International Journal of Signal Processing, Image Processing and Pattern Recognition
Vol.9, No.2 (2016), pp.95-106
http://dx.doi.org/10.14257/ijsip.2016.9.2.09

Action Recognition Based on Spatio-temporal Log-Euclidean Covariance Matrix

Shilei Cheng[2], Jiangfeng Yang[1], Zheng Ma[1], Mei Xie[2]

[1]School of Communication and Information Engineering
[2]School of Electronic Engineering
[1, 2]University of Electronic Science and Technology of China, Xiyuan Ave, No.2006, West Hi-Tech Zone, 61173
wallsonyang@163.com, 369322023@qq.com

Abstract

In this paper, we handle the problem of human action recognition by combining covariance matrices as local spatio-temporal (ST) descriptors and local ST features extracted densely from action video. Unlike traditional methods that separately utilizing gradient-based feature and optical flow-based feature, we use covariance matrix to fuse the two types of feature. 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, covariance features are transformed to log-Euclidean covariance matrices (LECM) by matrix logarithm operation. After encoding LECM by Locality-constrained Linear Coding method, in order to provide position information to ST-LECM features, spatial pyramid is used to partition the video frames, and the average-pooling-on-absolute-value function is implemented over each sub-frames. Finally, non-linear support vector machine is used as classifier. Experiments on public human action datasets show that the proposed method obtains great improvements in recognition accuracy, in comparison to several state-of-the-art methods.

Keywords: *image processing, action recognition, spatio-temporal covariance matrix*

1. Introduction

Human action recognition has received significant attention in several video analysis tasks, mainly because of its applications to content-based video analysis, visual surveillance, and human-computer interaction. Many methods have been proposed for reliable action recognition based on various feature detectors/descriptors to capture local motion patterns. Recently, dense spatial-temporal representation of action videos has been recently shown to be promising for the action classification task.

Covariance matrices as image local descriptors have been successfully applied in image classification. In [1], they were firstly proposed by Tuzel *et. al.*, and since then they have been employed successfully for pedestrian detection [2], non-rigid object tracking [2], face recognition [3], and analyzing diffusion tensor images [4]. Furthermore, a ST version of covariance matrix descriptors has shown superior performance for action/gesture recognition [5].

Using covariance matrix as a region descriptor has several advantages:

- Firstly, it captures second-order statistics of the local features.
- Secondly, it is straightforward approach to fusing various features.
- Thirdly, it is a low dimensional descriptor and is independent of the size of the region.

ISSN: 2005-4254 IJSIP

- Fourth, through the averaging process in its computation, the impact of the noisy samples is reduced.
- Finally, efficient methods for its fast computation in images and videos are available.

While the above advantages make covariance-based features attractive, using them for discrimination purposes can be challenging. Covariance matrices are Symmetric Positive Definite (SPD) matrices, the space of which is not a Euclidean space but a smooth Riemannian manifold. In the Log-Euclidean framework [6], the SPD matrices form a commutative Lie group which is equipped with a Euclidean structure. This framework inspires us to compute the logarithms of SPD matrices, which can then be flexibly and efficiently handled with common Euclidean operations.

Our contribution. In the paper, we utilize region covariance matrices as the local descriptors to capture the local motion information, and covariance matrices are treated as points on a Riemannian manifold. We firstly form ST covariance descriptors, which combine image intensity and motion optical flow information, from dense sampling motion-based features. The covariance descriptors are then encoded in a Log-Euclidean Bag-of-Features (LE-BoF) model. To achieve this, we employ a diffeomorphism and form the LE-BoF model by embedding the Riemannian manifold into a tangent vector space. The embedding is obtained with flattening the manifold into a corresponding tangent space, and Locality-constrained Linear Encoding (LLC) method is employed to encode the LE-BoFs. The proposed action recognition system was compared with the recent systems proposed by Wang *et. al.*, [15], Messing *et. al.*, [7], and Niebles *et. al,*. [8], and experiment results on two datasets (KTH [9], Activity of Daily Living [7]) show that the proposed approach obtains an impressed performance.

We organize the rest paper as follows: Section 2 presents ST Log-Euclidean Covariance Matrix (ST-LECM). Section 3 shows the framework of our system. In Section 4 we compare the performance of the proposed method with previous approaches on datasets.

