Journal of Computers Vol. 33 No. 1, February 2022, pp. 63-70
doi: 10.53106/199115992022023301007

Sports Motion Feature Extraction and Recognition Based on a Modified Histogram of Oriented Gradients with Speeded Up Robust Features

Jing Zhao*

Department of Art and Sports, Huanghe S&T University, Zhengzhou City, 450000 Henan Province, China
910675024@qq.com

Received 27 September 2021; Revised 15 October 2021; Accepted 27 October 2021

Abstract. Traditional motion recognition methods can extract global features, but ignore the local features. And the obscured motion cannot be recognized. Therefore, this paper proposes a modified Histogram of oriented gradients (HOG) combining speeded up robust features (SURF) for sports motion feature extraction and recognition. This new method can fully extract the local and global features of the sports motion recognition. The new algorithm first adopts background subtraction to obtain the motion region. Direction controllable filter can effectively describe the motion edge features. The HOG feature is improved by introducing direction controllable filter to enhance the local edge information. At the same time, the K-means clustering is performed on SURF to obtain the word bag model. Finally, the fused motion features are input to support vector machine (SVM) to classify and recognize the motion features. We make comparison with the state-of-the-art methods on KTH, UCF Sports and SBU Kinect Interaction data sets. The results show that the recognition accuracy of the proposed algorithm is greatly improved.

Keywords: sports motion feature extraction and recognition, HOG, SURF, direction controllable filter, K-means clustering

1 Introduction

Motion recognition has attracted extensive attention from researchers in the fields of computer vision, machine learning and pattern recognition [1,2]. In addition, it has a wide range of applications in intelligent surveillance, augmented reality, video annotation, human-computer interaction and motion-sensing games, etc, [3]. Therefore, using computer vision algorithms to automatically identify human behavior in the video has become a hot research topic in the field of computer vision in recent years.

Existing motion recognition methods can be divided into two categories: local feature methods and global feature methods. Local feature methods, such as optical flow, scale-invariant feature transform (SIFT), are generally used to extract local sub-regions or points of interest from video or images [4]. Global feature methods usually extract the information of the whole human body, such as motion history map (MHI), contour features, etc. The representation of each feature has its advantages and disadvantages. Nowadays, the more popular research methods mainly combine local and global features.

Reference [5] not only improved the speed of recognition but also maintained the accuracy of the algorithm by integrating space-time domain features at decision level, but the algorithm was more sensitive to occlusion factors. Reference [6] introduced the concept of motion context to obtain the spatial and temporal distribution of video words. Although these algorithms have achieved good performance, they all lack gradient feature information. Reference [7] proposed to use Harris corner detector to detect interest points in space-time 3D space, and use these interest points to describe related actions. In view of the defects in the reference [7], the original algorithm was improved by Gabor filtering in the time-space domain in the reference [8]. Aiming at the problem that the HOG feature was not sufficient to describe the image edge, reference [9] used the improved HOG feature to better extract the edge information of buildings, which was superior to the traditional method in multiple indexes. In reference [10], dense sampling was performed on behavioral video to obtain dense trajectory (DT) of behavioral actions. Then, motion boundary histogram, optical flow histogram and gradient direction histogram of DT were calculated respectively, and the three features were fused. This method had a strong ability to express infrared human behavior. In reference [11], the deep learning method was adopted to carry out weighted fusion of features extracted from each network, and the algorithm had a high recognition rate and strong generalization ability, but the model was complex.

* Corresponding Author

According to the above analysis, in order to overcome the problem of insufficient feature description by a single feature, this paper proposes a new feature characterization method for motion recognition, which integrates SURF and improved HOG features to recognize motions.

2 Proposed Motion Recognition Method

2.1 SURF Feature Extraction

SURF [12] algorithm mainly includes two parts: feature point localization and feature point description.

(1) Feature point localization.

