Human Action Recognition Based on Oriented Gradient Histogram of Slide Blocks on Spatio-Temporal Silhouette

Yaqing Li[1,2], Tanfeng Sun[1,2] and Xinghao Jiang[1,2,*]

[1] *School of Information Security Engineering Shanghai Jiao Tong University, Shanghai 200240, China*
[2] *National Engineering Lab on Information Content Analysis Techniques, GT036001, Shanghai 200240, China*
{tfsun, xhjiang}@ sjtu.edu.cn

Abstract

Video can be regarded as three dimensional spatio-temporal volume, in which human action is a three dimensional shape (3D shape) surrounded by the spatio-temporal silhouette surface. The type of human action depends on the shape of the silhouette surface. In this paper, we proposed a new feature called Oriented Gradient Histogram of Slide Blocks by building dense overlapping spatio-temporal slide blocks to detect the shape of the 3D silhouette surface of the human action. Sparse coding is adopted to represent videos based on the new feature and Random Forest is utilized to classify the types of human actions. Experiments on KTH and Weizmann human action datasets demonstrate that the new feature can describe the spatio-temporal silhouette surface correctly, accordingly recognize the human action types accurately.

Keywords: *Oriented gradient; Slide block; Spatio-Temporal silhouette; Sparse coding.*

1. Introduction

In recent years, public safety issues have become increasingly serious and more and more concerned by the whole society. With the frequent occurrence of critical issues such as public transport safety, crowd control events and school violence, more and more professional cameras and surveillance systems are installed in corners of public areas. How to recognize human actions in thousands of Gb videos has become a hot area in the field of computer vision. Generally, there are three main steps for human action recognition in videos: feature extraction, video representation, and human action recognition.

For feature extraction, the traditional method was to track the body and obtain the trajectories [1, 2]. This kind of method is based on a 2D or 3D shape model which has a limitation of establishing a complete model library. Interest point detection is a very popular feature extraction method. There are corner detection, SIFT interest point by Lowe [3], interest point based on a set of linear filter [4], Harris and Förstner interest points [5]. Video representation plays an important role in human action recognition and many researchers adopted Bag of Words method. This kind of method treats a video as a collection of unordered appearance descriptors extracted from local patches which is called "visual words", and the representation of a video is a compact histogram of these words centers. In earlier researches, K-means cluster algorithm was the most commonly used method [6, 7]. However, the quality of the vocabulary is over-reliance on the number of the cluster centers in K-means, and each data is

* Corresponding author: Xinghao Jiang; Email: xhjiang@sjtu.edu.cn

forced to distribute into only one visual word, which may introduce large errors. Sparse coding has been used in image classification in recent years [8], but randomly used in video processing. An improper classifier may lead to a big failure even though the feature and the video representation are both effective. Hence, it is important to select a proper classifier. SVM classifier has been used very commonly in human actions recognition [9]. However, SVM has three drawbacks: first, it is difficult to choose a proper kernel for feature; second, SVM is a non-probabilistic binary classifier which can't deal with multi-categories problems directly; third, the training and testing are slow and have limitation in speed. In recent years, latent topic models such as the probabilistic latent semantic analysis (PLSA) [10] and the Latent Dirichlet Allocation (LDA) [11] has been more and more popular. Method [12] adopted these two models to recognition human actions. However, there is a defect in PLSA: there is no natural way to use it to assign probability to a new testing observation. In addition, the number of parameters to be estimated in PLSA grows linearly with the number of training example, which suggests that this model is prone to overfitting [12].

Many researchers adopted human shape feature [13]. In Navneet Dalal and Bill Triggs's [14] method, HOG is originally proposed as the feature for human detection in images which is based on evaluating well-normalized local histograms of image gradient orientations in a dense grid. The basic idea of the method is that local object appearance and shape can often be characterized rather well by the distribution of local intensity gradients. However, it may cause interference when used in human action recognition since it capture the pixels' gradients of all the regions repeatedly by an overlapping block including those with none human regions. As a result, the accuracy of evaluating human posture and motion direction may be reduced. Method [15] use 3D HOG descriptor based on local spatio-temporal (ST) volume feature which is built by dense sampling extracting video blocks at regular positions and scales in space and time. However, the position of extracted blocks may not be in the regions with distinct motion features since it samples regularly.

Considering the existing problems above, we want to find a feature that can not only avoid the non-humanoid area interference with the evaluation of human actions but also contain most obvious motion characters. Human action is the combination of human posture and motion direction, and obviously human body edge is the region with most distinct motion characters. In this paper, we build three dimensional spatio-temporal sliding blocks choosing human body edge as interested regions and accordingly proposed a new feature called oriented gradient histogram of slide blocks on spatio-temporal silhouette which can describe both human posture and human action directions. A BOW (Bags of Words) model is used based on this new feature combining sparse coding to represent videos taking advantage of the max spatial pooling which is more robust to local spatial translations and more biological plausible [4] and Random Forests as the classifier since it processes quickly.

The rest of the paper is organized in the following way. In Section 2, we describe the new proposed Oriented Gradient Histogram of Slide Blocks on Spatio-Temporal Silhouette feature extraction in details. In Section 3, we present the whole human action recognition scheme based on the new feature. In Section 4, we evaluate our algorithm and compare the performance of our method with other related methods. Finally, Section 5 concludes the paper.

2. Oriented Gradient Histogram of Slide Blocks

This section describes the proposed new oriented gradient histogram of slide blocks features which can describe both the shape of the object and the direction of motion.

2.1. 3D Gradient Definition

The 3D gradient of a pixel is defined:

$$\begin{cases} d_x(x,y,t) = I(x+1,y,t) - I(x-1,y,t) & \forall x,y,t \\ d_y(x,y,t) = I(x,y+1,t) - I(x,y-1,t) & \forall x,y,t \\ d_t(x,y,t) = I(x,y,t+1) - I(x,y,t-1) & \forall x,y,t \end{cases} \quad (1)$$

Where $I(x,y,t)$ denotes the pixel intensity at position (x,y) in frame t. d_x and d_y denote x and y components of the image gradient and d_t denotes the discrimination of the pixel intensity between the front frame and the next frame in the same position.

