J Electr Eng Technol.2015; 10(4): 1851-1863
http://dx.doi.org/10.5370/JEET.2015.10.4.1851

ISSN(Print) 1975-0102
ISSN(Online) 2093-7423

Multiscale Spatial Position Coding under Locality Constraint for Action Recognition

Jiang-feng Yang†, Zheng Ma* and Mei Xie**

Abstract – In the paper, to handle the problem of traditional bag-of-features model ignoring the spatial relationship of local features in human action recognition, we proposed a Multiscale Spatial Position Coding under Locality Constraint method. Specifically, to describe this spatial relationship, we proposed a mixed feature combining motion feature and multi-spatial-scale configuration. To utilize temporal information between features, sub spatial-temporal-volumes are built. Next, the pooled features of sub-STVs are obtained via max-pooling method. In classification stage, the Locality-Constrained Group Sparse Representation is adopted to utilize the intrinsic group information of the sub-STV features. The experimental results on the KTH, Weizmann, and UCF sports datasets show that our action recognition system outperforms the classical local ST feature-based recognition systems published recently.

Keywords: Action recognition, Action representation, Multiscale spatial position coding under locality constraint

1. Introduction

Human action recognition has attracted significant interest in the computer vision community in the recent decade and has spurred the development of a wide variety of applications, including video surveillance human-computer interaction and the analysis of sporting events. However, automatic human action recognition is highly challenging due to the non-stationary background of most video content, the ambiguity of the human body shape among different actions, and the existence of intra-class variations in the appearance, physical characteristics, and motion style of different human subjects.

Generally speaking, action recognition is composed of two components: action representation and action classify-cation. Action representation is the process of modeling human behaves, encoding the extracted features from action video. The quality of the learnt model decides discriminative power usually results in good classification result. There are two representation methods: holistic action representations [1-7] and local action representations [8-18].

Commonly, holistic action representation is derived from silhouette or body sketch. It requires fine background subtraction or body part tracking. Thus, it is sensitive to noise, variation in viewpoint, and partial occlusion. On the country, local action representation bases on the local spatio-temporal features, and it usually is utilized together with bag-of-features (BoF) to model human action. Local action representation is robust to viewpoint changes, environment noise and partial occlusion [19].

In many literatures, BoF-based human action recognition consists of extracting local features from videos, obtaining action representation vectors via these local features, and classifying action videos with a classifier upon the vectors. To obtain action representation vector, several feature coding and pooling methods are provided. *k*-means and vector quantization (VQ) are used to encode features, next, action representation histogram is computed.

Recently, data locality was observed to be a key role in clustering, dimension reduction [20, 21], density estimation [22], anomaly detection [23], and image classification [24-27]. In pattern recognition, the k-NN (Nearest Neighbor) classifier can be considered as a recognition algorithm using data locality, since it considers the locality information of training data for performing classification. To be more precise, k-NN assigns the class label for a test input according to the majority of the nearest training data of the same class. Motivated by the importance of data locality, we proposed a Multiscale Spatial Position Coding under Locality Constraint (MSPC-LC), which considers the spatial position relationship of local features, to encode local features. Compared to the standard sparse representation based classification (SRC) proposed by Wright *et al.* [28] or recently proposed locality-constrained linear coding (LLC), our algorithm improves classification performance.

To reduce the quantization error caused by *k*-means and VQ, rather than assigning one codeword for a feature only, soft vector quantization (SVQ) and sparse coding (SC) [29]

† Corresponding Author: School of Communication and Engineering Information, University of Electronic Science and Technology of China (UESTC), P.R. China. (369322023@qq.com)
* School of Communication and Engineering Information, UESTC, P.R. China. (wallsonyang@163.com)
** School of Electronic Engineering, UESTC, P.R. China. (369322023@qq.com)
Received: June 26, 2014; Accepted: March 17, 2015

are adopted to encode features in action recognition tasks. However, the local features usually reside on nonlinear manifolds [30, 31]. Neither SVQ nor SC can preserve the nonlinear manifold structure. The manifold is nonlinear and not Euclidean in its whole space, but linear and Euclidean in a local region [32]. Because SVQ uses all bases to encode each feature and generates dense codes, it can not precisely represent the nonlinear manifold structure with a global way. Due to the overcomplete dictionary, SC tends to choose the code words which are distant to the input features [31]. Thus, it cannot correctly represent manifold structure. Hereafter, we consider these limitations on both quantization error and loss manifold structure in feature coding as representation error. For this issue, Yu *et al.* [30] provided a Local Coordinate Coding (LCC) to encode feature with locality-constrained; Wang *et al.* [31] introduced an improved version of LCC named Locality-constrained Linear Coding (LLC) to reduce computational cost, and Wei *et al.*[33] proposed a local sensitive dictionary learning method for image classification.

In classification stage, support vector machine (SVM) has been widely used as action classifier. Recently, inspired by the impressive success of Sparse Representation based Classification (SRC) in face recognition [34], some authors [18] explored SRC for human action recognition and achieved better performance than SVM. Nevertheless, these representation methods suffer one major limitation: the spatio-temporal (ST) relationship among local features is ignored, such as temporal order and spatial arrangement [35]. For instance, in Fig. 1, same local features from two different actions are projected on the XY plane, but their spatial configurations are different. Due to the same histograms generated by BoF model, they are incorrectly considered as one action. Recently, some researchers exploited ST context information [35], local feature distribution [36], and spatial pyramid matching (SPM) [23, 27] to handle this problem.

In this paper, we developed MSPC-LC algorithm to address this limitation and reduce the representation error.

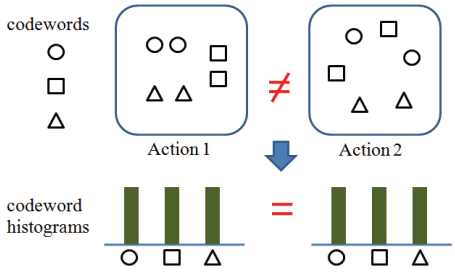

Fig. 1. The illustration of traditional BoF model ignoring the spatial relationship between local features. Same features from two different actions are projected on XY plane, and have different spatial configuration. However, they can not be correctly classified with BoF model, because it contains two same codeword histograms due to ignoring their spatial relationship.

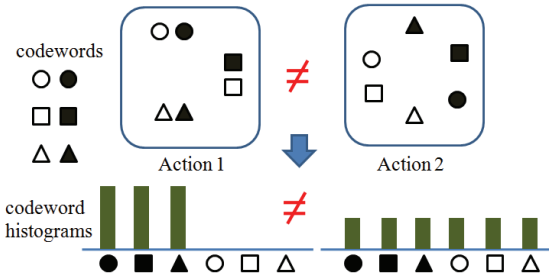

Fig. 2. The illustration of our proposed approach. The shape of codewords indicates their appearance information, while the color indicates their spatial position. For example, the codewords ■, □ have same appearance but different spatial position. They are considered as two different codewords during feature coding. Then, two different codeword histograms are obtained for two actions upon these codewords. Hence, Actions 1 and 2 which can not be distinguished in BoF (Fig. 1) can be correctly classified with our method.

Specifically, to reduce the representation error (quantization error and manifold structure loss), we adopted locality constraint on codebook learning and feature coding with respect to manifold learning [32]. To model the ST relationship of features, spatial position information between them is taken into account in learning codebook and coding features. In such way, the spatial relationship is embedded into the coded features. In addition, to handle the different action styles, multi-temporal-scale pooled features are built. In experiment, local features are firstly projected on the XY plane for obtaining their spatial configuration. Then codebook learning and feature coding are implemented under locality position constraint. To classify one action video (see Fig. 2), which can be treated as a spatiotemporal volumn (STV), this STV is then represented as a group of multi-temporal-scale sub-STVs obtained by dense sampling. Next, multi-temporal-scale pooled features are built with max-pooling method. Finally, Locality-constraint Group Sparse Representation (LGSR) [26] is adopted for action classification upon these sub-STV pooled features.

