Human motion recognition using three-dimensional skeleton model based on RGBD vision system

Qing Ye 1, Zexian Tan 1, Chang Qu 1 and Li Zhang 1
1 School of Computer Science and Technology, North China University of Technology, Beijing, 100144, P.R.China
Corresponding author’s e-mail: yeqing@ncut.edu.cn

Abstract. Human action recognition is one of the most important topics in computer vision. It has a lot of applications in intelligent monitoring, human-computer interaction, virtual reality, motion analysis and video annotation, etc. However, the effects of recognition are usually affected by human occlusion, self-occlusion, ambiguity and so on. In view of the above problems, this article describes an algorithm to recognize human motion using three-dimensional skeleton model based on RGBD vision system. Firstly, the human body three-dimensional skeleton model is established with the Kinect. Next, the normalized relative positions of skeleton joints and joint angles are adopted as features, and the action database containing the color image, the depth image, the skeletal image and the coordinates of the joints is built. Then, the features are trained and classified by multi-class support vector machine classifier. Finally, human motion recognitions are carried out. Experimental results indicate that the algorithm can recognize 12 kinds of actions behaved by different humans effectively.

1. Introduction
Research on human action recognition has great theoretical and practical values, which involving subjects of pattern classification, image processing, computer vision, etc. It has penetrated into people's lives, such as security monitoring, virtual reality, medical rehabilitation and human-computer interaction and so on[1]. Despite of the certain development of human action recognition technology, there are still some constraints for its large scale application, such as the problems of non-rigid bodies, occlusion, illumination, backgrounds and so on. In recent years, much research has been done on human action recognition. S. Win et al.[2] in this paper, a human motion understanding system based on real-time video sequence is proposed, which is based on bone model recognition based on deep learning framework, It broadens the use scene of 3D human motion recognition, but its accuracy is easily affected by external illumination and other conditions; J. sedmidubsky et al.[3] proposed the enhancement technology of manually expanding the existing 3D human skeleton sequence set, which significantly improves the recognition accuracy, but requires a lot of manual means to enhance the data; W Che et al.[4] proposed a new three-dimensional dual path network(3-DPN), which makes the network sensitive to spatio-temporal input at different scales, but the computational complexity increases because attention mechanism is added to the spatio-temporal channels of different scales. In general speaking, there are still some defects in the previous human action recognition algorithm. This paper proposes a method of human motion recognition using three-dimensional skeleton model based
on RGBD vision system. The Kinect depth sensor is used to obtain the human skeleton information, reducing impact caused background and light effectively. And the normalized relative positions of skeleton joints and joint angles are combined as human action features, increasing the recognition accuracy greatly. In addition, the data is trained and classified by multi-class SVM classifier based on radial basis function kernel, achieving a better recognition effect.

2. Human motion recognition using three-dimensional skeleton model based on RGBD vision system

Base on the research of human motion recognition in RGBD vision, this paper raise a new method using three-dimensional skeleton model. The color image, the depth image and the skeletal image are obtained with the Kinect to establish human body three-dimensional skeleton model. Next, the normalized relative positions of skeleton joints and joint angles are selected as human motion features, and the action database containing the color image, the depth image, the skeletal image and the coordinates of the joints is built. Then, these feature data is trained and classified by multi-class support vector machine classifier. Finally, we are able to identify human action. The block diagram of this algorithm is represented in Figure 1.

![Algorithm Block Diagram](image)

2.1 The extraction of human skeleton model

2.1.1 Kinect sensor. Kinect depth sensor is mainly composed of color camera\(^5\), infrared camera and infrared projector. In this paper, the acquired depth information acquired by Kinect for Xbox sensor is used to identify the action. The color image, the depth image and the three-dimensional skeleton image are shown in Figure 2.

![Images obtained by Kinect](image)

(a) Color image. (b) Depth image. (c) Skeleton image.

2.1.2 The three-dimensional human skeleton model. In this paper, the skeletal information is used to build three-dimensional human skeleton model, and the relative positions of skeleton joints and joint angles and the motion changes are chosen as action characteristics. Here, the 20 skeletal joints provided by Kinect are enough to describe the process of general human movements.

2.2 Feature Extraction and Descriptor

Feature extraction and descriptor are important in action recognition, which determine the results of classification. There will be significant differences among different figures, even the same movement. Meanwhile, the angles of view also affect the coordinates of the joint. So, it isn't effective enough with
the naive initial coordinates of the joint point collected. In order to eliminate the effects caused by the
limb lengths, different bodily forms and changes of view, we combine normalized relative positions of
skeleton joints and joint angles as features.

