (c) (d)

Fig. 1 A feature representation example in a local region. (a) STIP detector; (b) All ST-LECM Features in the local region depicted in red in (a); (c), (d) are representation results produced by SR [10] and LRS.

Manuscript received August 9, 2017.
Manuscript revised October 24, 2017.
Manuscript publicized November 24, 2017.

[†]The authors are with School of Electronic Engineering, University of Electronic Science and Technology of China, P. R. China.
[††]The author is with Department of Aircraft Maintenance Engineering, Chengdu Aeronautic Polytechnic, Chengdu, P. R. China.
a) E-mail: slcheng1986@foxmail.com

DOI: 10.1587/transinf.2017EDL8176

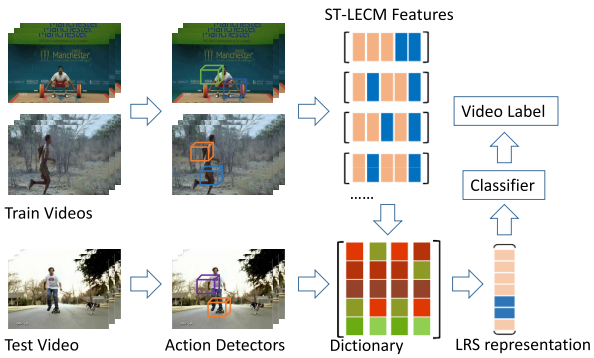


Fig. 2 Our LRS framework for building the local and global representation and training action classifiers.

spatial domain. Our method propose two components that are spatial features which mainly capture the discriminative appearance features for action understanding, meanwhile, temporal features are also included which obtain the motion characteristic.

In this paper, we firstly extract Spatio-Temporal Interest Points (STIPs) by STIPs detector [11] and introduce new descriptors called spatio-temporal log-Euclidean covariance matrix features as Fig. 1 (a), (b) proposed respectively. Figure 2 shows the flowchart of our framework. For the representation model, we further employ low rank regularizer to the traditional sparse coding objective function. The low rankness enforces similar descriptors to have similar sparse codes, which considers the global geometrical structure of the data. Our last work is to design a classification scheme for query action video, which is based on the idea of joint low-rank and sparse representation classification (JLRSRC).

2. Spatio-Temporal Log-Euclidean Covariance Matrix (ST-LECM) FEATURES

We first introduce the form of the spatio-temporal (ST) covariance descriptors, which has shown outstanding performance in action and gesture recognition [12]. More specifically, for a given $3D$ volume R, we extract the raw feature vector $f(x, y, t)$ from pixel position (x, y, t) inside R, and $f(x, y, t) = [g, o]^T$, while g and o are defined as:

$$g = [|I_x|, |I_y|, |I_{xx}|, |I_{yy}|, \sqrt{I^2{}_x + I^2{}_y}, \arctan \frac{|I_y|}{|I_x|}] \quad (1)$$

$$o = [u, v, w, \frac{\partial u}{\partial t}, \frac{\partial v}{\partial t}, \frac{\partial w}{\partial t}] \quad (2)$$

In Eq. (1), according to our previous work [13], $|\cdot|$ denotes the absolute value, I_x, I_{xx} denote the first and second-order partial derivative with respect to x at position (x, y), respectively. I_y, I_{yy} have similar definitions, the last two gradient features represent gradient magnitude and gradient orientation.

