Multiple-Instance Support Vector Machine Based on a New Local Feature of Hierarchical Weighted Spatio-Temporal Interest Points

Chun Shan[1], Liyuan Liu[1], Jingfeng Xue[1], Zhaoliang Sun[1,2], Tingping Ma[1]
[1] School of Software, Beijing Institute of Technology, China
[2] China Minsheng Bank Science and Technology Department, China
sherryshan@bit.edu.cn, 597098022@qq.com, xuejf@bit.edu.cn, sunzhaoliang@cmbc.com.cn, 610115255@qq.com

Abstract

Human action recognition is a hot research topic. However, in actual scene such as house intelligent monitoring, the background is disordered, many external factors harden the automatic recognition of human action. In this paper, we mainly paid attention to finding a feature to describe human actions efficiently and meanwhile deal well with intra-class and inter-class changes of human bodies, and also solve the problems that external factors cause. Thus, We proposed a new kind of feature, the Local Feature of Hierarchical Weighted Spatio-Temporal Interest Points, which fused different features in a specific way. To more accurately classify the presented features, based on Support Vector Machine, we introduced a new Multiple Instance Learning algorithm, forming the Multiple-Instance Support Vector Machine. Finally, we validated on the KTH public dataset and tested on the captured family activity video dataset. And we got a higher accuracy for human action recognition in home environment.

Keywords: Human action recognition, Local feature of hierarchical weighted spatio-temporal interest points (LFHWSTIPs), Multiple-instance learning, Support vector machine (SVM)

1 Introduction

Human action recognition has been researched in the last decades and achieved large development. It has been widely applied into intelligent surveillance and human-computer interaction. Until now, researchers have proposed a lot of methods and many perform well in simple scenes with single body. For example, for features to extract, Mori and Malik [1] proposed the shape context operator to describe the human silhouette. Laptev et al. proposed the Spatio-Temporal Interest Points (STIPs) [2] features to describe the key parts of the human action. Bobick and Davis proposed the Motion History Image (MHI) [3] and Motion Energy Image (MEI) [4] by combining the characteristics of temporal and spatial motion to describe human motion. In recent years, kinematic features [5] are quite popular which based on the optical flow of the pixel in time, and include features such as the divergence, curl (vorticity) among others. Zhang et al. presented a novel Local Surface Geometric Feature [6] which is extracted from each skeleton joint in point cloud space. Liang et al. proposed the Local Depth Map Feature [7] to describe local spatio-temporal details of human action, which characterizes the local temporal change of human motion and the local spatial structure (appearance) of an action. And for classifier, Decision Tree [8], K-Nearest Neighbor [9], Hidden Markov Models (HMM) [10], Dynamic Bayesian Network (DBN) [11], Support Vector Machine (SVM) [12] are four kinds of general recognition algorithm. SVM was proposed based on the statistical learning theory of machine learning methods by Vapnik et al. in 1995, it is popular for the good performance in solving nonlinear classification problems in terms of small sample, nonlinearity, high dimension and local minimum.

However, many external factors such as the disordered background, different attitude of human body, variable time interval of the movement cycle, rapid segmentation of motion in dynamic environment, changes in light, occlusion of objects and ambiguities of action and scene add difficulties to the recognition. As a result, the existing methods don't perform very well in actual house environment.

Motivated by these works, and to find a way to improve the current performance of human action recognition in home environment, we paid attention to find a new feature to describe human action more efficiently and meanwhile deal well with the problems caused by external factors. So we mixed appearance feature, motion feature, spatio-temporal interest points [2] feature and Histograms of Oriented Gradient (HOG) feature [13] and proposed a new kind of feature to extract, called the Local Feature of Hierarchical

*Corresponding Author: Jingfeng Xue; E-mail: xuejf@bit.edu.cn
DOI: 10.3966/160792642018051903029

Weighted Spatio-Temporal Interest Points (LFHWSTIPs). The idea is, weight the spatio-temporal interest points on MHI [3] by layer first, then calculate the HOG feature of the weighted MHI image and the binary image and mix them. Moreover, we improved SVM [12] classifier by introducing a new statistical algorithm called Multiple Instance Learning (MIL) [14] into it, forming the MI-SVM algorithm for action matching and recognition. Last we compared our recognition results with those got by existing methods, and we found that the feature and algorithm we proposed in this paper performed generally more accurately and applicably.

2 Multiple-Instance Support Vector Machine Based on Local Feature of Hierarchical Weighted Spatio-Temporal Interest Points

2.1 Summary

In this paper, we mainly aimed to improve the performance of human action recognition in dealing with problems caused by external factors in actual scenes. We did our research in the three standard steps.

Pretreat and object detection. Considering that high-resolution camera always got so much video information, so we did some pretreatment to reduce the amount of computation, including graying and changing the resolution. As the color space information does not make sense in this study, so the original image can be grayed out. We improved the classical background difference [15] method, we used the first pin to initialize the background model for simple-background KTH dataset, and used Gaussian Mixture Model (GMM) [16] to model and update dynamical background for complex-background video indoor.

Feature extraction. In order to adapt to the complex and occluded indoor environment, this paper proposed a new feature, Local Feature of Hierarchical Weighted Spatio-Temporal Interest Points (LFHWSTIPs), which make full use of the advantages of existing features, and improve the accuracy and robustness.