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

The space of Symmetric Positive Definite (SPD) matrices is not a vector space but a Riemannian manifold (an open convex half-cone). Hence, the conventional Euclidean operations, *e.g.*, the Euclidean distance, mean or the statistics do not apply. Two class of Riemannian framework have been presented for dealing with SPD matrices: the affine-invariant Riemannian framework [12] and the Log-Euclidean Riemannian framework [6]. The latter has almost the same good theoretical properties as the former, and in the meantime enjoys a drastic reduction in computational cost. In the following we first introduce briefly the Log-Euclidean Framework and then present the proposed ST-LECM features.

2.1. Log-Euclidean Framework on SPD Matrices

‘Matrix exponential and logarithm operations. The matrix exponential and logarithm are fundamental to the Log-Euclidean framework. Let $SPD(n)$ and $S(n)$ denote the space of $n \times n$ real SPD matrices and $n \times n$ real symmetric matrices, respectively. Any matrix $\mathbf{S} \in S(n)$ can be represented as the eigen-decomposition of the form $\mathbf{S} = \mathbf{U}\Lambda\mathbf{U}^T$, where $\mathbf{U}$ is an orthonormal matrix and $\mathbf{S} = \mathrm{Diag}(\lambda_1, \cdots, \lambda_n)$ is a diagonal matrix composed of the eigenvalues λ_i of $\mathbf{S}$.

Furthermore, if $\mathbf{S}$ is positive-definite, i.e., $\mathbf{S} \in SPD(n)$, then $\lambda_i \succ 0$ for $i = 1, \cdots, n$. The exponential map, exp: $S(n) \to SPD(n)$, is bijective, *i.e.*, one to one and onto. By eigen-decomposition, the exponential of a $\mathbf{S} \in S(n)$ can be computed as

$$\exp(S)=U.Diag(\exp(\lambda_1),\cdots,\exp(\lambda_n)).U^T \quad (1)$$

For any SPD matrix $\mathbf{S} \in SPD(n)$, it has a unique logarithm $\log(\mathbf{S})$ in $S(n)$:

$$\log(S)=U.Diag(\log(\lambda_1),\cdots,\log(\lambda_n)).U^T \quad (2)$$

Vector space structure on $SPD(n)$. The commutative Lie group $SPD(n)$ admits a bi-invariant Riemannian metric and the distance between two matrices $\mathbf{S}_1, \mathbf{S}_2$ is

$$d(\mathbf{S}_1, \mathbf{S}_2) = \left\| \log(\mathbf{S}_1) - \log(\mathbf{S}_2) \right\|_F \quad (3)$$

The desirable property of such a vector space structure of $SPD(n)$ is that, by matrix logarithm operation, the Riemannian manifold of SPD matrices is mapped to the Euclidean space. As such, in the logarithmic domain, the SPD matrices can be handled with simple Euclidean operations and, if necessary, the results can be mapped back to the Riemannian space via the matrix exponential.

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

We first present the form of the proposed ST covariance descriptors (Cov3D). Commonly used features for action and gesture recognition include intensity gradients and optical flow. Previous studies [11] have shown the benefit of combining both types of features. Therefore, we combine gradient and optical flow based features in building ST-LECM features. More specifically, for a given 3D volume R, we can extract the raw feature vector $\mathbf{f}(x, y, t)$ from pixel position (x, y, t) inside R, and $\mathbf{f}(x, y, t)$ has the following form:

$$\mathbf{f}(x, y, t) = [\mathbf{g}, \mathbf{o}]^T \quad (4)$$

$$\mathbf{g} = [|I_x|, |I_y|, |I_{xx}|, |I_{yy}|, \sqrt{I_x^2 + I_y^2}, \operatorname{artan} \frac{|I_y|}{|I_x|}] \quad (5)$$

$$\mathbf{o} = [u, v, w, \frac{\partial u}{\partial t}, \frac{\partial v}{\partial t}, \frac{\partial w}{\partial t}] \quad (6)$$

where the first four gradient based features in (5) represent the first and second order intensity gradients at pixel location (x, y, t). The last two gradient based features correspond to the gradient magnitude and gradient orientation. The optical-flow based features in (6) represent, in order: two horizontal component (u, v) and one vertical components (w) of the flow vector, and three first-order derivatives of the flow components $(\partial u/\partial t, \partial v/\partial t, \partial w/\partial t)$ with respect to t.