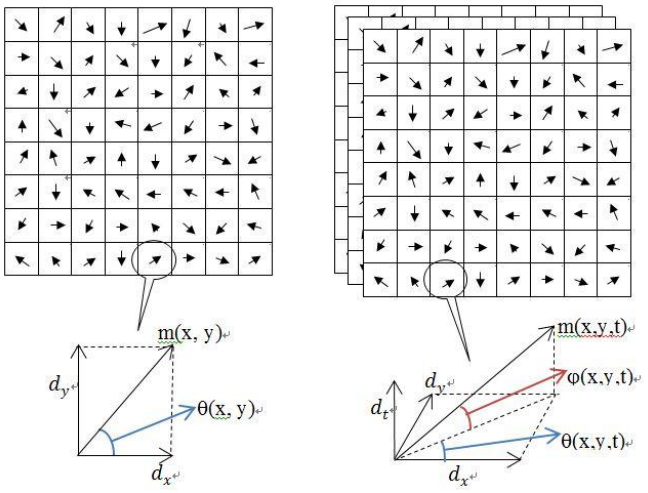


Figure 1. 2D and 3D Gradient. The left figure shows the magnitude and orientation of 2D gradient. The right figure shows the magnitude and the two orientations of 3D gradient.

The 3D magnitude $m(x,y,t)$ is:

$$m(x,y,t) = \sqrt{d_x(x,y,t)^2 + d_y(x,y,t)^2 + d_t(x,y,t)^2} \quad (2)$$

Thus there are two orientations $\theta(x,y,t)$ and $\varphi(x,y,t)$:

$$\theta(x,y,t) = \tan^{-1}\left(d_y(x,y,t) / d_x(x,y,t)\right) \quad (3)$$

$$\varphi(x,y,t) = \tan^{-1}\left(\frac{d_t(x,y,t)}{\sqrt{d_x(x,y,t)^2 + d_y(x,y,t)^2}}\right) \quad (4)$$

Where $\theta(x,y,t)$ denotes the spatial angle of the 3D gradient, $\varphi(x,y,t)$ denotes the temporal angle of the 3D gradient. Figure 1 shows the 2D and 3D gradient orientation.

2.2. Spatio-Temporal Silhouette

Video stream can be regarded as a three dimensional spatio-temporal volume in which human action is a continuous process. Hence each human action in video can be seen as a three dimensional shape (3D shape) and the slide area of human body edge can be seen as a spatio-temporal silhouette surface shown in Figure 2 (a) (b) (c). Actions of the same type generate similar 3D shapes, while there are great differences between the shapes of the different action types. For example, Figure 2 (a) shows the 3D shape of the action type of "waving hands", Figure 2 (b) and (c) respectively represents the 3D shape of "walking" and "running" and obviously these three shapes are distinctive from each other.

The type of action depends on the 3D shape which is surrounded by the spatio-temporal silhouette. For this reason, the type of human actions in videos can be predicted by detecting the shape of spatio-temporal silhouette surface which can be described by the collection of all the pixels' normal directions on the surface. Let 3D oriented gradient represent the normal direction, thus the shape of spatio-temporal silhouette surface can be represented by the collection of all the 3D oriented gradients on the surface.

2.3. Slide Blocks Definition

In order to detect the shape of the spatio-temporal silhouette, we define a series of 3D spatio-temporal overlapping blocks with a unique direction sliding on the spatio-temporal silhouette surface motivated by HOG which detects human in two dimensional image, extending two dimensional sliding blocks to three dimensional spatio-temporal blocks, narrowing the slide region to interested spatio-temporal silhouette surface shown in Figure 2 (b) (d).

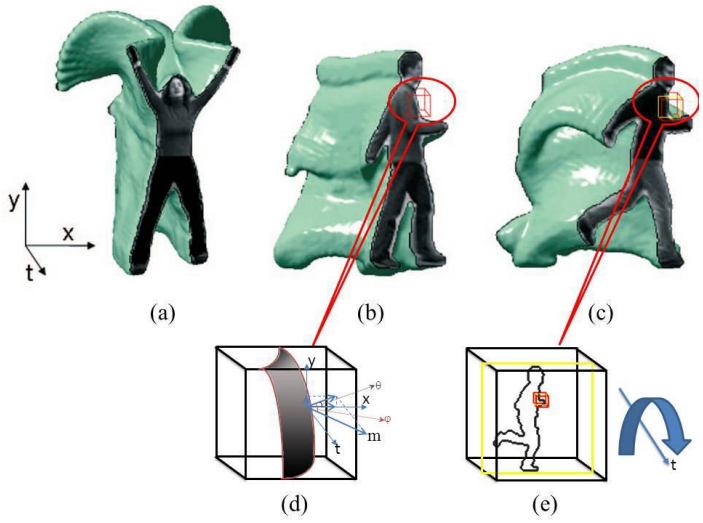


Figure 2. Figure (a) shows the three dimensional spatio-temporal contour surface of human action in video. Figure (b) and (d) show the three dimensional slide block on the spatio-temporal contour surface. Figure (c) and (e) show the slide block on a certain frame

The spatio-temporal block around the spatio-temporal silhouette surface in the direction of t from left to right, top to bottom, meaning that each time it slides around a section of silhouette surface perpendicular to t over and moves to the next one. Specifically, we consider a

section perpendicular to t, a frame. In fact, the intersection of a frame and the spatio-temporal silhouette surface is the edge of human body in this frame. Hence, the sliding of spatio-temporal block on the silhouette surface in three dimensional space is transformed to the sliding of a two dimensional block around the human body edge from left to right, top to bottom, shown in Figure 2 (c) (e). In this paper, we set the length of the spatio-temporal slide block in t dimension to 3, which means that when a two dimensional block sliding around the human body edge in a frame, the two blocks of the same position in both former and later frames should be taken into account at the same time, forming a three dimensional spatio-temporal block. Hence, the three dimensional spatio-temporal block can slide over the spatio-temporal silhouette surface integrally by controlling a two dimensional block sliding around the human body edge completely in a frame.