The idea is to weight the spatio-temporal interest points on the MHI by layer and calculate the HOG features of the weighted MHI image and the binary image, then combine the two features. The specific methods are as follows. The first step is to use the Motion History Image (the optical flow method can also get very good results) which can reduce the image noise and enhance the robustness. And then detect interest points, we used the interest points detection method proposed by Laptev. In order to further reduce the influence of noise and reduce the calculation, we used the Motion History Image as a mask to filter out non-kinetic or less-kinetic interest points. Then, weighted the local image around the interest points. The non-zero pixel feature region which is closer to the interest point is given a larger weight, and the one farther away from the interest point is assigned a smaller weight. If the feature region pixel value is zero, its weight value is zero.

The weight value was calculated as follows. Take the position and intensity values of each point as eigenvectors, for each point and the interest point, calculate the Mahalanobis distance [17] of its eigenvectors, and take the inverse of the Mahalanobis distance as the weight value, then calculate the HOG feature of the weighted local image. Last use the HOG feature of the local image as the input of the classification algorithm. To further improve the accuracy and robustness, HOG features of the binary and optical flow graphs could be spliced behind the aforementioned features.

Action recognition. To more effectively classify the new features we presented, in this paper, we focused on the basic theory and technology of SVM, and introduce MIL into SVM to solve the problem of optimal hyperplane selection in multiple-class classification problem to obtain better classification results. Last we tested our method and compared with the existing ones on KTH dataset and actual indoor videos, and got the results in Section 3.

2.2 Local Feature of Hierarchical Weighted Spatio-Temporal Interest Points (LFHWSTIPs)

2.2.1 Interest Points Detection

In this paper, we used Harris corner detection to compute the interest points in time-space domain on the motion history image, and then using the idea of HOG, we layered the detected interest points by their motion direction. The weights assigned to the point were determined by calculating the Mahalanobis distance of the other points in the neighborhood relative to it. Finally, we used the direction gradient histogram to describe the feature distribution of the motion sequence. Tian Yingli et al [18]. used structural similarity to determine the weight, but the structural similarity can not reflect the influence of the pixel intensity on the weight, so we adopted Mahalanobis distance. The method above can not only take the feature information on the spatial domain into account, but also gather the feature information on the time domain, which can fully describe action features. As is shown in Figure 1, the left is the method used herein, the right illustrates the method used by Laptev, and the red dot is the spatio-temporal interest points detected.

Figure 1. Interest point detection result in two methods

Harris proposed Harris Corner [19] detection method in 1988, which is the corner detection algorithm with great effect until now, not only overcomes the uncertainty of the camera angle, but also performs well in environment with many corner points and light influence.

The corner point is defined as the pixel point with the maximal curvature in the image edge curve or the pixels whose intensity value changes significantly in the image. The use of corner point not only reduces the amount of information extracted from the whole image, but also greatly preserves the important information of the original image. Harris algorithm principle is, if the gray scale value of a pixel point changes greatly after minimally deviating to pixels in any direction around, then the pixel is the Harris corner [20]. Here, we used the autocorrelation function in signal processing to describe the degree of gray-scale variation, and defined its variation as Eq. (1).

$$\begin{aligned} E_{x,y} &= \sum E_{u,v} \left| I_{x+u,y+v} - I_{u,v} \right| \\ &= \sum W_{u,v} \left| x\frac{\partial I}{\partial X} + y\frac{\partial I}{\partial Y} + o\left(\sqrt{u^2+v^2}\right) \right|^2 \end{aligned} \quad (1)$$

Traditional corner detection has too many corner points, which is difficult to calculate and has high redundancy. In this paper, the Principal Component Analysis (PCA) was used to obtain the most important corner points of the vector matrix, the samples with high correlation and little contribution to the result but causing high time consumption were removed.

However, using the Principal Component Analysis, some of the samples which should be closely related become scattered, so the accuracy of the algorithm has a relatively large impact. In this paper we used the motion feature to filter out the irrelevant corners, and tried to minimize the effect on the human body action area. This was done by first doing Harris corner detection on the original image, then calculating the Motion History Image (MHI), and next, using the MHI to filter the Harris corners. Since the MHI exists only in the regions where the human body is moving, we can filter out other corners in the image that are is not related to human action.

2.2.2 Global Filter

After the spatio-temporal interest points were obtained, the features required for classification could be calculated according to the position of interest points. In order to eliminate the errors caused by the calculation of MHI in complex scenes, we carried out morphological filtering operations on motion history image(MHI) and binary image (Mask) simultaneously. Dilated, and then Eroded, which can eliminate some of the isolated movement direction. The results are shown in Figure 2.

Figure 2. Morphological filter, the left is before filter, the right is after filter

The role of corrosion is to remove small bumps, boundaries, small isolated points, so that the boundary points can contract inside the object, so the image can express the main part more clearly. The effect of expansion is to merge all the background points connected with the object into the object and extend the boundary points of the object to outside. In this example, by using the morphological operation on the original image, we removed the back and other parts of human body that is irrelevant to main action, and the filtering effect is shown good.