The ST covariance descriptor $Cov3D_R$ is computed as follows:

$$Cov3D_R = \frac{1}{S}(\mathbf{F} - \mu)(\mathbf{F} - \mu)^T \quad (7)$$

$$\mathbf{F} = (\mathbf{f}(x_1, y_1, t_1), \ldots, \mathbf{f}(x_S, y_S, t_S)), \quad \mu = \frac{1}{S}\sum_{i=1}^{S} \mathbf{f}(x_i, y_i, t_i) \quad (8)$$

where $\mathbf{F} \in R^{12\times S}$ is a column vector matrix, and the column vector $\mathbf{f}(x_i, y_i, t_i) \in R^{12\times 1}$ is the feature extracted from pixel position (x_i, y_i, t_i) inside the volume R, μ is the mean of all features. Each descriptor is hence a 12×12 matrix, as $\mathbf{f}(x, y, t)$ has 12 dimensions.

We wish to exploit these covariance matrices as fundamental features for vision applications. It is known that the Affine-Riemannian framework involves intensive computations of matrix square root, matrix inverse, matrix exponential and logarithm. Hence, we utilize the Log-Euclidean framework: $Cov3D_R$ in the commutative Lie group $SPD(n)$ is mapped by matrix logarithm to $\log Cov3D_R$ in the vector space of $S(n)$. $\log Cov3D_R$ can then be handled with the Euclidean operations and the intensive computations involved in Affine-Riemannian framework are avoided. It also facilitates greatly further analysis or modeling of the SPD matrices

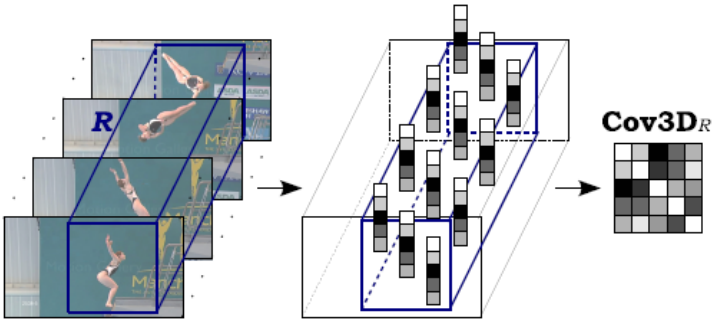

Figure 1. Conceptual demonstration for obtaining a Cov3D spatio-temporal covariance descriptor. A spatio-temporal volumn R is defined inside the input video. For each pixel in R a feature vector $\mathbf{f}(x_i, y_i, t_i)$ is calculated. The feature vectors are then used to compute the covariance matrix $Cov3D_R$.

From the tensor-valued image $Cov3D_R \in SPD(n)$, we compute the logarithm of the covariance matrix $Cov3D_R$ according to Eq. (2). $\log Cov3D_R$ is a symmetric matrix of Euclidean space. Because of its symmetry, we perform half-vectorization of $\log Cov3D_R$, denoted by $\mathrm{Vec}(\log Cov3D_R)$, i.e., we pack into a vector in the column order the upper triangular part of $\log Cov3D_R$. The final ST-LECM feature descriptor can thus be represented as

$$\mathrm{V}(\log Cov3D_R) = [\log Cov3D_R(1,1), \cdots, \log Cov3D_R(12,12)]^T \quad (9)$$

where $\mathrm{V}(\log Cov3D_R)$ is the proposed ST-LECM feature of volume R.

2.3. Encoding ST-LECM Features by LLC Method

In contrast to the previous coding schemes, LLC coding algorithm [12] has attracted much attention due to its impressive properties:

- Better reconstruction. In VQ (Figure 2.a), each descriptor is represented by a single basis in the codebook. Due to the large quantization errors the VQ code for similar descriptors might be very different. Besides, the VQ process ignores the relationships between different bases. Hence non-linear kernel projection is required to make up such information loss. On the other side, as shown in (Figure 2.c) in LLC, each descriptor is more accurately represented by multiple bases, and LLC code captures the correlations between similar descriptors by sharing bases.

