Computer Science and Information Systems 20(1):537–550 https://doi.org/10.2298/CSIS220131067G

Human Action Recognition Based on Skeleton Features

Yi Gao[1*], Haitao Wu[1*], Xinmeng Wu[1], Zilin Li[1], and Xiaofan Zhao[2*]

[1] College of Intelligence and Computing,
Tianjin University, Tianjin, China,
gaoyi_art@tju.edu.cn
[2] School of Information Technology and Cyber Security,
People's Public Security University of China, Beijing, China
zhaoxiaofan@ppsuc.edu.cn

Abstract. Based on human bone joints, skeleton information has clear and simple features and is not easily affected by appearance factors. In this paper, an improved feature of Gist, ExGist, is proposed to describe the skeleton information of human bone joints for human action recognition. The joint coordinates are extracted by using OpenPose and the thermodynamic diagram, and ExGist is used for feature extraction. The advantage of ExGist is that it can effectively characterize the local and global features of skeleton information while maintaining the original advantages of Gist feature. Compared with Gist, ExGist achieves better results on different classifiers. Additionally, compared with C3D and APTNet, our model also obtains better results with an accuracy rate of 89.2%.

Keywords: Human Action Recognition, Gist, OpenPose, Euclidean Distance, Thermodynamic Diagram.

1. Introduction

Human motion recognition which captures the changing process of human motion by transforming the original video sequence has been one of the research focuses in the field of computer vision for a long time. The key to motion recognition is extracting the features that can represent human motion information from the region where the moving object is located. Many researchers have proposed a large number of technologies and methods based on feature representation. According to the feature extraction methods, human action recognition can be classified into manual feature extraction and feature extraction based on deep learning.

Manual Feature Extraction: The human action recognition method based on manual feature extraction firstly samples the continuous frames of data and obtains the sampling points. According to the designed manual feature extraction method, the features of the sampling points are extracted, which are encoded into feature vectors. Then the encoded feature vectors are input into the behavior classifier for training. Finally, the manual feature vectors extracted from the test video are input into the trained classifier to obtain the classification results. By using the method, researchers from all over the world have proposed gait recognition 1, silhouette 2, human junction

[*] These authors contributed equally to this work and should be considered co-first authors.

34, space-time interest points 56, movement trajectory 78 and other human behavior recognition methods.

Deep Learning: The human action recognition method based on deep learning uses a trainable feature extraction model to automatically learn behavior representations from videos in an end-to-end manner to complete classification. Up to now, the network structure of action recognition methods based on deep learning mainly includes convolutional neural network (CNN) 91011, cyclic neural network (RNN) 1213, graph convolution neural network 1415 and hybrid network 161718. Other researchers have proposed Restricted Boltzmann Machine 19, recurrent neural network 20, independent subspace analysis 21, etc., which also got good results.

There have been some researches regarding to Gist. A static human behavior classification method that combined local constraint linear coding (LLC) and global feature descriptor Gist was proposed 23. This method was limited to the processing of static human behavior images, and did not apply to the field of videos. In addition, another new combined feature called global Gist feature and local patch coding was also proposed 24: Gist feature included the spectrum information of the action in the global view. Then, according to the frequency of the action variance, Gist feature located in different grids in the action center area was divided into four blocks and local patch coding was adopted. What's more, a new method was proposed for recognizing person-to-person interaction behavior based on the statistical features of key frame feature library 25. This method firstly extracted the global Gist and the regional HOG features, and then used K-means clustering algorithm to construct the key frame feature library corresponding to the action category. At the same time, according to the similarity measure, the frequency of each key frame in the feature library in the interactive video was counted, and a statistical histogram feature representation of the action video was obtained, which was trained for classification and recognition. The calculation of the above three methods was complex and the accuracy rate was relatively low.

