Study on Human Body Action Recognition

Dong Yin[1,2(✉)], Yu-Qing Miao[3,4(✉)], Kang Qiu[1,2], and An Wang[1]

[1] School of Information Science and Technology,
USTC, Hefei 230027, Anhui, China
{yindong, anwang}@ustc.edu.cn, qk0208@mail.ustc.edu.cn
[2] Key Laboratory of Electromagnetic Space Information
of CAS, Hefei 230027, Anhui, China
[3] School of Computer Science and Information Security,
GUET, Guilin 541004, Guangxi, China
miaoyuqing@guet.edu.cn
[4] Key Laboratory of Intelligent Processing of Image and Graphics,
GUET, Guilin 541004, Guangxi, China

Abstract. A novel human body action recognition method based on Kinect is proposed. Firstly, the key frame of the original data is extracted by using the key frame extraction technology based on quaternion. Secondly, the moving pose feature based on the motion information of each joint point is constituted for the skeleton information of each key frame. And, combined with key frame, online continuous action segmentation is implemented by using boundary detection method. Finally, the feature is encoded by Fisher vector and input to the linear SVM classifier to complete the action recognition. In the public dataset MSR Action3D and the dataset collected in this paper, the experiments show that the proposed method achieves a good recognition effect.

Keywords: Action recognition · Kinect · Support vector machine
Fisher vector

1 Introduction

In recent years, home service robots have developed rapidly. The human motion detection and recognition has become an important research topic in the field of robot application. Due to the expensive and complex design of robot equipment, it is a very successful way to study human action recognition by using cheap and good Kinect.

In action recognition technology, scholars have been studied a lot. Wang [1] proposed a combination representation called global Gist feature and local patch coding to identify actions reliably. Kwak [2] proposed an algorithm which could be efficiently applied to a real-time intelligent surveillance system. Vinagre [3] presented a geometric correspondence between joints called Trisarea feature. It was defined as the area of the triangle formed by three joints. He [4] proposed self-taught learning features and unsupervised learning pre-processing. Das Dawn [5] presented a comprehensive review on STIP-based methods. These methods had achieved good results in human motion recognition.

J. Zhou et al. (Eds.): CCBR 2018, LNCS 10996, pp. 87–95, 2018.
https://doi.org/10.1007/978-3-319-97909-0_10

In recent years, neural networks have been developed rapidly and applied widely in video image processing. Ijjina [6] proposed an approach using genetic algorithms (GA) and deep convolutional neural networks (CNN). Wu [7] gave a review of various state-of-the-art deep learning-based techniques. Sargano [8] presented a method based on a pre-trained deep CNN model for feature extraction & representation followed by a hybrid Support Vector Machine (SVM) and K-Nearest Neighbor (KNN) classifier for action recognition. Ma [9] addressed the problems of both general and also fine-grained action recognition in video sequences. Their work sought to improve fine-grained action discrimination, while also retaining the ability to perform general recognition.

To sum up, scholars have done a lot of work on human action recognition, and have achieved fruitful results. However, due to factors such as background, illumination and occlusion, recognition based on color video is still difficult to achieve satisfactory results. As the function of Kinect is enhanced and the price is low, people have developed and studied on its platform. It is gradually applied to robot, medical treatment, education and so on. Therefore, based on Kinect V2, we carried out the action recognition method of depth information and skeleton information. And we finally transplant the method to the robot.

2 Relevant Theoretical Basis

With the rise of artificial intelligence, the traditional human-computer interaction mode has been unable to meet the demands of people. The way to transfer information through action becomes a more friendly and natural choice. Human action recognition is a comprehensive subject, which involves many fields.

2.1 Device of Kinect

Kinect, which was launched in September 2010 by Microsoft, is a peripheral applied to the XBOX360 host. Its sensor contains a depth sensor, a color camera and a microphone array. Kinect uses optical coding to get deep data and provides information about body skeleton and joints. Each joint is represented by a three-dimensional coordinate. Figure 1 shows the sketch map of extraction skeleton.

(a) (b) (c) (d) (e)

Fig. 1. Sketch map of Kinect extraction skeleton information. (a) Kinect device (b) Input depth map (c) Distribution of body parts (d) 3D joint point (e) Output 3D skeleton

2.2 Action Features Extraction

Feature extraction is divided into feature-joint, feature-joint selection and dynamic features. The feature-joint point is the feature extracted from the skeleton information,

which can represent the relationship between the joint points of the human body, and can be divided into three subcategories: spatial, geometric and key attitude features. The feature-joint selection refers to extract the most influential parts of the body from all joints. Ofli [10] put forward a feature representation method called "maximum information" joint sequence to describe human action. The dynamic feature refers to the characteristics of skeleton sequence as a 3D trajectory and modeling the dynamics of time series.

2.3 Action Classification Recognition

Action classification algorithm can be divided into dynamic time warping algorithm, generative model and discriminant model. The warping algorithm is a dynamic programming based nonlinear regularization, which has been widely used in speech recognition early. The generative model, which is the dynamic classifier, is usually modeled on the joint probability distribution $P(x, y)$, and the $P(y_i|x)$ is obtained by the Bayes formula, and the largest y_i of $P(y_i|x)$ is selected as the recognition result. The discriminant model named static classifier models directly the probability $P(y|x)$.

3 Our Method

Based on Kinect and robot oriented applications, a human action recognition system framework is presented in this paper, as shown in Fig. 2. It includes three major modules: Data Acquisition, Continuous Action Segmentation and Action Recognition.