- Local smooth sparsity. Similar to LLC, SC also achieves less reconstruction error by using multiple bases. Nevertheless, the regularization term of ℓ_1 norm in SC is not smooth. As (shown in Figure 2.b), due to the over-completeness of the codebook, the SC

process might select quite different bases for similar patches to favor sparsity, thus losing correlations between codes. On the other side, the explicit locality adaptor in LLC ensures that similar patches will have similar codes.

- Analytical solution. Solving SC usually requires computationally demanding optimization procedures. Unlike SC, the solution of LLC can be derived analytically such that LLC can be performed very fast in practice.

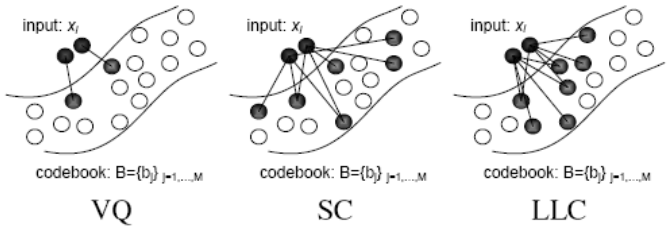

Figure 2. Comparison between VQ, SC and LLC. The Selected bases for Representation are Highlighted in Black

LLC can be formulated by

$$\mathbf{c}_i = \arg\min_{\mathbf{c}}(\|\mathbf{g}_i - \mathbf{B}\mathbf{c}\|_2^2 + \lambda\|\mathbf{d} \odot \mathbf{c}\|_2^2), \quad \text{s.t.}\, \mathbf{1}^T\mathbf{c} = 1, \tag{5}$$

$$\mathbf{d} = \exp(\frac{\text{dist}(\mathbf{g}_i, \mathbf{B})}{\sigma}), \quad \text{dist}(\mathbf{g}_i, \mathbf{B}) = [\text{dist}(\mathbf{g}_i, \mathbf{b}_1), \cdots, \text{dist}(\mathbf{g}_i, \mathbf{b}_M)]^T, \tag{6}$$

where the first term is reconstruction error; the second term is the locality constraint regularization on code $\mathbf{c}$, and λ is a regularization factor; in the second term, $\odot$ denotes the element-wise multiplication, and $\mathbf{d} \in R^M$ is the locality adaptor that gives different weight for each base vector proportional to its similarity to the input feature $\mathbf{f}_i$; and $\text{dist}(\mathbf{f}_i, \mathbf{b}_j)$ is the Euclidean distance between $\mathbf{f}_i$ and the j-th base $\mathbf{b}_j$. σ is used for adjusting the weight decay speed for the locality adaptor. $\mathbf{1}^T\mathbf{c} = 1$ is the shift invariant constraint according to [23].

LLC coding scheme bases on the hypothesis that descriptors approximately reside on a lower dimensional manifold in an ambient descriptor space; thus, it reduces the quantization error while preserving the consistent encoding ability.

Assuming that the motion information of video sequence V is represented as a set $\mathbf{G} = \{\mathbf{g}_i, i \in 1, \cdots, N_v\}$, $\mathbf{g}_i$ denotes the i-th ST-LECM feature. In the paper, to reduce quantization error and keep the consistent coding, LLC method and a codebook with M bases $\mathbf{B} = [\mathbf{b}_1, \mathbf{b}_2, \cdots, \mathbf{b}_M]$ are employed to encode the ST-LECM features $\mathbf{G}$, and obtain its reconstruction coefficients $\mathbf{C} = \{\mathbf{c}_i \in R^M, i \in 1, \cdots, N_v\}$.

3. The System Framework

In this section, we present the framework of our system. Basically, our system consists of five stages:

(1) Gradient and optical flow information $\{\mathbf{f}(x_i, y_i, t_i) = [\mathbf{g}, \mathbf{o}]^T\}$ on each pixel position in action video V is extracted.

(2) Each video is partitioned into several segments with a fixed-length along the temporal axis. Covariance matrices as local region descriptors are densely extracted over

the segments, and transformed to ST-LECM features $\mathbf{G}=\{\mathbf{g}_i, i\in 1,\cdots,N_v\}$ by matrix logarithm operation.

(3) Then, the ST-LECM features are encoded with LLC method, and obtain their LLC codes $\mathbf{C}=\{\mathbf{c}_i\in R^M, i\in 1,\cdots,N_v\}$.