In this paper, to describe the skeleton information of human bones and identify human behavior, a new feature descriptor based on the improved Gist 22 is proposed, named ExGist, which extends the classic feature descriptor Gist for scene understanding into the field of video processing. OpenPose 26 and thermodynamic diagram 27 are used to extract joint coordinates, then ExGist is used for feature extraction. Finally, the extracted features are input into the classifiers for classification. According to the experimental results, when SVM 28 is used as the classifier, the accuracy of this method is as high as 89.2%. Compared with C3D 29 and APTNet 30 models, our method has achieved a better result.

The main contributions are as follows:

1. A new feature ExGist based on Gist is proposed. After being compared on different classifiers, ExGist achieves better results than Gist.
2. A new human action recognition method based on the ExGist feature description is proposed. This method is not only superior to some current research methods of human action recognition using Gist, but also has a better recognition accuracy than C3D and APTNet models.
3. This method shows that the classical feature Gist can apply to not only the field of scene understanding but also the field of video processing. The improved version of Gist achieves good results, providing new ideas and methods for more related researches.

2. Related Work

2.1. Skeleton-based Human Action Recognition by the Integration of Euclidean distance

In our previous work, a skeleton-based model was presented for human action recognition 31. Firstly, Euclidean distance was combined with OpenPose and thermodynamic diagram to estimate human poses. Additionally, the attention mechanism was integrated into the human pose estimation process, to gain both the overall features and the partial features. Finally, MLP classifier was used for classification. This method could also be applied to real-time recognition of multi-person behavior. On the KTH and ICPR datasets, the accuracy of the mode verified by changing several parameters was tested. The highest accuracy rate of single-person behavior recognition was 0.821, and the highest accuracy rate of multi-person behavior recognition was 0.812. The high running speed enabled the mode to be a real-time model.

On the basis of previous work, the following improvements are made in this paper: Firstly, a new feature extraction method ExGist is proposed by fusing the global feature descriptor Gist. ExGist makes up for a lack of the local feature description of Gist and achieves unexpected good results in the performance comparison of different classifiers. Secondly, the data sets UCF10132 and HMDB5133 are selected for their more abundant action categories and wider application.

2.2. Multi-person Pose Estimation

There are two prominent algorithms for multi-person pose estimation. The first one is AlphaPose 34, where the human body on the graph is detected first, and then the key points and skeleton of each human body are obtained (top-down). Another algorithm is OpenPose 26, where the key position is gained firstly, and then different human skeletons are distinguished by the correlation between joint points(bottom-up). AlphaPose is more accurate, but as the number of people on the picture increases, the speed slows down. The accuracy rate of OpenPose is lower, but the speed is not affected by the number of people on the graph. In order to ensure real-time performance, OpenPose is adopted in our work. However, OpenPose is easy to introduce the interference of non-human objects, resulting in the confusion of detection results. Additionally, thermodynamic diagram 27 is used to obtain the number of people on the picture, and then establish the corresponding partition. In the image processing of target detection, thermodynamic diagram represents the key points by using two-dimensional Gaussian kernel.

2.3. Multi-person Pose Estimation

Support Vector Machine (SVM) 28is often used for classification problems and is widely used in pedestrian monitoring as a feature classifier. SVM classifies features by solving the maximum margin hyperplane. Kernel used in our experiment includes RBF, Linear and Poly.

Multilayer Perceptron (MLP) 35is a fully connected multi-layer neural network, which is a supplement of feed forward neural network. It generally consists of three types of layers: the input layer, output layer and hidden layer. The input layer receives the input signal to be processed. The required tasks such as prediction and classification are performed by the output layer. An arbitrary number of hidden layers that are placed between the input and output layer are the true computational engine of the MLP. If it has more than one hidden layer, it is called Artificial Neural Network (ANN). MLPs are designed to approximate any continuous function and can solve problems which are not linearly separable. The major applications of MLP are pattern classification, recognition, prediction and approximation.

K-Nearest Neighbors (KNN)36 firstly calculates the distance between the training set and the sample (L2 Norm is generally used). Then, k pieces of data closest to the sample are selected. The category that contains the most selected points is the predicted category. It is worth noting that the larger the k value is, the larger the approximate error will be. When the k value is small, the overfitting will occur.