The 3D optical-flow based feature in Eq. (2) represent in order: the horizontal component (u, v) and one vertical components (w) of the flow vector whose Gauss-Seidel iterative equations can be defined as:

$$\begin{aligned} u^{k+1} &= \overline{u^k} - \frac{I_x\left[I_x\overline{u^k} + I_y\overline{v^k} + I_z\overline{w^k} + I_t\right]}{\alpha^2 + I_x^2 + I_y^2 + I_z^2} \\ v^{k+1} &= \overline{v^k} - \frac{I_y\left[I_x\overline{u^k} + I_y\overline{v^k} + I_z\overline{w^k} + I_t\right]}{\alpha^2 + I_x^2 + I_y^2 + I_z^2} \\ w^{k+1} &= \overline{w^k} - \frac{I_z\left[I_x\overline{u^k} + I_y\overline{v^k} + I_z\overline{w^k} + I_t\right]}{\alpha^2 + I_x^2 + I_y^2 + I_z^2} \end{aligned} \quad (3)$$

where $\overline{u^k}$, $\overline{v^k}$, $\overline{w^k}$ are average velocities at iteration k, $\overline{u^0}$, $\overline{v^0}$, $\overline{w^0}$ are typically set to 0, I_x, I_y and I_z are the spatial intensity gradient. I_t is the temporal intensity derivative, which calculates difference between adjacent frames, α is a Lagrangian multiplier. The last three components Eq. (2) represents first-order derivatives of the flow components with respect to t.

Given N feature vectors $f(x, y, t)$, the ST covariance descriptor $Cov3D$ is computed as $Cov3D = \frac{1}{N-1}(\boldsymbol{D} - \mu)(\boldsymbol{D} - \mu)^T$, where $\boldsymbol{D} = (f(x_1, y_1, t_1), \ldots, f(x_n, y_n, t_n))$, $\mu = \frac{1}{N}\sum_{i=1}^{n} f(x_i, y_i, t_i)$. Concretely, $D \in \Re^{m\times n}$ with m dimensions and n sizes, is a column vector matrix, and each vector of column $f(x_i, y_i, t_i) \in \Re^{m\times 1}$ is the ST covariance descriptors from pixel position (x_i, y_i, t_i), μ is the mean of all features. Since covariance matrices are Symmetric Positive Definite (SPD) matrices, which form a special type of Riemannian manifold. To measure the distance of SPDs while avoid computing the geodesic distance between them, in this study, we further extend these types of feature to build ST-LECM features, which is referred to our previous work [14], we utilize the Log-Euclidean framework that is to map the $n \times n$ covariance descriptor in the commutative Lie group into $\log Cov3D$ in vector space by matrix logarithm. Besides, $\log Cov3D$ is a symmetric matrix of Euclidean space. Due to the symmetry property, $\log Cov3D$ has only $\frac{n\times(n+1)}{2}$ independent entries (elements on and above the main diagonal). Thus we use $V(\log Cov3D_R) \in \Re^m$, where $m = \frac{n\times(n+1)}{2}$ to denote the final ST-LECM feature descriptor.

3. Low-Rank Sparse Representation for Action Recognition

We consider p dimension action features which denoted as: $\boldsymbol{X} = [x_1, x_2, \ldots, x_p]$, $x_i \in \Re^m$, where $m = \frac{n\times(n+1)}{2}$ and x_i denotes the ST-LECM feature descriptor: $V(\log Cov3D_R)$. In the noiseless cases, each test sample x_i is represented as linear combination of templates which form dictionary such that $\boldsymbol{X} = \boldsymbol{DZ}$, where, $\boldsymbol{D} = [d_1, d_2, \ldots, d_m]$, The columns of Z denote the representation of the feature samples with respect to $\boldsymbol{D}$. In real visual scenarios, data often corrupted by noise or even grossly corrupted, thus we add a noise term E, that is $\boldsymbol{X} = \boldsymbol{DZ} + \boldsymbol{E}$. Since some of the interest points in one action have similar representations with respect to dictionary, the resulting representation matrix is expected to be low rank. Moreover, only a few dictionary templates are required to reliably represent the interest points, therefore the representation matrix has the property of sparseness. As we

analysed above, cuboid descriptors from one action of each person are represented by dictionary under both sparse regularizer and low-rank regularizer, thus we can obtain a low-rank sparse recovery to $\boldsymbol{X}$ by solving the following convex optimization problem: Eq. (4)