Regular edges is roughly divided into four types, vertical edge, horizontal edge, 45 degree edge, 135 degree edge as in Figure 3 (a) (b) (c) (d). Each none-directional edge can be categorized approximately as one kind of these four regular edges according to its extension direction. For example the non-directional edges in Figure 3 (e), the bottom left one can be classified as belonging to 135 degree edge, while the top right one belongs to horizontal edge. Accordingly, four types of sliding directions are defined: vertical bottom, horizontal right, bottom left 45 degree, bottom right 45 degree based on the four types of edges. In practice, it is implemented by making the following definitions in the two dimensional frames:

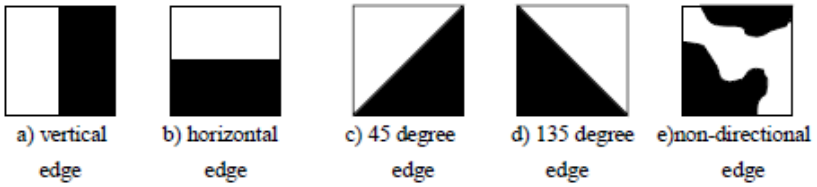


Figure 3. The Types of Edges

In this paper, several kinds of blocks are defined by dividing each image into equal-sized grids in video sequences first. Each grid region is called a "cell" shown in Figure 4 (a). A "block" is a larger spatial region composed of some cells shown in Figure 4 (b). A block is called a "full block" when edge is detected in this block, the block is called an "empty block". However, if the eight blocks around a specific full block are all empty blocks, this full block is forced to be changed to empty block since the edge of human body is considered to be a continuous boundary due to the connectivity of human body region.

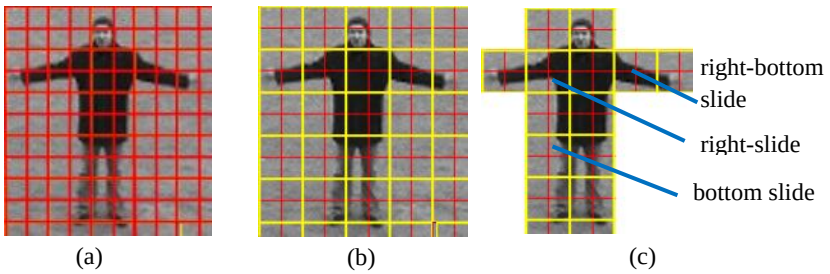


Figure 4. Figure (a) shows the definition of "cell". Figure (b) shows the definition of "block". Figure (c) shows all the full blocks. Each full block has a unique sliding orientation.

As a result, all the full blocks form the regions of human body edges. Then, straight line fitting is used on all the locations of edge pixels in the full block to estimate the extension direction of the edge. Accordingly, each full block is assigned a unique sliding direction by matching the straight line with the four types of regular edges. Figure 4 (c) shows all the full blocks and some sliding directions.

In order to slide the edge regions integrally, all the full blocks slide as the following steps:

Step0: Start from the top row;

Step1: Start from the left of the row;

Step2: Move right to detect a full block. If detected, go to Step3, otherwise, go to Step4;

Step3: Process the full block as described above, get the sliding direction. Slide the block from the former one in its sliding direction, meaning that the neighbour block in its sliding direction in the upper row, to the current one, as shown in Figure 5 (a), the sliding direction of block B is obvious bottom right 45 degree, therefor A is the former one in its sliding direction. If the detector detects B, the block will slide from A to B. Go to Step2;

Step4: If not the bottom row, move to the following row. Go to Step1; else, end.

When the sliding process is finished, all the edge of human body is slide over. For example, as shown in Figure 5 (b), when detected B, the block slides from A to B, similarly, when detected C, the block slides from B to C, and D is the same, the block slides from C to D as well. As a result, the block slide on the edge from A to D once.

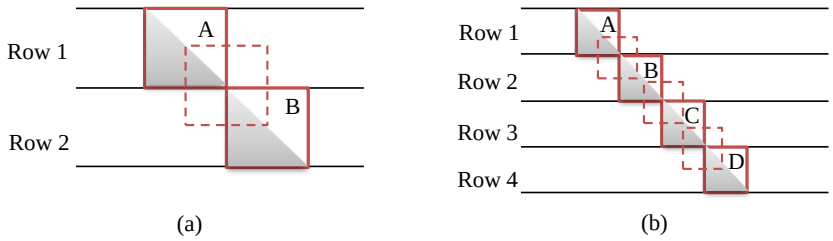


Figure 5. Slide Blocks. Figure (a) shows the slide way of a full block. Figure (b) shows the sliding on a continuous edge.

2.4. New Feature Definition

At each slide step, histogram calculation is processed. At each cell of the three dimensional spatio-temporal block, the 3D orientations of all pixels is quantized into some orientation bins, weighted by its magnitude to make a three dimensional histogram shown in Figure 6 (a). In practice, this is implemented by quantizing both orientations $\theta(x, y, t)$ and $\varphi(x, y, t)$ into an orientation bins, and the three dimensional histogram is built with the axes θ-bins and φ-bins shown in Figure 6 (b). The block histogram is obtained by combining all the cell histograms in this block. The two dimensional histogram is formed by splicing all the φ-bins shown in

Figure 6 (c). The value set of the two dimensional histogram can be represented by a vector V , the histogram vector.

For better invariance to illumination, shadowing, etc., each local histogram of a block is normalized:

$$V_i' = \frac{V_i}{\sqrt{\sum_{j=1}^{k} V_j^2}} \tag{5}$$

Where V denotes the histogram feature vector of a block, k denotes the length of vector V , V_j denotes the $j-th$ element of V , V_i denotes the un-normalized histogram value of the block in position i and V_i' denotes the corresponding normalized value.