Decision tree 37 and random forest 38 are also commonly used classification methods. Although decision trees are common supervised learning algorithms, they are prone to bias and overfitting. However, when multiple decision trees form an ensemble in the random forest algorithm, they will predict more accurate results, especially when the individual trees are not correlated with each other. The random forest algorithm is an extension of the bagging method in that it exploits bagging and feature randomness to create unrelated decision tree forests. Feature randomness, also known as feature bagging or the random subspace approach, generates random subsets of features, thus ensuring low correlation among decision trees. This is the key difference between decision trees and random forests. While decision trees consider all possible feature segmentation, random forests only select a subset of these features.

3. Model

3.1. Gist Feature Descriptor

Gist 2239is a low-dimensional signature vector of a scene, representing global feature information, and is often used for feature extraction of scene recognition and classification tasks. To extract the global features of Gist, the image needs to be divided into several grids of equal size, and then Gabor filters with different direction scales are used to correlate these grids. Finally, the calculation results of these grid regions are averaged ions are averaged to obtain the required feature information. The steps are as follows:

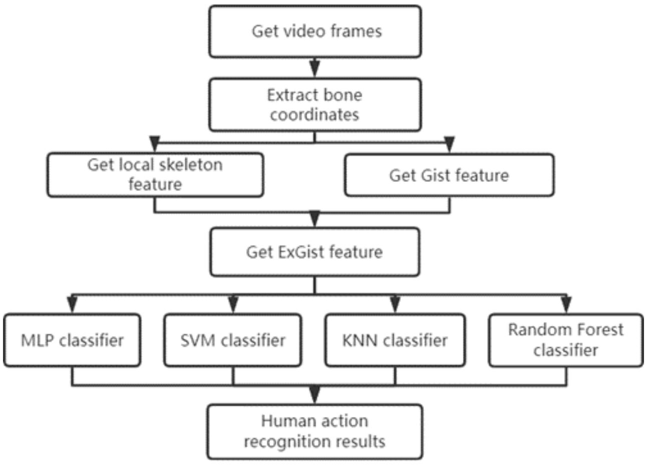

Fig. 1. Flow chart of feature extraction

Assume that the original gray-scale image to be processed is $I(x,y)$, and its size is $M \times N$. It is first divided into $n \times n$. Each net block represents an area, and $n_g = n_b \times n_b$ is used to record the total number of net blocks. Each of the mesh blocks after the image division is marked with B_i, where $I = 1 \ldots g$. In order to simplify the calculation and processing, each area is the same size, and its size is $M' \times N'$.

Gabor filter has great similarity with human visual perception function. By changing the mother wavelet of the filter, different Gabor filters can be obtained by some mathematical operations. The mother wavelet of Gabor filter is expressed as follows:

$$g(x,y) = \frac{1}{2\pi\sigma_x\sigma_y}\exp\left[-\left(x^2/\sigma_x{}^2 + x^2/\sigma_y{}^2\right)\right] * \cos(2\pi f_0 x + \varphi). \qquad (1)$$

Where, x and y represent coordinate information of pixel points, σ_x and σ_y represent Gaussian standard deviation of x-axis and y-axis respectively, f_0 represents center frequency and φ represents phase shift.

A group of Gabor filters with different scales and directions can be obtained by corresponding mathematical processing on the mother wavelet. The specific calculation formula is as follows:

$$g_{mn}(x,y) = a^{-m} g(x', y'), a > 1. \qquad (2)$$

$$x' = a^{-m}(x \cos(\theta + y \sin\theta). \qquad (3)$$

$$y' = a^{-m}(-x \sin(\theta + y \cos\theta). \qquad (4)$$

$\theta = \frac{n\pi}{n+1}$ represents the rotation angle, a^{-m} represents the scale factor. m represents the scale number, and n represents the direction number.