(4) Each frame is divided into K sub-regions with a spatial pyramid with different scales, and the average-pooling-on-absolute-value function is implemented over the LLC codes in each sub-region. The average-pooling-on-absolute-value function is defined as follows:

$$\mathbf{h}_k=\frac{1}{N_k}\sum_{i=1}^{N_k}|\mathbf{c}_i| \tag{7}$$

(5) where $\{\mathbf{c}_i : i=1,\cdots,N_k\}$ denotes the LLC codes in the k -th sub-region; $\mathbf{h}_k$ denotes the polled feature of the k-th sub-region. Next, all pooled features $\{\mathbf{h}_k\}_{k=1}^K$ are concatenated to form a high-dimensional feature $\mathbf{H}_V$ to represent video V .

$$\mathbf{H}_V=[\mathbf{h}_1,\mathbf{h}_2,\cdots,\mathbf{h}_K] \tag{8}$$

(6) Finally, non-linear support vector machines (SVM) with χ^2 kernel $K_{\text{chi2}}(.,.)$ and intersection kernel $K_{\text{inter.}}(.,.)$ are used as action classifiers.

$$K_{\text{chi2}}(\mathbf{H}_i,\mathbf{H}_j)=\sum_{s=1}^{S}\frac{2(\mathbf{H}_i(s).\mathbf{H}_j(s))}{(\mathbf{H}_i(s)+\mathbf{H}_j(s))} \tag{9}$$

$$K_{\text{inter.}}(\mathbf{H}_i,\mathbf{H}_j)=\sum_{s=1}^{S}\min\{\mathbf{H}_i(s),\mathbf{H}_j(s)\} \tag{10}$$

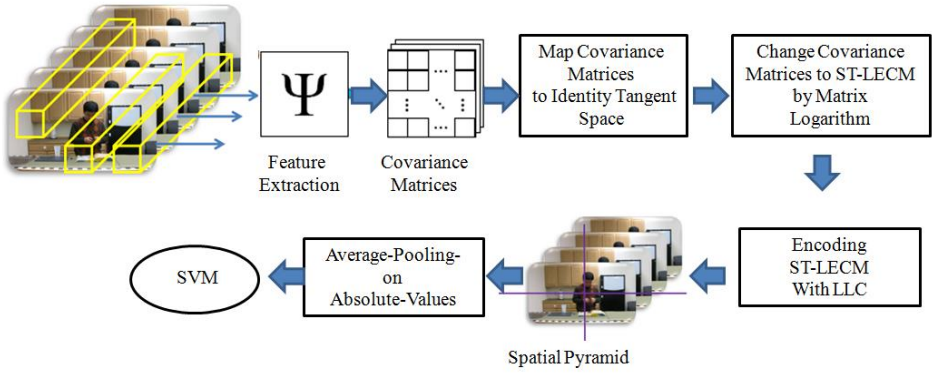

Figure 3. The Flowchart of Action Recognition System based on ST-LECM Features

4. Experiments

In this section, two public video datasets, the KTH [12] and Activity of Daily Living (ADL) [13] datasets, are used to evaluate the performance of our recognition system 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. Covariance matrices fuse two types of information: gradient vector and optical flow. The former describes the local appearance information in video, and the latter depicts the motion information in local ST region.

Figure 4. Example images from the datasets used in our experiments: (a) KTH (b) Activity of Daily Living

4.1. KTH Action Datasets

The KTH dataset contains six human action classes: boxing, hand-clapping, hand-waving, jogging, running, and walking, performed by 25 subjects in 4 scenarios: outdoors, outdoors with scale variation, outdoors with varying clothes, and indoors, see Figure 4(a) for examples. The videos are recorded with static and homogeneous background. However, the camera is not static, *i.e.*, vibration and unknown zooming exist.

On KTH dataset, Laptev *et. al.*, [14] proposed a system where ST interest points are detected and described using HOG/HOF descriptors. In order to classify a query video, BoF model is utilized in a multi-channel SVM classifier with χ^2 kernel. Gilbert *et. al.*, [13] proposed to use an over-complete set of simple 2D corners in ST area. The extracted points are first grouped spatially and temporally using a hierarchical process. The most distinctive and descriptive features are learned. And Wang *et. al.*, [15] tracked densely sampled points by a median filter kernel and extract aligned shape, appearance, and motion features. BoF model is utilized in a 30-channel (5 types of features and 6 channels) SVM classifier with χ^2 kernel for classification.