When an action is performed, a certain posture is generated followed by the direction of movement. The spatial gradient of the edge of human body just contains the posture information, while the temporal gradient can show the direction of movement. Since the histogram calculation of slide blocks is performed on overlapping block, the gradients of the pixels in the sliding regions are repeatedly obtained by the different blocks. Hence, the gradients of long continuous edge are strengthened, like arms and legs which are regions with spatially distinguishing characteristics undergoing a significant motion in human actions.

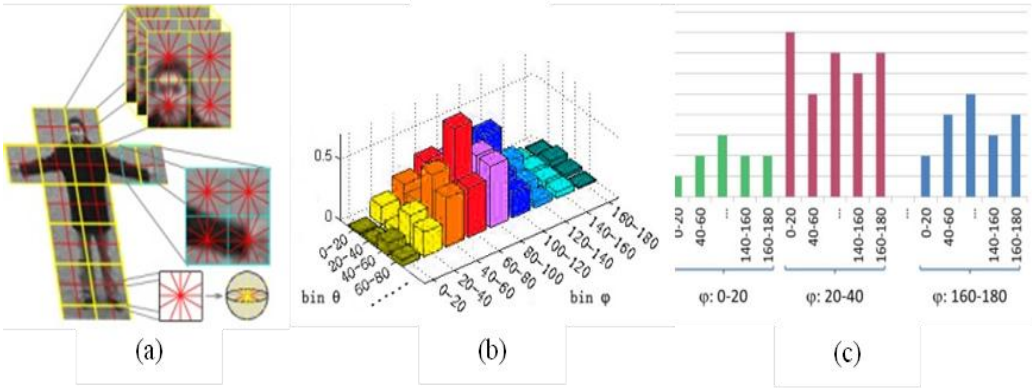


Figure 6. Figure (a) shows the building of the histogram of slide block. Figure (b) shows the three dimensional histogram. Figure (c) shows the two dimensional histogram.

3. Human Action Recognition Scheme based on New Feature

BOW model is used based on the new proposed feature. As shown in Figure 7, the main scheme contains four steps:

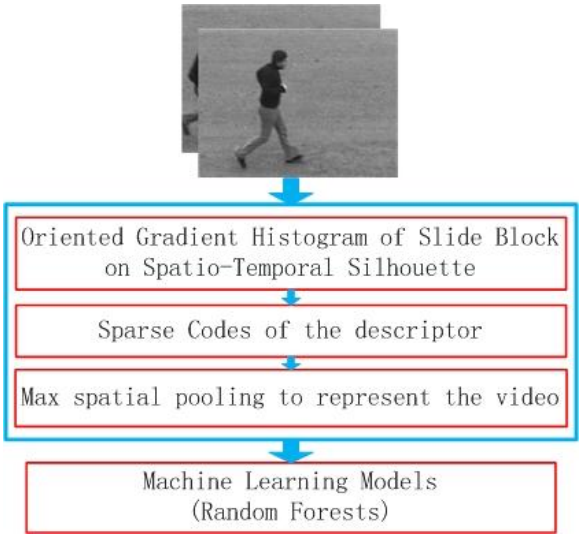


Figure 7. The Framework of the Proposed Human Action Recognition Scheme

3.1. New Feature Extraction

At first, we extract the feature vector of each frame in the video sequences by calculating the Oriented Gradient Histogram of Slide Blocks on Spatio-Temporal Silhouette as described in Section 2. Generally, a normalized block histogram is obtained at each slide of the slide block. The final feature vector of the frame is the accumulation of all the obtained normalized histograms. Hence, the length of the feature vector is equal to the length of the block histogram which depends on the number of orientation bins and the block size. Let the number of bins for θ and φ are bin_θ and bin_φ, the block size is $block_w$ (cells) × $block_h$ (cells), thus the histogram length of each cell is $bin_\theta \times bin_\varphi$, and the vector length of the block is $3 \times bin_\theta \times bin_\varphi \times block_w \times block_h$. Obviously, the vector length of the final feature of each frame is $3 \times bin_\theta \times bin_\varphi \times block_w \times block_h$.

3.2. Feature's Representation with Sparse Coding

When all the histograms of video frames are generated using the method recommended in last section, sparse coding is adopted to represent the feature in a sparse way. In practice, this is implemented by training an optimal dictionary first using some selected features from all the obtained features. Consider linear equations:

$$h \approx Ad \tag{6}$$

Where $A \in \mathbb{R}^{m \times n}$ is a full rank matrix called a dictionary and each column of it represents a basis vector $a_i \in \mathbb{R}^m$ $(1 \le i \le n)$, h is the original signal. In this case, h can be linear represented by the basis vectors of the dictionary. Consider that if h and A are both known, how to find a d to establish the equation. This problem is equal to that finding the coefficients of the linear combination. Obviously, there is infinite number of solutions. Although some decompositions provide an exact reconstruction of the data $(h = Ad)$, here we consider the approximate ones in nature.

Furthermore, we add some other constrains to the solution d which is hoped to be as sparse as possible, meaning that the number of nonzero elements as small as possible. This is an optimization problem and has been demonstrated to have a unique solution if the dictionary A satisfies certain conditions [16, 17]. Hence, the dictionary learning plays an important role in sparse representation. Consider a finite training set of signals $H=[h_1,h_2,...,h_k]$ in $\mathbb{R}^{m\times k}$ (n << k) used to optimize the empirical cost function [18]:

$$f_k(A)=\frac{1}{k}\sum_{i=1}^{k}\ell(h_i,A) \tag{7}$$

Where A is the dictionary, and $\ell(h,A)$ is a loss function that should be small if A is "good" at representing h in a sparse fashion, h_i is the $i-th$ element of h. Thus $\ell(h,A)$ can be defined as the optimal value of the ℓ_1 sparse coding problem:

$$\ell(h,A)=\min_{\alpha\in\mathbb{R}^k}\frac{1}{2}\|h-Ad\|_2^2+\lambda\|d\|_1 \tag{8}$$

