Hindawi Publishing Corporation
The Scientific World Journal
Volume 2013, Article ID 405645, 11 pages
http://dx.doi.org/10.1155/2013/405645

Research Article

Multi-Scale Locality-Constrained Spatiotemporal Coding for Local Feature Based Human Action Recognition

Bin Wang, Yu Liu, Wei Wang, Wei Xu, and Maojun Zhang

College of Information System and Manage, National University of Defense Technology, 109 Deya Road, Changsha, Hunan 410073, China

Correspondence should be addressed to Bin Wang; nudtwangbin@163.com

Received 2 July 2013; Accepted 21 August 2013

Academic Editors: R. Haber, P. Melin, and Y. Zhu

We propose a Multiscale Locality-Constrained Spatiotemporal Coding (MLSC) method to improve the traditional bag of features (BoF) algorithm which ignores the spatiotemporal relationship of local features for human action recognition in video. To model this spatiotemporal relationship, MLSC involves the spatiotemporal position of local feature into feature coding processing. It projects local features into a sub space-time-volume (sub-STV) and encodes them with a locality-constrained linear coding. A group of sub-STV features obtained from one video with MLSC and max-pooling are used to classify this video. In classification stage, the Locality-Constrained Group Sparse Representation (LGSR) is adopted to utilize the intrinsic group information of these sub-STV features. The experimental results on KTH, Weizmann, and UCF sports datasets show that our method achieves better performance than the competing local spatiotemporal feature-based human action recognition methods.

1. Introduction

Human action recognition in video has been widely studied over the last decade due to its widespread application prospects in the areas such as video surveillance [1, 2], action-based human computer interfaces [3], and video content analysis [4]. It is an important branch in the field of artificial intelligence. It has also been an increasingly active field of computer vision and pattern recognition. However, the action videos are affected by illumination changes, motion blur, occlusion, and other factors. They make human action recognition still a challenging task [5, 6].

Many human action recognition techniques have been proposed, and several reviews are devoted to this topic [5, 6]. There are two respects in this field [5]: video representation and classification. Video representation is the process of extracting features from videos and obtaining the behavior representation by encoding the features. Then an action model is learned from the final behavior representations and used to recognize new behaviors. In general, there are two representations methods: global representations [7–14] and local representations [15–25]. Common global representations are derived from silhouettes or body sketch. They need fine foreground segment or body part tracking. Thus, they are sensitive to noise, variations in viewpoint, and partial occlusion. Local representations are based on the local spatiotemporal features together with bag of features (BoF) model. Without foreground segment or body part tracking, they are less sensitive to viewpoint changes, noise, appearance, and partial occlusions [5].

There are three respects in BoF-based human action recognition: extracting local features in videos, obtaining video representation vector via these local features, and classifying action videos with a classifier upon the video representation vector [5]. To obtain video representation vector, several feature coding and pooling methods are provided. Many authors used K-means and vector quantization (VQ) for feature coding, as well as the avg-pooling [16] to group these feature codes to generate the video representation vector. To reduce the quantization error due to K-means and VQ, assign one code word for a feature, soft vector quantization (SVQ) [26] and sparse coding (SC) [27] are adopted to encode local features for action recognition tasks [24]. However, the local features usually reside on nonlinear manifolds [15, 28, 29]. Neither SVQ nor SC can preserve the nonlinear manifold structure. The manifold is nonlinear

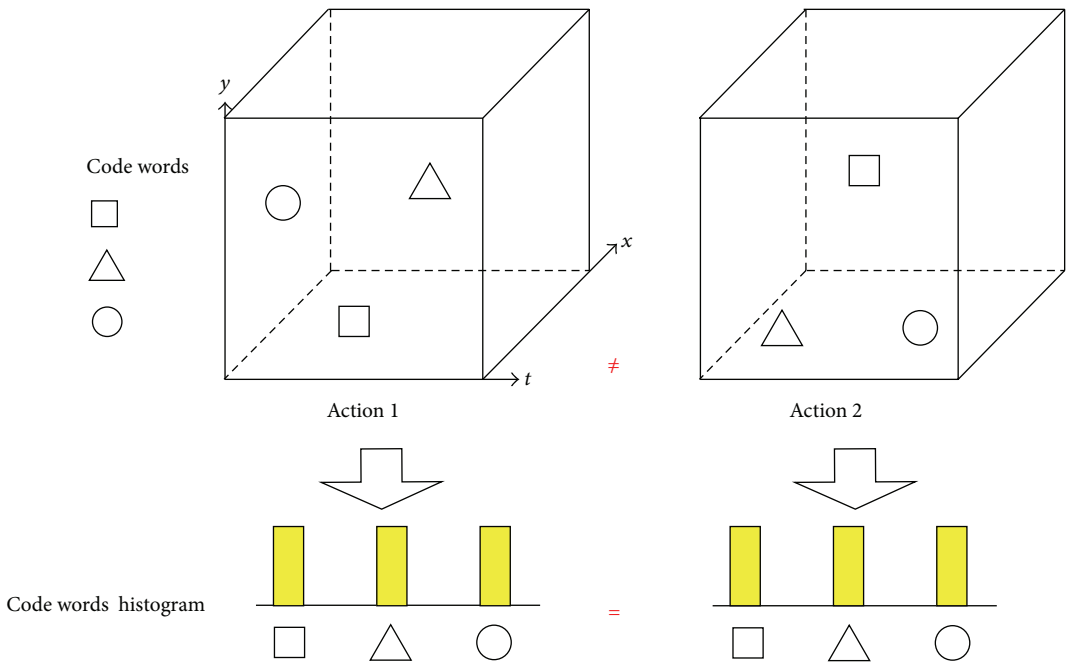


Figure 1: The illustration of that BoF ignores the spatiotemporal relationship of local features in STV. There are two groups of similar features with different spatiotemporal arrangements in STV. They are form Action 1 and Action 2, respectively. However, they cannot be correctly classified with BoF, because it obtains two same code words histograms due to ignoring feature spatiotemporal relationship.

and not Euclidean in its whole space, but linear and Euclidean in a local region [30, 31]. Because SVQ uses all bases to encode each feature and generates dense codes, it cannot precisely represent the nonlinear manifold structure with a global way. Due to the over complete dictionary, SC tends to choose the code words which are distant to the input features [29]. Thus, it cannot correctly represent manifold data. Hereafter, we consider these limitations both quantization error and loss manifold structure in feature coding as representation error. For this issue, Yu et al. [28] provided a Local Coordinate Coding (LCC) to encode feature with locality-constrained, Wang et al. [29] introduced an improved version of LCC named Locality constrained Linear Coding (LLC) to reduce computational cost, and Wei et al. [32] proposed a local sensitive dictionary learning method for image classification.