$$\begin{aligned} &\underset{Z,E}{\operatorname{argmin}} \|\boldsymbol{Z}\|_* + \beta\|\boldsymbol{W}\|_1 + \lambda\|\boldsymbol{E}\|_1 \\ &s.t. \quad \boldsymbol{X} = \boldsymbol{DZ} + \boldsymbol{E},\ \boldsymbol{Z} = \boldsymbol{W} \end{aligned} \tag{4}$$

where $\|\cdot\|_*$ is the nuclear norm, which approximates the rank of $\boldsymbol{Z}$, each column of $\boldsymbol{Z}$ is a particle in $\Re^d$, $\|\cdot\|_1$ is L_1 norm that represent sparse recovery of $\boldsymbol{X}$, $\boldsymbol{W}$ is an auxiliary variable that we introduce to make the objective function separable. β is a parameter that provides a trade-off between the low-rank and sparse, λ is a parameter that controls the effect of the noise. $\boldsymbol{D}$ is the dictionary whose atoms are constructed by clustering. Actually LRS has close relation with LRR [7], when we set $\beta = 0$, it transforms to LRR [7] approach exactly. To minimize the above problem, we employ the conventional Inexact Augmented Lagrange Multiplier (IALM) method which has extensively used in matrix rank minimization problems due to its quadratic convergence properties aiming at non-smooth optimization problem [15]. The augmented Lagrangian function of problem (4) is given as follow:

$$\begin{aligned} &L(\boldsymbol{Z}, \boldsymbol{W}, \boldsymbol{E}, \boldsymbol{Y_1}, \boldsymbol{Y_2}, \mu) \\ &= \|\boldsymbol{Z}\|_* + \beta\|\boldsymbol{W}\|_1 + \lambda\|\boldsymbol{E}\|_1 \\ &\quad + \langle \boldsymbol{Y_1}, \boldsymbol{X} - \boldsymbol{DZ} - \boldsymbol{E}\rangle + \langle \boldsymbol{Y_2}, \boldsymbol{Z} - \boldsymbol{W}\rangle \\ &\quad + \frac{\mu}{2}\left(\|\boldsymbol{X} - \boldsymbol{DZ} - \boldsymbol{E}\|_F^2 + \|\boldsymbol{Z} - \boldsymbol{W}\|_F^2\right) \end{aligned} \tag{5}$$

where $\boldsymbol{Y_1}$ and $\boldsymbol{Y_2}$ are Lagrangian multipliers, and $\mu > 0$ is a penalty parameter. Accordingly, the pseudo-code and optimization scheme are summarized in Algorithm 1.

4. Classification Approach

Sparse representation based classification (SRC) has been successfully used in face recognition and image classification, however as we observed in [5] sparse representation is inaccurate at capturing the global structures of data, while low-rank representation compensate deficiency of sparse representation. Inspired by this, we propose video-based joint low-rank and sparse representation-based classification (JLRSRC). Given a training data matrix $\boldsymbol{X} = [x_1, x_2, \ldots, x_m] \in \Re^{M\times N}$ and a test data matrix $\boldsymbol{Y} = [y_1, y_2, \ldots, y_C] \in \Re^{M\times D}$, we can find the low-rankest and sparsest representation matrix $\boldsymbol{Z} = [z_1, z_2, \ldots, z_D]$ and the error matrix $\boldsymbol{E} = [e_1, e_2, \ldots, e_D]$ by Algorithm 1. We then classify $\boldsymbol{Y}$ by assigning it to the expression class that minimizes the residual as follows:

$$r(y_j) = \underset{i}{\operatorname{argmin}} \|y_j - \boldsymbol{X}\delta_i(z_j) - e_j\|_2 \tag{12}$$

where $\delta(\cdot)$ is the Dirac delta function and $\delta_i(z_j)$ indicates a vector whose only nonzero entries are the entries in z_j that are associated with class i. Algorithm 2 summarizes the complete classification procedure.