Where λ is a regularization parameter which is set to be 0.15 in this paper. The dictionary columns $a_1,a_2,...,a_n$ are constrained to have an ℓ_2-norm less than or equal to one in order to avoid arbitrarily large values of A. The convex set of matrices C verifies the constraint:

$$C=\{A\in\mathbb{R}^{m\times n}\ \text{s.t.}\ \forall j=1,...,n,\ \ a_j^Ta_j\le 1\} \tag{9}$$

Then the problem of minimizing the empirical cost function $f_k(A)$ can be rewritten as a joint optimization problem with respect to the dictionary A and the coefficients $d=[d_1,d_2,...,d_k]$ in $\mathbb{R}^{n\times k}$ of the sparse decomposition:

$$\min_{A\in C,d\in\mathbb{R}^{n\times k}}\frac{1}{k}\sum_{i=1}^{k}\left(\frac{1}{2}\|h_i-Ad_i\|_2^2+\lambda\|d_i\|_1\right) \tag{10}$$

The optimizing process can be executed iteratively through alternating optimization over A or d while fixing the other until a specific minimum is obtained.

When the dictionary is generated, all the original feature vectors are transformed to the corresponding sparse codes which are the new feature vectors now. Accordingly, the length of feature vector is the number of the basis of the dictionary.

3.3. Reducing Dimension by Max Pooling

Max spatial pooling following sparse coding is used to represent video by first divide all the frames of the video sequences into a set of frame groups in which each frame generates a 3D orientation histogram. After all the descriptors are obtained, the sparse code of each histogram descriptor is calculated as described in Section B. Then Max spatial pooling function is applied on each feature groups generating the final sparse representation vector of each video:

$$p_j=\max\{|d_{1j}|,|d_{2j}|,...,|d_{kj}|\} \tag{11}$$

Where d_j denotes the $i-th$ frame descriptor of the frame group, k denotes the number of descriptors of the frame group, d_{ij} denotes the $j-th$ element of d_i, p_j denotes the $j-th$ element of p. At last, each p of a frame group is combined to form a sparse vector which represents the video.

Figure 8 shows the sparse representation of videos:

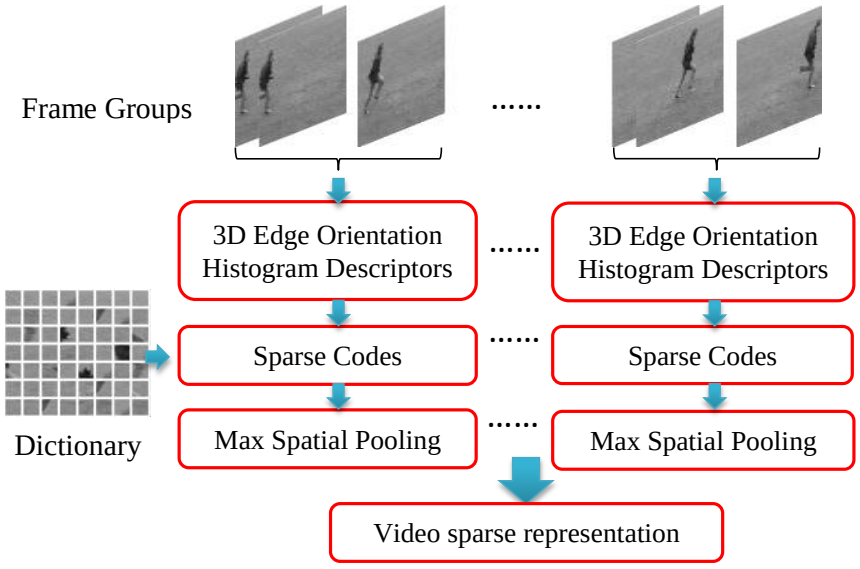


Figure 8. Video Sparse Representation using Max Spatial Pooling

3.4. Human Action Recognition by Classifiers

When all the videos get their sparse representations, Random Forest is used to recognize and classify human actions. Random forests classifier has several advantages, especially in speed. Random forests classifier is much faster in training and testing than traditional classifiers such as SVM, and it is as accurate as SVM. Random forest constructing each tree using a different bootstrap sample of the data is a kind of statistical learning theory.

4. Experiment and Result

In the experiments, KTH (Kungliga Tekniska Högskolan) and Weizmann human motion datasets are used to test the new proposed algorithm.

4.1. Experiment Setup

Feature extraction. The slide step coefficient is set to 0.5. The orientation of $\theta(x,y,t)$ and $\varphi(x,y,t)$ are both 0° ~180° .

Sparse coding. The regularization parameter λ is set to 0.15.

Random forests. The number of tree is set to 100 in the forest.

4.2. Evaluation and Performance of our Algorithm on KTH

KTH human motion dataset is the largest available video sequence dataset of human actions. There are six types of human action in the dataset: walking, jogging, running, boxing, hand waving and hand clapping shown in Figure 9. Each type of action is performed four times by 25 individuals in varied scenarios of outdoor and indoor environment with illumination diversification and scale changes. There are totally 599 short video sequences.

Figure 9. The example images of video sequences in KTH dataset. The dataset is available for download at http://www.nada.kth.se/cvap/actions/

In order to evaluating the performance of the proposed method, half of data is selected randomly for training leaving the left half as test data set.

Too big block may cause that the gradient of pixels outside the edge regions brings too much interfere, while too small block may bring huge computational cost. In order to study the influence of the block size, we test our algorithm using various cell size : 4 (pixels) × 4 (pixels), 6 (pixels) × 6 (pixels), 8 (pixels) × 8 (pixels) and block size: 1 (cell) × 1 (cell), 2 (cells) × 2 (cells), 3 (cells) × 3 (cells). Figure 10 shows the dependence of the recognition accuracy on the size of cells and blocks. It can be seen that the block size 2(cells)×2(cells) with the cell size 6(pixels)×6(pixels) achieves the best performance that the accuracy is 93.31%.

