International Journal of Signal Processing, Image Processing and Pattern Recognition
Vol. 8, No. 1 (2015), pp. 241-254
http://dx.doi.org/10.14257/ijsip.2015.8.1.21

Action Recognition Based on Multi-scale Oriented Neighborhood Features

Jiangfeng Yang[1], Zheng Ma[1] and Mei Xie[2]

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

Abstract

The spatio-temporal (ST) position information between local features plays an important role in action recognition task. To use the information, neighborhood-based features are built for describing local ST information around ST interest points. However, traditional methods of constructing neighborhood, such as sub-ST volumetric method and nearest-neighbor-based neighborhood method, ignore the orientation information of neighborhood. To make the neighborhood-based features more discriminative, we construct a novel, oriented neighborhood by imposing weights on the distance components. Specifically, in our scheme, firstly, local features are produced, and encoded by locality-constrained linear coding (LLC). Then, oriented neighborhoods are constructed by imposing weights on the distance components between features, and obtain single-scale oriented neighborhood features (SONFs). Next, multi-scale oriented neighborhood features (MONFs) are formed by concatenating SONFs. As a result, action video sequences are represented as a collection of MONFs. Finally, locality-constrained group sparse representation (LGSR) is used as classifier upon MONFs. Experimental results on the KTH and UCF Sports datasets show that our method achieves better performance than the competing local ST feature-based human action recognition methods.

Keywords: *action recognition, action representation, oriented neighborhood feature*

1. Introduction

Human action recognition in video has been widely studied over the past decade due to its widespread application prospects in many fields 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, and has also been an increasingly active field of computer vision and pattern recognition.

Numbers of human action recognition techniques have been proposed, and several reviews are devoted to this topic [5, 6]. There are two parts in this field [5]: action representation and classification. Action representation is the procedure of extracting features from videos and obtaining the human behavior representation by encoding the features. Then learning an action model from the final behavior representations and recognizing query behaviors with the learnt model. In general, there are two kinds of action representation: global representations [7, 8] and local representations [9-19]. Commonly, global representations are derived from silhouettes or body sketch. They require good background subtraction or body part tracking. So, they are sensitive to noise, variations in viewpoint, and partial occlusion. Local representations are based on the local ST features plus bag-of-features (BoF) model. Due to no requiring background subtraction or body part tracking, they are robust to viewpoint/appearance changes, noise, and partial occlusions [5].

ISSN: 2005-4254 IJSIP

Human action recognition usually consists of three stages: extracting local features from video sequences, learning action representation vectors via these local features, and classifying query action videos with a classifier upon the video representation vectors [5]. To obtain video representation vectors, several feature coding and pooling methods are developed. Many authors used K-means and vector quantization (VQ) for feature coding, as well as the average-pooling [10] to group these feature codes to generate the video representation vector. For reducing quantization error caused by K-means and VQ, soft vector quantization (SVQ) [20] and sparse coding (SC) [21] are proposed to encode local features for action recognition tasks [18]. However, the local features usually reside on nonlinear manifolds [9, 22, 23]. None of SVQ and 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 [24, 25]. Because SVQ uses all bases to encode each feature and generates dense codes, it cannot correctly represent the nonlinear manifold structure in a global way. Due to the overcomplete dictionary, SC tends to choose the codewords which are distant to the input features [23]. Therefore, it cannot correctly represent manifold data. For handling both of quantization error and loss manifold structure in feature coding, Yu *et al.* [22] provided a Local Coordinate Coding (LCC) to encode local features with locality-constrained, Wang *et al.* [23] introduced an improved version of LCC called Locality-constrained Linear Coding (LLC) to reduce computational cost, and Wei *et al.* [26] proposed a local sensitive dictionary learning method for image classification.

After encoding local features, action video sequence is represented as BoF model. To overcome the major limitation of BoF model that ignore the ST relationship between local features, many methods [27-33] are developed during the past years, and fall into two categories: temporal sequential approaches and ST volumetric approaches. For temporal sequential approaches, actions in digital videos are regarded as sequences of states (e.g. poses). Inspired by speech recognition, sequential probabilistic models [27-30] use dynamic probabilistic graphical models (e.g., Hidden Markov Models (HMM) [31]) to learn temporal transitions between hidden states. A limitation of temporal sequential approaches, however, is that they require a large amount of training examples in order to model all events might occur, and too many parameters are involved when complicated actions are modeled. In contrast to temporal sequential approaches, ST volumetric methods [32, 33] view actions as 3D (X-Y-T) objects in a spatio-temporal volume (STV), thus treating space and time in a unified manner. In order to utilize the ST position relationship between local features, STV usually is partitioned into sub-STVs, each of which surrounds a local feature. All local features within same sub-STV are merged into a feature by max-pooling or average-pooling. A limitation of ST volumetric methods is that one needs to decide the temporal boundary of sub-STV.

In the paper, we proposed a Multi-scale Oriented Neighborhood Feature (MONF) for making use of the ST position information between features. The MONF is based on the neighborhood formed by the nearest neighbors. Unlike the traditional nearest neighbor methods that are isotropic with respect to 3D space, our nearest neighbor algorithm is anisotropic, direction-selective (see Figure 3).