In action classification stage, support vector machine (SVM) has been widely used when the video representation vectors are provided. Recently, inspired by the major success of Sparse Representation-based Classification (SRC) in face recognition [33], some authors [25] explored SRC for human action recognition and achieved better performance than SVM. Nevertheless, these local representation methods suffer one important limitation. They largely ignore the spatiotemporal relationship among local features, such as temporal order and spatial arrangement [34–36]. For example, in Figure 1, two different actions in the left and right space-time-volume (STV) have the same local features and appear different spatiotemporal configurations. Due to the same histograms generated by BoF, they are incorrectly considered as one action. Recently, some researchers exploited some approaches to use spatiotemporal context information [34, 35], local feature distribution [34, 36], and spatial pyramid matching (SPM) [23, 37] with regard to this problem.

In this paper, we introduce a Multiscale Locality-Constrained Spatiotemporal Coding (MLSC) method to address this limitation and reduce the representation error simultaneously. To reduce the representation error (quantization error and loss manifold structure), we adopt locality-constraint into dictionary learning and feature coding from the respect of manifold learning [30, 31]. To model the spatiotemporal relationships of local features, we involve feature spatiotemporal positions into dictionary learning and feature coding. Then, the spatiotemporal relationship of local features can be obtained from the features codes. In addition, to handle with the different action styles (the space and time range variant of action), the multiscale spatiotemporal relationship is also modeled by MLSC. In practice, local features are firstly projected into sub space-time-volume (sub-STV) to obtain their spatiotemporal positions. Then dictionary learning and local features coding are implemented with locality and position constraint. To classify one action video (see Figure 2), a group of sub-STV are densely sampled, and a group of sub-STV descriptors are obtained with MLSC and max-pooling [29]. Then Locality-Constraint Group Sparse Representation [38] is adopted for action classification upon these sub-STV descriptors.

Compared to these methods which use spatiotemporal context information [34, 35] or feature distribution [36] to handle the limitations of BoF, MLSC is a more *fine* and *whole* method, because it records the whole elements (*where, when, who,* and *how*) of local features for human action recognition

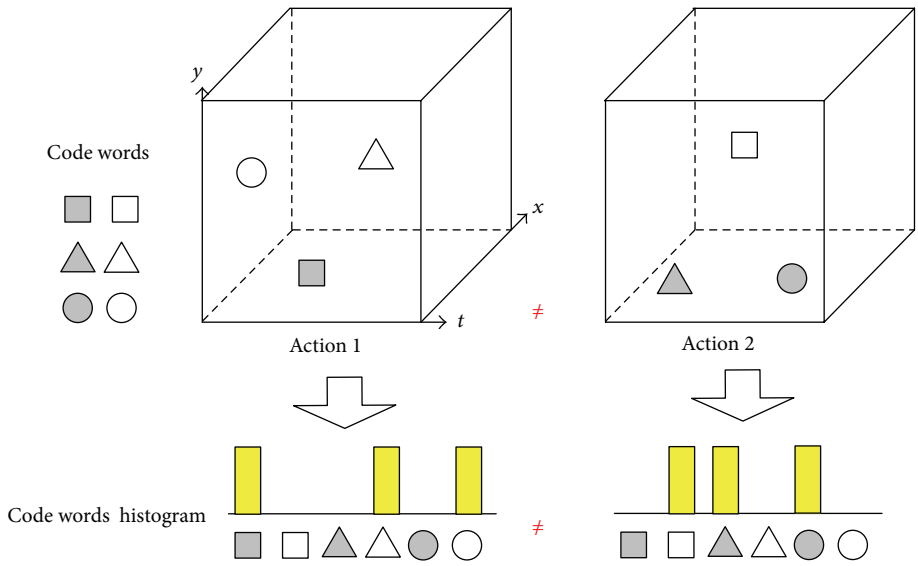


FIGURE 2: The illustration of our method. The shape of code words indicates appearance information, while the color indicates spatiotemporal position. For example, the code words ■ □ have same appearance but different position. They are considered as two different code words in feature coding. Then, two different code words histograms are obtained for two actions upon these code words. Hence, Actions 1 and 2 which cannot be distinguished in BoF (Figure 1) can be correctly classified with our method.

(detailed in Section 4.6). The experimental results on KTH, Weizmann, and UCF sports datasets show that our method achieves better performance than these methods [23, 34–36] and other local spatiotemporal feature-based methods.

There are three contributions in this paper. First, to solve the limitations of BoF, a novel feature coding method MLSC is proposed for modeling local feature spatiotemporal relationships, at the same time, reducing representation error. In addition, to deal with action style variant, the multiscale spatiotemporal relationship is also modeled by MLSC. Second, to effectively use MLSC, a novel human action recognition framework is proposed (detailed in Figure 2). It extracts the dense sub-STV descriptors from videos and classifies actions upon these descriptors. Third, in order to utilize the intrinsic group information from these sub-STV descriptors within one video, the Locality-Constrained Group Sparse Representation-(LGSR-) [38] based classifier is adopted for action classification.

The rest of this paper is organized as follows. MLSC is proposed in Section 2. The human action recognition framework with MLSC and LGSR is provided in Section 3. Then, experimental results and analysis are shown in Section 4. Finally, conclusions are drawn in Section 5.

2. Multiscale Locality-Constrained Spatiotemporal Coding

2.1. Modeling Spatiotemporal Relationship with Feature Position. In BoF model, both code words learning and local feature coding only use feature appearing information, but they discard feature position [16]. It is the reason why BoF ignores the feature spatiotemporal relationship. To solve this problem, the feature spatiotemporal positions are involved into dictionary learning and feature coding in this paper. It is inspiring from the work of Liu et al. [39]. They involved spatial locations into DCT-based local feature descriptors to model the spatial relationship of local features for face recognition. Their experimental results showed that feature locations can improve local feature-based face recognition accuracy.

In this paper, the feature descriptor and feature spatiotemporal location (x, y, t) are connected together to generate a new feature descriptor $\mathbf{f}^{\alpha,\beta}$:

$$\mathbf{f}^{\alpha,\beta} = \left[\mathrm{HOG}^T, \mathrm{HOF}^T, \alpha(x, y), \beta t\right]^T, \tag{1}$$

where α and β are the position weighting factors which represent the importance of the spatial and temporal position in feature matching, respectively. The histograms of gradient orientations (HOG) and the histogram of optic flow (HOF) [20] are adopted. Then the feature spatiotemporal relationship can be modeled with dictionary learning and feature coding upon the new feature descriptor $\mathbf{f}^{\alpha,\beta}$.

