Research on Human Action Recognition Algorithm Based on Sine Feature

Haiyun Zhang¹, a, Huazhou Dou¹ and Bo Li¹, *, b

¹Xi’an Jiao Tong University, Xi’an 710000, China

*a haiyunzhang@stu.xjtu.edu.cn; *, b bolley@xjtu.edu.cn

Abstract. Recent studies have shown that exploring features of the skeleton data is vital for human action recognition. Nevertheless, how to effectively extract discriminative features is still a challenging work. In this paper, we propose a novel method that extracts a sine feature for human action recognition from skeleton data. Kinect is used to extract human skeleton information (3D coordinates of each joint point) firstly, then two joint points are connected to define a skeleton vector according to the principle of human body structure, and the sine feature of each skeleton vector is calculated as a new pose description feature, SVM is used to classify the obtained features for human action classification. Experimental results on Cornell Activity Database are provided, and the results demonstrate the effectiveness of our approach.

1. Introduction

Computer vision has developed rapidly in recent years, and human action recognition, as one of its directions, has attracted much attention. It has been widely used in intelligent monitoring, human-computer interaction, virtual display, video annotation, and motion analysis.

Human action recognition research can be divided into non-visual mode and visual mode according to the data collection approach. Non-visual methods mainly use sensors to obtain human motion information [1]. Compared with the former, visual methods can collect richer motion information [2]. But both of them are inseparable from the choice of feature extraction and action recognition algorithms.

Microsoft Kinect can collect RGB images, depth images, and 3D skeleton data. Using Kinect to study human action recognition has also become a hot spot in the field of action recognition recently [3], such as Yao et al. [4] proposed a method based on Kinect to collect human 3D skeleton information to recognize actions. Pei [5] proposed a new spatiotemporal skeletal motion descriptor using Kinect joint point data, which combines three complementary features of relative geometric velocity features, relative joint positions, and joint angles based on Lie group and Lie algebra. Xu [6] proposed an interpolation method for skeletal data classification and interpolation, which can accurately fit the gesture of the interpolated frame position. Zhu [7] uses the skeleton data captured by Kinect to define skeletal vectors, and then obtains the direction cosine feature of these vectors for human motion recognition. Zhong [8] collected the relative depth information and relative position information by dividing the human skeleton node sequence into modules, and proposed a feature extraction method based on skeleton information. Shi et al. [9] extracted two features of the joint point position in the key frame and the skeleton angle between the rigid part of the human body, and selected the SVM classifier for the classification. Tian et al. [10] proposed that the combination
feature of the human skeleton angle and modulus ratio for human action recognition, but some actions have very similar skeletal angles, so their features do not perform well.

In this paper, we present a new skeleton feature for human action based on the position of human skeleton vector in three-dimensional space inspired by the direction cosine feature. We construct the skeleton vector and then calculate the sine feature of the skeleton vector in three-dimensional space. Due to the influence of the human body and environmental factors, the feature we proposed ignores the influence of the angle of the skeleton vector compared with the traditional skeletal features, and can have a perfect alignment with the position of human action in the real world.

From the above, we mainly have two contributions in this paper. The first one is that we make the first attempt to get the feature of skeleton vector itself instead of the angle of two vectors or center distances. The second one is that we simplify the calculation of skeleton features and our model can distinguish similar actions well.

2. Exaction of Features Vectors based on Skeleton Data

2.1. Kinect skeleton Joints

The position of the human body can be reflected by the three-dimensional coordinate data of the skeleton joints, and changes in human action can also be reflected by the positional relationship among the skeletons and the positional relationship between the skeleton and the world coordinate system. Therefore we select skeleton as an important feature.

The human body model is very complicated. It contains more than two hundred bone nodes, and each bone node can rotate a certain angle. This is one of the reasons why our human body can make many flexible and complex actions. However, such a human body model is not suitable for human action recognition, because constructing an overly complex human body model is not suitable for human action recognition because of requiring a lot of manpower and resources. Kinect can track the skeleton to obtain the change in the position of the skeleton during the change of human action. The display of skeleton information is represented by the coordinates of the human skeleton joints, and a simplified version of the above complex human body model was proposed. Kinect's detection has a certain range. Only in detection range can recognize the joint nodes of the human body, such as HEAD (X, Y, Z). Skeleton-based action recognition is the analysis and learning of these skeleton joints data. There are 20 skeleton joints captured by Kinect as shown in Figure 1.

![Figure 1. 20 skeleton joints captured by Kinect.](image1)

![Figure 2. 14 Skeleton vectors.](image2)

![Figure 3. Sine Features.](image3)

2.2. Skeleton Vector

By analysing human action, we construct vectors called skeleton vectors by selecting appropriate skeleton joints based on ergonomics. The construction of the skeleton vector is very simple. Suppose that the three-dimensional coordinates of the left shoulder are A \((x_1, y_1, z_1)\) and the left elbow are B \((x_2, y_2, z_2)\), the skeleton vector can be expressed as:

\[
\overrightarrow{AB} = \{x_2 - x_1, y_2 - y_1, z_2 - z_1\}
\]

\(\overrightarrow{AB}\) describes the posture of the left forearm. This posture contains many human actions which are related with the actual situation, such as making phone calls and drinking water.
In order to describe human action, we propose a method that divides human body into upper limb part and lower limb part. The model contains 14 skeleton vectors (refer to Figure 2) and their specific meaning is in Table 1.