In our system, above all, ST interest points (STIPs) are detected from action video; corresponding local features are produced, and encoded by LLC algorithm. Then, several single-scale oriented neighborhood features (SONFs) around each STIP are computed and formed MONFs. Thus, actions are represented as a group of MONFs. To utilize the intrinsic group information from MONFs within each video, we treat each test/training video as one group of MONFs and employ the Locality-constrained Group Sparse Representation (LGSR) [34]-based classifier as action classification.

The experiments on the KTH, UCF Sports datasets show that our system outperforms the methods in [18, 19, 44-46] and classical local ST feature-based methods.

This paper has two contributions as follows:

- To make use of ST position information between local features, we proposed a MONF features that are orientation-selection and robust to ST scale changes.
- Compared to traditional sub-STV volumetric approach that is required to determine the temporal scale, the MONF adapts to the ST scale of action video.
- In contrast to traditional neighborhood approach that equally treats each position element, our obtained neighborhoods are oriented by assigning different weights to position elements. Therefore, the resulting MONFs are more discriminative.

The rest of this paper is organized as follows. Extracting STIPs from action video is shown in Section 2. Encoding local features by LLC is provided in Section 3. Building MONFs by the proposed algorithm is offered in Section 4. Then, employing LGSR algorithm as action classifier is presented in Section 5. Experimental results and analysis are shown in Section 6. Finally, conclusion is drawn in Section 7.

2. Detecting Spatio-temporal Interest Points (STIPs)

The inputs to our recognition system are the STIP positions and their associated local descriptors. We utilize Dollar detector [10] to extract STIPs from video sequences. Dollar detector generally produces a high number of STIPs that is important for learning ST neighborhood feature. Action video sequence V is described by a set of position-descriptor tuples:

$$V = \{(x_i, y_i, t_i), \mathbf{f}_i : i = 1, \cdots, n_v\}, \tag{1}$$

where (x_i, y_i, t_i) records the ST position; $\mathbf{f}_i$ denotes a local feature vector of 3D support region around STIP i; n_v denotes the video's total number of STIPs.

3. Encoding Local Features by Locality-constrained Linear Coding (LLC)

Let $\mathbf{F} = \{\mathbf{f}_i \in R^D, i \in 1, \cdots, N\}$ be N local features with D-dimension. Given a codebook with M bases $\mathbf{B} = [\mathbf{b}_1, \mathbf{b}_2, \cdots, \mathbf{b}_M] \in R^{D \times M}$, for simplicity, codebook $\mathbf{B}$ is generated by k-means clustering over training samples with Euclidean distance as metric. Feature $\mathbf{f}_i$ is converted into a M-dimensional code denoted as $\mathbf{c}_i \in R^M$ by feature coding schemes, such as vector quantization (VQ), soft vector quantization (SVQ), sparse coding (SC), and LLC. Table 1 shows the performance comparison of coding schemes.

3.1. VQ, SVQ and SC

In VQ, its coding strategy assigns just a single base to a local feature; each local feature is assigned to the nearest visual codeword:

$$c_{i,j} = \begin{cases} 1, & \text{if } j = \arg\min_j \left\| \mathbf{f}_i - \mathbf{b}_j \right\|_2^2, \\ 0, & \text{otherwise,} \end{cases} \tag{2}$$

where code vector $\mathbf{c}_i = [c_{i,1}, c_{i,2}, \cdots, c_{1,M}]^T$. This coding is simple but, as reported in [35], suffers from the reconstruction error due to the reason that it only assigns a single code word to the descriptor.

To alleviate the quantization error of VQ, Gemert *et al.* [37] proposed SVQ on which a feature is encoded across several codewords instead of using one:

$$c_{i,j} = \frac{\exp(-\beta \left\| \mathbf{f}_i - \mathbf{b}_j \right\|_2)}{\sum_{m=1}^{M} \exp(-\beta \left\| \mathbf{f}_i - \mathbf{b}_m \right\|_2)} \tag{3}$$

where β is a parameter controlling how widely the assignment distributes the weight across all the code words. A small β gives a broad distribution, while a large β gives a peaked distribution, more closely approximating hard assignment. And this is further improved by Liu *et al.* [38], who used localized soft assignment (LSVQ). Their difference is that SVQ encodes the features with all the codewords, while LSVQ confines the soft assignment to a local neighborhood around the feature being coded.

Another way to reduce the quantization loss of VQ is SC [21] that encodes a local feature by using the coefficients of a linear combination of the codewords in $\mathbf{B}$, with a sparsity regularity term ℓ_1-norm:

$$\mathbf{c}_i = \arg\min_{\mathbf{c}} (\left\| \mathbf{f}_i - \mathbf{B}\mathbf{c} \right\|_2^2 + \lambda \left\| \mathbf{c} \right\|_1), \quad \lambda \in R, \tag{4}$$

where the first term represents the reconstruction error of $\mathbf{f}_i$ with respect to codebook $\mathbf{B}$. The second term denotes a sparse constraint regularization on code $\mathbf{c}$, and λ is a regularization factor to balance these terms. Although compared to VQ, SC significantly reduces the quantization loss, its computation complex is high, and not guarantee that same input features produce same encoding result.