Algorithm 1

input: the codebook $\boldsymbol{D}$, n local features for one video $\boldsymbol{X} = [x_1, x_2, \ldots, x_n]$, parameter λ, β

output: $\boldsymbol{Z}, \boldsymbol{W}, \boldsymbol{E}$

Initialize $\boldsymbol{Z} = \boldsymbol{W} = \boldsymbol{E} = 0$, $\boldsymbol{Y}_1 = \boldsymbol{E}$, $\boldsymbol{Y}_2 = \boldsymbol{Z}$, $\mu = 0.1$, $\mu_{max} = 10^{10}$

while not converged **do** We denote

$$p(\boldsymbol{Z}) = p(\boldsymbol{Z}_k) + \langle \nabla_z p(\boldsymbol{Z}_k), \boldsymbol{Z} - \boldsymbol{Z}_k\rangle + \frac{\eta\mu_k}{2}\|\boldsymbol{Z} - \boldsymbol{Z}_k\|_F^2 \tag{6}$$

$$p(\boldsymbol{W}) = p(\boldsymbol{W}_k) + \langle \nabla_w p(\boldsymbol{W}_k), \boldsymbol{W} - \boldsymbol{W}_k\rangle + \frac{\mu_k}{2}\|\boldsymbol{W} - \boldsymbol{W}_k\|_F^2 \tag{7}$$

Compute $\boldsymbol{Z}$

$$\boldsymbol{Z}_{k+1} = \mathscr{J}_{(\eta\mu_k)^{-1}}(\boldsymbol{Z}_k - \frac{1}{\eta\mu_k}\nabla_z\, p(\boldsymbol{Z}_k)) \tag{8}$$

where $\nabla_z p(\boldsymbol{Z}_k)$ is the partial differential of $P(\boldsymbol{Z})$, and we set $\eta = \|\boldsymbol{D}\|_2^2$.

Update $\boldsymbol{W}$

$$\boldsymbol{W}_{k+1} = \mathscr{S}_{\frac{\beta}{\mu_k}}(\boldsymbol{W}_k - \frac{1}{\mu_k}\nabla_w\, p(\boldsymbol{W}_k)) \tag{9}$$

where $\nabla_w p(\boldsymbol{W}_k)$ is the partial differential of $P(\boldsymbol{W})$

Estimate $\boldsymbol{E}$

$$\boldsymbol{E}_{k+1} = \mathscr{S}_{\frac{\lambda}{\mu_k}}(\boldsymbol{E}_k - \frac{1}{\mu_k}\nabla_w\, p(\boldsymbol{W}_k)) \tag{10}$$

In .8, .9, .10, $\mathscr{J}$ and $\mathscr{S}$ is the singular value thresholding [16]

Update $\boldsymbol{Y}_1$ and $\boldsymbol{Y}_2$

$$\begin{aligned} \boldsymbol{Y}_{1,k+1} &= \boldsymbol{Y}_{1,k} + \mu_k(\boldsymbol{X} - \boldsymbol{DZ}_k - \boldsymbol{E}_k) \\ \boldsymbol{Y}_{2,k+1} &= \boldsymbol{Y}_{2,k} + \mu_k(\boldsymbol{Z}_k - \boldsymbol{W}_k) \end{aligned} \tag{11}$$

Update the parameter μ by $\min(\rho\mu, \mu_{max})$

end while

Algorithm 2

input: the codebook $\boldsymbol{D}$, n local features for one test video $\boldsymbol{Y} = [y_1, y_2, \ldots, y_n]$, parameter λ, β.

output: the class label of the test sample.

1: Normalize the column of $\boldsymbol{Y}$ to have unit L_2-norm.

2: Solve the optimization problem 4 to obtain the representation matrix Z and the error matrix $\boldsymbol{E}$ by Algorithm 1.

3: Compute the residuals $r_i(\boldsymbol{Y})$ defined in 12.