The first step is to construct the Hessian matrix. The Hessian matrix of any pixel point $X(x, y)$ in the image I is:

$$H(X,\sigma)=\begin{bmatrix} L_{xx}(X,\sigma)L_{xy}(X,\sigma) \\ L_{xy}(X,\sigma)L_{yy}(X,\sigma) \end{bmatrix} \quad . \qquad \textbf{(1)}$$

Where σ is the scale. L_{xy} (X, σ) is the convolution of the Gaussian second-order differential $\frac{\partial^2 g(\sigma)}{\partial x \partial y}$ between pixel point X and image I. L_{xx} (X, σ) is similar to L_{yy} (X, σ). $g(\sigma)$ is the Gaussian function, its expression is:

$$g(\sigma)=\frac{1}{2\pi\sigma^2}\exp(-\frac{x^2+y^2}{2\sigma^2}) \quad . \qquad \textbf{(2)}$$

D_{xx}, D_{xy}, and D_{yy} are used to represent the convolution result of the box filter and the image in different directions, then the determinant of the simplified Hessian matrix can be expressed as:

$$\det(H_{approx})=D_{xx}D_{yy}-(\varepsilon\cdot D_{xy})^2 \quad . \qquad \textbf{(3)}$$

Where $\varepsilon=0.9$ is the empirical value, which is used to balance the error caused by the box filter approximation. All pixel points processed by the Hessian matrix are compared with 26 points in the 3×3×3 stereo neighborhood with non-maximum values. When the determinant of the Hessian matrix obtains a local maximum or minimum value, the extreme point is considered as a candidate feature point. Then, linear interpolation is performed in the scale space and image space to obtain the final stable feature points.

(2) Feature point description.

It creates a square area of 20s (s is the scale value of the feature point) near the point of interest and rotates them to the main direction. Then it divides the square region into 4×4=16 sub-regions, and calculates the Haar wavelet feature of 5×5=25 pixels for each sub-region. The Haar wavelet feature includes the sum of the horizontal Haar wavelet response values, the absolute value sum of the horizontal Haar wavelet response values, the sum of the vertical Haar wavelet response values, and the absolute value sum of the vertical Haar wavelet response values. So each sub-region has four values, each feature point is a vector with 16×4=64-dimension.

2.2 Constructing a Visual Dictionary

In order to generate the visual dictionary, the most commonly used K-means clustering algorithm (K represents the dictionary size) is used in this paper. The Euclidean distance is expressed as:

$$E_{ss}=\sum_{i=1}^{k}\sum_{p\in C_i}|\,p-m_i\,|^2 \quad . \tag{4}$$

In the formula, E_{ss} is the clustering error of all objects. p is a point in space. m_i is the average value of cluster C_i. In this paper, all samples are divided into the cluster represented by the nearest cluster center by Euclidean distance, and K clusters are formed. Then the K clusters are calculated again to get the new clustering center, and the sample categories are reclassified according to the new clustering center. The iteration will be stopped until the center position is unchanged. The center of each cluster is used as the word in the visual dictionary. Then the distance of each feature to these k visual words is calculated. They are mapped to the nearest visual word to construct the histogram.

2.3 Modified HOG Feature Extraction

Controllable filter was proposed by Freeman et al [13]. It is a special filter, which can be arbitrarily rotated. It performs well in image texture analysis. The controllable filter is composed of a set of basic filters, i.e.

$$f_\theta(x,y)=\sum_{i'=1}^{M}k_{i'}(\theta)f_{\theta'}(x,y) \ . \tag{5}$$

Where $f_\theta(x,y)$ is the filter in the direction of θ. $k_{i'}(\theta)$ is the interpolation function. $f_{\theta'}(x,y)$ is the basis function. Interpolation function and basis function exist in pairs. M is the number of basis functions.