The Gabor filters obtained by calculation firstly implement the same processing for the different regions divided in the original image, and then use cascade operation to obtain the block Gist features of the image, as follows:

$$G_i^B = cat\big(I(x,y) * g_{mn}(x,y)\big), (x,y) \in B_i. \quad (5)$$

Where, G^B represents the gist feature of the block, the dimension is $m \times n \times M' \times N'$, and cat⍰⍰represents the concatenation operation, and $*$ represents the convolution operation. For each different filter, average the obtained block gist features, and then integrate the calculation results by line to obtain the gist global features of the image, which are shown as follows:

$$G = \left(\overline{G_1^B}, \overline{G_2^B}, \ldots \overline{G_n^B}\right). \quad (6)$$

Gist global feature is to describe an image as a whole. We use the corresponding feature operator to extract the features of the image, and record the relevant category information with the calculated multi-dimensional features. In the whole process, there is no need to consider much complex local information, which can reduce the impact of some small noises on clustering and reduce the additional errors caused by unnecessary processing.

3.2. ExGist Feature Descriptor

In order to classify video actions, features from video frames should be extracted. Both static and dynamic features based on skeleton information are proposed to represent video actions. Firstly, skeleton information is extracted from these frames using OpenPose. Then, feature extraction is carried out for skeleton information, which is mainly divided into two categories, namely static feature and dynamic feature. After that, these features are normalized.

Firstly, the matrix X is defined to represent the coordinates of the nodes between different frames. Most methods directly use X as the feature, and some improvements have been made for the matrix X.

$$X_i = [(x_{i1}, y_{i1}), \ldots (x_{in}, y_{in})]. \quad (7)$$

$$X = \{x_1, x_2, \ldots x_n\}. \quad (8)$$

Then the matrix X are normalized to adapt to images of different sizes, so as to represent features effectively. Inspired by the spatial partitioning method of ST-GCN40, the nodes are divided into different subsets to represent height, arm span and stationary center of gravity, respectively. In ST-GCN, a domain of nodes is divided into three subsets. The first subset is a node further away from the whole skeleton than the root node. Another subset is a node closer to the center, and the third subset is the root node itself, which is used to represent the motion characteristics of centrifugal motion, centripetal motion and stationary motion respectively. This strategy is to add "weight" to the key parts of rapid and accurate recognition. In this paper, data refers to the coordinates of skeletons ranging from 50 to 400, which is larger than expected. In that case, the distance from node 1, the head, to the barycenter is used as the height H, which will be used in the normalization. Node 8 and node 11 are the two sides of human's waist, so the average coordinate of node 8 and node 11 is used as the coordinate of barycenter.

$$H = \sqrt{(x_1 - x_b)^2 + (y_1 - y_b)^2}. \tag{9}$$

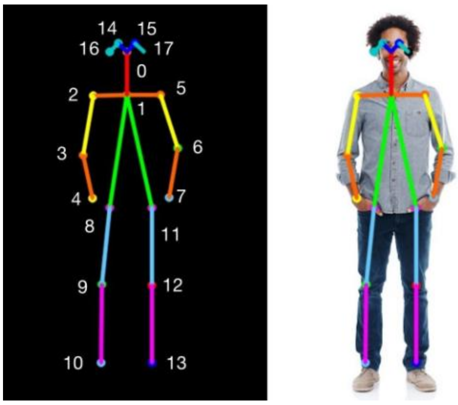

Fig. 2. Spanning tree of human posture features and mapping relation diagram of human joints.