To easily explain the role of involving feature position into dictionary learning and feature coding, we adopt K-means to learn dictionary and VQ to encode features, respectively. The representation error caused by them will be solved in Section 2.2. $\mathbf{D}^{\alpha,\beta} \in R^{N\times M}$ is a dictionary learnt with K-means clustering upon the features $\mathbf{F}^{\alpha,\beta} = [\mathbf{f}_1^{\alpha,\beta}, \ldots, \mathbf{f}_n^{\alpha,\beta}]$. In $\mathbf{D}^{\alpha,\beta}$, each visual words $\mathbf{b}^{\alpha,\beta}$ has three types information: visual words appearing information (HOG/HOF), spatial position (x, y), temporal position (t). The code $\mathbf{c}$ for feature $\mathbf{f}^{\alpha,\beta}$ is obtained with VQ:

$$\mathbf{c}_i = \begin{cases} 1, & \text{if } i = \arg\min_i \left\| \mathbf{b}_i^{\alpha,\beta} - \mathbf{f}^{\alpha,\beta} \right\|_2, \\ 0, & \text{otherwise,} \end{cases} \tag{2}$$

where $\mathbf{f}^{\alpha,\beta}$ is the input feature and is described with (1). $\mathbf{b}_i^{\alpha,\beta}$ is the ith base in dictionary $\mathbf{D}^{\alpha,\beta}$. $\mathbf{c} \in R^M$ is the code for $\mathbf{f}^{\alpha,\beta}$.

According to (1), the base $\mathbf{b}^{\alpha,\beta}$ which is chosen to encode $\mathbf{f}^{\alpha,\beta}$ must be the closest to $\mathbf{f}^{\alpha,\beta}$ in three respects: feature similarity, spatial distance, and temporal distance. Thence, the spatiotemporal position of $\mathbf{f}^{\alpha,\beta}$ form its code $\mathbf{c}$ can be obtained. Given a group of local features, their spatiotemporal relationship can be represented with their code words histogram:

$$\mathbf{H}_i = \frac{1}{n}\sum_{i=1}^{n}\mathbf{C}_i, \tag{3}$$

where $\mathbf{H} \in R^M$ is the code words histogram, n is the number of features, and $\mathbf{C}$ is the code of these features.

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

2.2. Reducing Representation Error with Locality Constraint. In Section 2.1, K-means and VQ are adopted in dictionary learning and feature coding. However, Yu et al. [28] discovered that VQ cannot handle nonlinear manifold structure well. Because it is a 0th order (constant) approximation of object functions from the view of function approximation. In addition, VQ causes nontrivial quantization error. They suggested that 1st-order (linear) approximation can solve these problems and introduced adding locality constraint into object function:

$$\mathbf{c} = \arg\min_{\mathbf{c}} \left\|\mathbf{f}^{\alpha,\beta} - \mathbf{D}^{\alpha,\beta}\mathbf{c}\right\|_2 + \lambda\|\mathbf{p}\odot\mathbf{c}\|_1, \quad \text{st: } 1^T\mathbf{c} = 1, \tag{4}$$

where the first term represents the reconstruction error of an input feature $\mathbf{f}^{\alpha,\beta}$ with respect to dictionary $\mathbf{D}^{\alpha,\beta}$, the second term is locality-constraint regularization on code $\mathbf{c}$, and λ is a regularization factor to balance these terms. In the second term, $\mathbf{p}_j = \|\mathbf{f}^{\alpha,\beta} - \mathbf{b}_j^{\alpha,\beta}\|_2$ is the distance between $\mathbf{f}^{\alpha,\beta}$ and jth code word $\mathbf{b}_j^{\alpha,\beta}$, $\odot$ is the element product, and $1^T\mathbf{c} = 1$ is the shift invariant constraint according to [28].

Equation (4) tends to choose the code words which are close to $\mathbf{f}^{\alpha,\beta}$for generating the code $\mathbf{c}$. Because $\mathbf{p}$ is fixed, to minimize $\|\mathbf{f}^{\alpha,\beta} - \mathbf{D}^{\alpha,\beta}\mathbf{c}\|_2 + \lambda\|\mathbf{p}\odot\mathbf{c}\|_1$, one needs to make the coefficient $\mathbf{c}_j$ corresponding to large $\mathbf{p}_j$ equals 0. In addition, $\|\|_1$ is spares regularization term and intends to obtain sparse solution. Sparsity indicates that many elements in $\mathbf{c}$ are zero, while only a few are nonzero. Thus only a few code words near to $\mathbf{f}^{\alpha,\beta}$ are selected to encode feature $\mathbf{f}^{\alpha,\beta}$. Obviously, the selected code words belong to the local neighbor of $\mathbf{f}^{\alpha,\beta}$.

However, an iterative optimization is needed to solve the l^1optimization problem in (4). To reduce the computational cost in (4), we use $\|\mathbf{p}\odot\mathbf{c}\|_2$ to replace $\|\mathbf{p}\odot\mathbf{c}\|_1$. Consider

$$\mathbf{c} = \arg\min_{\mathbf{c}} \left\|\mathbf{f}^{\alpha,\beta} - \mathbf{D}^{\alpha,\beta}\mathbf{c}\right\|_2 + \lambda\|\mathbf{p}\odot\mathbf{c}\|_2, \quad \text{st: } 1^T\mathbf{c} = 1. \tag{5}$$

In (5), $\mathbf{p}$ is fixed. To minimize $\|\mathbf{p}\odot\mathbf{c}\|_2$, the code words far from $\mathbf{f}^{\alpha,\beta}$ will be assigned zero in $\mathbf{c}$. In contrast, the code words near to $\mathbf{f}^{\alpha,\beta}$ will be assigned nonzero in $\mathbf{c}$. Therefore, similar to (4), the code words that belong to the neighbor of $\mathbf{f}^{\alpha,\beta}$ will be selected to encode $\mathbf{f}^{\alpha,\beta}$. From the respect of manifold learning [23, 25], although the whole data of a manifold are nonlinear and Euclidian, in a local region, they can be considered as linear [23–25]. Therefore, benefiting from the locality constraint, the problems of VQ can be solved.

The object function in (5) can be solved with an analytical solution according to [32]:

$$\begin{aligned}
\rho &= \left(\psi + \lambda\ \mathrm{diag}\left(\mathbf{p}_j\right)^2\right)^{-1}\mathbf{1},\\
\psi &= \left(\mathbf{f}^{\alpha,\beta}\mathbf{1}^T - \mathbf{D}\right)^T\left(\mathbf{f}^{\alpha,\beta}\mathbf{1}^T - \mathbf{D}\right),\\
\mathbf{c} &= \frac{\rho}{\mathbf{1}^T\rho}.
\end{aligned} \tag{6}$$

Similarly, the problems of K-means dictionary learning can also be solved with locality constraint. According to [35], the object function of our dictionary learning method is formulated as follows:

$$\min_{\mathbf{D}^{\alpha,\beta},\mathbf{C}} \left\|\mathbf{F}^{\alpha,\beta} - \mathbf{D}^{\alpha,\beta}\mathbf{C}\right\|_2 + \lambda\sum_{i=1}^{n}\|\mathbf{p}_i\odot\mathbf{c}_i\|_2, \quad \text{st: } 1^T\mathbf{c}_i = 1 \quad \forall i = 1,\ldots,n, \tag{7}$$

where $\mathbf{F}^{\alpha,\beta} = \{\mathbf{f}_1^{\alpha,\beta},\ldots,\mathbf{f}_n^{\alpha,\beta}\}$, n is the number of input local features, $\mathbf{c}_i \in R^M$ is the ith column of $\mathbf{C}$, and $\mathbf{p}_i \in R^M$ is the locality adaptor whose jth element is given by $\mathbf{p}_{ij} = \|\mathbf{f}_i^{\alpha,\beta} - \mathbf{d}_j^{\alpha,\beta}\|_2$. Equation (7) can be effectively solved with the Locality-Sensitive Dictionary Learning (LSDL) in [32].