![Image of Table 1. Skeleton Vector.](image)

### 2.3. Extract Sine Feature

Suppose that A \((x_1, y_1, z_1)\), B \((x_2, y_2, z_2)\), then we have a skeleton vector:

\[
\overrightarrow{AB} = \{x_2 - x_1, y_2 - y_1, z_2 - z_1\}
\]

The sine features (as shown in Figure 3) are calculated as followings:

\[
\sin \partial = (x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2 \]

\[
\sin \beta = (x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2 \]

\[
\sin \gamma = (z_2 - z_1)^2 + (x_2 - x_1)^2 + (y_2 - y_1)^2 \]

where

\[
[(x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2]^{-1/2} \neq 0
\]

Then we found that \(\sin \partial = \sin \beta\), so we just take one of them. Now we obtain 14 skeleton vectors’ sine feature.

### 3. SVM

Support vector machine is one of the supervised learning algorithms for machine learning. It is a learning model for two-class classification. As shown in Figure 4, the two-class recognition can be deemed as that there are two different points on a plane and the task of SVM is to separate the two different points with a straight line. This line may be difficult to find in the plane, so we think about whether we can find it in another space, so we map these points to other spaces through some mapping methods, and then look for this "dividing line" in this space. This is the essence of SVM.

Suppose that there are samples \(x_i, y_i, \ldots, x_k, y_k\), where \(x_i\) is a D-dimension vector , \(k\) is the number of samples, and \(y_i \in \{-1,1\}\) is the class label. We set the linear function as:

\[
f(x) = <w, x > + b
\]

Therefore, the hyperplane can be defined as:

\[
<w, x > + b = 0
\]

We can obtain the value of \(w\) and \(b\) through training where \(w, b\) satisfies the following inequality:

\[
<w, x_i > + b \geq 0, y_i = +1
\]

\[
<w, x_i > + b < 0, y_i = -1
\]

Then normalizing the formula (9) can get:

\[
Y_i[<w, x_i > + b] - 1 \geq 0, i = 1,2,\ldots,k
\]

Then we can find the maximum distance between the classes which is \(1/||w||\) where the \(||w||\) is minimized, so the minimized \(||w||\) can be calculated by solving the following problem:

\[
\min ||w||^2 \text{ subject to } Y_i[<w, x_i > + b] - 1 \geq 0, i = 1,2,\ldots,k
\]

For the above problem, we can use Lagrange method and get the optimal classification function:
\[ f(\mathbf{x}) = \text{sgn}[(\mathbf{w} \cdot \mathbf{x}) + b^*] = \text{sgn}\left[\sum_{i=1}^{n} \theta_i^{*} y_i \langle \mathbf{x}_i , \mathbf{x} \rangle + b^*\right] \quad (12) \]

Compared with other classifiers, SVM has the following three advantages: First, the required number of samples is smaller compared with other classifiers. The second one is that the error between the model built by the classifier and the real solution is small, which means the structural risk is low; the last advantage is that SVM can solve both linear and non-linear problems. It is the characteristics of SVM. However, there are only 29 features in this paper, which belong to the scope of small samples. Therefore, we choose SVM as the machine learning algorithm in our work.

**Figure 4.** SVM two-class recognition.  
**Figure 5.** Experimental flowchart.

4. Human Action Recognition Based on Sine Features

In this paper, we extract the skeleton joints from the 3D skeleton data firstly, then construct the skeleton vectors at last we calculate three sine features for each skeleton vector. The specific approach is shown in Figure 5.

5. Experiment

5.1. Cornell Activity Dataset-60

In order to verify the ability of the features proposed in this paper, we use the Cornell Activity Dataset-60(CAD60) [2], which is one of the commonly used human action video datasets. It contains 60 RGB-D videos captured by Kinect. This dataset was completed by 4 different experimenters (two males, two females, and one left-hander) in 5 different settings (office, kitchen, bedroom, bathroom, and living room), with a total of 12 activities: rinsing mouth with water, brushing teeth, wear contact lenses, talking on the phone, drink water, opening pill container, cooking (chopping), cook (stirring), talking on couch, relaxing on couch, writing on the whiteboard, working on computer.

5.2. Human action recognition experiment

In this paper, the Sine feature is used as the structural feature of human body to represent human action. We choose 6 action data of 3 subjects from the CAD60 dataset as the training set, and the same 6 action data of the remaining (left-handed) subject as the test set. The six most common human actions are talking on the phone, writing on whiteboard, drinking water, brushing teeth, talking on couch, and working on a computer.

| Features                         | Accuracy  |
|----------------------------------|-----------|
| skeleton angle [9]               | 74.75%    |
| skeleton angle and vector modulus ratio [10] | 81.25%    |
| angular velocity [11]            | 91%       |
| the center distance [12]         | 62.83%    |
| Ours                             | 98.5%     |

Table 2. Accuracy.
5.3. Results
The results are shown in the Table 2. By comparing the accuracy of each feature on the dataset, it is
shown that our feature has a higher accuracy compared with the skeleton angle feature [9], the
skeleton angle and vector modulus ratio feature [10], the angular velocity feature [11] and the center
distance feature [12].