5. Experiments

In this section, two public video datasets, KTH [11] and UCF Sports datasets [17], are used to evaluate the performance of our LRS representation and JLRSRC classification methods based on ST-LECM features. In all experiments, to generate covariance matrices, a set of overlapping ST blocks are extracted from the image sequence over a spatial grid with different scales. Next, the ST-LECM features of training videos are clustered by k-means clustering method. Finally, we compare our proposed method with Bag of Word (BoW) representation [3], Sparse Representation (SR) [10], Low Rank Representation (LRR) [4].

5.1 KTH Action Datasets

The KTH dataset contains six human action classes: boxing, hand-clapping, hand-waving, jogging, running, and walking, all of which performed by 25 subjects in 4 scenarios. KTH dataset is relatively complex w.r.t. view variations and can be considered as an important benchmark dataset to evaluate various human action recognition algorithms. Specifically, we extract overlapping ST blocks from the video segments over a spatial grid with spacing of 5 pixels, and the size of ST block is set as $14 \times 14 \times 15$. Then, we randomly select ST-LECM features of 24 videos belonging to the same class are clustered by k-means algorithm. In this experiment, we conduct experiments with different codewords (e.g. 0.5k, 1k, 1.5k). We use vector quantization method to represent the action for BoW, as Table 1 shown, other representation methods always outperform the baseline BoW+STIP in KTH dataset. Moreover, we can observe that ST-LECM+LRS seems to receive better performance, which improve the recognition rate about 1.1% compared with LMP+SR and STIP+LRR while codewords set to 1k, in the case of 1.5k of codewords, it improve 1.3% and 1.2 respectively. For all the methods compared, the accuracy increases as codewords increases.

5.2 UCF Sports Datasets

UCF sports dataset which contains 150 videos in total with 10 different action, is a challenging dataset with a wide range of scenarios and viewpoint variations. In this subsection, we mainly concern the parameters selection and explore how they affects the classification results. Specifically, we extract a set of overlapping ST blocks from the video segments over a spatial grid with spacing of 7 pixels and the size of ST block is set as $18 \times 16 \times 15$. Then, the ST-LECM features of 40 videos (4 action videos are randomly selected from each class) are clustered by k-means clustering method, likewise we evaluate different size of initial centroids, i.e. 1k, 2k, 4k. Figure 3 shows the evaluation results of ST-LECM+LRS with different values of β and λ. β is generally computed through $\beta = \frac{1}{\sqrt{\max(m,n)}}$, m, n is the dimension and size of the test sample respectively, in this work we set $m = 80$, $n = 400$, thus numerical calculations are performed for the values of parameters $\beta = [0, 0.05, 0.55, 1.05, 1.55]$, besides we empirically set parameter $\lambda = [0.1, 0.2, 0.3, 0.4]$. When initial cluster centers are set 1k, we find classification precision in the ranges of $[63.3\%, 85.13\%]$ and the best recognition result achieves at 85.3%, when $\beta = 0.05$ and $\lambda = 0.1$, while initial cluster centers are set 2k, compared with the case of 1k, the accuracy of recognition make a little progress which achieves in the ranges of $[73.35\%, 86.93\%]$, the highest accuracy of recognition rate achieves at 89.19%, when we set 4k cluster centers and $\beta = 1.05$, $\lambda = 0.3$.

Table 1 Comparison of different representation methods on KTH.

Method	Codewords	Classification	Accuracy
STIP+BoW	0.5k	linearSVM	81.7%
STIP+BoW	1.0k	linearSVM	87.7%
STIP+BoW	1.5k	linearSVM	**88.4**%
LMP+SR	1k	RSR	93.2%
STIP+LRR	1k	SRC	93.2%
Ours	1k	JRSRC	**94.3**%
LMP+SR	1.5k	RSR	93.4%
STIP+LRR	1.5k	SRC	93.5%
Ours	1.5k	JRSRC	**94.7**%