In our experiment on the KTH dataset, we locate the head position of subjects, and move the subject to the center of frames by trimming the frame width and preserving the frame height. Then, the trimmed frames are resized to the 120-pixel height while keep the ratio of height/width. Each resulting video is divided into several segments with 11-frame length. A set of overlapping ST blocks are extracted from the video segments over a spatial grid with spacing of 6 pixels, and the size of ST block is set as $16\times16\times11$ ($x\times y\times t$). Next, the ST-LECM features of 24 videos belonging to one subject are clustered by k-means clustering method, and obtain a codebook containing 250 codewords. In the stage of feature coding, the number of selected bases in LLC is set as 5. Next, frames are partitioned into sub-regions by a spatial pyramid with 1-by-1, 2-by-2, 1-by-4 and 4-by-1. All pooled features of a video are concatenated to form a high-dimensional feature with $(1+4+4+4)\times250=3250$ dimensions.

Leave-one-out cross-validation (LOOCV) strategy is used to evaluate the system performance. In each LOO run, we use the videos of 24 subjects for training, and the videos of the remaining subject for test, and the recognition rate is the average value of the 25 runs.

In Table 1, we compare our proposed system with the aforementioned systems on the KTH dataset. Our system is superior to the method proposed by Laptev *et. al.*, and Gilbert *et. al.*, The difference of the performance achieved by SVM on χ^2 and intersection is small, but the time-consumption on the intersection kernel is rather less than the χ^2 kernel, because at each vector entry, comparison operation is implemented for one time for the intersection kernel, and three operations (one addition, one multiplication and one division) are carried out for χ^2 kernel.

Table 1. (a) Comparison between the proposed method with previous methods on the KTH dataset. (b) The confusion table for the SVM classifier on χ^2 kernel. (c) The confusion table for the SVM classifier on Intersection kernel. S1 (boxing), s2 (hand-clapping), s3 (hand-waving), s4 (walking), s5 (jogging), s6 (running)

Method	Recognition rate(%)
Laptev et al. [14]	91.8
Gibert et al. [13]	94.5
Wang et al. [12]	95.3
Our ST-LECM (Chi-squared kernel)	97.1
Our ST-LECM (Intersection kernel)	96.9

(a)

	s1	s2	s3	s4	s5	s6
s1	98.1	0.90	1.00			
s2	0.80	98.4	0.80			
s3	0.50	0.50	99.0			
s4				96.4	1.40	2.20
s5				3.80	94.6	1.60
s6				4.00	1.60	95.4

(b)

	s1	s2	s3	s4	S5	s6
s1	98.1	0.80	1.10			
s2	1.00	98.4	0.60			
s3	0.35	0.65	99.0			
s4				95.2	1.60	3.20
s5				3.00	95.4	1.60
s6				3.40	2.20	94.4

(c)

4.2. ADL Datasets

The ADL dataset consists of 150 videos of 5 subjects performing a series of daily tasks in a kitchen environment, acquired using a stationary camera. Sample frames are shown in Figure 4(b).

We compare the proposed ST-LECM method against 3 state-of-the-art human action classification systems: Laptev *et. al.*, [14], Matikainen *et. al.*, [17], Messing *et. al.*, [18]. In [17], a method for augmenting quantized local features with relative ST relationships between pairs of features is proposed. Their discriminative classifier is trained by estimating all the cross probabilities for various local features of an action. In [18], Messing *et. al.*, tracked Harris3D interest points with a KLT tracker [16] and extract velocity history information along the trajectories. Appearance and location features are utilized in a mixture model to improve the recognition performance.

In our experiment on the ADL dataset, since the movement of subjects during performing action is small, it is not necessary to local the subjects. All frames are resized to the 180-pixel height while keep the ratio of height/width. And each resulting video is divided into segments with 25-frame length. A set of overlapping ST blocks are extracted from the video segments over a spatial grid with spacing of 4 pixels, and the size of ST block is set as $11\times11\times25\,(x\times y\times t)$. Then, the ST-LECM features of 20 videos (2 action videos selected from each class) are clustered by k-means clustering method, and obtain the codebook with 250 codewords. In the stage of feature coding, the number of selected bases is set as 5. Next, frames are partitioned into sub-regions by a spatial pyramid with scales 1-by-1, 2-by-2, 1-by-6 and 6-by-1.