Thirdly, the process of transforming the coordinates into the index of action label is challenging. In that case, the height H is used to converge the unwieldy data into normalized matrix X. The coordinates of some nodes in skeleton graph are less relevant to the accuracy of the action frame to be tested, such as node 0 and node 1. As a result, we decide to subtract the coordinates of these nodes.

$$X_i = [(x_{i1}/H, y_{i1}/H), \dots (x_{in}/H, y_{in}/H)]. \tag{10}$$

$$X_i = [(d_{i0} - d_{i0}, t_{i0} - t_{i0}), \dots (d_{in} - d_{i0}, t_{in} - t_{i0})]. \tag{11}$$

Then the normalized matrix X is obtained, which includes the coordinates of frame. Up to now, features from the normalized matrix X are extracted. Some parts of the skeleton features are gained at first by means of ST-GCN. The data of the previous frame in some parts of skeleton is subtracted from the data of the next frame, except for the first frame because there is no frame before the first one. As the frames are continuous, the data obtained by subtracting the normalized data can reflect an action. Additionally, it should be extracted as a symbol of one action. Now that the frames are continuous, this work should be done as soon as the next normalized data arrives.

$$Y = X_{i=1}^{n}[i + step][0:3] - X_{i=1}^{n}[i][0:3]. \tag{12}$$

Then the data of all nodes in the next frame is subtracted from the data in the previous frame.

$$Z = X_{i=1}^{n}[i + step][:] - X_{i=1}^{n}[i][:]. \quad (13)$$

Up to now, two parts of features have been gained. One part is coordinates, and another is the relationship among continuous frames. Matrix F is used to store all the features.

$$F = [X, Y(10), Z]. \quad (14)$$

Additionally, Euclidean distance D between joints is calculated and is stored as a characteristic.

$$F = [X, Y(10), Z, D]. \quad (15)$$

The compactness of skeleton C and the rate of change of compactness between different frames $\triangle C$ are defined.

$$F = [X, Y(10), Z, D, C, \Delta C]. \quad (16)$$

In addition, three-dimensional features are used to characterize the degree of skeleton integrity, which represent the degree of upper body integrity, lower body integrity and face integrity respectively.

$$S = [s_1, s_2, s_3]. \quad (17)$$

$$F = [X, Y(10), Z, D, C, \Delta C, S, G]. \quad (18)$$

ExGist adds some more detailed local features on the basis of Gist. The 18 human bone points are divided into three groups, namely the facial bone points, the upper body bone points and the lower body bone points. According to the coordinates of bone points from each group, the geometric characteristics of human movements and their changes in the dimension of time are obtained.

3.3. Defective Skeleton Graph

When extracting a skeleton feature, the model collects skeleton coordinates, partial features, overall features and Euclidean distances, all of which need a complete skeleton graph. Actually, in real-world applications, the skeleton graphs are more likely to be defective. In that case, firstly, a standard skeleton graph is prepared, from which all the joint coordinate pairs can be acquired. Then, when the picture of human body is incomplete, previous joint coordinates can fill the vacancy. As a result, all the obtained skeleton graphs are integrated.

3.4. Multi-person Pose Estimation

MLP Classifier, a supervised learning algorithm, is adopted in the model. It’s divided into three types of layers named input layer, hidden layer and output layer. LBFGS and stochastic gradient descent are used to optimize the logarithmic loss function. MLP Classifier performs iterative training, because the partial derivative of the loss function is calculated at each step when the parameters are updated. The model uses the sigmoid function and the tanh function to activate.

$$G(a) = softmax(a). \tag{19}$$

$$f(x) = G\left(b^{(2)} + W^{(2)}\left(s\left(b^{(1)} + W^{(1)}x\right)\right)\right). \tag{20}$$

Different strides and learning-rate are used to obtain a better model. Weights can be initialized in the model, and the function is used to calculate the probability when testing the action label of a picture.

4. Experiments

4.1. Datasets

HMDB51. HMDB5133 is a large collection of realistic videos from various sources, including movies and web videos. The dataset is composed of 6,849 video clips from 51 action categories (such as “jump”, “kiss” and “laugh”), with each category containing at least 101 clips. The original evaluation scheme uses three different training/testing splits. In each split, each action class has 70 clips for training and 30 clips for testing. The average accuracy over these three splits is used to measure the final performance.