2.2.3 Hierarchical Weighting

In the Motion History Image, we first determined the square area of x * x pixels centered on the point of interest. This area was the subject of local filtering. As shown in Figure 3, the red and green blobs are far away from the interest point P0, so for P0 to belong to what type of action, their impact is less than the other two blobs in the dashed box.

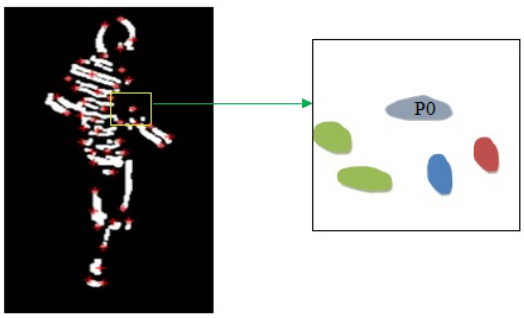

Figure 3. Extract local from global

When calculating the Histograms of Oriented Gradient (HOG), we needed to divide the image according to the motion direction. In this paper, the MHI was filtered in eight directions, with each direction ranging from $n*(+22.5°)(n=1,2,...,8)$. In each region represented by a interest point, in each direction the weighting value between each pixel point and the interest point was calculated to determine whether strengthening or weakening the effect. The range of the region was the same as the range when calculating the HOG features of the MHI. Next introduce how to calculate the weight value of each block. P0 represents the interest point, B represents the P0 block, and d (P0, B) represents the minimum distance between P0 and points in B block. As shown in the Eq. (2).

$$d(p0,B)=\min_{p\in B} d(p_0,p) \quad \textbf{(2)}$$

Use Wx, Wy represent local window size, then the maximum distance between any points and P0 are $\sqrt{W_x^{\ 2}+W_y^{\ 2}}/2$. Then for any pixel points $p \in B$, define its weight value in the block as Eq. (3). s (p) has a value between 0 and 1, and if a pixel does not belong to any block, then define its weight value as 0. This value is used in the normalization step in calculating directional gradient histogram for MHI.

$$s(p)=1-\frac{2d(p_0,B)}{\sqrt{w_x^2+w_y^2}} \quad \textbf{(3)}$$

2.2.4 Feature Extraction

The Histograms of Oriented Gradient (HOG) features show good accuracy and robustness in pedestrian detection and motion recognition and are widely used in this field. In this paper, based on HOG features, we improved its calculation method, and expected to achieve a more accurate result on human action recognition.

At the CVPR conference in 2005, Navneet Dalal and Bill Triggs presented the HOG feature [13] for the first time. The main principle of the HOG feature is that the gradient or the edge direction of local object can describe object's appearance and shape, so we can statistic the local object gradient information, which mainly exists in the edge region. Its performance is much better than single appearance feature, motion feature and spatio-temporal interest point feature, and the behavior description ability is stronger and more robust in practical application.

The calculation steps for HOG are as follows:

Normalization of color space and gamma space. We can convert the image into a gray-scale one first, then normalize it. Normalized compression processing can effectively reduce local shadows and illumination variations.

Gradient calculation. Calculate the gradient of each pixel of the image, including the size and direction of the gradient, which can not only capture the contours, silhouette and some texture information, but also further weaken the impact of light. The gradient size of the (x, y) pixel in the image is calculated as Eq. (4).

$$\Re(x,y)=\sqrt{(I(x+1,y)-I(x-1,y))^2+(I(x,y+1)-I(x,y-1))^2} \quad \textbf{(4)}$$

I (x, y) denotes the pixel value of the image at (x, y), and the direction of the gradient is shown in Eq. (5).

$$Ang(x,y)=\arccos(\ I(x+1,y)-I(x-1,y)/\Re(x,y)) \quad \textbf{(5)}$$

Gradient histogram calculation in cell. The purpose of this step is to calculate gradient histograms within each cell, that is, mapping the gradient information for each pixel into the corresponding direction. In this paper, we used 8 bin histograms to compute the gradient information of 3 * 3 pixels in each cell, and each pixel in the cell voted for the bin of this direction based on weight, generally the weight is the value of the gradient. We used the Mahalanobis distance to determine the voting weight, and the voting results formed a histogram, which is the feature descriptor of the cell. So the number of feature vectors in the cell is 3 * 3. According to the direction of the gradient, 8 bin was divided as shown in Figure 4.

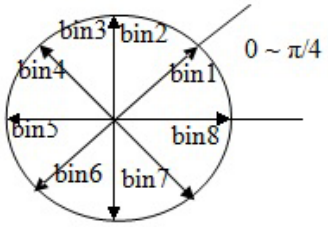

Figure 4. 8 bin divided by gradient direction

Histogram normalization within the block. Because of the influence of local illumination change, the variation range of the gradient value in each cell is large. So normalization operation is needed to avoid the emergence that large features dominate the entire feature.

The specific operation is combining some small cells into a relatively large, spatially connected block. We used 3*3 cells, the feature vector of the block is obtained by concatenating the HOG eigenvectors of 3 * 3 cells. The number of eigenvectors in each block is 72 (8 * 3 * 3), and then normalized the feature vector.

HOG feature collection. Collect the HOG signatures of all blocks in the detection window.