The second derivative of the Gaussian function approximates the boundary and can be expressed by the product of a circularly symmetric window function and a polynomial. The controllable filter is composed of the second derivative of the Gaussian function and can be expressed as:

$$G_2^{\theta}=k_1(\theta)G_2^{0^\circ}+k_2(\theta)G_2^{60^\circ}+[1-k_1(\theta)]G_2^{120^\circ} \ . \tag{6}$$

Where $k_1(\theta)$ and $k_2(\theta)$ are interpolation functions. $G_2^{0^\circ}$, $G_2^{60^\circ}$ and $G_2^{120^\circ}$ represent the second derivatives of the Gaussian function at angles 0, 60 and 120. The interpolation function and the basis function are:

$$\begin{cases} k_1(\theta)=\cos^2\theta \\ k_2(\theta)=-\sin 2\theta' \end{cases} \ . \tag{7}$$

$$\begin{cases} G_2^{0^\circ}=0.92(2x^2-1)\exp[-(x^2+y^2)] \\ G_2^{60^\circ}=1.84xy\exp[-(x^2+y^2)] \\ G_2^{120^\circ}=0.92(2y^2-1)\exp[-(x^2+y^2)] \end{cases} \ . \tag{8}$$

HOG feature was first proposed by Dalal et al and applied to pedestrian detection. The idea of HOG method is to calculate the normalized local directional gradient histogram in a dense grid. The improved HOG feature extraction process is as follows.

First of all, in order to reduce the impact of shadow and light factors, the image needs to be normalized.

Compute the gradients in directions 0° and 90°. Assuming that the coordinate of any pixel point in image I is (x,y), then the gradients $G_x(x,y)$ and $G_y(x,y)$ in 0° and 90° are,

$$\begin{cases} G_x(x,y) = I \times f_{0^\circ} \\ G_y(\mathrm{x,y}) = I \times f_{90^\circ} \end{cases} . \quad \textbf{(9)}$$

The gradient amplitude and gradient direction at the pixel point can be respectively expressed as:

$$\begin{cases} G(x,y) = \sqrt{G_x(x,y)^2 + G_y(x,y)^2} \\ \theta(x,y) = \arctan[G_x(x,y) / G_y(x,y] \ ^2 \end{cases} . \quad \textbf{(10)}$$

(3) In order to maintain a weak sensitivity to the motion in the image, the image is divided into several cells. Each cell is divided into 9 directional blocks. Then, a histogram of gradient direction is constructed for each cell.

(4) Combining the cells obtained in Step 3 into large blocks, and conducting L2-norm normalization for the histogram within the interval. The formula is as follows:

$$L_2(v) = \frac{v}{\sqrt{\| v \|_2^2 + e^2}} \ . \quad \textbf{(11)}$$

Where e is a constant parameter, and its value is smaller. v represents the histogram vector within the interval.

(5) Connecting the histogram vectors within all the intervals, finally obtaining the HOG feature of the whole image.

(6) In addition, the calculation equation of HOG feature dimension is:

$$D_{HOG} = D_{cell} \times \frac{S_{block}}{S_{cell}} \times (\frac{h - S_{block}}{S'_{block}} + 1) \times (\frac{w - S_{block}}{S'_{block}} + 1) \ . \quad \textbf{(12)}$$

Where D_{cell} represents the dimension of cell. S_{block} is the size of the Block. S_{cell} is the size of the cell. S'_{block} is the step size of the Block. h and z represent the height and width of one image respectively.

The improved HOG feature has good robustness to illumination and geometric changes, and it can effectively extract edge texture information.

In this paper, SVM is used to classify data. SVM uses hyperplane in high dimensional space to divide data with maximum margin. The mathematical expression is:

$$\begin{aligned} &\min \frac{\| \omega \|^2}{2} + C \sum_{j}^{n} \xi_j , \\ &s.t. \ y_j(\omega x_j + b) \geq 1 - \xi_j, j \in \{1, \cdots, n\} \end{aligned} \ . \quad \textbf{(13)}$$

Where ω is the normal vector of the hyperplane. C is the penalty factor. n is the number of sample points. ξ is the relaxation variable. b is the offset.