Compared to these methods which use ST context information [35] or feature distribution [36] to handle the limitation of BoF, MSPC-LC is a more fine and complete method, because it records most fundamental information (*where, how*) of local features for human action recognition. The experimental results on the KTH, Weizmann, and UCF sports datasets show that our method achieves better performance than these methods [35-38] and other local ST feature-based methods.

The paper has two contributions as follows:

- To describe spatial relationship between local features and reduce representation error, a mixed feature combining spatial configuration and motion information was proposed, and was encoded by MSPC-LC

algorithm.

- In order to utilize the intrinsic group information from these sub-STV pooled features, LGSR based classifier is employed for action classification.

The rest of this paper is organized as follows: Algorithm MSPC-LC is proposed in Section 2. The action recognition framework in MSPC-LC and LGSR is provided in Section 3. Then, experimental results and analysis are shown in Section 4. Finally, the conclusions are drawn in Section 5.

2. Multiscale Spatial Position Coding under Locality Constraint

2.1 Modeling position relationship between local features

Two facts motivate us to propose MSPC-LC for action representation. The first fact is that the relative position relationship between features in BoF model is ignored, and the spatial cue often plays an important role for visual appearance modeling. Moreover, inspired by the work of Liu *et al.* [40], spatial position relationship between features was considered during dictionary learning and feature coding in this paper to handle this problem. They involved spatial locations into DCT based feature descriptors to model the spatial relationship of local features for face recognition. Their experimental results showed that feature position can improve local feature-based face recognition accuracy. The second fact motivate us is that local features from the same part of body are often repeatedly and similar, and usually aggregate in a local region. For instance, local features generated by hand motion usually locate in the upper part of human body at high probability; those from leg motion are most inclined to fall into the lower part of human body. Therefore, the quantization error can be alleviated by incorporating feature position and appearance information.

In this paper, local spatio-temporal oriented energy (STOE) features $\{\mathbf{v}_g\}_{g=1}^{G}$ and their spatial position $\{rel\_x, rel\_y\}$ are connected together to build a mixed feature descriptor $\mathbf{f}^{\alpha,\beta}$

$$\mathbf{f}^{\alpha,\beta} = [\mathbf{v}_1, \cdots, \mathbf{v}_g, \cdots \mathbf{v}_G, \alpha(rel\_x), \beta(rel\_y)]^T \quad (1)$$

Fig. 3. The local feature position distribution map of actions in the KTH dataset.

where row vectors $\mathbf{v}_g$ denote ℓ_2 STOE features along direction g ; and α, β are the spatial scale parameter and temporal scale parameter, respectively; (rel_x, rel_y) is the normalized coordinate of feature $\mathbf{f}^{\alpha,\beta}$, and obtained by (see Fig. 3)

$$rel\_x = \frac{(abs\_x) - (body\_cen\_x)}{body\_length} \quad (2)$$

$$rel\_y = \frac{(abs\_y) - (body\_cen\_y)}{body\_length} \quad (3)$$

$$body\_cen\_y = (head\_y) - \frac{(body\_length)}{2} \quad (4)$$

$$body\_cen\_x = head\_x \quad (5)$$

where (abs_x, abs_y) is the coordinate of $\mathbf{f}^{\alpha,\beta}$ on the XY plane; $body\_length$ is body height of action performer. $(body\_cen\_x, body\_cen\_y)$ is the position of body center; $(head\_x, head\_y)$ is the position of performer's head in XY plane.

Going through the above process, we can see that all local features are relocated around the origin. The purpose of introducing head position and body length is make sure that the features from different parts of body are correctly relocated with respect to head position, for example, the features from hand motion should be located in the upper of body, and those from leg motion tend to appear in the lower of body. In the paper, we manually locate head position and body length. Fortunately, there are just few videos (few of the videos in UCF Sports dataset) need to do this.

It can be seen from (1) that the distance between two features $\mathbf{f}_i^{\alpha,\beta}, \mathbf{f}_j^{\alpha,\beta}$ consists of two parts: the difference of two STOE feature descriptors $(\mathbf{v}_i^1 - \mathbf{v}_j^1, \cdots, \mathbf{v}_i^G - \mathbf{v}_j^G)$ and the difference of their spatial positions $(\alpha(rel\_x_i - rel\_x_j), \beta(rel\_y_i - rel\_y_j))$. The former is fixed value, but the latter changes by setting various α, β. With a decrease in α, β , the influence of feature position on clustering result drops. As a result, *k*-means on features with Euclidean distance as metric gradually becomes traditional BoF clustering that ignores ST position relationship between features. Fig. 4 illustrates the different *k*-means results by various α, β values.

To depict the spatial position relationship between features, codebook building and feature coding upon the feature $\mathbf{f}$ are implemented. The computation of STOE features and the construction of corresponding STOE descriptor are given in Section 2.4.

To explain the role of feature position during codebook building and feature coding, we realized clustering algorithm *k*-means for dictionary learning and encoded local features by VQ. The representation error caused by them will be solved in Section 2.2. Assuming that $\mathbf{D}^{\alpha,\beta} \in R^{N \times M}$, $\mathbf{D}^{\alpha,\beta} = \{\mathbf{d}_1^{\alpha,\beta}, \cdots, \mathbf{d}_M^{\alpha,\beta}\}$ is a codebook built via *k*-means upon the features $\mathbf{F}^{\alpha,\beta} = \{\mathbf{f}_1^{\alpha,\beta}, \cdots, \mathbf{f}_n^{\alpha,\beta}\}$, $\mathbf{f}_j^{\alpha,\beta} \in R^N$,

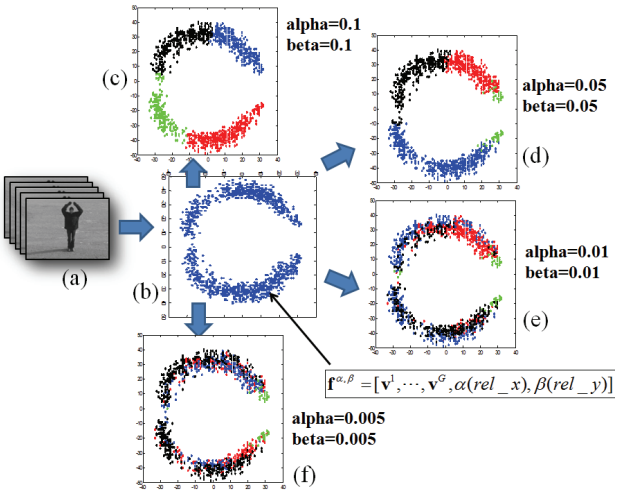

Fig. 4. The local feature position distribution map of actions in the KTH dataset. Spatial scale α,β influences on the *k*-means clustering results; (a) Input action data; (b) Local STOE features extracted from the video. (c-f) Local features containing both STOE and spatial configuration are clustered into four types (denoted by four different colors) using *k*-mean with Euclidean distance as metric. It can be seen that with a decrease in α,β values, the impact from feature position over the clustering results gradually drops. In (c), the large values of α,β make feature position dominates the clustering result, while the information of STOE hardly influences on the clustering result. In (f), the information of STOE significantly controls the clustering result, since the small values of α,β depress the influence of feature position.

$(j=1,\cdots,n)$ with Euclidean distance as metric. Each codeword $\mathbf{d}_i^{\alpha,\beta}$ contains two types information: ST oriented motion information $(\{\mathbf{v}^g\}_{g=1}^G)$ and feature spatial position $(\alpha(rel\_x),\beta(rel\_y))$. The class label vector $\mathbf{c}_j^{\alpha,\beta}=\{\mathbf{c}_{j,1}^{\alpha,\beta},\cdots,\mathbf{c}_{j,M}^{\alpha,\beta}\}$ for feature $\mathbf{f}_j^{\alpha,\beta}$ is obtained with VQ:

$$\mathbf{c}_{j,i}^{\alpha,\beta}=\begin{cases}1, & \arg\min_{i\in\{1,\cdots,M\}}\left\|\mathbf{f}_j^{\alpha,\beta}-\mathbf{d}_i^{\alpha,\beta}\right\|_2\\ 0, & \text{otherwise}\end{cases} \tag{6}$$

Where $\mathbf{f}_j^{\alpha,\beta}$ is the input feature and described in (1); $\mathbf{d}_i^{\alpha,\beta}$ is the *i*th codeword in codebook $\mathbf{D}^{\alpha,\beta}$.