UCF101. UCF10132 is an extension of UCF50 and consists of 13,320 video clips, which are classified into 101 categories. These 101 categories can be classified into 5 types (Body motion, Human-human interactions, Human-object interactions, Playing musical instruments and Sports). The total length of these video clips is over 27 hours. All the videos are collected from YouTube and have a fixed frame rate of 25 FPS with the resolution of 320 × 240.

Fig. 3. HMDB51 dataset presentation

Fig. 4. UCF101 dataset presentation

Due to the large amount of data, the sub-set of HMDB51 and UCF101 is selected for the experiment, and the training set and test set are randomly divided according to 7:3.

It can be seen that the accuracy of ExGist proposed by us is much higher than the accuracy of Gist on all the classifiers. Especially, when ExGist feature is extracted and put into SVM classifier, the accuracy can reach to 0.892. In conclusion, the experimental results show that ExGist can effectively represent human actions, thus improving the accuracy of action classification task.

Compared with C3D and APTNet, our model also obtains better results with a higher accuracy rate. In that case, the feature ExGist and SVM classifier can achieve relatively good human action recognition results.

The location information of bone joint extracted by OpenPose is clear and simple, which can make up for the shortage of global feature description in Gist. When using Gist to extract the global features of bone joint nodes, the method inevitably ignores the local features and connections of joints in the extraction of bone joint information. Therefore, it is necessary to improve and supplement more local features on the basis of global features of Gist. The coordinate information of joint nodes, the position relation of different joint nodes and the change rate between different frames are taken as local information to supplement Gist, so as to obtain ExGist. ExGist can not only give full play to the advantages of global feature description of Gist, but also greatly avoid the problems of high global feature dimensions and large computation.

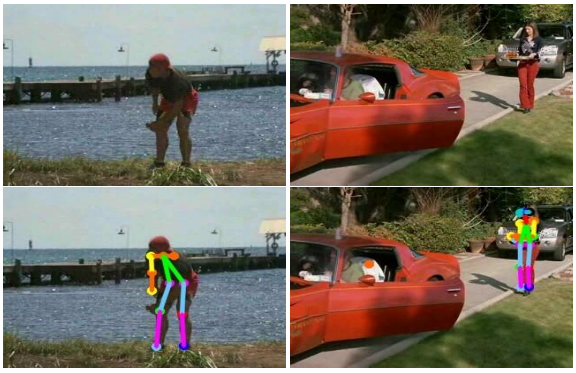

Fig. 5. HMDB51 dataset performance of the model

Table 1. Results for different classifiers

Classifier	Feature	Accuracy
SVM(poly)	Gist	0.771
SVM(poly)	ExGist	0.892
SVM(rdf)	Gist	0.751
SVM(rdf)	ExGist	0.847
SVM(linear)	Gist	0.747
SVM(linear)	ExGist	0.851
MLP(lbfgs)	Gist	0.755
MLP(lbfgs)	ExGist	0.871
MLP(adam)	Gist	0.747
MLP(adam)	ExGist	0.871
KNN	Gist	0.712
KNN	ExGist	0.847
Random Forest	Gist	0.723
Random Forest	ExGist	0.863
Decision Tree	Gist	0.591
Decision Tree	ExGist	0.755
AdaBoost	Gist	0.703
AdaBoost	ExGist	0.795
Gaussian Naive Bayes	Gist	0.618
Gaussian Naive Bayes	ExGist	0.6797
Linear Discriminant	Gist	0.651
Linear Discriminant	ExGist	0.759

Table 2. Results for different models

Model	Accuracy
C3D	0.838
APTNet	0.872
ExGist+SVM	0.891

5. Conclusions

ExGist, an improved feature of Gist, is proposed to describe the skeleton information for human action recognition. Firstly, OpenPose is combined with thermodynamic diagram to estimate human poses and get skeleton coordinates. Additionally, ExGist is used to gain both the global features and the local features. Finally, MLP, KNN, SVM and other classifiers are used for classification. This method can also be applied to real-time recognition of multi-person behavior. After being tested on two data sets, our improvement proves to be of great help to improve the accuracy rate of behavior recognition.