2.2.1. Joint angles. The human limb actions are achieved by the movement of elbow and wrist and
other joints. To some extent, the human movements based on skeletal model can be seen as rigid
motion which depends on angles of the joint.

Assuming that the space coordinate of skeletal joint point \( J_{i,1} \) in the i-th frame of the action
sequence is \((x_{i,1}, y_{i,1}, z_{i,1})\), and the coordinates of \( J_{i,2} \) is \((x_{i,2}, y_{i,2}, z_{i,2})\), We can get one bone joint
point as the starting point of the vector and other bone joint points as the end point of the vector, to
form a new vector \( V_{i,2} = (x_{i,2} - x_{i,1}, y_{i,2} - y_{i,1}, z_{i,2} - z_{i,1}) \). According to the mathematical formula, we can
calculate the angle between the corresponding vectors.

2.3 Multi-class support vector machine classifier

The multi-class SVM classifier based on radial basis function kernel is taken to conduct training. SVM
can solve the problem of small sample classification[6].

Here, we need to achieve classifications of multiple behaviors. There will be L identically distributed
sample sets of k class and \( x_i \) is L-dimensional feature vector, \( y_i \in \{1, 2, \ldots, k\} \). Multi-class SVM
classifier [7] is based on binary classification which achieved by one-to-one, one-to-many, DAG,
decision tree or other classification strategy. One versus one classification strategy is used in this paper.
Select two in k different classes samples randomly to form a two class classifier and design a SVM
classification model. There are \( k(k - 1)/2 \) SVM models should be design. The training samples of
each classifier are composed of two groups of samples correspondingly, and the unknown action
samples are predicted by the voting method that the class with the most votes is the action category to
which the unknown sample belongs.

3. Results

In the laboratory environment, 12 different actions completed by 12 different genders and height
bodies are collected. Just to be clear, as shown in Figure 3, A1, A2, A3, A4, A5, A6, A7, A8, A9, A10,
A11 and A12 are used to represent bending, raising one hand, arms sideways, arms raised, clapping,
arms akimbo, kicking, drinking, standing, hand laced behind head, sitting and boxing, respectively.
Each action frame is represented by a skeletal depth map which has a resolution of 320x240, and the
colored lines are used to connect the joint nodes. The feature data of each action is saved to the text
document to create the action database. In order to verify the validity of the selected features and the
universality of the algorithm, 6 different height and sexes bodies are collected for data collection in
this paper.

Figure 3. actions collected by Kinect
from left to right and from top to bottom, the actions are bending, raising one hand, arms sideways, arms raised, clapping, arms akimbo, kicking, drinking, standing, hand laced behind head, sitting and boxing.

As seen in Table 1, the recognition rate of the joint angle is slightly higher than the relative position of the skeleton, and the combination of both as action characteristic obtains a higher recognition rate. Therefore, this paper combines relative positions of skeleton joints and joint angles (33-dimensional) as the action characteristics.

| Table 1 action recognition accuracy of different features |
|----------------------------------------------------------|
| feature types                                           | recognition accuracy |
| relative positions of skeleton joints                    | 89.94%               |
| joint angles (degree)                                   | 91.08%               |
| joint angles(cosine value)                              | 90.28%               |
| relative positions of skeleton joints and joint angles (degree) | 93.92%               |
| relative positions of skeleton joints and joint angles (cosine value) | 93.03%               |

Compared with the existing UTKinect Action 3D data set from a single action, the accuracy rate of "sitting" in UT data set is only 95% in the research of Wu et al.[8], and the recognition rate of "sitting" in the data set made by Wu et al. is only 98.9%, while that of this study is 99.7%. The action of "standing" is 0.9% higher than that of Wu et al. In the same single action accuracy rate is higher than the public data set and the data set made by Wu et al.

| Table 2 Recognition rates of different motions |
|-----------------------------------------------|
| Action types recognition rate                 | Action types recognition rate |
| A1 98.9%                                      | A7 91.2%                      |
| A2 100%                                       | A8 97.7%                      |
| A3 100%                                       | A9 100%                       |
| A4 99.3%                                      | A10 99.1%                     |
| A5 99.3%                                      | A11 99.7%                     |
| A6 100%                                       | A12 100%                      |

As shown in Table 2, Compared with other papers, Liu et al.[9] selected CNN + softmax method to obtain 82% average recognition rate on UTKinect-Action3D data set. Kim y et al.[10] used the method of three stream CNN + LSTM in UTKinect-Action3D data set, and the recognition accuracy rate was 81.1%. In this paper, the normalized relative positions of skeleton joints and joint angles are selected as human motion features, removing the impacts caused by different individual shapes. And with the help of the multi-class SVM classifier, the satisfying right identifying rate is gained in this paper. It can be seen in Table 2 and Table 3 that the recognition method proposed in this paper has a high recognition accuracy rate for 12 different actions, indicating the fact that the method is feasible for the classification and recognition of different action types.