In this paper, the HOG feature would be extracted based on the mask and motion history map (MHI). We mainly used the mixed features of HOG-Mask and HOG-MHI feature descriptors to recognize human

action, which made full use of the robustness of HOG features and the abundant temporal information in MHI. In Mask and MHI, gradient histogram the appearance feature and the motion feature of a partial detection window (W_x, W_y) whose center point is interest point. The region would be further divided into many (n_x, n_y) grids, normalize the histograms of all the grids, and spliced them up to be HOG feature and HOG-MHI eigenvector. In the calculation of HOG, regardless of direction makes its effect more robust. However, direction must be taken into account for it is very important information in moving images while calculating the HOG-MHI.

After global filtering the MHI, the HOG feature can be calculated. Firstly, decompose the filtered MHI image into 8 different directions (8 bin in the histogram). As shown in Figure 5, each bin falls within the interval of $\left((n-2)*22.5,n*22.5\right)(n=1,2,...,8)$.

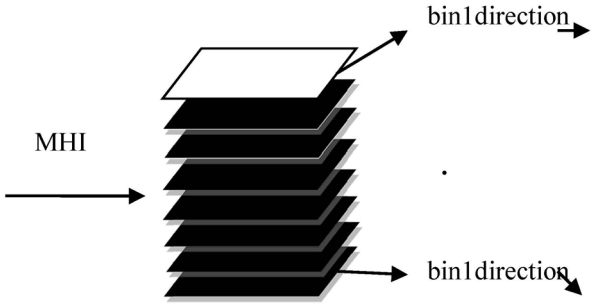

Figure 5. The filtered MHI is decomposed into an 8-layer image

Then local filtering operation was carried out on each level of image to eliminate the changes caused by shadows, illumination and small noises. The specific way was to strengthen or weaken, or even remove the local area, to do which operation was determined by calculating the Mahalanobis distance between the interest points and each pixel in the detection window in each direction. For example, in the detection window of an interest point in bin1 hierarchical graph, calculate the structural similarity of between each point in the window and the interest point, and take the structural similarity value as the weight of the normalized operation when calculating the HOG features. While calculating, a Connected component analysis operation will be performed on the detection window area to obtain the connected domain (called Blob) in the detection window. Figure 6 shows the connected domain (blob) of a detection window in the bin 1 hierarchical graph.

The connected domains that are closer to the interest point are more likely to belong to the same area as the interest point, so these connected domains (the blue area in Figure 6) should be strengthened; Similarly, the regions father away (red and green regions in Figure 6) should be weakened. For how to strength or weaken, we compared two methods in this paper, generally, although the computational complexity of Mahalanobis distance is greater than structural similarity, the performance of Mahalanobis distance is superior to structural similarity, and structural similarity is calculated by Euclidean distance, it can not fully reflect a variety of information (distance and intensity). In this paper, we calculated the weight for 1000 blobs of 5*5 in two methods, and found that the time when using Mahalanobis distance is 20% more than using the structural similarity, but the weighted effect is more advisable, so we adopted Mahalanobis distance.

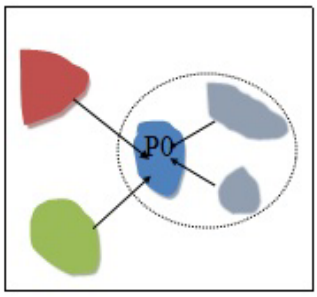

Figure 6. Blob analyzation in a detection window that interest point correspond

After the above steps, we could get the weight value used when normalized HOG-MHI features. Firstly, extract the small window corresponding to each Harris corner in the MHI image after global filtering. In this paper, we made the size of the window 27 * 27 pixels as the benchmarks, and divided the window into 3 * 3 pixels in a cell. In order to further improve the performance, 3 * 3 cells could form a large domain (block), as shown in Figure 7, do the contrast normalization operation in the domain, which can get a good result with light and shadows.

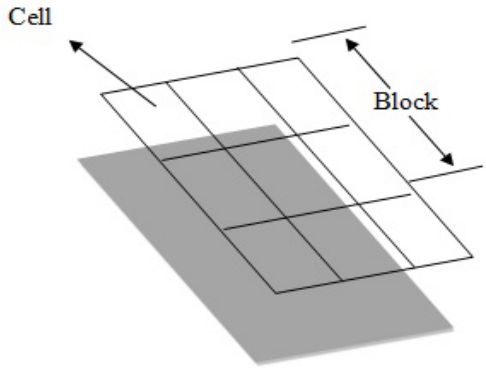

Figure 7. 3*3 cells form a block

Then constructed a gradient histogram for each cell, each pixel in the cell did weighted voting for some direction-based histogram, the weight was the one calculated by the Mahalanobis distance above. After this operation, we obtained an 8-dimensional HOG eigenvector X = (x1, x2, ..., x8). After the histogram of each cell was obtained, 3 * 3 cells formed a block, then the eigenvectors became 72 dimensions X = (x1, x2, ..., x72),

then do the normalization operation to the HOG features of the block.