According to (1) and (6), the codeword $\mathbf{d}_i^{\alpha,\beta}$ chosen for encoding $\mathbf{f}_j^{\alpha,\beta}$ should be the most closest to $\mathbf{f}_j^{\alpha,\beta}$ with respect to two aspects: feature similarity and spatial distance. Therefore, the spatial position of $\mathbf{f}_j^{\alpha,\beta}$ can be obtained from its code $\mathbf{c}_j^{\alpha,\beta}$. Given a group of local features, their spatial relationship can be represented with their codeword histogram:

$$\mathbf{H}^{\alpha,\beta}=\frac{1}{n}\sum_{j=1}^{n}\mathbf{c}_j^{\alpha,\beta} \tag{7}$$

where $\mathbf{H}^{\alpha,\beta}\in R^M$ is the codeword histogram with scale factors α,β; n is the number of features; and $\mathbf{c}_j^{\alpha,\beta}$ is the code of $\mathbf{f}_j^{\alpha,\beta}$.

For example, as illustrated in Fig. 2, these two actions in Fig. 1 can be distinguished with their new histograms. Benefiting from involving feature position into codewords, two different codeword histograms are provided for Actions 1 and 2. Actions that have similar features but different spatial relationship can be correctly classified by this method. Therefore, fusing spatial position into codebook building and feature coding is a feasible way to model the spatial relationship of features for human action recognition.

2.2 Reducing representation error with locality constraint

In Section 2.1, *k*-means and VQ are applied in dictionary learning and feature coding. However, Yu *et al.* [28] discovered that VQ can not handle nonlinear manifold structure well, because it is a *0*th-order (constant) approximation of object functions from the view of function approximation, and that VQ causes nontrivial quantization error. They suggested that *1*st-order (linear) approximation can solve these problems and introduced locality constraint into object function:

$$\begin{aligned}&\mathbf{c}_j=\arg\min_{\mathbf{c}_j}\left(\left\|\mathbf{f}_j-\mathbf{D}.\mathbf{c}_j\right\|_2^2+\lambda\left\|\mathbf{p}_j\odot\mathbf{c}_j\right\|_1\right)\\&\text{subject to }\mathbf{1}^T.\mathbf{c}_j=1,(\forall j=1,\cdots,n)\end{aligned} \tag{8}$$

where $\mathbf{p}_j=[p_{j,1},\cdots,p_{j,M}]$ is the locality adaptor that gives different freedom for each base. The *k*th element in $\mathbf{p}_j$ is obtained by $p_{j,k}=\left\|\mathbf{f}_j-\mathbf{d}_k\right\|_2$ and $p_{j,k}$ is the distance between local feature $\mathbf{f}_j$ and *k*th atom $\mathbf{d}_k$; sparse coefficient $\mathbf{c}_j=[c_{j,1},\cdots,c_{j,M}]$. In (8), the first term represents the reconstruction error of an input feature $\mathbf{f}_j$ with respect to codebook $\mathbf{D}$; the second term is locality constraint regularization on $\mathbf{c}_j$; and λ is a regularization factor to balance these terms; $\odot$ is the element-wise multiplication; $\mathbf{1}^T.\mathbf{c}_j=1$ is the shift-invariant constraint according to [28].

Eq. (8) tends to choose the atoms near $\mathbf{f}_j$ for generating coefficient $\mathbf{c}_j$. Because distance vector $\mathbf{p}_j$ is fixed, to minimize $(\left\|\mathbf{f}_j-\mathbf{D}.\mathbf{c}_j\right\|_2^2+\lambda\left\|\mathbf{p}_j\odot\mathbf{c}_j\right\|_1)$, one needs to make the coefficient $c_{j,k}$ corresponding to large $p_{j,k}$ equals 0. In addition, $\mathbf{1}^T.\mathbf{c}_j=1$ in (8) is sparse regularization term and intends to obtain sparse solution. Sparsity indicates that many elements in $\mathbf{c}_j$ are zero, while only a few are nonzero. Thus only a few codewords in $\mathbf{D}$ are selected to encode feature $\mathbf{f}_j$. Obviously, the selected codewords belong to the local neighborhood of $\mathbf{f}_j$.

However, an iterative optimization is applied to solve the ℓ_1 optimization problem in (8). To reduce the computational cost in (8), we replace the term $\|\mathbf{p}_j \odot \mathbf{c}_j\|_1$ with $\|\mathbf{p}_j \odot \mathbf{c}_j\|_2$. Hence, (8) can be rewritten as follows:

$$\begin{aligned} &\mathbf{c}_j = \arg\min_{\mathbf{c}_j}(\|\mathbf{f}_j - \mathbf{D}.\mathbf{c}_j\|_2^2 + \lambda\|\mathbf{p}_j \odot \mathbf{c}_j\|_2^2) \\ &\text{subject to } \mathbf{1}^T.\mathbf{c}_j = 1, (\forall j = 1, \cdots, n) \end{aligned} \tag{9}$$

In (9), feature $\mathbf{f}_j$ is fixed. To minimize $\|\mathbf{p}_j \odot \mathbf{c}_j\|_2^2$, the codewords far from $\mathbf{f}_j$ are assigned zero in $\mathbf{c}_j$, meanwhile, the codewords close to $\mathbf{f}_j$ are assigned nonzero in $\mathbf{c}_j$. Therefore, similar to (8), the codewords belonging to the neighborhood of $\mathbf{f}_j$ are selected to generate coefficient $\mathbf{c}_j$. From the respect of manifold learning [16, 18], although the data on a manifold are nonlinear, in a local region, they can be considered as linear [16]. Therefore, using the locality constraint, the problems of VQ can be solved.

To achieve good classification performance, the coding scheme should generate similar codes for similar descriptors. Following this requirement, the locality constraint term $\|\mathbf{p}_j \odot \mathbf{c}_j\|_2^2$ in (9) presents several attractive properties:

- Better reconstruction. In VQ, each descriptor is represented by a single basis in the codebook, as illustrated in Fig. 5. 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 Fig. 5, in LLC, each descriptor is more accurately represented by multiple bases, and LLC code captures the correlations between similar descriptors by sharing bases
- 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 Fig. 5, 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. For instance, the Feature Sign algorithm utilized by Yang et al. [43] has a computation complexity of $O(M \times K)$ in the optimal case [44], where K denotes the number of non-zero elements; M denotes the number of bases in dictionary. Unlike SC, the object function in (5) can be solved with an analytical solution according to [33]:

$$\begin{aligned} &\rho = (\psi + \lambda.\mathrm{diag}(\mathbf{p}_j)^2)^{-1}.\mathbf{1} \\ &\psi = (\mathbf{f}_j.\mathbf{1}^T - \mathbf{D})^T.(\mathbf{f}_j.\mathbf{1}^T - \mathbf{D}) \\ &\mathbf{c}_j = \rho / (\mathbf{1}^T \rho) \end{aligned} \tag{10}$$

Similarly, the problems of k-means codebook building can also be solved with locality constraint. According to [38], the object function of our codebook building method is formulated as follows:

$$\begin{aligned} &\arg\min_{\mathbf{D},\mathbf{C}} \|\mathbf{F} - \mathbf{D}.\mathbf{C}\|_F^2 + \lambda\sum_{j=1}^{n}\|\mathbf{p}_j \odot \mathbf{c}_j\|_2^2 \\ &\text{subject to } \mathbf{1}^T.\mathbf{c}_j = 1, (\forall j = 1, \cdots, n) \end{aligned} \tag{11}$$

where $\mathbf{F} = \{\mathbf{f}_1, \cdots, \mathbf{f}_n\}$, n is the number of input local features, $\mathbf{c}_j \in R^M$ is the jth column of $\mathbf{C}$, and $\mathbf{p}_j = R^M$ is the locality adaptor whose kth element is given by $p_{j,k} = \|\mathbf{f}_j - \mathbf{d}_k\|_2$. Eq. (11) can be solved by the Locality-Sensitive Dictionary Learning (LSDL) in [33].