Leave-one-out cross-validation (LOOCV) strategy is used to evaluate the system performance. In each LOO run, we use the videos of 4 subjects for training, and the videos of the remaining subject for test, and the recognition rate is the average value of the 5 runs.

As recommended by [18], we evaluate our results on this dataset using 5-fold LOOCV. In each fold, videos from four subjects are considered for training and the fifth for testing.

Table 2 shows that the proposed ST-LECM method outperforms the state-of-the-art methods. The difference between ADL and KTH datasets is that the actions in the ADL generate much less amount of motion information than the actions in KTH. For example, the actions 'dialPhone','answerPhone' in the ADL are implemented in a small area, the range of them is small; in the KTH dataset, actions 'hand-waving','walking' generates great amount of motion information. As a result, classifying the ADL dataset is more difficult than classifying the KTH dataset. Another reason why the ADL dataset is more challenging is that the actions in the ADL is not periodic action, and the time duration is much longer than the actions in KTH. As we known, classifying periodic actions is much easier than classifying non-periodic ones, due to periodic action produces rich amount of motion information for building the good action model.

Table 2 (a) Comparison between the proposed method with previous methods on the ADL dataset. (b) The confusion table for the SVM classifier on χ^2 kernel. (c) The confusion table for the SVM classifier on Intersection kernel. S1 (answer Phone), s2 (chop Banana), s3 (dial Phone), s4 (drink Water), s5 (eat Banana), s6 (eat Snack), s7 (lookup In phonebook), s8 (peel Banana), s9 (use Silverware), s10 (write On whiteboard)

Method	Recognition rate(%)
Laptev et al. [14]	80.2
Gibert et al. [13]	89.3
Wang et al. [12]	70.7
Our ST-LECM (Chi-squared kernel)	90.7
Our ST-LECM (Intersection kernel)	90.3

(a)

	s1	s2	s3	s4	s5	s6	s7	s8	s9	s10
s1	**83.4**	16.6								
s2		**100**								
s3			**100**							
s4				**83.4**					16.6	
s5					**94.7**	5.3				
s6						**88.7**		5.30	6.00	
s7							**100**			
s8			15.3					**74.7**	10.0	
s9			11.3						**88.7**	
s10		5.20								**94.8**

(b)

	s1	s2	s3	s4	s5	s6	s7	s8	s9	s10
s1	**82.6**	17.4								
s2		**100**								
s3			**100**							
s4				**82.7**					17.6	
s5					**94.4**	5.7				
s6						**88.4**		5.40	6.20	
s7							**100**			
s8			15.7					**74.3**	10.0	
s9			11.7						**88.3**	
s10		5.60								**94.4**

(c)

5. Cc

In the paper, we combine covariance matrices, dense sampling over video and the spatial pyramid method to solve the problem of action recognition. We extend the popular BoF model to a special class of non-Euclidean spaces, the space of Symmetric Positive Definite (SPD) matrices formed by covariance descriptors of ST features. In doing so, we elaborate on how ST-LECM features can be obtained for covariance matrices and devise Log-Euclidean BOF, an extrinsic extension of conventional BoF using Riemannian geometry of SPD matrices. Benefiting from the good property of the proposed ST-LECM features and dense sampling method, our system outperforms the classical system published recently.

Acknowledgments

This research is supported by National Nature Science Foundation of China (Grant no. 61271288), and the National High Technology Research and Development Program (Grant no. 2012AA011503).