The calculation process of HOG-Mask was similar to that of HOG-MHI, HOG-Mask was to extract HOG feature from binary image, and for the binary map is less sensitive to subtle changes of directions than MHI, so here we only constructed 6 direction-based histogram channels. And the weight is different from the HOG-MHI calculation, it is the amplitude of the pixel gradient. The gradient magnitude of the (x, y) pixel in the image is shown in Eq. (6).

$$G(x,y)=\sqrt{(H(x+1,y)-H(x-1,y))^2+(H(x,y+1)-H(x,y-1)^2)} \quad \textbf{(6)}$$

After voting, the 6-dimensional HOG eigenvector $X2 = (x1, x2, ..., x6)$ was obtained. Similarly, in order to further improve the performance of the algorithm, to reduce the impact of light changes an the shadow on the detection results, 3 * 3 formed a block, and after normalization, the HOG eigenvector of 54 (3 * 3 * 6) dimension $X2 = (x1, x2, ..., x54)$ was obtained.

By combining the HOG-MHI and the HOG-Mask features, a new HOG feature was obtained. The feature vectors were 126 dimensions, $X = (x1, x2, ..., x126)$.

In order to eliminate the error caused by the size changing of detected objects, we could make a multiple-scale processing to the detection window. In this paper, we increased the detection window size from 27 pixels, and increased 3 pixels each time, until the window size was 54 pixels. The disadvantage of multiple-scale scaling is that the computational complexity would be greatly increased by scale changing.

2.3 Multiple-Instance Support Vector Machine (MI-SVM)

SVM model is a supervised learning model. One of the shortcomings of the supervised learning model is that in many cases training can not fully know the classification of each sample [21]. Multiple Instance Learning (MIL) is an overview of the supervised classification of classes in a collection of samples or bags. It provides a new modeling method to overcome the shortcomings of supervised learning model, which divides the training sample data into multiple pattern sets to construct the classification model, rather than the individual model. By defining sets with all positive patterns as positive samples, and defining sets as negative while there is one pattern is negative in it, it strictly limits the boundaries between the classification faces and improves the accuracy of the classification. Therefore, the most important challenge in MIL is to deal with ambiguities caused by not knowing which samples in the sample bags are positive and negative.

In 1997, T. G. Dietterrich proposed the concept of Multiple Instance Learning when studying drug activity. By learning about molecules that are known to be applicable and not applicable to the manufacture of drugs, it is possible to predict whether other new molecules are suitable for manufacturing drugs. However, there would be a lot of noise using traditional supervised learning algorithm.

Although SVM has good classification performance in small sample size, the classification effect is not ideal when the feature distribution is dense and the distance between classes is not obvious. So we focused on the basic theory and technology of SVM, and introduce MIL into SVM to solve the problem of optimal hyperplane selection in multiple-class classification problem to obtain more correct classification results.

Next we would introduce two kinds of modified SVM classification algorithm using MIL theory.

Multiple Instance Learning is a derivative of supervised learning classification, which models training sample data as a collection of patterns (commonly called Bag) rather than a separate pattern. Although each schema can have the same related class, it is assumed that the class that fetches each schema can only be accessed indirectly through its associated Bag. When in a Bag at least one pattern has a class, the Bag is marked for that class. For example, for a given set of eigenvectors $x1, x2, ..., xN$, group them into bags B_1, B_2, B_m, $Bm = \{Xi: i \in Im\}$ where $Im \in \{1, ..., N\}$. And each bag Bi corresponds to a class Ym, rather than each pattern corresponds to a class. $Ym = 1$ indicates that the bag is a positive sample and at least one pattern in the packet is a positive sample. Similarly, $Ym = -1$ means that the packet is negative and all the patterns in the packet are negative. If y_i is used to denote the class of each pattern, it can be represented by the mathematical formula in (7). SVM based on MIL can be described as a minimization of the objective function in Eq. (8) under constraint in Eq. (7).

$$\sum_{i\in I}\frac{y_i+1}{2}\geq 1, \forall I_{s.t}\ Y_I=1, y_i=-1, \forall I_{s.t.}, Y_I=-1. \quad \textbf{(7)}$$

$$\min_{yi}\ \min_{w,b,\xi}\ \frac{1}{2}||w||^2+c\sum_i \xi_i \quad \textbf{(8)}$$

Under the constraint condition in Eq. (9).

$$y_i(w^T+b)\geq 1-\xi_i, \quad \xi_i\geq 0, y_i(-1,1) \quad \textbf{(9)}$$

The identity y_i of the sample xi in the positive sample bag is treated as an unknown integer. In the mi-SVM, the soft-margin must be maximized, and the corresponding optimal hyperplane is found as the classification surface.

Another SVM based on multiple instance is to extend finding the interval between separate samples to finding the interval between bags, aiming at directly maximizing the interval between bags. Define the hyperplane classification equation between the bags as

Eq. (10).

$$\gamma_I = Y_I \max_{i \in I} (< w, xi > + b) \quad \textbf{(10)}$$

As can be seen from the above expression, the function used in predicting the identity of a sample bag is Eq. (11).

$$Y_I = \operatorname{sgn} \max_{i \in I} (< w, x_i > + b) \quad \textbf{(11)}$$

Using the above classification interval, we define the classification interval equation based on MIL as Eq. (12).

$$\min \min(\frac{1}{2} \|w\|^2 + C \sum \xi_I) \quad \textbf{(12)}$$

Under the constraint in Eq. (13), find the minimum of Eq. (12).

$$\begin{aligned} &\forall I : Y_I = -1 \wedge -(w, x_i) - b \geq 1 - \xi_I, \forall i \in I \\ &or \quad Y_I = 1 \wedge (w, x_{s(I)}) + b \geq 1 - \xi_I, \xi_I > 0 \end{aligned} \quad \textbf{(13)}$$

If you are more concerned about the classification of new test bags, MI-SVM is more effective. If more concerned with the classification of a new separate sample, then the traditional SVM is more suitable.