Table 1. Comparison between Coding Schemes

Coding schemes	Quantization error	Nonconsistent coding	Computational cost
VQ	High	Low	Low
SVQ	Low	Low	Median
SC	Low	High	High
LLC	Low	Low	Low

3.2. Locality-constrained Linear Coding (LLC)

In contrast to the previous coding schemes, LLC coding algorithm [23] has attracted much attention due to its impressive properties:

- Better reconstruction. In VQ (Figure 1.a), each descriptor is represented by a single basis in the codebook. Due to the large quantization errors the VQ code for similar descriptors might be very different. Besides, the VQ process ignores the relationships between different bases. Hence non-linear kernel projection is required to make up such information loss. On the other side, as shown in (Figure 1.c) in LLC, each descriptor is more accurately represented by multiple bases, and LLC code captures the correlations between similar descriptors by sharing bases.
- Local smooth sparsity. Similar to LLC, SC also achieves less reconstruction error by using multiple bases. Nevertheless, the regularization term of ℓ_1 norm in SC is not smooth. As (shown in Figure 1.b), due to the over-completeness of the codebook, the SC process might select quite different bases for similar patches to favor sparsity, thus losing correlations between codes. On the other side, the explicit locality adaptor in LLC ensures that similar patches will have similar codes.
- Analytical solution. Solving SC usually requires computationally demanding optimization procedures. Unlike SC, the solution of LLC can be derived analytically such that LLC can be performed very fast in practice.

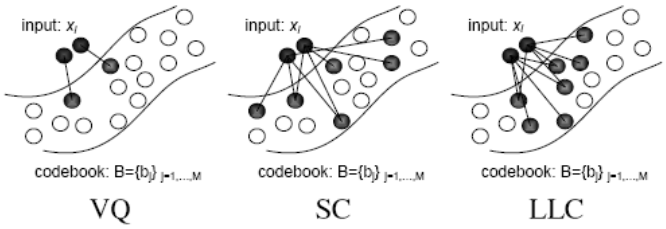


Figure 1. Comparison between VQ, SC and LLC. The Selected bases for Representation are Highlighted in Black

LLC can be formulated by

$$\mathbf{c}_i = \arg\min(\left\|\mathbf{f}_i - \mathbf{Bc}\right\|_2^2 + \lambda \left\|\mathbf{d} \odot \mathbf{c}\right\|_2^2), \quad \text{s.t.}\, \mathbf{1}^T \mathbf{c} = 1, \tag{5}$$

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

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

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

In the paper, to reduce quantization error and keep the consistent coding, both LLC algorithm and a codebook with M bases $\mathbf{B} = [\mathbf{b}_1, \mathbf{b}_2, \cdots, \mathbf{b}_M] \in R^{D \times M}$ are employed to encode local feature $\mathbf{F} = \{\mathbf{f}_i \in R^D, i \in 1, \cdots, N\}$, and obtain corresponding reconstruction coefficient vectors $\mathbf{C} = \{\mathbf{c}_i \in R^M, i \in 1, \cdots, N\}$.

4. Building Multi-scale Oriented Neighborhood Features (MONFs)

After encoding local features with LLC method, action video sequence V is represented as a group of reconstruction coefficient vectors $\mathbf{C} = \{\mathbf{c}_i \in R^M, i \in 1, \cdots, N\}$, and equation (1) is rewritten as

$$V = \{\mathbf{g}_i\} = \{(x_i, y_i, t_i), \mathbf{f}_i, \mathbf{c}_i : i = 1, \cdots, n_v\}, \tag{7}$$

where (x_i, y_i, t_i) records the ST position; $\mathbf{f}_i$ denotes a local feature at position (x_i, y_i, t_i); $\mathbf{c}_i$ is the code related to $\mathbf{f}_i$; n_v denotes the video's total number of STIPs.

Due to the different styles of human action, it is difficult to model the ST relationship of local features in a single space-time scale. The actions with different styles appear in different motion range (different spatial scale) and speed (different temporal scale). To utilize the ST position relationship between local features, many methods were developed, such as the most popular sub-STV method and local neighborhood method.

In sub-STV, an action is viewed as a STV, and partitioned into multi-temporal-scale segments called sub-STVs, each of which surrounds a local feature (see Figure 2(a)-(b)). A major limitation of this method is that temporal length of sub-STV is determined experientially. In contrast to sub-STV method, the idea of local neighborhood method is that the ST relation around a feature can be described by its nearest neighbors. Traditional method of choosing nearest neighbors directly calculate the distance between two features, that is,

given two feature positions $(x_1, y_1, t_1), (x_2, y_2, t_2)$, their distance is $\|(x_1 - x_2, y_1 - y_2, t_1 - t_2)\|_2$ and this method suffers from isotropic, or non-direction-selection (see Figure 2(c)-(d)).

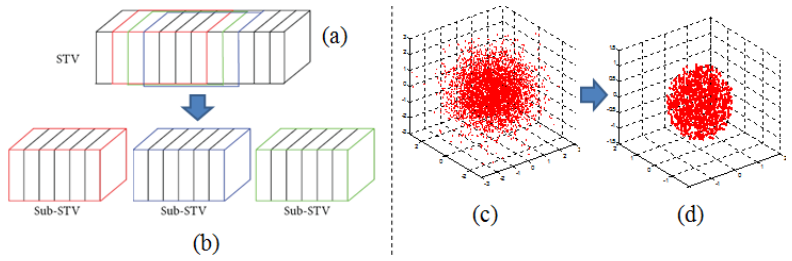


Figure 2. The Sub-STV Method and Traditional Neighborhood Method. In (a)-(b), a STV is Divided into Three sub-STVs with a Temporal Scale. In (c), 10000 Points with Positions $\{(x_i, y_i, t_i) : i = 1, \cdots, 10000\}$ are Generated in 3D Space, and $x_i, y_i, t_i \square N(0,1)$. In (d), 1000 Nearest Points to the Origin are Selected, and Forms a Sphere Neighborhood that is Non-direction Selection