2.4 Specific Implementation of the Proposed Algorithm

The proposed motion recognition process is shown in Fig. 1.

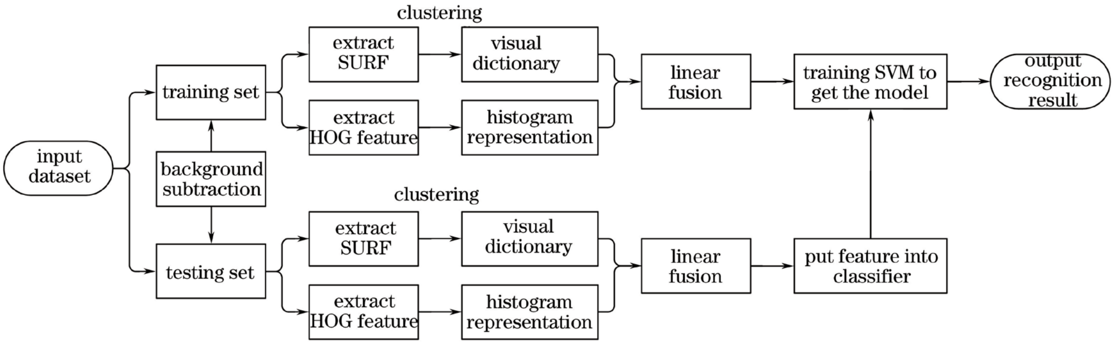


Fig. 1. Flow chart of proposed method

The specific implementation steps of the new algorithm are as follows.

1. Background subtraction is applied to the input video frame. Only the region where the motion is located is extracted from each frame of the video, and this region is called the motion region.

2. Interest-point detection is carried out in the motion area, and SURF is used to describe it. Bag of words (BOW) model is used to construct the histogram of features. The dimension of each feature point is 64.

3. HOG feature extraction is carried out in horizontal and vertical directions for each extracted motion region. The dimension of HOG feature is D_{HOG}, which is 3780 in this paper.

4. It fuses the histogram features constructed by BOW model and HOG features to encode video information. The final motion features are input to the trained SVM classifier to realize motion recognition.

3 Experiments and Analysis

The experimental software environment is Windows10 64-bit, Intel core i5-760 CPU processor, 64G memory, MATLAB R2017a, Python3.6. In this paper, the widely used KTH data set [14], UCF Sports data set [15] and SBUKinect Interaction data set [16] are adopted. KTH data set has six types of motion, and each motion is completed by 25 people. The six motions contain Walking, Jogging, Running, Boxing, and Handwaving and Handclapping. The background of this data set is relatively simple. The UCFSPORTS dataset consists of 150 video sequences from various broadcast sports channels on the BBC and ESPN. This data set contains 10 different types of motions, each motion is performed by different persons. It includes Diving, GolfSwing, Kicking, Lifting, Riding Horse, Running, SkateBoarding, Swing Bench, SwingSide Angle and Walking. The shooting background of this data set is relatively complex with different scales and perspectives, so it brings some challenges to motion recognition. The Sbukinect Interaction data set contains8 types of motions, which are Approaching, Departing, Exchanging, Hugging, Kicking, Punching, Pushing and Shaking hands. Each motion is performed by seven different persons in the same lab.

We take UCF Sports dataset as an example. Fig. 2 shows SURF extraction of some video frames, and Fig. 3 shows the comparison of HOG features before and after improvement for some video frames. The improved HOG feature can fully describe the edge details of the image.