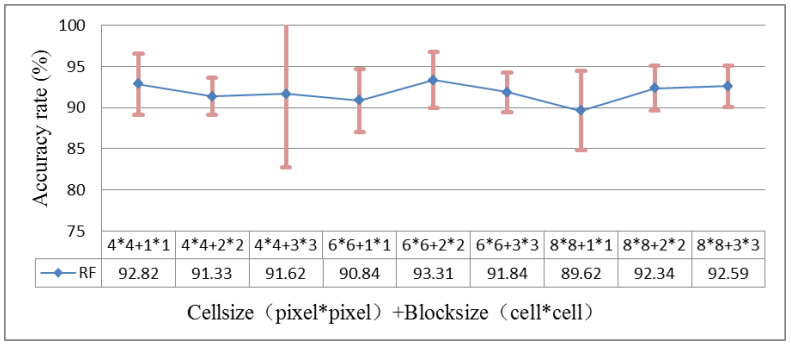


Figure 10. Evaluation of Block Sizes

Too many bins may lead to a very long histogram vector which will reduce the efficiency of the algorithm. However, too few bins may cause low recognition and classification accuracy since large angle interval reduces the precision of the orientation. In the evaluation experiment, 5 different numbers of bins are chose to test evaluate performance: 6, 8, 10, 12, 14. Figure 11 (a) shows the dependence of the accuracy rate on the number of bins for both θ and φ. As shown in the figure, it reaches the highest accuracy rate 93.56% when the number of bins is set to 8 for both θ and φ.

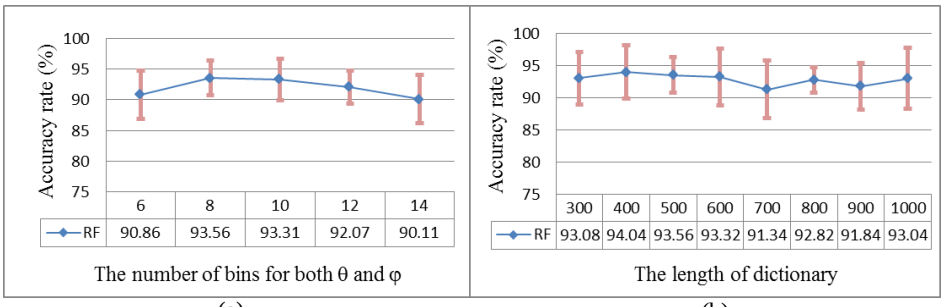


Figure 11. Figure (a) shows the evaluation of the number of bins. Figure (b) shows the evaluation of the length of dictionary

Dictionary plays an important role in sparse representation. Too long dictionary causes huge amount of computation, while the recognition of feature vector may be week if the dictionary is too small. The influence of the dictionary length on recognition accuracy has been tested on 8 different sizes: 300, 400, 500, 600, 700, 800, 900, and 1000 shown in Figure 11 (b). It is shown that the most appropriate dictionary size is 400 when the accuracy rate reaches 94.04%.

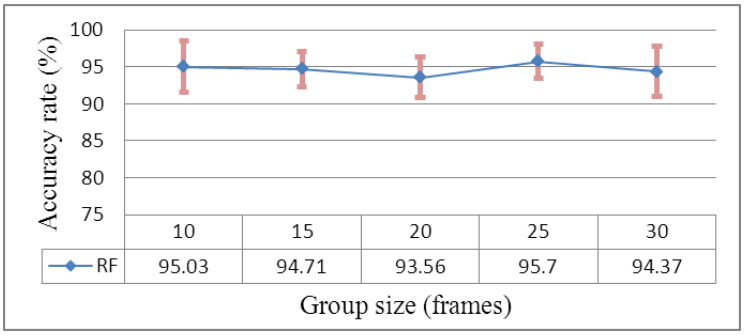


Figure 12. The Evaluation of Group Size

In the step of max spatial pooling, all the frames are divided into several groups and the feature descriptor is processed by each group. Too small groups generate long histogram vector at a very high computational cost. While too big groups reduce the ability of characteristic expression, result in a lower recognition accuracy rate. In this paper, five sizes of groups are chosen to test the dependence of recognition accuracy on the group size, which are 10 frames, 15 frames, 20 frames, 25 frames and 30 frames as shown in Figure 12. It is shown that the best performance is the accuracy of 95.7% obtained by group of 25 frames. Note that the recognition accuracy by group of 10 frames is also high reaching 95.03, however, the computational cost is relatively higher. For this reason, we choose the group of 25 frames as the most appropriate group size.

The following experiments are performed under the most appropriate parameters: the cell size is set to 6 (pixels) × 6 (pixels), the block size is set to 2 (cells) × 2 (cells), the number of bins is set to 8, and the dictionary length is set to 400, the number of frames in each group is set to 25.

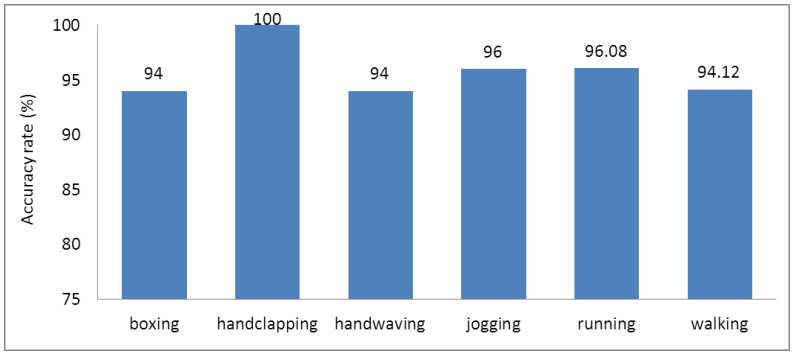


Figure 13. The Recognition Performance on Six Types of Human Actions

The recognition accuracy on the six types of actions is shown in Figure 13. It is shown that the best performance reaches 100% obtained at types of "handclapping". The recognition accuracy of "running" and "jogging" is also high reaching 96.08% and 96%. Specially, note that similar action types get close accuracy values such as "jogging" and "running", "boxing" and "handwaving", and this is consist with our intuition. The performance of "boxing" and "handwaving" is slightly lower than the other types obtaining both 94% since it may cause incorrect prediction due to the similarity between "boxing" and "handwaving". Overall, all the six types of actions get an accuracy rate higher than 90%, and the mean accuracy reaches 95.7%. From the result of this experiment, it can be seen that the new proposed oriented gradient histogram of slide blocks feature captures the character of human motion effectively, and the model based on this feature performs good on human action recognition, getting high recognition accuracy rate.