5.3 Compared with Deep Learning Approaches

We compare the proposed method with a few deep learning based baselines: 3D ConvNets [8] and Sequential Deep Model [18]. 3D ConvNet is one of the best performing deep learning architectures which is pretrained on Caffe's Sports1M datasets [9], from this framework, we can obtain the current popular-used features namely C3D, yield from fc6 and fc7. For another, Sequential Deep Model (SDM) is also a neural-based model to classify sequences of human actions, without a priori modeling, but only relying on automatic learning from training examples, by combining with LSTM [24], SDM leads to significant performance improvement on some challenging datasets. In this experiment, we extract 3D ConvNet's fc6 features for each frame, average these frame features to make video descriptors. Due to 3D ConvNet is not an end-to-end network, it is necessary to use a multi-class SVM to classify test videos. Moreover, to evaluate the performance of SDM, we propose to use Recurrent Neural Network architecture with LSTM cells, we have tested several network configuration, varying the number of hidden LSTM, a configuration of 50 LSTM has been found to be good for KTH dataset. Finally, we consider evaluating these two approaches on KTH dataset and comparing with our best result yield from our proposed method.

We report the result of LRS and compare with deep learning methods in Table 2, from which we can clearly see that our proposed method performs best among 3D ConvNets and SDM described previously. C3D using one net which has only 4,096 dimensions feature descriptor, obtains an accuracy of 90.1%, although it has been proved that C3D is capable of learning appearance and motion simultaneously, this indicates our method can better capture both appearance and motion information. The performance gap between SDM and LRS, however, is small (about 0.3%), due

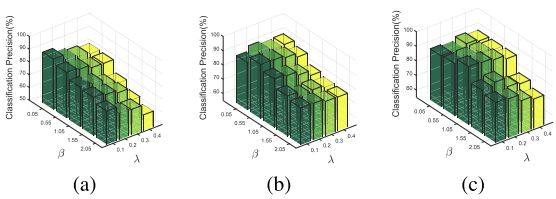


Fig. 3 Effects of parameter selection of β and λ with different size of initial centroids on the classification accuracy on UCF Sports dataset, (a), (b), (c) represent the accuracy with 1k, 2k, 4k centroids respectively.

Table 2 Action recognition results on KTH. Our methods compared with baselines with current competitive deep learning methods.

Method	Accuracy
C3D+linearSVM	90.1%
SDM+LSTM	94.39%
Ours	**94.7**%

to SDM need not any pretrained model, we generate vertically flipped and mirrored versions of each training sample to increase the number of examples, the total training samples contains 2271 sequences (with length between 8 and 59 seconds), which is about 4.5 times larger than our method. All the experiments have been carried out on the platform which mainly contains two NVIDIA GTX1080Ti GPUs, Intel core i7-6700K CPU and 32GB RAM.

6. Conclusion

In this paper, we present a low-rank sparse representation approach to encode local and global spatial-temporal Log-Euclidean covariance matrix features. The low-rank sparse representation is proposed to find the lowest rank and sparse representation jointly when a set of local features is given. Moreover, joint low-rank and sparse representation-based classification is devised to evaluate the performance of recognition. Experimental results demonstrate the robustness of the proposed approach on synthetic video datasets. In future, we will focus on the construction of deep discriminative spatial-temporal descriptors set through dictionary learning approach in order to improve the accuracy of action recognition with large datasets.