The purpose of SVM is to find a minimum of W. The traditional SVM is a quadratic programming problem, in this paper, Eq. (7) include the optimization process under condition in Eq. (8), so the classification ability of SVM is improved.

In order to verify the performance of MI-SVM is higher than that of traditional SVM actually, this paper used LibSVM and MI-SVM to test the KTH dataset respectively. The experimental results are shown in Table 1. We can see from the experimental results, the results using MI-SVM are much more accurate. Therefore, this paper would use MI-SVM classification algorithm to classify all the features and innovative features.

Table 1. Comparison of performance of mi-SVM and MI-SVM on KTH dataset

Method	Accuracy
mi-SVM	84.2%
MI-SVM	86.7%

3 Experiments and Results

Figure 8 showed our complete experimental procedure.

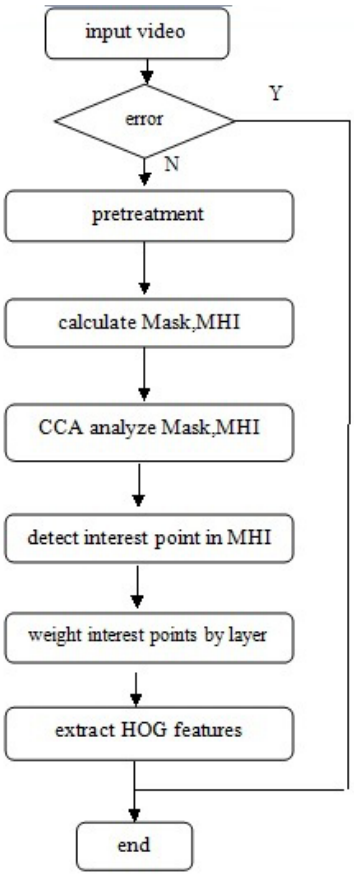

Figure 8. Experimental procedure of local feature of hierarchically weighted spatio-temporal interest points.

In order to verify the reliability and practicability of the proposed method, we designed four different experimental groups and compared with the proposed in this paper. The four schemes and the proposed scheme were tested on the KTH dataset [22]. The performance results are shown in Table 2.

Table 2. Comparison of recognition results of four schemes

Scheme	Features	Classification Algorithm
Scheme 1	Appearance features	SVM
Scheme 2	Motion features	SVM
Scheme 3	spatio-temporal interest points features	SVM
Scheme 4	Appearance features + Motion features + spatio-temporal interest points features + HOG features	SVM
Our scheme	HOG features of layered weighted spatio-temporal interest points	SVM

Next we tested the four schemes on a unified standard test set, Table 3 shows the preliminary results obtained on the current test set.

Table 3. Accuracy of five schemes

Scheme	Accuracy
Scheme 1	82.32%
Scheme 2	80.70%
Scheme 3	85.56%
Scheme 4	86.21%
Our scheme	88.19%

And the testing results schematic of the five schemes are shown one by one in Figure 9.

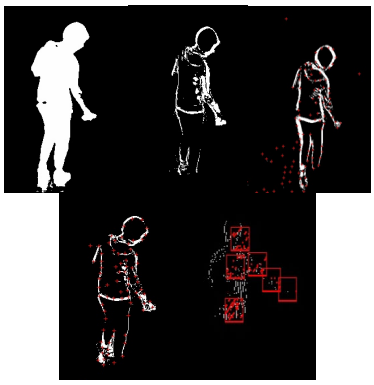

Figure 9. Five schemes' testing results, from left to right is from scheme 1 to our scheme

In order to show the effect of the five features more intuitively and detailedly, we used the PR curve to describe the recognition performance, as shown in Figure 10.

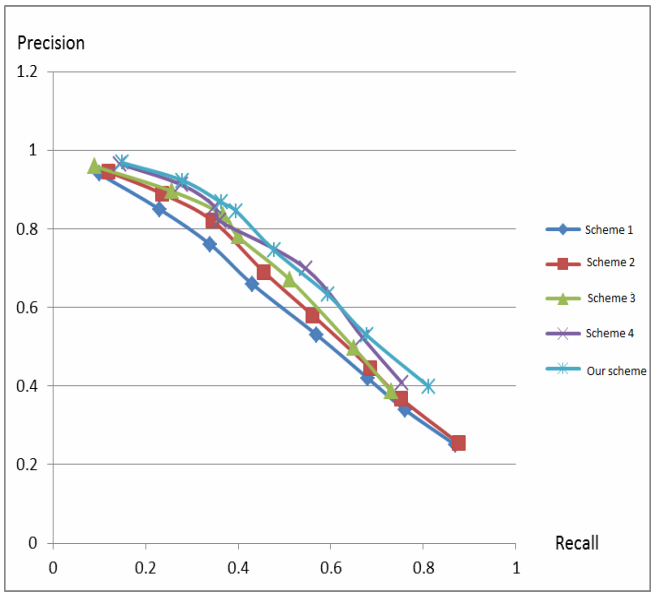

Figure 10. The PR curve of five schemes

It can be seen that the expression ability of the features used in our scheme is superior to that of the other four schemes, which indicates that the feature expression method we proposed is not only feasible but also has improved performance.