2.3. Modeling the Multiscale Spatiotemporal Relationship of Local Features. Due to the different styles of human action, it is difficult to model the spatiotemporal relationship of local features in a single space-time scale. The actions with different styles appear in different motion range (spatial scale is different) and speed (temporal scale is different). Therefore, it is necessary to capture their multiscale spatiotemporal relationship in feature coding. In implementation, instead of building spatial or temporal pyramid structures, we use position weighting factors α and β to control the spatial and time scales, respectively. According to (1), a large (small) α or β intends to select the code words from a small (large) spatial or temporal neighbor. Thus we can adjust α or β to obtain the multiscale feature descriptor $\mathbf{f}^{\mathrm{ms}}$:

$$\mathbf{f}^{\mathrm{ms}} = \left\{\mathbf{f}^{\alpha(1),\beta(1)},\ldots,\mathbf{f}^{\alpha(i),\beta(j)},\ldots\right\}, \tag{8}$$

where $i \in [1, n_s]$, $j \in [1, n_t]$, and n_s and n_t are the numbers of spatial and time scales, respectively. For example, we set $\alpha = \{4, 3, 2, 1\}$, $\beta = \{4, 3, 2, 1\}$.

Then the code is given as

$$\mathbf{c}^{\text{ms}} = \left[\left(\mathbf{c}^{\alpha(1),\beta(1)}\right)^T, \ldots, \left(\mathbf{c}^{\alpha(i),\beta(j)}\right)^T, \ldots\right]^T,$$
$$\mathbf{c}^{\alpha(i),\beta(j)} = \arg\min_{\mathbf{c}} \left\|\mathbf{f}^{\alpha(i),\beta(j)} - \mathbf{D}^{\alpha(i),\beta(j)}\mathbf{c}\right\|_2 + \lambda\|\mathbf{p} \odot \mathbf{c}\|_2, \tag{9}$$

where $\mathbf{D}^{\alpha(i),\beta(j)}$ is the dictionary learnt by LSDL with the features descriptors at ith spatial and jth temporal scales.

3. Human Action Recognition with MLSC and LGSR

3.1. Framework. The spatiotemporal positions (x, y, t) of local features play a key role in our method. Intuitively, we can construct spatial coordinate system with human ROI and build time coordinate system with a complete action cycle. Profiting from the existing methods of extracting human ROI from videos, the space coordinate system is easy to be set up. Because it is difficult to estimate action cycles in videos, the time coordinate system is difficult to establish. Fortunately, the feature spatiotemporal relationships can be locally modeled by sub-STV. We propose a novel framework without estimating action cycles as follows: (1) it densely samples several sub-STV from one video; (2) it carries out MLSC in each sub-STV to obtain a sub-STV descriptor; and (3) it classifies action upon these sub-STV descriptors with LGSR.

The proposed framework is illustrated in Figure 3. In the first step (Figure 3(a)), the local features such as space time interest points (STIP) and human ROI are extracted in each action video. In the second step (Figure 3(b)), it aligns ROI to build STV and extract sub-STV with multitime-scale densely sampling (detailed in Section 3.2). Then, many sub-STV in one video will be collected. In the third step (Figure 3(c)), it obtains a group of sub-STV descriptors with MLSC and max-pooling (detailed in Section 3.3). In the last step (Figure 2(d)), it utilizes these sub-STV descriptors to classify action with LGSR (detailed in Section 3.4).

3.2. Extract Sub-STV with Multi-Time-Scale Densely Sampling. The feature spatiotemporal relationships in STV are locally captured with the multi-time-scale densely sampling (MTDS) method. First, a set of time scales is defined for MTDS according to the possible action cycle lengths. Several sub-STVs are then densely sampled by a sliding window operation with one frame step (Figure 3(b)). Finally, several space time coordinate systems are established based on these sub-STVs. After that, a group of multiscale feature descriptors are obtained with (8) with setting $\beta = \{a, a, a, a\}$ (a is a constant), because the multi-time-scale information has been considered in MTDS.

The advantage of MTDS is that it is not necessary to consider whether the time coordinate system is aligned with the human action cycles in each sub-STV. Because if only the training samples are sufficient, then any tests sub-STV can always find a matching sub-STV in the training samples. Usually, this condition can be satisfied in real applications.

3.3. Describe Sub-STV with MLSC and Max-Pooling. Given a sub-STV, we take it as the space-time coordinate system (x, y, t) to generate a group of multiscale feature descriptors:

$$\mathbf{F}^{\text{ms}} = \left\{\mathbf{f}_1^{\text{ms}}, \ldots, \mathbf{f}_n^{\text{ms}}\right\}. \tag{10}$$

Then, we use MLSC to encode each feature and obtain multiscale codes:

$$\mathbf{C}^{\text{ms}} = \left[\mathbf{c}_1^{\text{ms}}, \ldots, \mathbf{c}_n^{\text{ms}}\right]. \tag{11}$$

After coding each feature, we use the max-pooling [29] to get the sub-STV descriptor:

$$\mathbf{s}(j) = \max\left\{\mathbf{c}_1^{\text{ms}}(j), \ldots, \mathbf{c}_n^{\text{ms}}(j)\right\}, \tag{12}$$

where $\mathbf{s}(j)$ is jth element in this sub-STV descriptor $\mathbf{s}$ and $\mathbf{c}_i^{\text{ms}}$ is the MLSC coefficient vector for ith feature.

3.4. LGSR-Based Action Videos Classification. To utilize the intrinsic group information from these sub-STV descriptors within one video for action classification, we adopt the locality-Constrained Group Sparse Representation (LGSR) to classify actions. LGSR was proposed in [39] for human gait recognition. It is an extended sparse representation-based classifier (SRC). The pioneering work of SRC was proposed in [33] and used to classify face images by minimizing the norm-regularized reconstruction error. There are three advantages of LGSR comparing with SRC: (1) SRC is designed for single image classification and cannot directly classify a group of samples, while LGSR is designed for sample group classification; (2) the locality constraint in LGSR is more reasonable than sparsity constraint in SRC, especially for representing manifold data [29, 32]; (3) LGSR is a block sparse constraint classifier. It is better than SRC in classification task when the used features are discriminative. The comparison experiment in Section 4.5 also proves that LGSR is more suitable than SRC for our task.