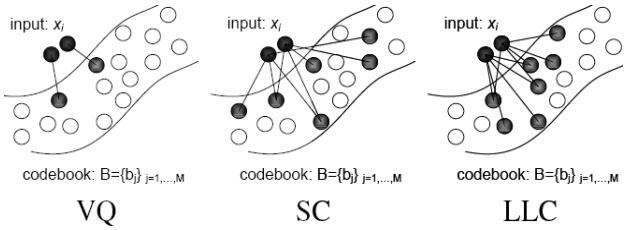

Fig. 5. Comparison between VQ, SC (sparse coding) and LLC (locality-constrained linear coding). The selected bases for representation are highlighted in black.

2.3 Constructing multiscale spatial relationship of features

Due to the different styles of human action, it is difficult to model the ST relationship of local features in a single ST scale. The actions with different styles appear in different motion range (different in spatial scale) and speed (different in temporal scale). To handle this problem produced by different action styles, multiscale is taken into account during feature coding for the purpose of capturing ST relationship between local features.

In our system, spatial and temporal information between features are taken into account, respectively. Specifically, the spatial information is fused into feature coding by MSPC-LC, and the temporal information is considered by multi-temporal-scale sub-STV.

In reality, instead of building spatial pyramid structure, position weight factors α, β are used to control the spatial scales. According to (1), a large (small) α or β intends to select the codewords within a small (large) spatial neighborhood. Thus we can use multiple α or β to obtain the multiscale feature descriptor $\mathbf{f}^{\mathrm{ms}}$

$$\mathbf{f}^{\mathrm{ms}} = [\mathbf{f}^{\alpha(1),\beta(1)}, \cdots, \mathbf{f}^{\alpha(i),\beta(j)}, \cdots] \tag{12}$$

where $i \in [1, n_s], j \in [1, n_t]$, and n_s, n_t are the number of

spatial scales. For example, we set $\alpha = \beta = [3,2,1]$. Then the final coefficient $\mathbf{c}^{\text{ms}}$ is obtained by concatenating all coefficients in multiscale

$$\begin{aligned}\mathbf{c}^{\text{ms}} &= [(\mathbf{c}^{\alpha(1),\beta(1)}),\cdots,(\mathbf{c}^{\alpha(i),\beta(j)}),\cdots]^T \\ \mathbf{c}^{\alpha(i),\beta(j)} &= \arg\min_{\mathbf{c}}(\left\|\mathbf{f}^{\alpha(i),\beta(j)} - \mathbf{D}^{\alpha(i),\beta(j)}.\mathbf{c}\right\|) + \\ &\lambda\left\|\mathbf{p}^{\alpha(i),\beta(j)}.\mathbf{c}^{\alpha(i),\beta(j)}\right\|_2^2)\end{aligned} \tag{13}$$

where $\mathbf{D}^{\alpha(i),\beta(j)}$ is the codebook learnt by LSDL over the feature descriptors at ith, jth temporal scales. Locality adaptor $\mathbf{p}^{\alpha(i),\beta(j)} = [p_1^{\alpha(i),\beta(j)}, \cdots, p_M^{\alpha(i),\beta(j)}]$, whose kth element $p_k^{\alpha(i),\beta(j)}$ is the distance between local feature $\mathbf{f}^{\alpha(i),\beta(j)}$ and codeword $\mathbf{d}_k^{\alpha(i),\beta(j)}$, and defined as $p_k^{\alpha(i),\beta(j)} = \left\|\mathbf{f}^{\alpha(i),\beta(j)} - \mathbf{d}_k^{\alpha(i),\beta(j)}\right\|_2$

2.4 Computing STOE features and their descriptors

In the paper, local STOE features are obtained by filtering using a set of Gaussian derivative filters and their corresponding Hilbert transform filters, point-wise squaring and summation over each 3D cuboid that is associated with a detected STIP. We use the approach in [45] to obtain local STOE feature. Specifically, in the first step, a small ST cuboid $C(\mathbf{x}_i)$ centering on a STIP $p(\mathbf{x}_i)$ is filtered by the directionally selective filters G_{2,θ_j} and H_{2,θ_j} at direction cosines $\theta_j = (\alpha_j, \beta_j, \gamma_j)$, $j = \{1,\cdots,J\}$, J is the number of orientations. Next, the filters are taken in quadrature to eliminate phase sensitivity in the output of each filter. Local

STOE energy $E_{\theta_j}(\mathbf{x}_i)$ in cuboid $C(\mathbf{x}_i)$ at orientation θ_j can be computed according to

$$E_{\theta_j}(\mathbf{x}_i) = (G_{2,\theta_j} * C(\mathbf{x}_i))^2 + (H_{2,\theta_j} * C(\mathbf{x}_i))^2 \tag{14}$$

where $\mathbf{x}_i = (x_i, y_i, t_i)$ denotes the ith STIP coordinate; $*$ denotes convolution; G_2 is 3-D steerable, separable filter based on Gaussian second derivative; H_2 is their corresponding Hilbert transform. The method of constructing ST filters G_2 and H_2 in [45] was adopted in the paper. Then, STOE features at each STIP is normalized as follows [46] :

$$\hat{E}_{\theta_j}(\mathbf{x}_i) = \frac{E_{\theta_j}(\mathbf{x}_i)}{\sum_{k=1}^{J} E_{\theta_k}(\mathbf{x}_i) + \varepsilon} \tag{15}$$

where ε is a constant introduced as a noise floor and to avoid instabilities at points where the overall energy is small. In our system, motion information within each cuboid is decomposed into local STOE features along axes X, Y and T, respectively, corresponding direction cosines $\theta_1 = (1,0,0)$, $\theta_2 = (0,1,0)$, $\theta_3 = (0,0,1)$. Finally, normalized STOE features are converted into feature descriptors using the proposed algorithm as follows:

The cuboid containing normalized response is divided into non-overlapping cells. Histogram features with variable bins are then constructed by summing up normalized response within each cell.

The entropy of each histogram feature is calculated, and normalized by corresponding maximum entropies (for example, for k-bin histogram, its maximum entropy equals to $(-1/k)\ln(1/k)$. Feature of a cell is formed by assembling the normalized entropies of it.

Final feature descriptor is generated by concatenating all cell features.

3. Action Recognition with LGSR and MSPC_LC

3.1 System framework

Extraction and representation of motion information plays a crucial role in human action recognition in video sequence. To obtain sufficient motion information, the steerable filters are employed to decompose the raw dynamic information into STOE features along several directions.

In the paper, to depict ST motion information between STIPs in action video, a novel system framework is proposed, and its basic idea is illustrated in Fig. 6. Firstly, spatio-temporal interest points (STIPs) are extracted from action videos, and local STOE features are calculated (Fig. 6(a), (b)). Secondly, both of STOE features and multi-spatial-scale (MSS) positions of STIPs are connected by the proposed mixed features, and the related MSS codebooks are constructed by k-means over the mixed features (Fig. 6(c), (d)). Then, the mixed features are encoded by MSPC-LC (Fig. 6(e), (f)). Next, multi-temporal-scale (MTS) sub-STVs are built for making use of the temporal information between STIPs, and their MTS coefficient histograms are obtained with max-pooling (Fig. 6(g), (h)). Finally, LGSR is used as classifier over the MTS coefficient histograms (Fig. 6(i)).