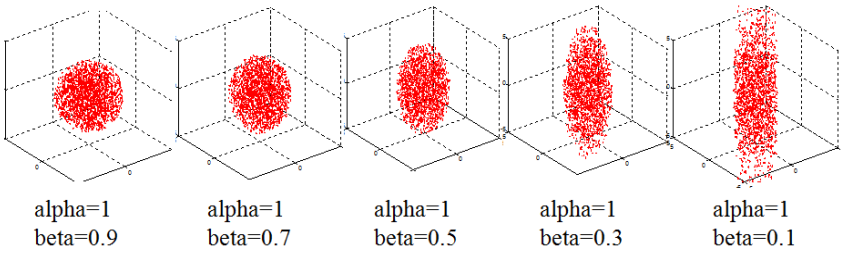


Figure 3. Our ST Oriented Neighborhoods at Various (α, β) are Direction-Selective. 1000 Points (red points) Selected from 10000 Points in Figure 2(C) to Form an Oriented Neighborhood, According to Equation (8)

To make the neighborhood feature more discriminative power, we impose ST weights (α, β) upon the distance metric between features. Specifically, we replace (x_i, y_i, t_i) with $(\alpha(x_i, y_i), \beta(t_i)), \alpha, \beta \succ 0$ during selecting the nearest neighbors, and the resulting neighborhoods are anisotropic, or direction-selective (see Figure 3), where α and β are the position weighting factors which reflect the importance of the spatial and temporal position in distance calculation, respectively, e.g. with the increase of β, the resulting neighborhood tends to select the STIPs that close to the central STIP with respect to the temporal direction.

For feature $\mathbf{g}_i$, its nearest neighbors with ST scales (α, β) can be obtained by ranking the Euclidean distance between $\mathbf{g}_i$ and all other features, where the Euclidean distance between $\mathbf{g}_i$ and $\mathbf{g}_j$ is defined as

$$\|\alpha(x_i - x_j), \alpha(y_i - y_j), \beta(t_i - t_j)\|_2 \tag{8}$$

Let $N^{\alpha,\beta}(\mathbf{g}_1) = \{\mathbf{g}_1, \cdots, \mathbf{g}_{n_e}\}$ denote the oriented neighborhood of $\mathbf{g}_1$ with (α, β), $\{\mathbf{g}_1, \cdots, \mathbf{g}_{n_e}\}$ are the nearest neighbors of $\mathbf{g}_1$, then its SONF $\mathbf{h}_1^{\alpha,\beta}$ is computed by average-pooling over their LLC codes,

$$\mathbf{h}_1^{\alpha,\beta} = (1/n_e) \sum_{i=1}^{n_e} |\mathbf{c}_i| \tag{9}$$

where $\mathbf{c}_i$ denotes the LLC code of $\mathbf{g}_i$, and $|\mathbf{c}_i|$ denotes the absolute value of $\mathbf{c}_i$.

Next, we can adjust α or/and β to obtain SONF $\mathbf{h}_1^{\alpha(i),\beta(j)}$, and its MONF $\mathbf{h}_1$ of $\mathbf{g}_1$ is obtained by concatenating all SONFs

$$\mathbf{h}_1 = [\mathbf{h}_1^{\alpha(1),\beta(1)}, \cdots, \mathbf{h}_1^{\alpha(i),\beta(j)}, \cdots] \tag{10}$$

where $i \in [1, n_s], j \in [1, n_t]$, and n_s and n_t are the numbers of spatial and temporal scales, respectively.

After computing the MONF of each local feature by (10) and (11), video sequence V is represented as group of MONFs

$$V = \{\mathbf{h}_i : i = 1, \cdots, n_v\} \tag{11}$$

where $\mathbf{h}_i$ denotes the MONF feature of $\mathbf{g}_i$.

5. Classifying with LGSR

To utilize the intrinsic group information from these MONFs within one video for action classification, we employed the Locality-constrained Group Sparse Representation (LGSR) to classify actions. LGSR was proposed in [34] for human gait recognition. It is an extended sparse representation-based classifier (SRC). The pioneering work of SRC was proposed in [40] and used to classify face images by minimizing the norm-regularized reconstruction error. There are three advantages of LGSR compared with SRC:

- SRC is designed for single image classification and cannot directly classify a group of samples, while LGSR is designed for sample group classification.
- The locality constraint in LGSR is more reasonable than sparsity constraint in SRC, especially for representing manifold data [23, 26].
- LGSR is a block sparse constraint classifier. It is better than SRC in classification task when the used features are discriminative.