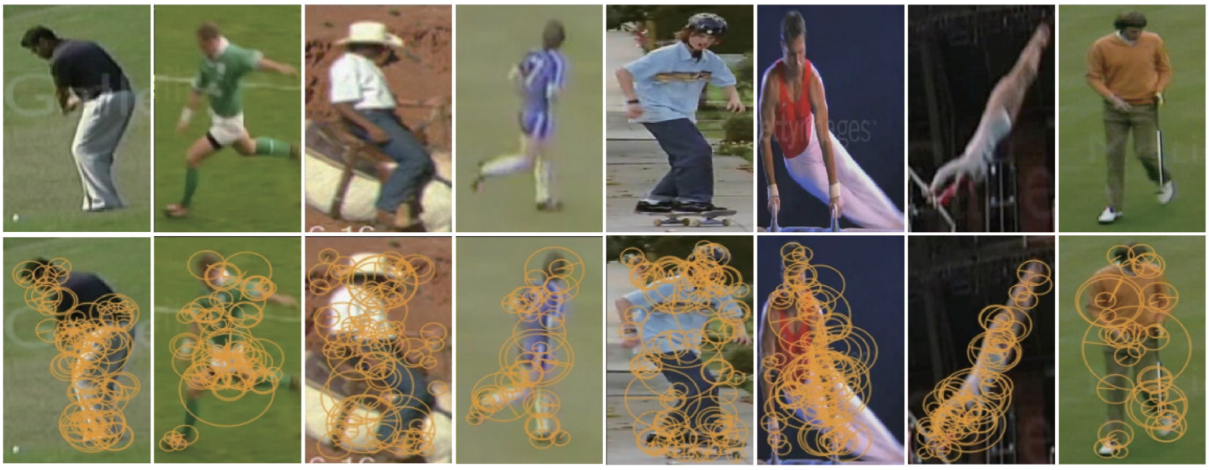


Fig. 2. SURF extraction

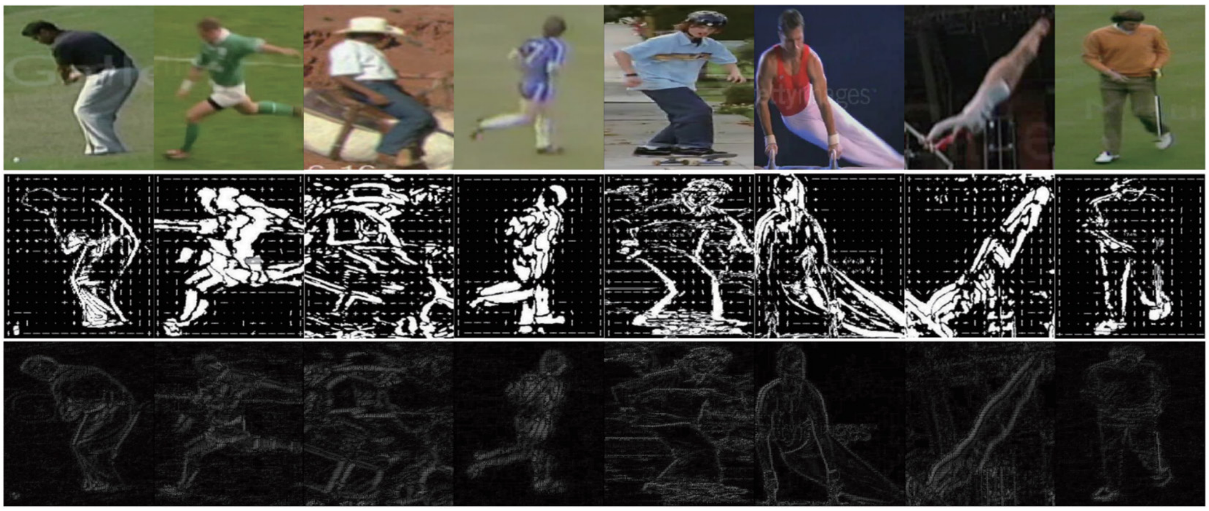
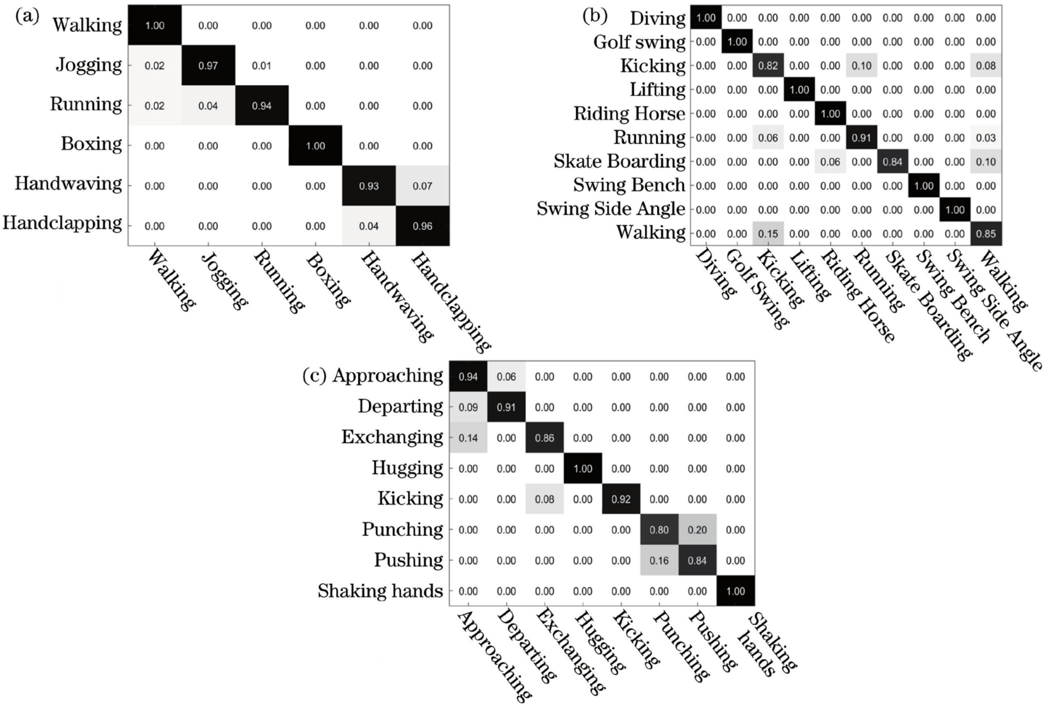