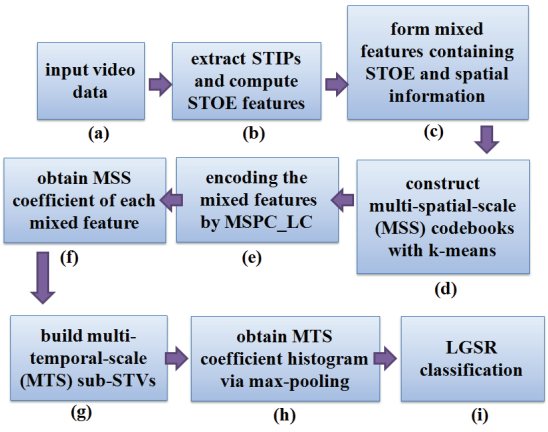

Fig. 6. The flowchart of the proposed action recognition with MSPC-LC and LGSR.

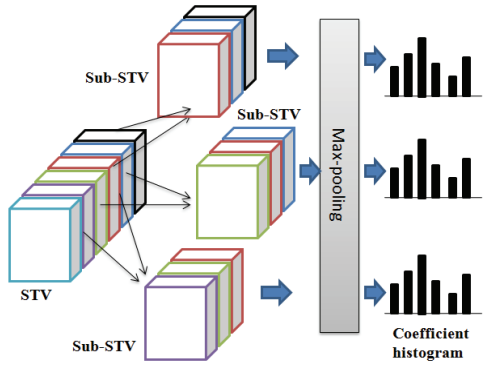

Fig. 7. The illustration of step(g) and step(h) in Fig. 6.

3.2 Building Sub-STVs by Multi-temporal-scale Dense Sampling (MTDS)

Using multi-temporal-scale dense sampling (MTDS) method, the temporal relationship between features can be captured. Specifically, a set of time scales is defined for MTDS depending on the possible action cycle length. Several sub-STVs are then constructed by a sliding window operation. Finally, a group multi-spatial-scale feature descriptors as (12) is built by setting $[\alpha_1,\alpha_2,\alpha_3,\alpha_4]$, $[\beta_1,\beta_2,\beta_3,\beta_4]$, hence the multi-temporal-scale information is included in MTDS.

The advantage of MTDS is that no necessary to consider the problem of aligning the time coordinate system with human action cycles. Because if only the training samples are sufficient, any test sub-STV can always match a training sub-STV. In many cases, this assumption is true in real applications.

3.3 Depicting Sub-STVs by MSPC-LC and Max-pooling Method

Given a sub-STV, there are several local features within it. Each local feature is converted into a multi-spatial-scale feature as (12), and generates a group of multi-spatial-scale feature descriptors:

$$\mathbf{F}^{\text{ms}} = \{\mathbf{f}_1^{\text{ms}},\cdots,\mathbf{f}_n^{\text{ms}}\} \tag{16}$$

Then, MSPC-LC is used to encode each feature descriptor and obtain multiscale class label coefficients:

$$\mathbf{C}^{\text{ms}} = \{\mathbf{c}_1^{\text{ms}},\cdots,\mathbf{c}_n^{\text{ms}}\} \tag{17}$$

sum pooling [31]:

$$\mathbf{s}_{out}(j) = (\mathbf{c}_1^{\text{ms}}(j)+\cdots+\mathbf{c}_n^{\text{ms}}(j)) \tag{18}$$

max pooling [31]:

$$\mathbf{s}_{out}(j) = \max\{\mathbf{c}_1^{\text{ms}}(j)+\cdots+\mathbf{c}_n^{\text{ms}}(j)\} \tag{19}$$

These pooled features can then normalized by sum normalization:

$$\mathbf{s} = \mathbf{s}_{out}\Big/\sum\nolimits_j \mathbf{s}_{out}(j) \tag{20}$$

ℓ_2 normalization

$$\mathbf{s} = \mathbf{s}_{out}\Big/\left\|\mathbf{s}_{out}(j)\right\|_2 \tag{21}$$

where $\mathbf{s}_{out}(j)$ is jth element in sub-STV un-normalized descriptor $\mathbf{s}_{out}$, and $\mathbf{c}_i^{\text{ms}}$ is the MSPC-LC coefficient vector for feature $\mathbf{f}_i^{\text{ms}}$.

3.4 Action classification based on LGSR

To utilize effectively the intrinsic group information from sub-STV descriptors for action classification, we treat each video as one group of sub-STV descriptors and employ Locality-constrained Group Sparse Representation (LGSR) [39] as human action classifier. LGSR was proposed in [39] for human gait recognition. It is an extended version of sparse representation-based classifier (SRC). The pioneering work of SRC was proposed by Wright *et al.* [34] and used to classify face images by minimizing the ℓ_1 norm-regularized reconstruction error, in which it seeks a sparse representation for only a single test image.

Compared with traditional SRC, there are three advantages for LGSR:

1) SRC is designed for single image classification and fails to classify a group of samples, while LGSR is developed for sample group classification.
2) Locality constraint on LGSR is more reasonable than sparsity constraint on SRC, especially for representing manifold data [31, 33].
3) LGSR is a block sparse constraint classifier, and it is suitable than SRC in classification task when the features are discriminative.

The experimental result in Section 4.5 demonstrates the third advantages. The LGSR enforces both of the group sparsity and local smooth sparsity constraint by minimizing the weighted $\ell_{1,2}$ mixed-norm-regularized reconstructed error, and LGSR is defined as

$$\mathbf{A}^* = \arg\min_{\mathbf{A}}\left((\frac{1}{2}\left\|\mathbf{S}-\mathbf{D}\mathbf{A}\right\|_F^2)+\lambda\sum\nolimits_{k=1}^{K}\left\|\mathbf{P}^k\odot\mathbf{A}^k\right\|_F\right) \tag{22}$$

where the first term represents the reconstruction error of the test action video $\mathbf{S}$ with respect to all the training videos $\mathbf{D}$. The second term is the weighted $\ell_{1,2}$ mixed-norm based regularization on the reconstruction coefficient $\mathbf{A}$, and $\lambda \geq 0$ is the regularization parameter to balance these terms. $\mathbf{D}$ is the classification codebook constructed

by concatenating K class-specific codebooks $[\mathbf{D}^1,\cdots\mathbf{D}^K]$. Each class-specific codebook $\mathbf{D}^k$ is learnt with LSDL algorithm [33] over the sub-STV descriptors corresponding to the kth action class. $\mathbf{S}$ is the group of sub-STV descriptors for one test action. Coefficient vector $\mathbf{A}^k$ is one part of $\mathbf{A}$ and corresponds to $\mathbf{D}^k$. $\mathbf{P}^k$ is the distance matrix between $\mathbf{S}$ and $\mathbf{D}^k$, and the element p_{ij}^k in $\mathbf{P}^k$ is defined as $\left\|\mathbf{s}_i-\mathbf{d}_j^k\right\|^2$. Since $\mathbf{A}^k$ values are independent to each other, we can separately update each $\mathbf{A}^k$ using its subgradient [36]. To solve (24), the active set-based subgradient descent algorithm [39, 40] was adopted.

Once we obtain the optimal reconstruction coefficient $\mathbf{A}^*$, two classification methods [39] based on different criteria can be used to classify the test video.

1) Minimum Reconstruction Error (minRE) Criterion: We compute the reconstruction error for each class as follows:

$$R_k((\mathbf{A}^k)^*)=\frac{1}{2}\left\|\mathbf{S}-\mathbf{D}^k(\mathbf{A}^k)^*\right\|_F \tag{23}$$

where the reconstruction coefficient $(\mathbf{A}^k)^*$ is from $\mathbf{A}^k$ that corresponds to the kth training video. Then, we classify the test video to $k^*=\arg\min_k R_k((\mathbf{A}^k)^*)$.