The object function of LGSR is defined as

$$\mathbf{C}^* = \arg\min_{\mathbf{C}} \frac{1}{2}\left\|\mathbf{Y} - \mathbf{B}\mathbf{C}\right\|_F^2 + \lambda \sum_{k=1}^{K} \left\|\mathbf{D}^k \square \mathbf{C}^k\right\|_F, \tag{12}$$

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 $\mathbf{C}$. λ is the regularization parameter to balance these terms. $\mathbf{B}$ is the classification dictionary constructed by connecting K class-special dictionaries $[\mathbf{B}^1, \cdots, \mathbf{B}^K]$. Each class-special dictionary $\mathbf{B}^k$ is learnt with K-means algorithm from the MONF descriptors corresponding to the k-th action. $\mathbf{Y}$ is the group of MONF descriptors for one query action. $\mathbf{C}^k$ is one part of $\mathbf{C}$ and corresponds to $\mathbf{B}^k$. $\mathbf{D}^k$ is the distance matrix between $\mathbf{Y}$ and $\mathbf{B}^k$. The i-th and j-th element in $\mathbf{D}^k$ is calculated as $\mathbf{D}_{ij}^k = \left\|\mathbf{Y}_i - \mathbf{B}_j^k\right\|_2$ Since $\mathbf{C}^k$ values are independent of each other, we can separately update each $\mathbf{C}^k$ using its subgradient [41]. To solve (12), the active set-based subgradient descent algorithm [34, 42] was employed.

Once we obtain the optimal reconstruction coefficient $\mathbf{C}^*$, two classification methods [38] based on different criteria can be used to classify the test video.

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

$$R_k((\mathbf{C}^k)^*) = \frac{1}{2}\left\|\mathbf{Y} - \mathbf{B}^k(\mathbf{C}^k)^*\right\|_F \tag{13}$$

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

- 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 Forbenius norm of the reconstruction coefficient. Specifically, we define the following weighted inverse reconstruction error

$$Q_k((\mathbf{C}^k)^*) = \frac{\left\|(\mathbf{C}^k)^*\right\|_F}{\left\|\mathbf{Y} - \mathbf{B}^k(\mathbf{C}^k)^*\right\|_F} \tag{14}$$

Then, we classify the test video to $k^* = \arg\min_k Q_k((\mathbf{C}^k)^*)$.

In the paper, we use maxWIRE criterion as human video action classifier

6. Experiments

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

6.1. Experimental Setup

The experimental setup is as follows:

- For the two datasets, Dollar detector based on multiple ST scales is used to extract STIPs from action videos, and its spatial scale $\tau =$ [1.2,1.3,1.4,1.5] and temporal scale $\omega =$ [0.4,0.45,0.5,0.55]. (HOG+HOF) is employed as local features (e.g., feature $\mathbf{f}_i$ in (1)).
- To obtain the dictionaries for LLC coding, features from 24 videos (4 videos per action, 6 actions) of one subject are clustered by k-means for the KTH; and features from 20 videos (2 videos selected from each action, 10 actions) are clustered by k-means for the UCF Sports. The resulting dictionary size is set to 250 for KTH, and 400 for UCF Sports.
- The number of selected bases is set to be 5 during LLC encoding local features for both of the KTH and UCF Sports.
- In building MONFs, the ST scales are set to be α =[1,1,1,1], β = [1,0.75,0.50,0.25] for the KTH, and for the UCF sports, α =[1,1,1,1], β = [0.9,0.7,0.5,0.3].
- Since there are 4 ST scales for either of HOG/HOF, the dimension of a MONF, which is concatenated by 4 SONFs, is $4 \times 250 \times 2 = 2000$ for the KTH, and $4 \times 400 \times 2 = 3200$ for the UCF Sports. In order to guarantee that the class-special dictionaries in LGSR are over-complete, random projection in dimension reduction [43] is adopted to reduce the dimension of the MONF to 250, 400 for the KTH and UCF Sports, respectively.
- In LGSR classification, the size of each class-special dictionary is set to 300 for the both datasets, the parameter λ are evaluated by 4-fold cross-validation.
- In experiment, leave-one-out cross-validation (LOOCV) strategy is used to evaluate the system performance. Specifically, for the KTH, in each LOO run, we use the videos of 24 subjects for training, and the videos of the remaining subject for test, and the recognition rate is the average value of the 25 runs. For the UCF sports, in each LOO, one video of each class is randomly selected as test data, the other videos are treated as training data, 100 LOO runs are carried out, and the recognition rate is the average value of the 100 runs.