4. Conclusions

Researchers on human motion recognition, an algorithm to recognize human motion using three-dimensional skeleton model based on RGBD vision system is raised in this paper. The Kinect depth sensor is used to obtain the human skeleton information, which reduces the impact caused by variety of external factors, such as illumination change, shadow and so on. Then the normalized relative positions of skeleton joints and joint angles are used as human action features, which is conducive to minimizing the impact of individual differences on recognition. Furthermore, the multi-class SVM classifier based on radial basis function kernel is applied to train date. Experimental results demonstrate that the assignment of recognizing 12 kinds of actions behaved by different humans is achieved by the method and the algorithm has higher recognition rate and higher robustness.

In the course of the experiment, this paper only studies the single action recognition, and adds some
hypothetical conditions, also. In the later research, we will focus on the description of the action features and complex or multi-person action recognition.

Acknowledgements
This paper is supported by Technology Project of Beijing Municipal Education Commission (No.SQKM20181009002), National Natural Science Foundation of China (No.61371143), National Natural Science Foundation of China (No.61806008), Beijing Natural Science Foundation(No.4192022), Ministry of Education Science and Technology Development Center Project (No. 2018A03029). The authors also gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the presentation.

References
[1] C. Pham et al., "SensCapsNet: Deep Neural Network for Non-Obtrusive Sensing Based Human Activity Recognition," in IEEE Access, vol. 8, pp. 86934-86946, 2020, doi: 10.1109/ACCESS.2020.2991731.
[2] S. Win and T. L. L. Thein, "Real-Time Human Motion Detection, Tracking and Activity Recognition with Skeletal Model," 2020 IEEE Conference on Computer Applications (ICCA), Yangon, Myanmar, 2020, pp. 1-5, doi: 10.1109/ICCA49400.2020.9022822.
[3] J. Sedmidubsky and P. Zezula, "Augmenting Spatio-Temporal Human Motion Data for Effective 3D Action Recognition," 2019 IEEE International Symposium on Multimedia (ISM), San Diego, CA, USA, 2019, pp. 204-2043, doi: 10.1109/ISM46123.2019.00044.
[4] W. Che and S. Peng, "3D Dual Path Networks and Multi-scale Feature Fusion for Human Motion Recognition," 2019 IEEE 4th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC), Chengdu, China, 2019, pp. 698-704, doi: 10.1109/IAEAC47372.2019.8997745.
[5] A. Bilesan, S. Behzadipour, T. Tsujita, S. Komizunai and A. Konno, "Markerless Human Motion Tracking Using Microsoft Kinect SDK and Inverse Kinematics," 2019 12th Asian Control Conference (ASCC), Kitakyushu-shi, Japan, 2019, pp. 504-509.
[6] H. Liu, X. Xiao, Y. Li, Q. Mi and Z. Yang, "Effective Data Classification via Combining Neural Networks and SVM," 2019 Chinese Control And Decision Conference (CCDC), Nanchang, China, 2019, pp. 4006-4009, doi: 10.1109/CCDC.2019.83242.
[7] A. M. Nuruddin Pk, X. Ding and T. Page, "An Integrated Approach for Face Recognition Using Multi-class SVM," 2020 IEEE 5th International Conference on Cloud Computing and Big Data Analytics (ICCCBDA), Chengdu, China, 2020, pp. 398-402, doi: 10.1109/ICCCBDA49378.2020.9095692.
[8] X. Wu, R. Li and S. Wu. (2017) Behavior recognition algorithm based on CNN and bidirectional LSTM. J. Computer engineering and design,41(02):361-366.
[9] Z. Liu, X. Feng, J. Zhang. (2018) Human behavior recognition based on deep convolution neural network and depth video. J. Journal of Chongqing University,40(11):99-106
[10] Y. Kim, D. Lee and S. Lee, "First-Person Activity Recognition Based on Three-Stream Deep Features," 2018 18th International Conference on Control, Automation and Systems (ICCAS), Daegwallyeong, 2018, pp. 297-299.