In order to further evaluate the performance of our scheme, we proposed two compromise schemes for the training process to compare. The new scheme combines the high efficiency of scheme one and the high performance of our scheme. One is to carry out the serial classification of scheme one and our scheme, the other is parallel classification. Serial scheme is using scheme one first, and then use our scheme to deal with the samples that can't be classified by scheme one. The detailed operations and results are shown in Figure 11. Parallel program is using scheme one and our scheme for all samples at the same time, and the do the Or operation to the results of the two schemes. The detailed are shown in Figure 12.

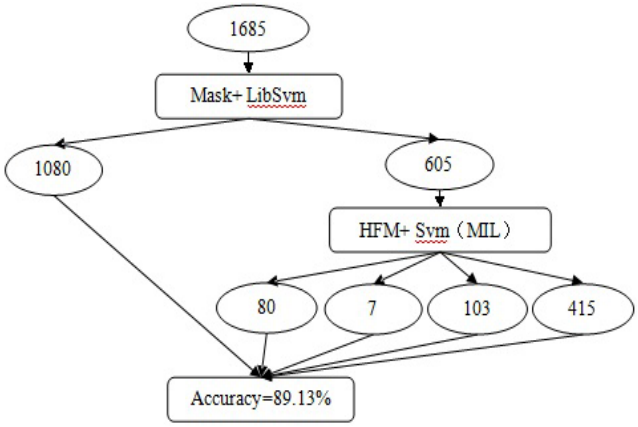

Figure 11. The serial flow and result

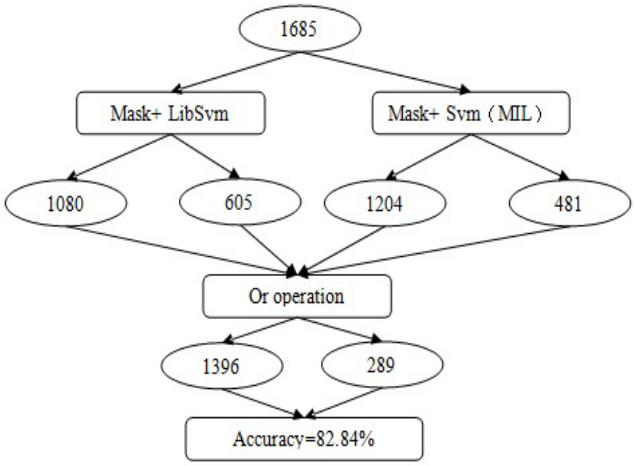

Figure 12. The parallel flow and result

As can be seen from the classification and recognition results of the above two schemes, the serial scheme is superior to the parallel scheme in general. Because most of the samples in the serial scheme are classified using the appearance feature, the time complexity is low. While the parallel scheme must do two operations to all the samples, which makes the time complexity very high. So in this paper, we finally chose the serial scheme for human action recognition.

4 Conclusion

This paper focused on the commonly used features and classification learning algorithm of human action recognition at home and abroad. Based on the local features of spatio-temporal interest points and multiple instance classification algorithm, we studied how to extract the feature expression which has stronger ability of expression and higher robustness. The main innovations of this paper are:

Local feature of hierarchically weighted spatio-temporal interest points. By analyzing the commonly used image features, we compared the performance of different features on the same test data set. On this basis, we presented a new feature extraction algorithm, layered weighting the spatio-temporal interest points and fusing various features. And the experimental results showed its availability and efficiency.

MI-SVM. We improved the traditional SVM classification algorithm by introducing MIL into it. MIL is to improve the classification ability by maximizing the classification interval between different feature spaces. Because the false alarm rate is required to be very low in the home intelligent monitoring, the MIL theory can be used to minimize the missed alarm rate.

The research we did in this paper is human action recognition based on family intelligent monitoring, mainly used in individual families and nursing homes for the elderly. Based on the dangerous degree of detected action, the system determines whether to perform an immediate alarm, and then notifies the intelligent robot or human to help. This would eliminate the need to guard the old and disabled all the time and save a lot of manpower costs.

Acknowledgments

This work was supported by Scientific Research Project of Beijing Institute of Technology (Grant No. 2017 CX02029).

References

[1] G. Mori, J. Malik, Recovering 3D Human Body Configurations using Shape Contexts, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 28, No. 7, pp. 1052-1062, July, 2006.

[2] I. Laptev, On Space-time Interest Points*, International Journal of Computer Vision,* Vol. 64, No. 2-3, pp. 107-123, September, 2005.

[3] M. A. R. Ahad, J. K. Tan, H. Kim, S. Ishikawa, Motion History Image: Its Variants and Applications, *Machine Vision and Applications*, Vol. 23, No. 2, pp. 255-281, March, 2012.