2) Maximum Weighted Inverse Reconstruction Error (maxWIRE) Criterion: In the above criterion, the reconstruction coefficient is not used directly for classification. Intuitively, if the reconstruction errors of the test video with respect to two training videos are the same, we should choose the class label of the training video that is associated with the larger Frobenius norm of the reconstruction coefficient. Specifically, we define the following weighted inverse reconstruction error as follows:

$$Q_k((\mathbf{A}^k)^*)=\frac{\left\|(\mathbf{A}^k)^*\right\|_F}{\left\|\mathbf{S}-\mathbf{D}^k(\mathbf{A}^k)^*\right\|_F} \tag{24}$$

We classify the test video to $k^*=\arg\min_k Q_k((\mathbf{A}^k)^*)$. In the paper, we use maxWIRE criterion as human video action classifier, because it is more reasonable than minRE criterion [33].

4. Experiment and Analysis

In this section, the effectiveness of our MSPC-LC is evaluated on three public datasets: the KTH, Weizmann, and UCF sports datasets. Leave-one-out cross-validation (LOOCV) strategy is used to evaluate the performance of our algorithm.

4.1 Experimental setup

In all experiments, Dollar detector based on multiple ST scales is used to extract STIPs form action videos, and spatial scale $\tau=[1.2,1.3,1.4,1.5]$ and temporal scale $\varpi=[0.4,0.45,0.50,0.55]$. For actions performed by subjects whose bodies do not move (i.e., boxing, handclapping and handwaving in the KTH dataset), STIPs extracted from each action video are coordinate normalized and moved to the region around the origin. For actions performed by moving human bodies and with static background (i.e., actions in the Weizmann dataset), we directly use background subtraction to obtain the center position of body, and centralize the extracted STIPs coordinate. In addition, for actions performed by moving human bodies and with dynamic background (i.e., actions in the UCF Sports dataset), the annotation bounding boxes are used to locate STIPs (for the UCF Sports datatset, the annotation bounding boxes are provided).

To capture multi-temporal-scale relationship of local features, the length of sub-STV is set as 5, 10, 25 and 50 frames. And to describe multi-spatial-scale relationship between local features, four spatial scale parameters are set as $\alpha=[0.1,0.05,0.01,0.005],\beta=[0.1,0.05,0.01,0.005]$ respectively. In MSPC-LC, the codebook size is set to 500. Since there are 4 spatial scales, the dimension of a sub-STV descriptor is $500\times4=2000$. In order to guarantee that the class-specific dictionaries in LGSR are over-complete, random projection in dimension reduction [47] is employed to reduce the dimension of the sub-STV descriptor to 400. In LGSR, the size of each class-specific codebook is set to 600. The other parameters in our methods (i.e., λ) and the parameters of other methods are evaluated by 5-fold cross-validation.

4.2 Action datasets

The KTH dataset contains six classes of human action (i.e., boxing, hand clapping, hand waving, jogging, running, and walking). The actions are performed by 25 different subjects. Each subject performs four action videos in each class. Therefore, the KTH dataset includes $25\times4\times6=600$ low-resolution video clips (160×120 pixels). Each action is performed in four scenarios: indoors, outdoors, outdoors with scale variation, and outdoors with different clothes. Examples of this datasets can be seen in Fig. 8.

The UCF sports dataset includes 150 action videos,

Fig. 8. Example images from the KTH dataset.

Fig. 9. Examples from the UCF Sports dataset

Fig. 10. Examples from the Weizmann dataset

which are collected from various broadcast sports channel such as BBC and ESPN. It contains 10 different actions: diving, golf swing, horse riding, kicking, lifting, running, skating, swing bar, swing floor, and walking. This dataset is challenging with a wide range of scenarios and viewpoints. Examples of this dataset can be seen in Fig. 9.

The Weizmann dataset contains 93 low-resolution video clips (180×144 pixels) from nine different subjects, each of whom performs 10 different actions including walking (walk), running (run), jumping (jump), galloping sideways (side), bending (bend), one-hand-waving (wave one), two-hand-waving (wave two), jumping in place (pjump), jumping jack (jack), and skipping (skip). One of the subjects performs walking, running, and skipping twice. The camera setting is fixed and there is no occlusion or viewpoint change. Besides, each subject performs under similar plain background. Some examples are demonstrated in Fig. 10.

4.3 Action datasets performance comparison with BoF model

In this section, the performance of our MSPC-LC algorithm is compared with existing local feature coding methods VQ, SC and LLC [31]. k-nearest neighbor (k-NN) is used as action classifier, and parameter k is set to 5. Keeping the same codebook size with MTDS, K-means clustering is utilized to build codebook with VQ and LLC, and the software in [41] is utilized for SC. In LLC, the locality constraint parameter k is set to 5. In our method, a group of sub-STV descriptors are extracted from a test video and classified with k-NN. Next, the vote score of these sub-STVs decides the label of the test video. In feature pooling phase, sum-pooling is used for VQ, while max-pooling is applied in SC, LLC, and MSPC-LC. And all select sum normalization as normalization method.

In addition, to evaluate the above locality coding algorithms which are used for improving BoF of feature coding in Section 2, another experiment is carried out. Firstly, considering the feature position constraint in Section 2.1, the coding method in (2) is treated as the basic spatio-temporal coding (StC). Secondly, considering the locality constraint in Section 2.2, the coding method in (5) is regarded as the locality-constrained spatio-temporal coding (LSC). In this experiment, the codebook size is still set to 500. K-means clustering is adopted to build codebook for StC. LSDL is used to build codebook for LSC. Sum-pooling is used for StC, and max-pooling is adopted for LSC. The parameter for k-NN is set to 5. The spatial control factors are set as $\alpha = [0.1, 0.05, 0.01, 0.005]$, $\beta = [0.1, 0.05, 0.01, 0.005]$. The length of sub-STV is set as 5, 10, 25 and 50 frames, for capturing multi-temporal-scale relationship of local features.

Table 1 shows the result of performance comparison. The recognition rates are the average values on three datasets. It can be seen that StC method achieves better performance than VQ, SC, and LLC. This demonstrates that the ST relationship is important for human video action recognition, and the locality-constrained ST coding is better than StC. In addition, the locality-constraint is useful to handle the manifold of local features. Benefiting from modeling the multiscale ST relationship of local features, MSPC-LC achieves the highest average recognition accuracy on the three datasets.

4.4 MSPC-LC versus SPM

The spatial pyramid matching (SPM) model is employed to capture the spatial relationship of local ST features [16]. Here, a 4-level SPM (Fig. 11) is used for evaluation. MSPCLC, SPM, LLC and max-pooling are used to describe sub-STV, respectively. Then, classifying sub-STVs is based on k-NN criterion. Finally, the vote score-based classifier (similar with Section 4.3) decides the class label of test video. Table 2 shows the average recognition accuracies on the three datasets. It is clear that k-NN+ MSPC-LC achieves better performance than only SPM on all datasets. Different from SPM [42] that only considers the spatial relationship of local features, MSPCLC

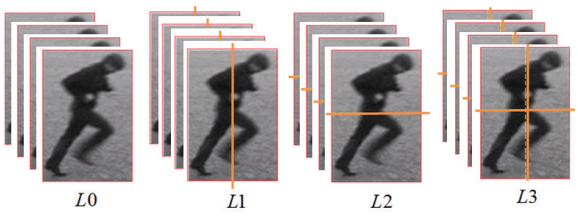

Fig. 11. SPM with 4-level structure is employed in performance evaluation.

Table 2. Performance comparison between MSPC-LC and SPM

Methods	KTH(%)	UCF Sp.(%)	Weiz.(%)
k-NN+SPM	91.7	82.8	93.5
k-NN+MSPC-LC	**94.4**	**85.6**	**96.5**

Table 3. Performance comparison between LGSR and SRC

Methods	KTH(%)	UCF Sp.(%)	Weiz.(%)
SRC+MSPC-LC	96.5	92.1	98.5
LGSR+MSPC-LC	**98.5**	**93.5**	**98.7**

simultaneously makes use of the spatial and temporal relationships. In addition, compared with the fixed grids in SPM, MSPC-LC is a more flexible representation.