References

1. Vinay Kukreja, Deepak Kumar, and Amandeep Kaur. Deep learning in human gait recognition: An overview. In 2021 International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE), pages 9–13. IEEE, 2021.
2. Lena Gorelick, Moshe Blank, Eli Shechtman, Michal Irani, and Ronen Basri. Actions as space- time shapes. IEEE transactions on pattern analysis and machine intelligence, 29(12):2247– 2253, 2007.
3. Xiaodong Yang and YingLi Tian. Effective 3d action recognition using eigenjoints. Journal of Visual Communication and Image Representation, 25(1):2–11, 2014.
4. Faisal Mehmood, Enqing Chen, Muhammad Azeem Akbar, and Abeer Abdulaziz Alsanad. Human action recognition of spatiotemporal parameters for skeleton sequences using mtln feature learning framework. Electronics, 10(21):2708, 2021.
5. Ivan Laptev. On space-time interest points. International journal of computer vision, 64(2):107–123, 2005.
6. Konstantinos Rapantzikos, Yannis Avrithis, and Stefanos Kollias. Dense saliency-based spatiotemporal feature points for action recognition. In 2009 IEEE Conference on Computer Vision and Pattern Recognition, pages 1454–1461. Ieee, 2009.
7. Heng Wang and Cordelia Schmid. Action recognition with improved trajectories. In Proceedings of the IEEE international conference on computer vision, pages 3551–3558, 2013.
8. Jogendra Nath Kundu, Maharshi Gor, Phani Krishna Uppala, and Venkatesh Babu Radhakrishnan. Unsupervised feature learning of human actions as trajectories in pose embedding manifold. In 2019 IEEE winter conference on applications of computer vision (WACV), pages 1459–1467. IEEE, 2019.
9. Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, pages 1725–1732, 2014.
10. Kyo-Min Hwang and Sang-Chul Kim. A study of cnn-based human behavior recognition with channel state information. In 2021 International Conference on Information Networking (ICOIN), pages 749–751. IEEE, 2021.
11. SH Basha, Viswanath Pulabaigari, and Snehasis Mukherjee. An information-rich sampling technique over spatio-temporal cnn for classification of human actions in videos. Multimedia Tools and Applications, pages 1–19, 2022.
12. Ashesh Jain, Amir R Zamir, Silvio Savarese, and Ashutosh Saxena. Structural-rnn: Deep learning on spatio-temporal graphs. In Proceedings of the ieee conference on computer vision and pattern recognition, pages 5308–5317, 2016.