Fig. 3. Comparison of HOG features before and after improvement

The parameters in the simulation are as follows: D_{cell} = 9, S_{block} = 32 × 32, S_{cell} = 16 × 16, S'_{block} = 16, C=1. When k is 1500, 3000 and 1000 respectively, the recognition effect of the three data sets is the best. To verify the performance of the presented method in this paper, the data sets mentioned above are tested. In the experiment, cross-validation is adopted for testing. 70% of the data sets are selected as the training set and the rest as the test set. Finally, the confusion matrixes of recognition accuracy for KTH dataset, UCF Sports dataset and SBU Kinect Interaction dataset are obtained as shown in Fig. 4.

It can be seen from the confusion matrix that there is a certain similarity between “walking”, “jogging” and “running” on the KTH data set, leading to a certain misidentification rate. There is also 4%~7% error between waving and clapping. In the UCF Sports data set, due to the complex background, multiple perspective changes, and occlusion and fast speed between motions, there is 8%~15% misidentification rate between “kicking” and “running” motions. For other motions with occlusion, the recognition rate is higher. In the SBU Kinect Interaction data set, due to the certain similarity between motions, “striking” and “pushing” have high misidentification rates. “approaching” and “exchanging” also have 14% misidentification rates, because the human body needs to be close to each other when exchanging objects. For other obvious features, the recognition rate is higher and the misrecognition rate is lower in this paper.