4.5 LGSR versus SRC

To evaluate the classification performance based on LGSR, the standard SRC [31] is also employed. The object function of SRC is defined as

$$\mathbf{a}_i^* = \arg\min_{\mathbf{a}_i} \|\mathbf{s}_i - \mathbf{D}\mathbf{a}_i\|_2^2 + \lambda_{SRC} \cdot \|\mathbf{a}_i\|_1 \quad (25)$$

where $\mathbf{s}_i$ is the ith sub-STV descriptor in $\mathbf{S}$, $\mathbf{a}_i$ is its corresponding coefficient. Similar to LGSR, the maxWIRE criterion is also used in SRC. As mentioned in Section 3.4, compared with SRC, there are three advantages of LGSR. In particular, if the features are not shared with other classes, the block sparse constraint is more suitable for the classification than sparse constraint. Hence LGSR is relatively better than SRC for classification task when using less shared features. The comparison of average accuracy (Table 3) shows that LGSR outperforms than SRC on the KTH and UCF sports datasets. It is worth to note that Guha and Ward [19] suggested that sparse constraint is more important than block sparse constraint in human action recognition based on local ST features. Compared with local ST features, the obtained sub-STV descriptors with MSPCLC are less shared with other actions. Hence, it is plausible to utilize block sparse constraint than sparse constraint for action classification together with MSPC-LC.

4.6 Performance comparison with other systems

Table 4 shows the performance comparison between the proposed algorithm and the classical systems. In experiment setting, action performers are randomly selected as training data and the rest as test data. The competing methods include local representation-based methods [8, 14], and global representation-based methods [14]. Specifically, SC was used for feature coding together with BoF in [17]; and a novel local feature detector was proposed for human action recognition in [14]; local feature distribution information was taken into consideration in [36]; ST context feature was employed in [34]; a ST context constraint coding method was utilized in [38]; sparse representation-based classification methods was applied in [18]; and the holistic action representation method was adopted in [5]. It demonstrates from Table 5 that our algorithm outperforms these classical algorithms. The confusion matrices on the KTH and UCF sports datasets are shown in Table 4 and 5, respectively.

Some reasons are responsible for our good classification performance. Firstly, by fusing spatial position information between local features into codebook construction and feature coding, our method performs better than these methods [8, 17, 18] that only use the feature appearance information to represent human action. Secondly, compared with the feature distribution information [36] and ST context methods [35, 38], our method is a fine and complete method. For example, as illustrated in Fig. 2, each local feature has two types fundamental information (*where, how*) in STV. More specifically, the coordinate (x, y) indicates *where* the body part locates. And the motion information (described as STOE feature) shows *how* the body part moves. It is noted that our feature do not contain appearance information (for example, histogram of oriented gradient (HOG) feature), because appearance information

Table 4. Confusion matrix on the KTH dataset with our method LGSR+MSPC-LC. s1(boxing), s2(hand-clapping), s3(hand-waving), s4(walking), s5 (jogging), s6(running).

	s1	s2	s3	s4	s5	s6
s1	**1.00**					
s2		**1.00**				
s3			**1.00**			
s4				**0.95**	0.02	0.03
s5				0.02	**0.96**	0.02
s6						**1.00**

Table 5. Performance comparison with other systems

methods	year	setting	KTH(%)	UCF(%)	Weiz.
Zhu *et al.*[18]	2010	Split	94.9	84.3	---
Wu *et al.* [35]	2011	LOOCV	94.5	91.3	---
Escobar *et al.* [15]	2012	Split	90.6	---	99.2
Guha *et al.* [19]	2012	LOOCV	---	91.1	98.9
Bregonzio *et al.* [36]	2012	LOOCV	94.3	---	96.6
Zhang *et al.* [38]	2012	LOOCV	95.6	87.3	---
Saghafi *et al.* [16]	2012	LOOCV	92.6	---	100
Deng *et al.* [9]	2012	LOOCV	96.9	88.4	100
LGSR+MSPC-LC		LOOCV	**98.5**	**93.5**	100

Table 6. Confusion matrix on the UCF Sports dataset with 4-level spatial pyramid match (SPM). s1(diving), s2(golf), s3(kicking), s4(lifting), s5(horse-riding), s6(running), s7(skating), s8(swing-bench), s9 (swing-high bar), s10 (walking)

	s1	s2	s3	s4	s5	s6	s7	s8	s9	s10
s1	**1.00**									
s2		**0.91**	0.06							0.03
s3		0.05	**0.90**			0.05				
s4				**1.00**						
s5		0.06			**0.92**		0.02			
s6			0.03	0	0.05	**0.91**	0.01			
s7				0.03		0.05	**0.91**			
s8						0.08		**0.92**		
s9									**1.00**	
s10								0.12		**0.88**

is prone to be disturbed by the various clothes of subjects. In MSPC-LC, all these information (*where, how*) is represented via (x, y) and STOE feature, then, fused into feature coding. Moreover, the spatial information (*when*) between features is utilized by multi-temporal-scale sub-STV. However, these methods [35, 38, 36] ignore some one of these information (*where, when* and *how*) in action representation processing.

5. Conclusion

In this paper, to capture the ST relationship of local features for human action recognition, we built the mixed features that combines STOE feature and spatial position information with spatial scale parameters. Then, the mixed features are encoded by the proposed MSPC-LC algorithm. The experimental results on the public datasets show that (1) feature spatio-temporal position information effectively improves the performance of action recognition (2) by changing spatial parameters, the mixed features can provide more useful information to the action classifier (3) combining feature spatial position into feature coding is a beneficial alternative way for this task. In particular, combining feature spatial position into feature coding is a better approach than feature distribution [36], spatio-temporal context [35], and SPM-based methods [23], when using a multiscale version.

The major limitations of our system is that (1) In the preprocessing stage, it is difficult in locating the human head and measuring the human body length, when the environment is complicated, such as crowed street (2) The values of spatial scale parameters can greatly influence on the performance of action recognition. However, in many cases, their values are set empirically. It is valuable to explore new methods to capture the spatiotemporal location of local features in our future work.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (No.61301270, No.61271288), the Foundation Research Funds for the Central Universities of China (No.ZYGX2013J025), and the Research Fund for the Doctoral Program of Higher Education of China (No.20130185120014), the National High Technology Research and Development Program (No.2012AA011503).

References

[1] X. Wu, J. Lai, "Tensor-based projection using ridge regression and its application to action classification," *IET Image Processing*, vol. 4, no. 6, pp. 486-493, 2010.

[2] A. A. Chaaraoui, P. C. Perez, "Silhouette-based human action recognition using sequences of key poses," *Pattern Recognition Letters*, vol. 34, no. 15, pp. 1799-1807, 2013.

[3] K. N. Tran, I. A. Kakadiaris, S. K. Shah, "Modeling motion of body parts for action recognition," *in Proceedings of the British Machine Vision Conference*, pp.1-12, 2011.

[4] B. Huang, G. Tian, F. Zhou, "Human typical action recognition using gray scale image of silhouette sequence," *Computers and Electrical Engineering,* vol. 38, no. 5, pp. 1177-1185, 2012.

[5] S. A. Rahman, M. K. H. Leung, S. Y. Cho, "Human action recognition employing negative space features," *Journal of Visual Communication and Image Representation*, vol. 24, no. 3, pp. 217-231, 2013.

[6] B. Saghafi, D. Rajan, "Human action recognition using pose-based discriminant embedding," *Signal Processing*, vol. 27, no. 1, pp. 96-111, 2012.

[7] S. M. Yoon and A. Kuijper, "Human action recognition based on skeleton splitting," *Expert Systems with Applications*, vol. 40, no. 17, pp. 6848-6855, 2013.