13. Pankaj Khatiwada, Matrika Subedi, Ayan Chatterjee, and Martin Wulf Gerdes. Automated human activity recognition by colliding bodies optimization-based optimal feature selection with recurrent neural network. arXiv preprint arXiv:2010.03324, 2020.
14. Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In Thirty-second AAAI conference on artificial intelligence, 2018.
15. Maosen Li, Siheng Chen, Xu Chen, Ya Zhang, Yanfeng Wang, and Qi Tian. Actional-structural graph convolutional networks for skeleton-based action recognition. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 3595–3603, 2019.
16. Niall McLaughlin, Jesus Martinez Del Rincon, and Paul Miller. Recurrent convolutional network for video-based person re-identification. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 1325–1334, 2016.
17. Nitish Srivastava, Elman Mansimov, and Ruslan Salakhudinov. Unsupervised learning of video representations using lstms. In International conference on machine learning, pages 843–852. PMLR, 2015.
18. Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, pages 1725–1732, 2014.
19. Graham W Taylor and Geoffrey E Hinton. Factored conditional restricted boltzmann machines for modeling motion style. In Proceedings of the 26th annual international conference on machine learning, pages 1025–1032, 2009.
20. Wan-Jin Yu, Zhen-Duo Chen, Xin Luo, Wu Liu, and Xin-Shun Xu. Delta: A deep dual-stream network for multi-label image classification. Pattern Recognition, 91:322–331, 2019.
21. Daniel Maturana and Sebastian Scherer. Voxnet: A 3d convolutional neural network for real- time object recognition. In 2015 IEEE/RSJ international conference on intelligent robots and systems (IROS), pages 922–928. IEEE, 2015.
22. Antonio Torralba, Kevin P Murphy, William T Freeman, and Mark A Rubin. Context-based vision system for place and object recognition. In Computer Vision, IEEE International Conference on, volume 2, pages 273–273. IEEE Computer Society, 2003.
23. Ende Wang, Qiaoying Liu, and Li Yong. Classification of static human behaviors based on llc and gist features. Computer Engineering, 44(8):268–272, 2018.
24. Yangyang Wang, Yibo Li, and Xiaofei Ji. Human action recognition based on global gist feature and local patch coding. International Journal of Signal Processing, Image Processing and Pattern Recognition, 8(2):235–246, 2015.
25. Xiaofei Ji and Xinmeng Zuo. Couple interaction behavior recognition based on static features of key-frame feature library. Computer Application, 36(8):2287–2291, 2016.
26. Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 7291–7299, 2017.
27. Kazumasa Tsutsui and Koji Moriguchi. A computational experiment on deducing phase diagrams from spatial thermodynamic data using machine learning techniques. Calphad, 74:102303, 2021.
28. Shan Suthaharan. Support vector machine. In Machine learning models and algorithms for big data classification, pages 207–235. Springer, 2016.
29. Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In Proceedings of the IEEE international conference on computer vision, pages 4489–4497, 2015.
30. Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, pages 6450–6459, 2018.

31. Yi Gao, Zhaokun Liu, Xinmeng Wu, Guangyuan Wu, Jiahui Zhao, and Xiaofan Zhao. Skeleton- based human action recognition by the integration of euclidean distance. In 2021 The 9th International Conference on Information Technology: IoT and Smart City, pages 47–51, 2021.
32. Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:1212.0402, 2012.
33. Hueihan Jhuang, Juergen Gall, Silvia Zuffi, Cordelia Schmid, and Michael J Black. Towards understanding action recognition. In Proceedings of the IEEE international conference on computer vision, pages 3192–3199, 2013.
34. Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. Rmpe: Regional multi-person pose estimation. In Proceedings of the IEEE international conference on computer vision, pages 2334–2343, 2017.
35. Ilya O Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Andreas Steiner, Daniel Keysers, Jakob Uszkoreit, et al. Mlp- mixer: An all-mlp architecture for vision. Advances in Neural Information Processing Systems, 34:24261–24272, 2021.
36. T. Abeywickrama, M. A. Cheema, and D. Taniar. k-nearest neighbors on road networks: A journey in experimentation and in-memory implementation. Proceedings of the VLDB Endowment, 9(6), 2016.
37. Anthony J Myles, Robert N Feudale, Yang Liu, Nathaniel A Woody, and Steven D Brown. An introduction to decision tree modeling. Journal of Chemometrics: A Journal of the Chemometrics Society, 18(6):275–285, 2004.
38. Jean-Franc¸ois Le Gall. Random trees and applications. Probability surveys, 2:245–311, 2005.
39. Liangmin Pan. Research on clustering algorithm of phishing websites based on gist global feature. PhD thesis, Central South University Of Forestry And Technology, 2018.
40. Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In Thirty-second AAAI conference on artificial intelligence, 2018.

Yi Gao undergraduate student at the Intelligence and Computing Department of Tianjin University.

Haitao Wu undergraduate student at the Intelligence and Computing Department of Tianjin University.

Xinmeng Wu undergraduate student at the Intelligence and Computing Department of Tianjin University.

Zilin Li undergraduate student at the Intelligence and Computing Department of Tianjin University.

Xiaofan Zhao undergraduate student at the Intelligence and Computing Department of Tianjin University.

Received: January 31, 2022; Accepted: December 10, 2022.