The object function of LGSR is defined as

$$\mathbf{A}^* = \arg\min_{\mathbf{A}} \left(\frac{1}{2}\|\mathbf{S} - \mathbf{D}\mathbf{A}\|_F^2 + \lambda\sum_{k=1}^{K}\left\|\mathbf{P}^k \odot \mathbf{A}^k\right\|_F\right), \tag{13}$$

where the first term represents the reconstruction error of the test action with respect to all the actions. The second term is the weighted mixed-norm-based regularization on the reconstruction coefficient $\mathbf{A}$. λ is the regularization parameter to balance these terms. $\mathbf{D}$ is the classification dictionary constructed by connecting K class-special dictionaries $[\mathbf{D}^1, \ldots, \mathbf{D}^K]$. Each class-special dictionary $\mathbf{D}^k$ is learnt with LSDL [32] from the sub-STV descriptors corresponding to the kth action. $\mathbf{S}$ is the group of sub-STV descriptors for one test action. $\mathbf{A}^k$ is one part of $\mathbf{A}$ and corresponds to $\mathbf{D}^k$. $\mathbf{P}^k$ is the distance matrix between $\mathbf{S}$ and $\mathbf{D}^k$. The ith

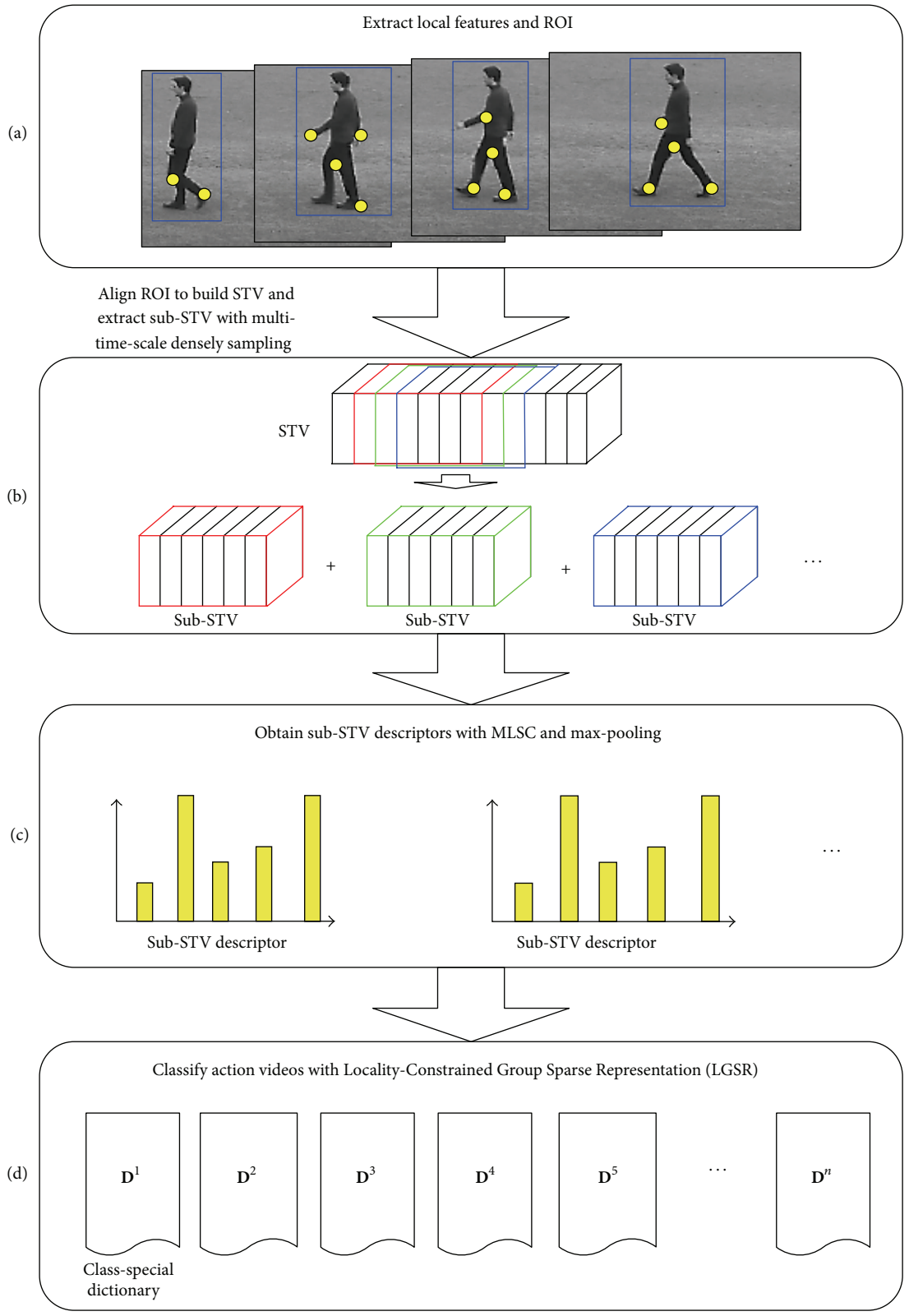


Figure 3: The flow chart of human action recognition with MLSC and LGSR.

and jth element in $\mathbf{P}^k$ is calculated as $\mathbf{P}_{ij}^k = \|\mathbf{S}_i - \mathbf{D}_j^k\|_2$. Since $\mathbf{A}^k$ values are independent of each other, we can separately update each $\mathbf{A}^k$ using its subgradient [36]. To solve (14), the active set-based subgradient descent algorithm in [38, 39] was employed.

Once the optimal reconstruction coefficient $\mathbf{A}$ is obtained, maximum weighted inverse reconstruction error (maxWIRE) criterion [38] is adopted for action classification. It is better than the original minimum reconstruction Error (minRE) criterion in [33].

WIRE is defined as

$$\text{WIRE}(k) = \frac{\left\|\mathbf{A}^{*k}\right\|_F}{\left\|\mathbf{S} - \mathbf{D}^k\mathbf{A}^{*k}\right\|_F}. \quad (14)$$

The action video label L is decided with the maximum WIRE:

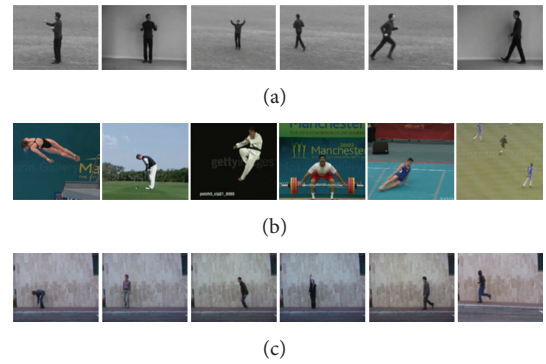
$$L^* = \arg\max_k \left(\text{WIRE}(k)\right). \quad (15)$$

4. Experiment and Analysis

In this section, the effectiveness of our MLSC is evaluated on three public datasets: Weizmann, KTH, and UCF sports. The leave-one-out cross-validation (LOOCV) is used to evaluate the performance of our algorithm. It employs actions from some people as the test samples, meanwhile leaving the remaining actions from other people as the training samples.