[8] L. Shao, L. Ji, Y. Liu, J. Zhang, "Human action segmentation and recognition via motion and shape analysis," *Pattern Recognition Letters*, vol. 33, no. 4, pp. 438-445, 2012.

[9] X. Deng, X. Liu, M. Song, "LF-EME: local features with elastic manifold embedding for human action recognition," *Neurocomputing*, vol. 99, no. 1, pp. 144-153, 2013.

[10] P. Dollar, V. Rabaud, G. Cottrell, S. Belongie, "Behavior recognition via sparse spatio-temporal features," *Proceedings of the 2nd Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance,* pp. 65-72, October, 2005.

[11] A. Klaser, M. Marszalek, C. Schmid, "A spatio-temporal descriptor based on 3D-gradients," *in Proceedings of the British Machine Vision Conference*, 2008.

[12] P. Scovanner, S. Ali, M. Shah, "A 3-dimensional sift descriptor and its application to action recognition," *in Proceedings of the 15th ACM International Conference on Multimedia,* pp. 357-360, September 2007.

[13] G. Willems, T. Tuytelaar, L. Van Gool, "An efficient dense and scaleinvariant spatio-temporal interest point detector," *in Proceedings of the Europen Conference on Computer Vision*, pp. 650-663, 2008.

[14] I. Laptev, M. Marszaek, C. Schmid, B. Rozenfeld, "Learning realistic human actions from movies," *in Proceedings of the 26th IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1-8, June 2008.

[15] M. J. Escobar, P. Kornprobst, "Action recognition via bioinspired features: the richness of center-surround interaction," *Computer Vision and Image Under-*

standing, vol. 116, no. 5, pp. 593-605, 2012.
[16] X. Zhu, Z. Yang, J. Tsien, "Statistics of natural action structures and human action recognition," *Journal of Vision*, vol. 12, no. 9, pp. 834-834, 2012.
[17] B. Chakraborty, M. B. Holte, T. B. Moeslund, J. Gonzalez, "Selective spatio-temporal interest points," *Computer Vision and Image Understanding*, vol. 116, no. 3, pp. 396-410, 2012.
[18] Y. Zhu, X. Zhao, Y. Fu, "Sparse coding on local spatialtemporal volumes for human action recognition," *in Proceedings of the Computer Vision*, pp. 660-671, Springer, Berlin, Germany, 2010.
[19] T. Guha, R K. Ward, "Learning sparse representations for human action recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 8, pp. 1576-1588, 2012.
[20] S. T. Roweis, L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, pp. 2323-2326, 2000.
[21] J. Tenenbaum, V. DeSilva, J. Langford, "Aglobal geometric framework for nonlinear dimensionality reduction", *Science*, vol. 290, pp. 2319-2323, 2000.
[22] A. Elgammal, R. Duraiswami, L. Davis, "Efficient kernel density estimation using the fast gauss transform with applications to color modeling and tracking," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 25, pp. 1499-1504, 2003.
[23] M. Breunig, H. P. Kriegel, R. T. Ng, J. Sander, "LOF: identifying density-based local outliers," *in Proceeding software 2000 ACM SIGMOD International Conference on Management of Data*, 2000.
[24] K. Yu, T. Zhang, Y. Gong, "Nonlinear learning using local coordinate coding," *in Advances in Neural Information Processing Systems,* vol. 22, pp. 2223-2231, 2009.
[25] J. Wang, J. Yang, K. Yu, F. Lv, "Locality-constrained linear coding for image classification," *in Proceedings of the IEEE International Conferenceon Computer Vision and Pattern Recognition*, pp. 3360-3367, 2010.
[26] Y. W. Chao, Y. R. Yeh, Y. W. Chen, "Locality-constrained group sparse representation for robust face recognition," *in Proceeding sof the IEEE International Conference on ImageProcessing (ICIP),* pp. 761-764, 2011.
[27] M. Zheng, J. Bu, C. Chen, "Graph regularized sparse coding for image representation," *IEEE Transactionson Image Processing*, vol. 20, pp. 1327-1336, 2011.
[28] J. A. Tropp, "Greed is good: algorithmic results for sparse approximation," *IEEE Transactions on Information Theory*, vol. 50, pp. 2231-2242, 2004.
[29] B. A. Olshausen, D. J. Field, "Emergence of simple-cell receptive field properties by learning a sparse code for natural images," *Nature*, vol. 381, no. 6583, pp. 607-609, 1996.
[30] K. Yu, T. Zhang, Y. Gong, "Nonlinear learning using local coordinate coding," *in Proceedings of the 23rd Annual Conference on Neural Information Processing Systems*, pp. 2223-2231, December, 2009.
[31] J. Wang, J. Yang, K. Yu, F. Lv, "Locality-constrained linear coding for image classification," *in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition,* pp. 3360-3367, June, 2010.
[32] S. T. Roweis, L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 5500, pp. 2323-2326, 2000.
[33] C.P. Wei, Y.W. Chao, Y. R. Yeh, "Locality-sensitive dictionary learning for sparse representation based classification," *Pattern Recognition*, vol. 46, no. 5, pp. 1277-1287, 2013.
[34] J. Wright, A.Y. Yang, A. Ganesh, S.S. Sastry, Y. Ma, "Robust face recognition via sparse representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 2, pp. 210-227, 2009.
[35] X. Wu, D. Xu, L. Duan, J. Luo, "Action recognition using context and appearance distribution features," *in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* pp. 489-496, June 2011.
[36] M. Bregonzio, T. Xiang, S. Gong, "Fusing appearance and distribution information of interest points for action recognition," *Pattern Recognition*, vol. 45, no. 3, pp. 1220-1234, 2012.
[37] B. Chakraborty, M. B. Holte, T. B. Moeslund, J. Gonzalez, "Selective spatio-temporal interest points," *Computer Vision and Image Understanding*, vol. 116, no. 3, pp. 396-410, 2012.
[38] Z. Zhang, C. Wang, B. Xiao, "Action recognition using context constrained linear coding," *Signal Processing Letters*, vol. 19, no. 7, pp. 439-442, 2012.
[39] D. Xu, Y. Huang, Z. Zeng, X. Xu, "Human gait recognition using patch distribution feature and locality-constrained group sparse representation," *IEEE Transactions on Image Processing*, vol. 21, no. 1, pp. 316-326, 2012.
[40] M. Liu, S. Yan, Y. Fu, T. S. Huang, "Flexible X-Y patches for face recognition," *in Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing,* pp. 2113-2116, April 2008.
[41] 2013, http://spams-devel.gforge.inria.fr/.
[42] S. Lazebnik, C. Schmid, J. Ponce, "Beyond bags of features: spatial pyramid matching for recognizing natural scene categories," *in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 2169-2178, June, 2006.
[43] J. Yang, K. Yu, Y. Gong, T. Huang, "Linear spatial pyramid matching using sparse coding for image classification," *in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2009.
[44] H. Lee, A. Battle, R. Raina, A. Ng, "Efficient sparse

coding algorithms," *Advances in Neural Information Processing Systems*, MIT Press, pp. 801-808, 2007.
[45] K. G. Derpanis, J. M. Gryn, Three-dimensional nth derivative of Gaussian separable steerable filters, *IEEE Int. Conf. on Image Processing*, vol. 3, 2005.
[46] K.G. Derpanis, M. Sizintsev, K. Cannons, R. P. Wildes, "Efficient Action Spotting based on a Spacetime Oriented Structure Representation," *In Proceedings of the IEEE Conference Computer Vision and Pattern Recognition*, 2010.
[47] R. baraniuk, M. Wakin, "Random projection of smooth manifold," *foundation of computational mathmaematics*, vol. 9, pp. 51-77, 2009.

Jiang-feng Yang 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 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 is a professor in School of Electronic Engineering, University of Electronic Science and Technology of China. 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. Her research interests include signal processing, machine vision and Internet security.