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] C. Deng, X. Cao, H. Liu, and J. Chen, "A global spatio-temporal representation for action recognition," 2010 20th International Conference on Pattern Recognition (ICPR), pp.1816–1819, IEEE, 2010.
[3] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse spatio-temporal features," 2nd Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance, pp.65–72, IEEE, 2005.
[4] F. Liu, J. Tang, Y. Song, X. Xiang, and Z. Tang, "Local structure based sparse representation for face recognition with single sample per person," 2014 IEEE International Conference on Image Processing (ICIP), pp.713–717, IEEE, 2014.
[5] G. Liu, Z. Lin, and Y. Yu, "Robust subspace segmentation by low-rank representation," Proc. 27th International Conference on Machine Learning (ICML-10), pp.663–670, 2010.
[6] J. Wei, X. Qi, and M. Wang, "Study on representation methods for classification," Advanced Science and Technology Letters, vol.53 (ISI 2014), pp.145–151, 2014.
[7] X. Zhang, Y. Yang, H. Jia, H. Zhou, and L. Jiao, "Low-rank representation based action recognition," 2014 International Joint Conference on Neural Networks (IJCNN), pp.1812–1818, IEEE, 2014.
[8] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," 2015 IEEE International Conference on Computer Vision (ICCV), pp.4489–4497, 2015.
[9] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," 2014 IEEE Conference on Computer Vision and Pattern Recognition, pp.1725–1732, 2014.
[10] T. Guha and R.K. Ward, "Learning sparse representations for human action recognition," IEEE Trans. Pattern Anal. Mach. Intell., vol.34, no.8, pp.1576–1588, 2012.
[11] I. Laptev, "On space-time interest points," International Journal of Computer Vision, vol.64, no.2-3, pp.107–123, 2005.
[12] A. Sanin, C. Sanderson, M.T. Harandi, and B.C. Lovell, "Spatio-temporal covariance descriptors for action and gesture recognition," 2013 IEEE Workshop on Applications of Computer Vision (WACV), pp.103–110, IEEE, 2013.
[13] S. Gu, J. Wang, L. Pan, S. Cheng, Z. Ma, and M. Xie, "Figure/ground video segmentation via low-rank sparse learning," IEEE International Conference on Image Processing, pp.864–868, 2016.
[14] S. Cheng, J. Yang, Z. Ma, and M. Xie, "Action recognition based on spatio-temporal log-Euclidean covariance matrix," International Journal of Signal Processing, Image Processing and Pattern Recognition, vol.9, no.2, pp.95–106, 2016.
[15] T. Zhang, S. Liu, N. Ahuja, M.-H. Yang, and B. Ghanem, "Robust visual tracking via consistent low-rank sparse learning," International Journal of Computer Vision, vol.111, no.2, pp.171–190, 2015.
[16] J.-F. Cai, E.J. Candès, and Z. Shen, "A singular value thresholding algorithm for matrix completion," SIAM Journal on Optimization, vol.20, no.4, pp.1956–1982, 2010.
[17] M.D. Rodriguez, J. Ahmed, and M. Shah, "Action MACH a spatio-temporal maximum average correlation height filter for action recognition," 2008 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2008, pp.1–8, IEEE, 2008.
[18] M. Baccouche, F. Mamalet, C. Wolf, C. Garcia, and A. Baskurt, "Sequential deep learning for human action recognition," Human Behavior Unterstanding, Lecture Notes in Computer Science, vol.7065, pp.29–39, Springer, Berlin, Heidelberg, 2011.
[19] J.C. Niebles, C.-W. Chen, and L. Fei-Fei, "Modeling temporal structure of decomposable motion segments for activity classification," Computer Vision – ECCV 2010, Lecture Notes in Computer Science, vol.6312, pp.392–405, Springer, Berlin, Heidelberg, 2010.
[20] J. Yang, W. Yin, Y. Zhang, and 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.
[21] Z. Lin, R. Liu, and H. Li, "Linearized alternating direction method with parallel splitting and adaptive penalty for separable convex programs in machine learning," Machine Learning, vol.99, no.2, pp.287–325, 2013.
[22] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," IEEE Trans. Pattern Anal. Mach. Intell., vol.29, no.12, pp.2247–2253, Dec. 2007.
[23] I. Laptev, B. Caputo, C. Schüldt, and T. Lindeberg, "Local velocity-adapted motion events for spatio-temporal recognition," Computer Vision and Image Understanding, vol.108, no.3, pp.207–229, 2007.
[24] F.A. Gers and N.N. Schraudolph, "Learning precise timing with LSTM recurrent networks," Journal of Machine Learning Research, vol.3, no.Aug, pp.115–143, 2002.