4.1. Experiment Setup. In all experiments, cuboids [16] is adopted to extract spatiotemporal local features, and HOG/HOF [20] is adopted to describe these features. According to [16], the standard space scale value 3 and time scale value 2 are used in cuboids detector. To extract ROI, we label a bounding box for the actor that locates at the first frame in each split and then track actor to obtain the ROI for KTH dataset; the annotation bounding boxes are used for extracting ROI for UCF sports dataset, and a rotation operation is used to obtain oriented ROI; and the background subtraction results are used for the Weizmann dataset. To capture multiscale temporal relationship of local features, the lengths of sub-STV are set as 5, 10, 25, and 50 frames. To capture multiscale spatial relationship of local features, four spatial scales are used. The parameters are set as $\alpha = \{4, 3, 2, 1\}$ and $\beta = \{1, 1, 1, 1\}$. In MLSC, the dictionary size is set to 1000. Since there are 4 spatial scales, the dimension of a sub-STV descriptor is 4000. In order to guarantee that the class-special dictionaries in LGSR are over complete, PCA is adopted to reduce the dimension of the sub-STV descriptor to 400. In LGSR, the size of each class-special dictionary is set to 800. The other parameters in our methods (for example σ and λ) and the parameters of other methods are evaluated by 5-fold cross-validation.

4.2. Datasets. The KTH dataset contains six types of human action examples (i.e., boxing, hand clapping, hand waving, jogging, running, and walking) featuring 25 different subjects. Each action is performed in four scenarios: indoors, outdoors, outdoors with scale variation, and outdoors with different clothes. Overall it has 599 low-resolution video clips (160×120 pixels), for one of the videos is missing. Examples of this datasets can be seen in Figure 4(a).



(a)

(b)

(c)

Figure 4: Examples from the public datasets: (a) KTH dataset; (b) UCF sports dataset; (c) the Weizmann dataset.

UCF sports dataset includes a set of 150 videos, 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 datasets can be seen in Figure 4(b).

Weizmann: this 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-hands-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 Figure 4(c).

4.3. Comparing with BoF. BoF-based action representation methods together with existing local feature coding methods VQ, SC, LLC [29], and our MLSC are further compared under the same condition that K-nearest Neighbor (KNN) classifier is used in classification stage. In KNN, k is set to 5. Keeping the same dictionary size with MTSC, K-means clustering is used to learn dictionary for VQ and LLC, and the software in [40] is adopted for SC. In LLC, the locality constraint parameter k is set to 5. In our methods, a group of sub-STVs descriptors are extracted from one test video and classified with KNN. After that the vote scores of these sub-STVs are used to label this test video. In feature pooling phase, avg-pooling is used for VQ, while max-pooling is adopted for SC, LLC, and MLSC.

In addition, to evaluate these factors which are used to improve BoF from feature coding in Section 2, another comparison is carried out. First, considering the feature position constraint in Section 2.1, the coding method in (2) is considered as the basic spatiotemporal coding (StC).

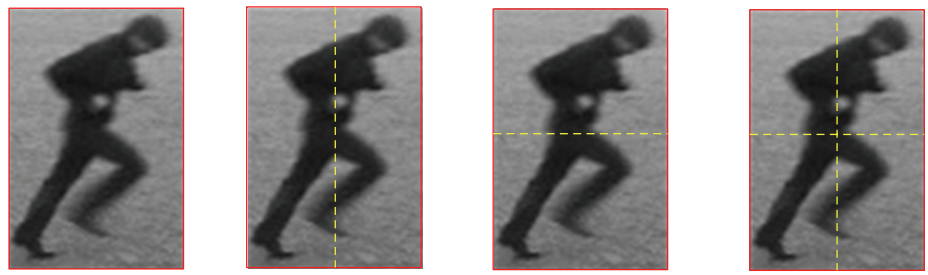


Figure 5: Confusion matrix for KTH dataset of our methods LGSR and MLSC.

Table 1: The average accuracy of BoF together with VQ, SC, LLC and our methods on KTH, Weizmann, UCF sports datasets (unit: %).

Methods	KTH	UCF sports	Weizmann
KNN + BoF + VQ	86.5	75.5	88.1
KNN + BoF + SC	88.1	78.1	89.6
KNN + BoF + LLC	89.5	79.7	90.8
KNN + StC ($\alpha = 4$)	89.1	78.6	**91.7**
KNN + StC ($\alpha = 3$)	**90.2**	79.6	90.1
KNN + StC ($\alpha = 2$)	88.4	**81.4**	89.5
KNN + StC ($\alpha = 1$)	87.9	79.8	88.7
KNN + LSC ($\alpha = 4$)	90.5	80.5	92.8
KNN + LSC ($\alpha = 3$)	**91.8**	82.1	**92.1**
KNN + LSC ($\alpha = 2$)	91.1	**83.5**	91.8
KNN + LSC ($\alpha = 1$)	90.9	82.5	91.2
KNN + MLSC	**94.4**	**85.6**	**94.9**

Second, considering the locality constraint in Section 2.2, the coding method in (5) is considered as the locality-constrained spatiotemporal coding (LSC). In this comparison experiment, the dictionary size is still set to 1000. K-means clustering is adopted to learn dictionary for StC. LSDL is adopted to learn dictionary for LSC. Avg-pooling is used for StC, and max-pooling is adopted for LSC. The parameter for KNN is set to 5. The spatial and temporal control factors are set as $\alpha = \{4, 3, 2, 1\}$ and $\beta = \{1, 1, 1, 1\}$.

The results of comparison are shown in Table 1. There are the average recognition accuracies on three datasets. The basic spatiotemporal coding method (StC) achieves better performance than VQ, SC, and LLC. This demonstrates that considering the spatiotemporal relationship is important for human action recognition in video. The locality-constrained spatiotemporal coding is better than StC. In addition, the locality constraint is useful to handle the manifold of local features. Finally, benefiting from modeling the multiscale spatiotemporal relationship of local features, MLSC achieves the highest average recognition accuracy on each dataset.