[4] C. V. Hutchinson, T. Ledgeway, Responses of First-order Motion Energy Detectors to Second-order Images: Modeling Artifacts and Artifactual Models, *Perception*, Vol. 34, No. 1_Supplement, pp. 123-124, August, 2005.

[5] A. Arinaldi, M. I. Fanany, Kinematic Features For Human Action Recognition Using Restricted Boltzmann Machines, *4th International Conference on Information and Communication Technology,* Bandung, Indonesia, 2016, pp. 1-6.

[6] E. Zhang, W. Chen, Z. Zhang, Y. Zhang, Local Surface Geometric Feature for 3D Human Action Recognition, *Neurocomputing*, Vol. 208, pp. 281-289, October, 2016.

[7] C. Liang, E. Chen, L. Qi, L. Guan, Improving Action Recognition Using Collaborative Representation of Local Depth Map Feature, *IEEE Signal Processing Letters*, Vol. 23, No. 9, pp. 1241-1245, September, 2016.

[8] A. Holzinger, Data Mining with Decision Trees: Theory and Applications, *Online Information Review*, Vol. 39, No. 3, pp. 437-438, 2015.

[9] H. B. Jaafar, N. B. Mukahar, D. A. B. Ramli, A Methodology of Nearest Neighbor: Design and Comparison of Biometric Image Database, *2016 IEEE Student Conference on Research and Development (SCOReD),* Kuala Lumpur, Malaysia, 2016, pp. 1-6.

[10] L. R. Rabiner, A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition, *Proceedings of the IEEE*, Vol. 77, No. 2, pp. 257-286, February, 1989.

[11] K. P. Murphy, *Dynamic Bayesian Networks: Representation, Inference and Learning*, University of California, 2002.

[12] S. Andrews, I. Tsochantaridis, T. Hofmann, Support Vector Machines for Multiple-Instance Learning, *Advances in Neural Information Processing Systems 15 (NIPS 2002)*, Vancouver, British Columbia, Canada, 2002, pp. 561-568.

[13] N. Dalal, B. Triggs, Histograms of Oriented Gradients for Human Detection, *IEEE Computer Society Conference on Computer Vision & Pattern Recognition,* San Diego, CA, 2005, pp. 886-893.

[14] B. Babenko, M.-H. Yang, S. Belongie, Robust Object Tracking with Online Multiple Instance Learning, *IEEE Transactions on Pattern Analysis & Machine Intelligence*, Vol. 33, No. 8, pp. 1619-1632, August, 2011.

[15] V. Thanikachalam, K. K. Thyagharajan, Human Action Recognition using Accumulated Motion and Gradient of Motion from Video, *Third International Conference on Computing Communication & Networking Technologies*, Coimbatore, India, 2012, pp. 1-6.

[16] B. Jian, B. C. Vemuri, Robust Point Set Registration Using Gaussian Mixture Models, *IEEE Transactions on Pattern Analysis & Machine Intelligence*, Vol. 33, No. 8, pp. 1633-1645, August, 2011.

[17] R. G. Brereton, The Mahalanobis Distance and Its Relationship to Principal Component Scores, *Journal of Chemometrics*, Vol. 29, No. 3, pp. 143-145, March, 2015.

[18] Y. L. Tian, A. Hampapur, Robust Salient Motion Detection with Complex Background for Real-Time Video Surveillance, *Application of Computer Vision*, Breckenridge, CO, 2005, pp. 30-35.

[19] C. J. Harris, A Combined Corner and Edge Detector, *Proc*

Alvey Vision Conf, Vol. 23, No. 3, pp. 147-151, February, 1988.

[20] Y. Han, P. Chen, T. Meng, Harris Corner Detection Algorithm at Sub-pixel Level and Its Application, *International Conference on Computational Science and Engineering (ICCSE 2015)*, Shandong, China, 2015, pp. 133-137.

[21] C.-C. Chang, C.-J. Lin, LIBSVM: A Library for Support Vector Machines, *ACM Transactions on Intelligent Systems and Technology*, Vol. 2, No. 3, Article No. 27, April, 2011.

[22] Z. Gao, H. Zhang, A. A. Liu, G. Xu, Y. Xue, Human Action Recognition on Depth Dataset, *Neural Computing & Applications*, Vol. 27, No. 7, pp. 2047-2054, October, 2016.

Biographies

Chun Shan received her Ph.D. Degree in Software Engineering from Beijing Institute of Technology. She is a lecturer and master tutor of school of software of Beijing Institute of Technology. Her research interests include Artificial Intelligence and Network Security.

Liyuan Liu is a master student of school of software of Beijing Institute of Technology. Her research interests are Artificial Intelligence and Software Security.

Jingfeng Xue received her Ph.D. Degree in Software Engineering from Beijing Institute of Technology. He is the vice dean, professor and doctoral supervisor of school of software of Beijing Institute of Technology. His research interests include Artificial Intelligence and Network Security.

Zhaoliang Sun received his master's degree in Software Engineering from Beijing Institute of Technology in 2017, he is a software engineer at the China Minsheng Bank Science and Technology Derpartment. His research interests are Pattern Recognition and Data Analysis.

Tingping Ma received his master's degree in Software Engineering from Beijing Institute of Technology in 2013. His research interest is Pattern recognition.