(a)

	Walking	Jogging	Running	Boxing	Handwaving	Handclapping
Walking	1.00	0.00	0.00	0.00	0.00	0.00
Jogging	0.02	0.97	0.01	0.00	0.00	0.00
Running	0.02	0.04	0.94	0.00	0.00	0.00
Boxing	0.00	0.00	0.00	1.00	0.00	0.00
Handwaving	0.00	0.00	0.00	0.00	0.93	0.07
Handclapping	0.00	0.00	0.00	0.00	0.04	0.96

(b)

	Diving	Golf Swing	Kicking	Lifting	Riding Horse	Running	Skate Boarding	Swing Bench	Swing Side Angle	Walking
Diving	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Golf swing	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Kicking	0.00	0.00	0.82	0.00	0.00	0.10	0.00	0.00	0.00	0.08
Lifting	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
Riding Horse	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00
Running	0.00	0.00	0.06	0.00	0.00	0.91	0.00	0.00	0.00	0.03
Skate Boarding	0.00	0.00	0.00	0.00	0.06	0.00	0.84	0.00	0.00	0.10
Swing Bench	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00
Swing Side Angle	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00
Walking	0.00	0.00	0.15	0.00	0.00	0.00	0.00	0.00	0.00	0.85

(c)

	Approaching	Departing	Exchanging	Hugging	Kicking	Punching	Pushing	Shaking hands
Approaching	0.94	0.06	0.00	0.00	0.00	0.00	0.00	0.00
Departing	0.09	0.91	0.00	0.00	0.00	0.00	0.00	0.00
Exchanging	0.14	0.00	0.86	0.00	0.00	0.00	0.00	0.00
Hugging	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00
Kicking	0.00	0.00	0.08	0.00	0.92	0.00	0.00	0.00
Punching	0.00	0.00	0.00	0.00	0.00	0.80	0.20	0.00
Pushing	0.00	0.00	0.00	0.00	0.00	0.16	0.84	0.00
Shaking hands	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00

Fig. 4. Recognition rate of proposed algorithm in different datasets: (a) Confusion matrix on KTH dataset; (b) confusion matrix on UCF Sports dataset; (c) confusion matrix on SBU Kinect Interaction dataset

In order to evaluate the performance of the proposed algorithm in this paper, the recognition accuracy rate (RAA) is taken as the index to compare with other algorithms SSED [17], MCSM-Wri [18], PGCN-TCA [19], AR3D [20]. The comparison results are shown in Table 1.

Table 1. Recognition rate with different methods/%

Method	KTH dataset	UCF Sports dataset	SBU Kinect Interaction dataset
SSED	95.5	88.5	80.3
MCSM-Wri	95.5	92.7	90.8
PGCN-TCA	96.7	94.4	91.4
AR3D	97.8	95.8	94.9
Proposed	98.2	97.6	98.5

It can be seen from Table 1, the RAA of proposed method is 98.2%, which improves by 2.7%, 2.7%, 1.5% and 0.4% than that of SSED, MCSM-Wri, PGCN-TCA, AR3D respectively in KTH dataset. It is similar to the other two datasets, which shows that the proposed method has the better recognition results with the approach of global and local feature extraction.

4 Conclusion

This paper presents an improved motion recognition algorithm based on global and local features. SURF is used to maintain the invariance of occlusion, perspective transformation and rotation, and the improved HOG feature is integrated for motion recognition. The advantage is that it not only retains the global feature of HOG feature and local information of SURF feature, but also reduces the influence of occlusion, perspective transformation and other factors on motion recognition. The evaluation of the KTH dataset, UCF Sports dataset and SBU Kinect Interaction dataset proved that the proposed method can recognize various motions in the video. The experimental results show that the recognition rates on three data sets are 98.2%, 97.6% and 98.5%, respectively. Future work will focus on saliency object detection methods to detect saliency objects in video frames. And it only extracts the features of these objects to recognize motions.

5 Acknowledgement

The authors thank the anonymous reviewer for their excellent comments.

References

[1] J.-A, S.-L. Yin, A New Feature Fusion Network for Student Behavior Recognition in Education, Journal of Applied Science and Engineering 24(2)(2021).

[2] H.-Y. Zhao , B.-X. Jia, Human Action Recognition Using Image Contour, Computer Science 42(2)(2013) 312-315.