4.4. MLSC versus SPM. The spatial pyramid match (SPM) model has been adopted to capture the spatial relationships of local spatiotemporal features [23]. Here, a 4-level SPM (detailed in Figure 5) is used for evaluation. MLSC and SPM, LLC, and Max-pooling are employed to describe sub-STV, respectively. KNN classifier is also used to classify sub-STVs.

Table 2: Comparison results between MLSC with SPM (UNIT: %).

Methods	KTH	UCF sports	Weizmann
KNN + SPM	91.7	82.8	93.5
KNN + MLSC	94.4	85.6	96.5

Table 3: Comparison results between LGSR and SRC (UNIT: %).

Methods	KTH	UCF sports	Weizmann
SRC + MLSC	96.5	92.1	100
LGSR + MLSC	**98.5**	**93.5**	**100**

The vote score-based classifier (similar with Section 4.3) is adopted to label a test video. Table 2 shows the average recognition accuracies. MLSC achieves better performance than SPM on all datasets. Different from SPM [37] which only considers the spatial relationship of local features, MLSC simultaneously considers the spatial and temporal relationships. In addition, comparing with the fixed grids used in SPM, MLSC is a more flexible representation.

4.5. LGSR versus SRC. To prove the ability of using LGSR for action classification, the standard SRC [29] is also evaluated. The object function of SRC is defined as

$$\mathbf{A}_i^* = \arg\min_{\mathbf{A}_i} \|\mathbf{S}_i - \mathbf{D}\mathbf{A}_i\|_2 + \omega\|\mathbf{A}_i\|_1, \quad (16)$$

where $\mathbf{S}_i$ is the ith sub-STV descriptor in $\mathbf{S}$, $\mathbf{A}_i$ is its corresponding code. Similar to LGSR, the maxWIRE criterion is also used in SRC. As mentioned in Section 3.4, there are three advantages of LGSR comparing with SRC. 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 results with average accuracy (Table 3) show that LGSR achieves better performance than SRC on KTH and UCF sports datasets. It is worth to note that Guha and Ward [25] suggested that sparse constraint is more important than block sparse constraint in human action recognition based on local spatiotemporal features. Comparing with local spatiotemporal features, the obtained sub-STV descriptors with MLSC are less shared with other actions. Hence, it is better to utilize block sparse constraint than sparse constraint for action classification together with MLSC.

Table 4: Comparison results with other methods (unit: %).

Methods	Year	Experiment setting	KTH	UCF sports	Weizmann
Zhu et al. [24]	2010	Split	94.92	84.33	—
Wu et al. [34]	2011	LOOCV	94.5	91.3	—
Escobar and Kornprobst [21]	2012	Split	90.56		99.26
Guha and Ward [25]	2012	LOOCV	—	91.1	98.9
Bregonzio et al. [36]	2012	LOOCV	94.33	—	96.66
Zhang et al. [35]	2012	LOOCV	95.06	87.33	—
Saghafi and Rajan [12]	2012	LOOCV	92.6	—	100
Deng et al. [15]	2013	LOOCV	96.91	88.4	100
LGSR + MLSC		LOOCV	98.5	93.5	100

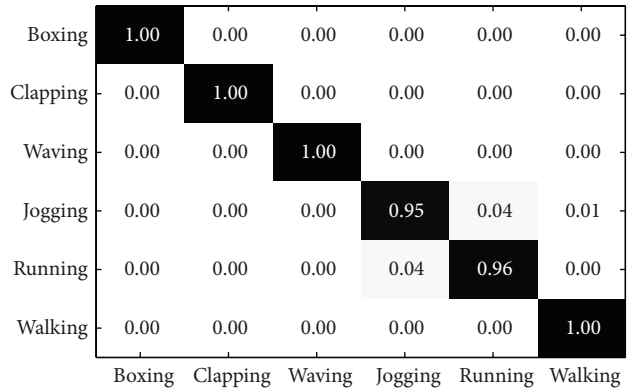


Figure 6: Confusion matrix for UCF sports dataset of our methods LGSR and MLSC.

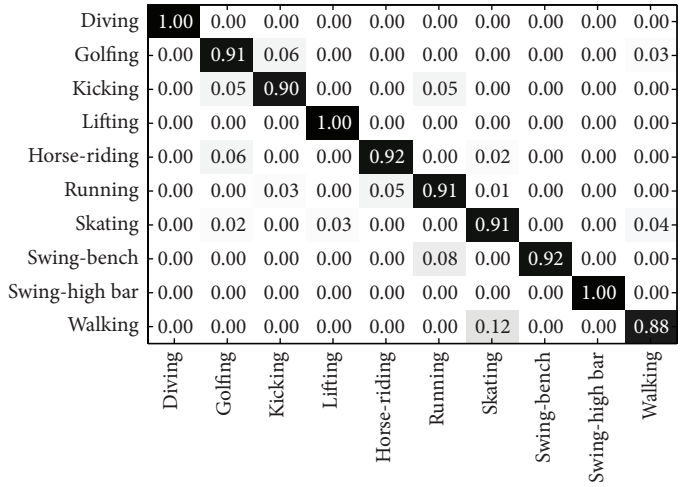


Figure 7: 4-level Spatial Pyramid Match (SPM).

4.6. Comparing with Other Methods. The present and some previously published results are compared in Table 4. The experiment setting "split" means that it randomly selects some people for training and leaves others for testing. The competing methods include local representation-based methods [15, 21, 24, 25, 34–36], global representation-based methods [12]. In detail, SC was used for feature coding together with BoF in [24], and a new local feature detector was proposed for human action recognition in [21], local feature distribution information was used in [36], spatiotemporal context feature was employed in [34], a spatiotemporal context constraint coding method was utilized in [35], sparse representation-based classification methods was applied in [25], and the global representation method was adopted in [12]. It demonstrates that our method achieves better performance than the competing methods. The confusion matrices for KTH and UCF sports datasets of our method LGSR+MLSC are shown in Figures 6 and 7, respectively.

First, benefiting from involving spatiotemporal locations into code words learning and feature coding, our method performs better than these methods [15, 24, 25] which only use the feature appearance information to represent human action. Second, comparing with the feature distribution feature [36] and spatiotemporal context methods [34, 35], our method is a *fine* and *whole* method. For example, as illustrated in Figure 2, each local feature has four types of information (*where, when, who,* and *how*) in STV. First, the coordinate (x, y) and (t) indicates *where* and *when* the body part appears respectively. Second, the feature appearance (described as HOG) indicates *who* (which human body part). Third, the motion information (described as HOF) indicates *how* this body part moves. In MLSC, all of these pieces information (*where, when, who,* and *how*) have been modeled with involving (x, y), (t), HOG, and HOF into feature coding. However, these methods [34–36] ignore some one of these information (*where, when, who,* and *how*) in action representation processing. Hence, it implies that our method records the whole elements (*where, when, who,* and *how*) of local features for human action recognition.