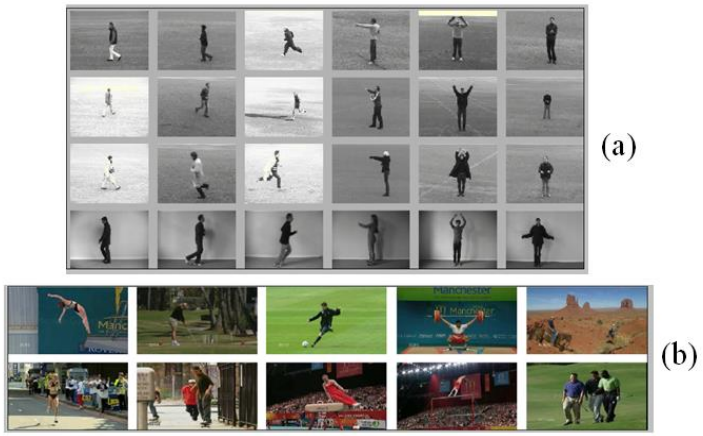


Figure 4. (a) Example Images from the KTH Dataset. (b) The UCF Sports Dataset

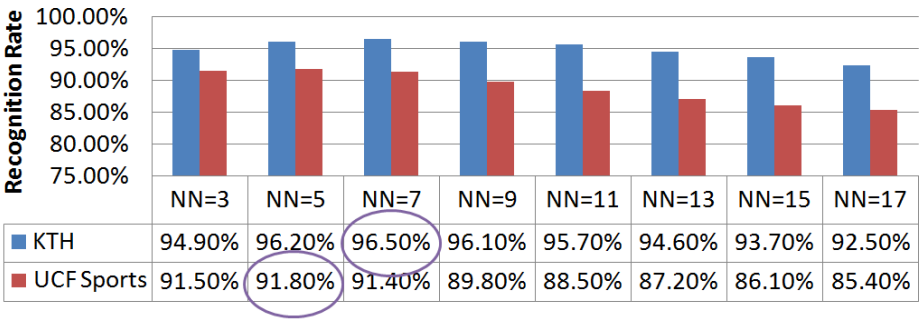
6.2. Human Action Datasets

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 Figure 4(a).

The UCF sports dataset includes 150 action 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 dataset can be seen in Figure 4(b).

6.3. Experimental Results and Analysis

To evaluate the influence of the number of nearest neighbors (NNs) on recognition accuracy during computing SONFs, various neighborhood sizes is used, and the results are shown in Figure 6. It can be seen that for the KTH, when the number of NNs is set to be 7 (NN=7), the highest recognition rate 96.5% is achieved; for the UCF Sports, the highest recognition rate 91.8% is obtained, when NN=5. Moreover, we found that when NN= (5, 7, 9) and (3, 5, 7), the recognition performance drops slightly for the KTH and UCF Sports, respectively.



	NN=3	NN=5	NN=7	NN=9	NN=11	NN=13	NN=15	NN=17
KTH	94.90%	96.20%	96.50%	96.10%	95.70%	94.60%	93.70%	92.50%
UCF Sports	91.50%	91.80%	91.40%	89.80%	88.50%	87.20%	86.10%	85.40%

Figure 5. The Relationship between the Number of Nearest Neighbors (NNs) of Oriented Neighborhood and the Recognition Rates on the KTH, UCF Sports Datasets, Respectively

Table 2 shows the performance comparison between our system and some classical system published recently over both the KTH and UCF Sports. The competing methods [18, 19, 44-46] include local representation-based, and global representation-based approaches. It is clear that our method achieves better performance than the competing methods. The confusion matrices for KTH and UCF sports datasets of our method are shown in Tables 3 and 4, respectively.

Table 2. The Performance Comparison between our System and the Recent Classical Systems

Methods	Year	Action Model	KTH(%)	UCF Spo.(%)
Zhu *et al.* [18]	2010	SC was used for feature coding together with BoF	94.9	84.3
Wu *et al.* [44]	2011	Spatiot-temporal context feature was employed	94.5	91.3
Guha *et al.* [19]	2012	Sparse representation-based classification methods, and sub-STV model	—	91.1
Bregonzio *et al.* [45]	2012	Local feature distribution information was utilized	94.3	—
Saghafi *et al.* [46]	2012	the global representation method was adopted	92.6	—
Our method		MSNFs based on oriented neighborhood and LGSR as classifier	96.5	91.8

Figure 6 presents the recognition accuracies on the KTH and UCF Sports under varied combinations on temporal scales. It can be found that when all temporal scales are considered in classification, the system performance achieves the best. Furthermore, with taking into account more spatial scales, the discriminative power of recognition system gets strong, the recognition rate increases.

Table 3. Confusion Matrix on the KTH Datasets. s1 (boxing), s2 (hand-waving), s3 (hand-clapping), s4 (walking), s5 (jogging), s6 (running)

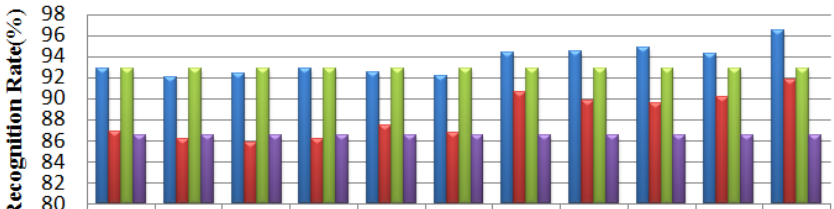
	s1	s2	s3	s4	s5	s6
s1	**100%**	0	0	0	0	0
s2	0	**100%**	0	0	0	0
s3	0	0	**100%**	0	0	0
s4	0	0	0	**95%**	2%	3%
s5	0	0	0	2%	**96%**	2%
s6	0	0	0	0	0	**100%**

Table 4. Confusion Matrix on the UCF Sports Dataset. S1 (diving), S2 (golfing), 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	**100%**	0	0	0	0	0	0	0	0	0
s2	0	**91%**	6%	0	0	0	0	0	0	3%
s3	0	5%	**90%**	0	0	5%	0	0	0	0
s4	0	0	0	**100%**	0	0	0	0	0	0
s5	0	6%	0	0	**92%**	0	2%	0	0	0
s6	0	0	3%	0	5%	**91%**	1%	0	0	0
s7	0	2%	0	3%	0	0	**91%**	0	0	4%
s8	0	0	0	0	0	8%	0	**92%**	0	0
s9	0	0	0	0	0	0	0	0	**100%**	0
s10	0	0	0	0	0	0	0	12%	0	**88%**