[3] J.-Cai, G.-Feng, X.-Tang, et al., Human Action Recognition Based on Local Image Contour and Random Forest, Acta Optica Sinica (10)(2014) 204-213.

[4] A.-Azeem, M.-Sharif, J.-H. Shah, et al., Hexagonal scale invariant feature transform (H-SIFT) for facial feature extraction, Journal of Applied Research and Technology 13(3)(2015) 402-408.

[5] Y.-Li, X.-Xu, Human Action Recognition by Decision-Making Level Fusion Based on Spatial-Temporal Features, Acta Optica Sinica, 38(8)(2018).

[6] Z.-Zhang, Y.-Hu, S.-Chan, L.-T. Chia, Motion Context: A New Representation for Human Action Recognition. In: Forsyth D., Torr P., Zisserman A. (eds) Computer Vision-ECCV 2008. Lecture Notes in Computer Science, vol 5305. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-88693-8_60

[7] M.-Bregonzio, S.-Gong, X.-Tao, Recognising action as clouds of space-time interest points, in: Proc. 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 2009.

[8] P. Dollar, V. Rabaud, G. Cottrell and S. Belongie, Behavior recognition via sparse spatio-temporal features, in: Proc. 2005 IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance, 2005.

[9] S.-Yang, L.-I. Sheng, Y.-Shao, et al., Building recognition method based on improved HOG feature, Computer Engineering and Applications 54(7)(2018) 196-200.

[10]Y.-H. Shao, Y.-C. Guo, C.-Gao, Infrared human action recognition using dense trajectories-based feature, Guangdianzi Jiguang/Journal of Optoelectronics Laser 26(4)(2015) 758-763.

[11]Y.-Huang, C.-Wan, H.-Feng, Multi-Feature Fusion Human Behavior Recognition Algorithm Based on Convolutional Neural Network and Long Short Term Memory Neural Network, Laser & Optoelectronics Progress 56(7)(2019) 071505.

[12]F. Schweiger, G. Schroth, R. Huitl, Y. Latif and E. Steinbach, Speeded-up SURF: Design of an efficient multiscale feature detector, in: Proc. 2013 IEEE International Conference on Image Processing, 2013.

[13]W.-T. Freeman, E.-H. Adelson, The design and use of steerable filters, IEEE Transactions on Pattern Analysis and Machine Intelligence 13(9)(1991) 891-906.

[14]C. Schuldt, I. Laptev, B. Caputo, Recognizing human actions: a local SVM approach, in: Proceedings of the 17th International Conference on Pattern Recognition, 2004.

[15]K.-Soomro, A.-R. Zamir, Action Recognition in Realistic Sports Videos, In: Moeslund T., Thomas G., Hilton A. (eds) Computer Vision in Sports. Advances in Computer Vision and Pattern Recognition. Springer, Cham. (2014) https://doi.org/10.1007/978-3-319-09396-3_9

[16]K. Yun, J. Honorio, D. Chattopadhyay, T. L. Berg, D. Samaras, Two-person interaction detection using body-pose features and multiple instance learning, 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, 2012.

[17]W. Yang, Q. Du, J. Cui, et al., Motion Recognition Based on Sum of the Squared Errors Distribution, IEEE Access 9(2021) 37116-37130.

[18]S.-Y. Ma, T.-P. Huang, S.-B.Li, et al., MCSM-Wri: A Small-Scale Motion Recognition Method Using WiFi Based on Multi-Scale Convolutional Neural Network, Sensors 19(19)(2019) 4162.

[19]H.-Y. Yang, Y.-Z. Gu, J.-C. Zhu, et al., PGCN-TCA: Pseudo Graph Convolutional Network With Temporal and Channel-Wise Attention for Skeleton-Based Action Recognition, IEEE Access 8(2020) 10040-10047.

[20]M.-Dong, Z.- Fang, Y.-Li, et al., AR3D: Attention Residual 3D Network for Human Action Recognition, Sensors 21(5)(2021) 1656.