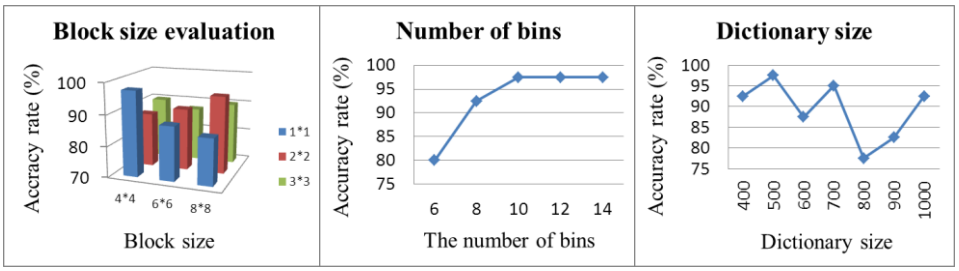
4.3. Evaluation and Performance of our Algorithm on Weizmann

Weizmann dataset contains 10 action categories containing 90 videos in total, each type of actions is performed by 9 subjects. Example frames of the action categories are shown in Figure 14. We adopt the Leave-One-Out (LOO) training method to perform the recognition of each motion type. For each run we learn a model from the videos of eight subjects, and test those of the remaining subject. The result is reported as the average of the nine runs.

Figure 14. The Example Images of Video Sequences in Weizmann Dataset

We test the performance dependence on block size, number of bins, and the dictionary size shown in Figure 15. When cell size is set to 4(pixels) × 4(pixels), block size is set to 1(cell) ×

1(cell), the number of bins is set to 10, and the length of dictionary is set to 500, the algorithm gets the highest recognition accuracy rate which is 97.5%.



(a) (b) (c)

Figure 15. Parameters Evaluation

Figure 16 shows the recognition accuracy of each type of Weizmann dataset. It can be seen that our recognition model achieves nine 100% accuracy rate of ten. The action type of "skip" gets the lowest performance which is 75%. Some recognition mistakes happen on this type of action because that it may cause recognition confusion since there are many similarities between "skip" and "jump", "skip" and "run", reducing the recognition accuracy.

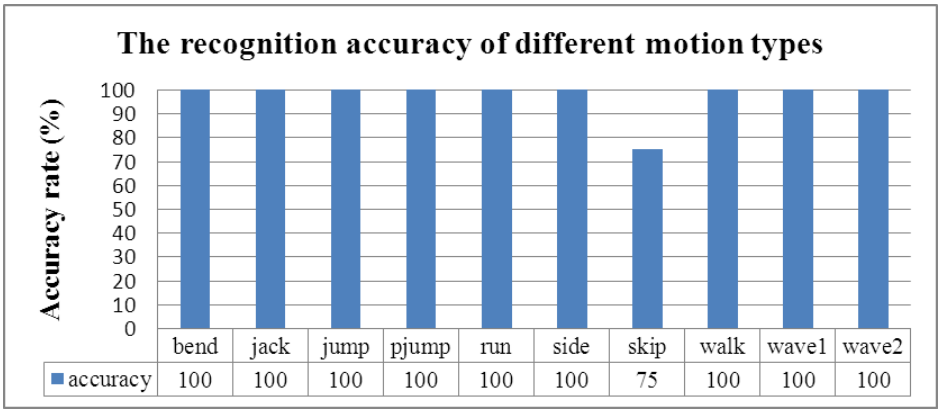


Figure 16. The Recognition Accuracy for Each Type on Weizmann

4.4. Comparison among Three Algorithms

The comparison of recognition accuracy on KTH between our algorithm and the state-of-art is show in Table 1. Two related methods are selected: Gang Yu et. al., [5] and Feng Shi et .al., [15]. Both of the two methods choose the videos of 16 persons of each motion type randomly as training set, leaving the other 9 persons as test set. In order to compare with them, we change our training method the same with theirs.

Gang Yu et. al., [5] utilizes Random Forest model similar to our method, however, the feature, which is based on interest points, is very different from ours. As can be seen from the table, this new proposed recognition feature combined with sparse coding gets better recognition accuracy which is 96.76% compared with 91.8% of method [5] since this new feature can describe both the shape of the object and the direction of motion by detecting the 3D spatio-temporal silhouette surface of human action. While the position of interest points may not in

the regions containing obvious movement characters, and the feature of interest points may not show the movement characters to the maximum.

Feng Shi et. al., [15] uses HOG3D feature which is also motivated by HOG feature. However, the local spatio-temporal (ST) volume feature is built by dense sampling extracting video blocks at regular positions and scales in space and time, very different from our human body edge regions. It can be seen that the recognition result of our algorithm is 96.76%, better than 91.77% of method [15]. This is because that the new feature we proposed chooses human edge regions as interested regions which contain distinct characters of human motion.

Table 1. Comparison with State-of-art

Methods	Recognition accuracy (%)
Our method	96.76
Gang Yu et al. [5](2011)	91.8
Feng Shi et al. [15](2011)	91.77

The comparison between our method and methods [5, 15] shows that 3D human action shape represented by spatio-temporal silhouette surface is an effective way to recognize human action type. Our new feature oriented gradient histogram of slide blocks which contains both the shape of the object and the direction of motion can describe the spatio-temporal silhouette surface correctly, accordingly recognize the human action type accurately. The way of building three dimensional spatio-temporal sliding blocks on the spatio-temporal silhouette surface of human action in our new feature can grab the prominent features of human action, making great contribution to recognizing human action accurately.