5. Conclusion

In this paper, in order to capture the spatiotemporal relationships of local spatiotemporal features for human action recognition task, we encode feature appearance and spatiotemporal position information together with locality constraint. The experimental results show that (1) feature spatiotemporal position is effective for action recognition (2) involving feature position into feature coding is a beneficial alternative way for this task. In particular, it is a better approach than feature distribution [36], spatiotemporal context [35], and SPM-based methods [23], when using a multiscale version.

The major limitation is that human ROI is required to construct STV. Although, dissimilar to the global representation methods which need the fine foreground segmentation, the coarse human box is enough in our method. It is valuable to explore new methods to capture the spatiotemporal location of local features in our future work.

Acknowledgments

This work was supported by the National Natural Science Foundation (NSFC) of China under projects no. 61175015, and no. 61175006.

References

[1] Y. Wang, Y. Qi, and Y. Li, "Memory-based multiagent coevolution modeling for robust moving object tracking," *The Scientific World Journal*, vol. 2013, Article ID 793013, 13 pages, 2013.

[2] T. H. Thi, L. Cheng, J. Zhang, L. Wang, and S. Satoh, "Structured learning of local features for human action classification and localization," *Image and Vision Computing*, vol. 30, no. 1, pp. 1–14, 2012.

[3] J. Baek and B.-J. Yun, "A sequence-action recognition applying state machine for user interface," *IEEE Transactions on Consumer Electronics*, vol. 54, no. 2, pp. 719–726, 2008.

[4] G. Zhu, M. Yang, K. Yu, W. Xu, and Y. Gong, "Detecting video events based on action recognition in complex scenes using spatio-temporal descriptor," in *Proceedings of the 17th ACM International Conference on Multimedia (MM '09)*, pp. 165–174, October 2009.

[5] D. Weinland, R. Ronfard, and E. Boyer, "A survey of vision-based methods for action representation, segmentation and recognition," *Computer Vision and Image Understanding*, vol. 115, no. 2, pp. 224–241, 2011.

[6] J. K. Aggarwal and M. S. Ryoo, "Human activity analysis: a review," *ACM Computing Surveys*, vol. 43, no. 3, article 16, 2011.

[7] X. Wu and 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.

[8] A. A. Chaaraoui and P. Climent-Pérez, "Silhouette-based Human action recognition using sequences of key poses," *Pattern Recognition Letters*, vol. 34, no. 15, pp. 1799–1807, 2013.

[9] K. N. Tran, I. A. Kakadiaris, and S. K. Shah, "Modeling motion of body parts for action recognition," in *Proceedings of the British Machine Vision Conference (BMVC '11)*, pp. 1–12, 2011.

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

[11] S. A. Rahman, M. K. H. Leung, and 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.

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

[13] 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.

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

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

[16] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse spatio-temporal features," in *Proceedings of the 2nd Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance (VS-PETS '05)*, pp. 65–72, October 2005.

[17] A. Klaser, M. Marszalek, and C. Schmid, "A spatio-temporal descriptor based on 3D-gradients," Proceedings of the British Machine Vision Conference (BMVC '08), 2008.

[18] P. Scovanner, S. Ali, and M. Shah, "A 3-dimensional sift descriptor and its application to action recognition," in *Proceedings of the 15th ACM International Conference on Multimedia (MM '07)*, pp. 357–360, September 2007.

[19] G. Willems, T. Tuytelaars, and L. Van Gool, "An efficient dense and scale-invariant spatio-temporal interest point detector," in *Proceedings of the Europen Conference on Computer Vision (ECCV '08)*, pp. 650–663, 2008.

[20] I. Laptev, M. Marszałek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *Proceedings of the 26th IEEE Conference on Computer Vision and Pattern Recognition (CVPR '08)*, pp. 1–8, June 2008.

[21] M.-J. Escobar and P. Kornprobst, "Action recognition via bio-inspired features: the richness of center-surround interaction," *Computer Vision and Image Understanding*, vol. 116, no. 5, pp. 593–605, 2012.

[22] X. Zhu, Z. Yang, and J. Tsien, "Statistics of natural action structures and human action recognition," *Journal of Vision*, vol. 12, no. 9, pp. 834–834, 2012.

[23] B. Chakraborty, M. B. Holte, T. B. Moeslund, and J. Gonzàlez, "Selective spatio-temporal interest points," *Computer Vision and Image Understanding*, vol. 116, no. 3, pp. 396–410, 2012.

[24] Y. Zhu, X. Zhao, Y. Fu et al., "Sparse coding on local spatial-temporal volumes for human action recognition," in *Proceedings of the Computer Vision (ACCV '10)*, pp. 660–671, Springer, Berlin, Germany, 2010.

[25] T. Guha and 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.

[26] J. C. Van Gemert, C. J. Veenman, A. W. M. Smeulders, and J.-M. Geusebroek, "Visual word ambiguity," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 7, pp. 1271–1283, 2010.

[27] B. A. Olshausen and 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.

[28] K. Yu, T. Zhang, and Y. Gong, "Nonlinear learning using local coordinate coding," in *Proceedings of the 23rd Annual Conference on Neural Information Processing Systems (NIPS '09)*, pp. 2223–2231, December 2009.

[29] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong, "Locality-constrained linear coding for image classification," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '10)*, pp. 3360–3367, June 2010.

[30] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 5500, pp. 2323–2326, 2000.

[31] J. Wang, "Locally linear embedding," in *Geometric Structure of High-Dimensional Data and Dimensionality Reduction*, pp. 203–220, Springer, Berlin, Germany, 2011.

[32] C. P. Wei, Y. W. Chao, and Y. R. Yeh, "Locality-sensitive dictionary learning for sparse representation based classification," *Pattern Recognition*, vol. 46, no. 5, pp. 1277–1287, 2013.

[33] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 2, pp. 210–227, 2009.

[34] X. Wu, D. Xu, L. Duan, and J. Luo, "Action recognition using context and appearance distribution features," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR '11)*, pp. 489–496, June 2011.

[35] Z. Zhang, C. Wang, B. Xiao et al., "Action recognition using context-constrained linear coding," *Signal Processing Letters*, vol. 19, no. 7, pp. 439–442, 2012.

[36] M. Bregonzio, T. Xiang, and S. Gong, "Fusing appearance and distribution information of interest points for action recognition," *Pattern Recognition*, vol. 45, no. 3, pp. 1220–1234, 2012.

[37] S. Lazebnik, C. Schmid, and 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 (CVPR '06)*, pp. 2169–2178, June 2006.

[38] D. Xu, Y. Huang, Z. Zeng, and 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.

[39] M. Liu, S. Yan, Y. Fu, and T. S. Huang, "Flexible X-Y patches for face recognition," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '08)*, pp. 2113–2116, April 2008.

[40] 2013, http://spams-devel.gforge.inria.fr/.