The confusion matrix of human action recognition based on different features is shown in Figure
6,7,8,9,10.

![Figure 6. The Confusion matrix of the center distances feature.](image1)

![Figure 7. The Confusion matrix of the skeleton angle feature.](image2)

![Figure 8. The Confusion matrix of skeleton angle and vector modulus ratio feature.](image3)

![Figure 9. The Confusion matrix of our feature.](image4)

The center distance reflects the change in the distance of the skeletal vector in the action. However,
the center distance does not change in some cases, such as rotation and the range of human action is
small. As shown in Figure 11(a), for talking on couch and working on computer, the distance of
head_to_root and limb_to_root does not change, so the center distance cannot distinguish the two
actions. The skeleton angle feature and the skeleton angle and vector modulus ratio feature reflect the
differences among actions, which are more effective for actions with obvious differences. However,
there are some actions which have similar skeleton angle, for example, for the action in Figure 11(b),
the included angle of elbow and arm is almost the same, so the two actions cannot be distinguished
from the above two features. In essence, the angular velocity is also based on the angle difference of
the same structure vectors between the adjacent frames, so it is also hard to distinguish actions which
have similar skeleton angle. But our feature can reflect the relationship of the bone vector in spatial
position, which can not only ignore the constant center distance, but also ignore the influence of the
similar skeleton angle, so it can well distinguish the two groups of similar actions in Figure 11.

![Figure 11. (a)Talking on couch and Working on computer. (b) Drinking water and Talking on phone.](image5)
In order to verify the performance of this algorithm, experiments are conducted to compared it with other methods (as shown in the Table 3). The table lists the comparison between the algorithm proposed in this paper and the K-Nearest Neighbor (KNN), Boosting, Bagging and Random Forest (RF) recognition methods. It can be seen from the results that the overall performance of ours is higher than KNN, Boosting, Bagging and RF.

Table 3. Comparison of different action recognition methods.

| Algorithm | Accuracy of ours feature | Accuracy of features [10] |
|-----------|--------------------------|---------------------------|
| SVM       | 98.5%                    | 81.25%                    |
| RF        | 95.42%                   | 78.3%                     |
| KNN       | 91.41%                   | 71.08%                    |
| Boosting  | 72.41%                   | 76.83%                    |

6. Conclusions

In this paper, we propose a new skeleton feature for human action recognition through the analysis of human action characteristics. First we construct the skeleton vector according to the relationship between the upper and lower limbs of human body, then calculate the sine value of the skeleton vector in three dimensions, finally use SVM to train. The experimental results demonstrate that ours feature has a higher performance for human action recognition.

Although the method we proposed has achieved good experimental results, there are still many problems that need to be further studied. In future work, we will consider combining skeletal features with other features.

References

[1] WU, J. A scalable approach to activity recognition based on object use[J]. Iccv, 2007:1-8.
[2] Enea C, Samuele G, Ennio G, et al. A Human Activity Recognition System Using Skeleton Data from RGBD Sensors[J]. Computational Intelligence and Neuroscience, 2016, 2016:1-14.
[3] Ling J, Tian L, Li C. 3D Human Activity Recognition Using Skeletal Data from RGBD Sensors[C]/ International Symposium on Visual Computing. Springer International Publishing, 2016.
[4] Yao H, Jiang X, Sun T, et al. 3D human action recognition based on the Spatial-Temporal Moving Skeleton Descriptor[C]. IEEE International Conference on Multimedia and Expo, 2017:937-942.
[5] Qicheng Pei. The research on the Kinect-based action recognition[D]. Nanjing University of Posts and Telecommunications, 2018.
[6] Haiyang Xu. Research on Human Action Recognition Algorithm Based on Depth and Skeleton Information [D]. Jiangnan University, 2018. [6]
[7] Guo-Gang Z , Lin C. Human Motion Recognition Based on Skeletal Information of Kinect Sensor[J]. computer simulation, 2014.
[8] Ping Zhong. The Research of Human Action Recognition with RGBD data [D]. Hunan University, 2017.
[9] X.Shi, S.Liu, D.Zhang . (2015). Human Action Recognition Method Based on Key Frames. XiTongFangZhenXueBao/Journal of System Simulation, 27(10),2401-2408
[10] Tian, G., Yin, J., Han, X., & Yu, J. (2014). A novel human activity recognition method using joint points information. Jiqiren/Robot, 36(3), 285-292.
[11] Mei, Y., Wang, Y., Qin, Q., Yin, Z., & Zhang, S. (2017). A human behavior recognition method based on key frames. Guangxue Jishu/Optical Technique, 43(4), 323-328.
[12] Shu-juan, P. (2012). Key frame extraction using central distance feature for human motion data. Journal of System Simulation, 24(3), 565-569.