7. Conclusion

To utilize the ST position information between local features for action recognition task, many authors used either sub-STV method or traditional NN-based neighborhood method. However, the former requires ones to decide the temporal length of sub-STV empirically, and the latter lacks of considering orientation information within neighborhood. To overcome their limitations, we construct the oriented neighborhood by introducing ST scales to the Euclidean distance between local features. Then LGSR is employed as classifier upon the resulting MONFs based on the oriented neighborhoods. The experimental results on the KTH and UCF Sports datasets show that (1) In contrast to traditional methods based on either sub-STVs or sphere-shape neighborhood, the recognition system based on the oriented neighborhoods achieves better recognition accuracy. (2) Compared to sphere-shape neighborhood that is non-direction-selective, the proposed MONFs can provide more discriminative information for boosting the system performance.



	AB	AC	AD	BC	BD	CD	ABC	ABD	BCD	ACD	ABCD
OM KTH(%)	92.8	92	92.4	92.9	92.5	92.2	94.4	94.5	94.9	94.2	96.5
OM UCF(%)	86.8	86.1	85.8	86.1	87.5	86.7	90.6	89.8	89.6	90.2	91.8
TM KTH(%)	92.9	92.9	92.9	92.9	92.9	92.9	92.9	92.9	92.9	92.9	92.9
TM UCF(%)	86.5	86.5	86.5	86.5	86.5	86.5	86.5	86.5	86.5	86.5	86.5

Our method	KTH (alpha, beta)	UCF (alpha, beta)
Channel A	**(1.0, 1.0)**	**(1.0, 0.9)**
Channel B	**(1.0, 0.75)**	**(1.0, 0.7)**
Channel C	**(1.0, 0.50)**	**(1.0, 0.5)**
Channel D	**(1.0, 0.25)**	**(1.0, 0.3)**

Traditional method	KTH (alpha, beta)	UCF (alpha, beta)
Channel A	**(1.0, 1.0)**	**(1.0, 1.0)**
Channel B	**(1.0, 1.0)**	**(1.0, 1.0)**
Channel C	**(1.0, 1.0)**	**(1.0, 1.0)**
Channel D	**(1.0, 1.0)**	**(1.0, 1.0)**

Figure 6. The Performance Comparison between our Method (OM) and Traditional Method (TM) in the Case of Different Combinations of Temporal Scales. Note that Traditional Method is Referred to as Non-Direction-selective Neighborhood, and its ST scales $(\alpha = 1, \beta = 1)$ in all Cases, Hence, its Recognition rate is Same in all Combinations

Acknowledgements

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

References

[1] H. Zhang, A. Berg, M. Maire and J. Malik, "SVM-knn: discriminative nearest neighbor classification for visual category recognition", Proceeding of Computer Vision and Pattern Recognition (CVPR), **(2006)**, pp. 23-34.
[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, **(2012)**, pp.1-14,
[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, **(2008)**, pp. 719-726.
[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," Proceedings of the 17th ACM International Conference on Multimedia, **(2009)** October, pp.165-174.
[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, **(2011)**, pp. 224-241.
[6] J. K. Aggarwal and M. S. Ryoo, "Human activity analysis: a review," ACM Computing Surveys, vol. 43, no. 3, article 16, **(2011)**, pp. 234-253.
[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, **(2010)**, pp. 486-493.
[8] A. A. Chaaraoui and P. Climent-Perez, "Silhouette-based Human action recognition using sequences of key poses," Pattern Recognition Letters, vol. 34, no.15, **(2013)**, pp.1799-1807.
[9] X. Deng, X. Liu and M. Song, "LF-EME: local features with elastic manifold embedding for human action recognition," Neurocomputing, vol.99, no.1, **(2013)**, pp.144-153
[10] P. Dollar, V. Rabaud, G. Cottrell and 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, **(2005)** October, pp.65-72.
[11] A. Klaser, M. Marszalek and C. Schmid, "A spatio-temporal descriptor based on 3D-gradients," Proceedings of the British Machine Vision Conference, **(2008)**, pp. 210-221.
[12] P. Scovanner, S. Ali, and M. Shah, "A 3-dimensional sift descriptor and its application to action recognition," Proceedings of the 15th ACM International Conference on Multimedia, **(2007)** September, pp. 357-360.
[13] G. Willems, T. Tuytelaars and L. Van Gool, "An efficient dense and scale-invariant spatio-temporal interest point detector," Proceedings of the European Conference on Computer Vision (ECCV), **(2008)**, pp.650-663.
[14] I. Laptev, M. Marszałek, C. Schmid and B. Rozenfeld, "Learning realistic human actions from movies," in CVPR, **(2008)** June, pp.1-8.
[15] M.-J. Escobar and P. Kornprobst, "Action recognition via bioinspired features: the richness of center-surround interaction," Computer Vision and Image Understanding, vol. 116, no. 5, **(2012)**, pp. 593-605.
[16] X. Zhu, Z. Yang and J. Tsien, "Statistics of natural action structures and human action recognition," Journal of Vision, vol.12, no. 9, **(2012)**, pp. 834-834.
[17] B. Chakraborty, M. B. Holte, T. B. Moeslund and J. Gonzalez, "Selective spatio-temporal interest points," Computer Vision and Image Understanding, vol. 116, no. 3, **(2012)**, pp. 396-410.
[18] Y. Zhu, X. Zhao and Y. Fu, "Sparse coding on local spatial-temporal volumes for human action recognition," Proceedings of the Computer Vision, **(2010)**, pp. 660-671, Springer, Berlin, Germany.
[19] T. Guha and R. K. Ward, "Learning sparse representations for human action recognition," IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), vol. 34, no. 8, **(2012)**, pp.1576-1588.
[20] J. C. Van Gemert, C. J. Veenman, A. W. M. Smeulders and J.- M. Geusebroek, "Visual word ambiguity," in PAMI, vol. 32, no. 7, **(2010)**, pp.1271-1283.
[21] 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, **(1996)**, pp. 607-609.
[22] K. Yu, T. Zhang and Y. Gong, "Nonlinear learning using local coordinate coding," Proceedings of the 23rd Annual Conference on Neural Information Processing Systems (NIPS), **(2009)**, pp. 2223-2231.
[23] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang and Y. Gong, "Locality-constrained linear coding for image classification," in CVPR, **(2010)** June, pp.3360-3367.
[24] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," Science, vol. 290, no. 5500, **(2000)**, pp. 2323-2326.
[25] J. Wang, "Locally linear embedding," Geometric Structure of High-Dimensional Data and Dimensionality Reduction, **(2011)**, pp. 203-220, Springer, Berlin, Germany.