References

[1] O. Tuzel, F. Porikli and P. Meer, "Region covariance: A fast descriptor for detection and classification", Proceedings of European Conference on Computer Vision (ECCV), **(2006)**, pp. 589-600.
[2] F. Porikli, O. Tuzel and P. Meer, "Covariance tracking using model update based on Lie algebra," Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), **(2006)**, pp. 728-735.
[3] Y. Pang, Y. Yuan and X. Li, "Gabor-based region covariance matrices for face recognition", IEEE Transactions on Circuits and Systems for Video Technology, vol.18, **(2008)**, pp. 989-993.
[4] X. Pennec, "Intrinsic statistics on Riemannian manifolds: Basic tools for geometric measurements", Journal of Mathematical Imaging and Vision, vol. 25, **(2006)**, pp. 127-154.
[5] A. Sanin, C. Sanderson, M. T. Harandi and B. C. Lovell, "Spatio-temporal covariance descriptors for action and gesture recognition", Proceedings of IEEE Workshop on the Applications of Computer Vision, **(2013)**, pp. 103-110.
[6] S. Arsigny, V. Fillard, X. Pennec and N. Ayache, "Geometric means in a novel vector space structure on symmetric positive-definite matrices", SIAM J. Matrix Anal. Appl., **(2006)**.
[7] R. Messing, C. Pal and H. Kautz, "Activity recognition using the velocity histories of tracked keypoints", In Proc. Int. Conference on Computer Vision, **(2009)**, pp. 234-256.
[8] J. C. Niebles, C. W. Chen and L. Fei-Fei, "Modeling temporal structure of decomposable motion segments for activity classification", In Proc. European Conference on Computer Vision, **(2010)**, Springer, pp. 392-405.
[9] C. Schuldt, I. Laptev and B. Caputo, "Recognizing human actions: a local SVM approach", In Proc. Int. Conference on Pattern Recognition, vol. 3, **(2004)**, pp. 32-36.
[10] X. Pennec, P. Fillard and N. Ayache, "A Riemannian framework for tensor computing", International Journal of Computer Vision, **(2006)**, pp. 41-66.
[11] V. Arsigny, P. Fillard, X. Pennec and N. Ayache, "Geometric means in a novel vector space structure on symmetric positive-definite matrices", SIAM Journal on Matrix Analysis and Applications, vol. 29, **(2007)**, pp. 328-347.
[12] W. Jinjun, Y. Jianchao, Y. Kai, L. Fengjun and T. Huang, "Locality-constrained Linear Coding for image classification", CVPR, **(2010)**, pp. 3360-3367.
[13] A. Gilbert, J. Illingworth and R. Bowden, "Action recognition using mined hierarchical compound features", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.33, **(2011)**, pp. 883-897.
[14] I. Laptev, M. Marszalek, C. Schmid and B. Rozenfeld, "Learning realistic human actions from movies", In Proc. IEEE Conference on Computer Vision and Pattern Recognition, **(2008)**, pp. 1-8.
[15] H.Wang, A. Klaser, C. Schmid and C.-L. Liu, "Dense trajectories and motion boundary descriptors for action Recognition", International Journal of Computer Vision, vol. 103, **(2013)**, pp. 60-79.
[16] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision", In Proc. Int. Joint Conference on Artificial Intelligence, **(1981)**, pp. 674-679.
[17] P. Matikainen, M. Hebert and R. Sukthankar, "Representing pairwise spatial and temporal relations for action Recognition", In Proc. European Conference on Computer Vision, Springer. **(2010)**, pp. 508-521.
[18] R. Messing, C. Pal, and H. Kautz, "Activity recognition using the velocity histories of tracked keypoints", In Proc. Int. Conference on Computer Vision, **(2009)**, pp. 104-111.

Authors

Shilei Cheng, he is a Ph.D student in the School of Electronic Engineering, University of Electronic Science and Technology of China. He received his Master degree from Chengdu University of Technology in 2014.His current research interests include computer vision, pattern recognition, motion detection.

Jiangfeng Yang, he is a Ph.D student in the School of communication and information engineering, University of Electronic Science and Technology of China. He received his Master degree from Kunming University of Science and Technology in 2009. His current research interests include computer vision, human action recognition, motion detection.

Zheng Ma, he is a professor in School of Communication and Information Engineering, University of Electronic Science and Technology of China. His current research interests include image processing, computer vision.

Mei Xie, she received her B.S. in 1981 from Chengdu Institute of Telecommunication, and her M.S. in 1990 and Ph.D. in 1996 from University of Electronic Science and Technology of China, China. From 1997-1998 in School of Electronic Engineering, University of Hong Kong, Hong Kong, and 1998-1999 in School of Electronic Engineering, University of Texas at Austin, USA, she studied as a postdoctor. She is currently a professor in School of Electronic Engineering, University of electronic science and technology of China, China. Her research interests include signal processing, machine vision and Internet security.