5. Conclusions

In this paper, a new feature called oriented gradient histogram of slide blocks on spatio-temporal silhouette is proposed for human action recognition. This feature is processed by defining a series of 3D spatio-temporal slide blocks of which each has a unique sliding direction on the spatio-temporal silhouette surface. A BOW model is used firstly combining sparse coding with max spatial pooling to represent videos based on the new feature and Random Forests as the classifier. Experiments show that detecting 3D human action shape represented by spatio-temporal silhouette surface is an effective way to recognize human action type. The new proposed feature containing both the shape of the object and the direction of motion can describe the spatio-temporal silhouette surface correctly, accordingly recognize the human action type accurately. The way of building three dimensional spatio-temporal sliding blocks on the spatio-temporal silhouette surface of human action in the new feature can grab the prominent features of human action, making great contribution to recognizing human action accurately.

Acknowledgments

The work of this paper is sponsored by the National Natural Science Foundation of China (No. 61071153, 61272439, 61272249), the National New Century Excellent Talents Support Plan of Ministry of Education, China (No. NECT-10-0569), and Shanghai Rising-Star Program (10QA1403700).

References

[1] D. Ramanan and D. A. Forsyth, "Automatic annotation of every-day movements", Advances in neural information processing systems, vol. 16, **(2004)** December 13-18; Vancouver, Canada.
[2] Y. Song, L. Goncalves and P. Perona, "Unsupervised learning of human motion", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 25, **(2003)**, pp. 1-14.
[3] D. G. Lowe, "Distinctive image features from scale-invariant key points", International Journal of Computer Vision, vol. 60, **(2004)**, pp. 91-110.
[4] P. Dollár, V. Rabaud, G. Cottrell and S. Belongie, "Behavior recognition via sparse spatio-temporal features", In 2nd joint IEEE international workshop on visual surveillance and performance evaluation of tracking and surveillance, **(2005)** October 15-16, pp. 65-72, Beijing, China.
[5] G. Yu, N. A. Goussies, J. Yuan and Z. Liu, "Fast Action Detection via Discriminative Random Forest Voting and Top-K Subvolume Search", IEEE Transaction on Multimedia, vol. 13, **(2011)**, pp. 507–517.
[6] Z. Jiang, Z. Lin and L. S. Davis, "Recognizing human actions by learning and matching Shape-Motion prototype trees", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 34, **(2012)**, pp. 533-547.
[7] L. Ballan, M. Bertini, A. D. Bimbo, L. Seidenari and G. Serra, "Effective Codebooks for human action categorization", IEEE 12th International Conference on Computer Vision Workshops, **(2009)** September 27 – October 4, pp. 506–513, Xi'an, China.
[8] J. Yang, K. Yu, Y. Gong and T. Huang, "Linear spatial pyramid matching using sparse coding for image classification, IEEE Conference on Computer Vision and Pattern Recognition, **(2009)** June 20-25, pp. 1794-1801; Miami, USA.
[9] C. Schuldt, I. Laptev and B. Caputo, "Recognizing human actions: A local SVM approach", International Conference on Pattern Recognition, **(2004)** August 23-26, pp. 32-36, Cambridge, England.
[10] T. Hofmann, "Probabilistic latent semantic indexing", ACM, **(1999)**, pp. 50-57.
[11] D. Blei, A. Ng and M. Jordan, "Latent dirichlet allocation", The Journal of Machine Learning Research, vol. 3, **(2003)**, pp. 993-1022.
[12] J. C. Niebles, H. Wang and F. Li, "Unsupervised learning of human action categories using spatial-temporal words", International Journal of Computer Vision, vol. 79, **(2008)**, pp. 65–72.
[13] L. Gorelick, M. Blank, E. Shechtman, M. Irani and R. Basri, "Actions as space-time shapes", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 29, **(2007)**, pp. 2247-2253.
[14] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection", IEEE Computer Vision and Pattern Recognitio, **(2005)** June 20-26, pp. 886-893; San Diego, USA.
[15] F. Shi, E. M. Petriu and A. Cordeiro, "Human action recognition from local part model", IEEE International Workshop on Haptic Audio Visual Environments and Games, **(2011)** October 14-17, pp. 35-38; Qinhuangdao, China.
[16] D. L. Donoho and M. Elad, "Optimally Sparse Representation in General (nonorthogonal) Dictionaries via L1 Minimization", Proceedings of the National Academy of Sciences, vol. 100, **(2003)**, pp. 245-365.
[17] E. Candes and T. Tao, "Near Optimal Signal Recovery From Random Projections: Universal Encoding Strategies?", IEEE Transactions on Information Theory, **(2006)**, pp. 5406-5425.
[18] J. Mairal, F. Bach, J. Ponce and G. Sapiro, "Online Learning for Matrix Factorization and Sparse Coding", Journal of Machine Learning Research, vol. 11, **(2010)**, pp. 19-60.

Authors

Yaqing Li

Female, born in 1988. Yaqing Li received the B.Eng. degree in School of Information Security at Shanghai Jiao Tong University, Shanghai, P. R. China in 2006. Now she is pursuing M.E. degree in School of Information Security at Shanghai Jiao Tong University, Shanghai, P. R. China. Her current research interests include video retrieval, video classification, and human action recognition.

Tanfeng Sun

Male, born in 1975. Tanfeng Sun received the Ph.D. degree in Information and Communication Engineering from Jilin University, Jili, P. R. China in 2003. He is a lecturer of the School of Information Security Engineering at Shanghai Jiao Tong University, Shanghai, P. R. China. His current research interests include multimedia security and image retrieval, information hiding and watermarking.

Xinghao Jiang

Male, born in 1976. Xinghao Jiang received the Ph.D. degree in Electronic Science and Technology from Zhejiang University, Hangzhou, P. R. China in 2003. He is a professor of the School of Information Security Engineering at Shanghai Jiao Tong University, Shanghai, P. R. China. His current research interests include multimedia security and image retrieval, intelligent information processing, cyber information security, information hiding and watermarking. Dr. Jiang is an IEEE member.