[26] 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, **(2013)**, pp. 1277-1287.
[27] C. C. Chen and J. K. Agarwal, "Modeling Human Activities as Speech," in CVPR, **(2011)**, pp.532-540.
[28] Q. Shi, L. Cheng, L. Wang and A. Smola, "Human action segmentation and recognition using discriminative semi-Markov models," International Journal of Computer Vision (IJCV), vol.3, **(2011)**, pp.890-921.
[29] M. Hoai, Z. Z. Lan and F. De la Torre, "Joint segmentation and classification of human actions in video," in CVPR, **(2011)**, pp.329-336.
[30] K. Tang, L. Fei-Fei and D. Koller, "Learning latent temporal structure for complex event detection," in CVPR, **(2012)**, pp.512-520.
[31] L. R. Rabiner and R.W. Schafer, "Introduction to digital speech processing," Foundations and trends in signal processing, **(2007)**, pp.471-482.
[32]T. Guha and R. K. Ward, "Learning sparse representation for human action recognition," in PAMI, vol. 20, **(2011),** pp.489-510.
[33]B. Wang, Y. Liu, W. Wang, W. Xu and M. Zhang. "Multi-Scale Locality-constrained Spatiotemporal Coding for Local Feature Based Human Action Recognition", the Scientific World Journal, **(2013)**, Article ID 405645, 11 pages.
[34] D. Xu, Y. Huang, Z. Zeng and X. Xu, "Human gait recognition using patch distribution feature and locality-constrained group sparse representation", IEEE transaction on image processing, vol. 21, no. 1, **(2012)** January**,** pp.510-521.
[35] J. Yang, K. Yu, Y. Gong and T. Huang, "Linear spatial pyramid matching using sparse coding for image classification," in CVPR, **(2009),** pp.1794-1801, Miami, USA.
[37] J. C. Gemert, J. Geusebroek, C. J. Veenman and A. W. M. Smeulders, "Kernel codebooks for scene categorization," in ECCV, **(2008)** October, pp. 696-709, Marseille, France.
[38] L. Liu, L. Wang and X. Liu, "In defense of soft-assignment coding," in Proceedings of IEEE International Conference on Computer Vision (ICCV), **(2011)** November, pp. 2486-2493, Barcelona, Spain.
[39] Y. Huang, Z. Wu, L. Wang, and T. Tan, "Feature coding in image classification: a comprehensive study," in PAMI, vol.23, no. 5, **(2013)**, pp.586-603.
[40] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry and Y .Ma, "Robust face recognition via sparse representation," in PAMI, vol. 31, no. 2, **(2009)**, pp. 210-227.
[41] M. Bregonzio, T. Xiang and S. Gong, "Fusing appearance and distribution information of interest points for action recognition," Pattern Recognition, vol. 45, no. 3, **(2012),** pp.1220-1234.
[42] 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, **(2008)** April, pp. 2113-2116.
[43] R. Baraniuk and M. Wakin, "Random projections on smooth manifolds," Foundations of computational mathematics, vol. 9, **(2004),** pp.91-110.
[44] X. Wu, D. Xu, L. Duan and J. Luo, "Action recognition using context and appearance distribution features," in CVPR, **(2011)** June, pp. 489-496.
[45] M. Bregonzio, T. Xiang and S. Gong, "Fusing appearance and distribution information of interest points for action recognition," Pattern Recognition, vol. 45, no. 3, **(2012)**, pp.1220-1234.
[46] B. Saghafi and D. Rajan, "Human action recognition using Pose-based discriminant embedding," Signal Processing, vol. 27, no.1, **(2012)**, pp. 96-111.

Authors

Jiangfeng Yang. He is a Ph.D. student in University of electronic science and technology of China, China. He received his Master degree of Engineering from Kunming University of science and technology, China in computer software and theory in 2009. His research interests include machine vision and action recognition in video.

Zheng Ma. He has been working as a professor in University of electronic science and technology of China, China. His research interests include signal processing, machine vision and Internet security.

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