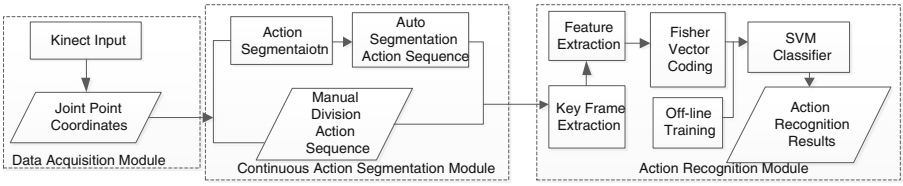

Fig. 2. Proposed system framework of human body action recognition.

Data Acquisition gets joint point coordinators from Kinect. Continuous Action Segmentation divides the start and end frames of the action. One is calibrated manually by user, another is auto. In Action Recognition, the current frame has been judged which belongs to the action. The extracted action will be encoded. The Fisher vector is input into the pre-trained model, and the recognition results are obtained.

3.1 Key Frame Processing

Key frames refer to the important frames that can be extracted from them and represent the motion characteristics. In this paper, we use a simple and efficient approximate algorithm. Firstly, use quaternion to describe the action information of the human body,

and based on it, define the distance between frames and frames. Secondly, decide the key frames whose distance is larger than a threshold.

As the original data obtained by Kinect is 3D coordinate sequence of each joint point, the Kinect V2 obtains 25 joint points per frame. We use 21 points shown as Fig. 3(a). In order to make full use of this chain structure, a tree model is used to represent the human body structure shown as Fig. 3(b). No.1 point is selected as the root node and the other 20 joints are used as children nodes. The distance between each node and root is a fixed value. While the distance between elbow and shoulder point is the length of arm, the space position of elbow is determined only by the rotation information of elbow relative to shoulder. So we can use a discrete time vector function to represent motion information at t time.

$$m(t) = [p(t),\ q_2(t),\ q_3(t),\ \ldots,\ q_{21}(t)]. \tag{1}$$

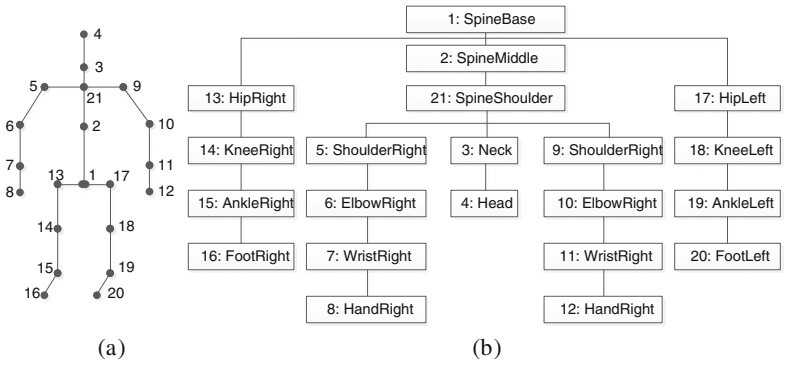

Fig. 3. Human skeleton stratification model. (a) Joint points (b) Hierarchical model

In (1), $p(t) \in R^3$ refers to the translation information of root at t time. $q_i(t)(i = 2,\ldots,21)$ represents the rotation information of i joint at t time relative to its parent node.

It is simple and fast to express the rotation information with the quaternion method. Given $s = w \in \boldsymbol{R}$ and $\boldsymbol{V}= (x, y, z) \in \boldsymbol{R}^3$, quaternion $q \in S^4$ is defined as:

$$\mathbf{q} = [s, -\mathbf{V}] = [w,\ x,\ y,\ z]. \tag{2}$$

In (2), s represents the scalar part of quaternion $\boldsymbol{q}$, $\boldsymbol{V}$ represents 3D vector part. The root node translation information is $p(t) = (x_t, y_t, z_t)$. We need to find out the rotation axis and angle. Sign $\boldsymbol{P}_p(t)$ to the coordinate of parent node and $\boldsymbol{P}_c(t)$ to the coordinate of child node at t time. Rotation axis $\boldsymbol{u}$ and angle $\boldsymbol{\theta}$ are defined as:

$$\mathbf{u} = P_p(t) \times P_c(t), \boldsymbol{\theta} = \arccos\left(\left(P_p(t) \bullet \cdot P_c(t)\right) / \left(\left|P_p(t)\right| \bullet \cdot \left|P_c(t)\right|\right)\right). \tag{3}$$

Using $\boldsymbol{u}$ and $\boldsymbol{\theta}$, we can get the rotational and action information in the form of quaternion. Then, we use moving pose (MP) which proposed by Zanfir [11] to extract

the feature of joint points. The MP feature is Taylor form in (4), δP is the first-order differential, and $\delta^2 P$ is the second-order.

$$X_t = \left[P_t, \delta P_t, \delta^2 P_t\right]. \tag{4}$$

Each node extracts 3D P_t, δP_t and $\delta^2 P_t$, which constitute 6D dynamic characteristics. So the feature of each joint is a 9D vector, and each frame can extract a feature of 21×9.

3.2 Feature Coding

The Fisher vector combines the advantages of generative and discriminant model. The first- and second- order statistical features are included in addition to the zero-order.

The feature is represented as $X = \{x_t,\ t = 1,2,...,T\}$. We will get the Fisher core based on gradient function shown as formula (5).

$$K(X, Y) = G_\lambda^{X\prime} F_\lambda^{-1} G_\lambda^{Y}. \tag{5}$$

In (5), G is the Fisher vector, F is information matrix. Since the final Fisher vector does not contain the time sequence information of the action, that is to say, any Fisher vector is the same, and more specifically, the two opposite actions, such as "stand up" and "sit down", have the same Fisher vector when the time sequence information is not considered. To distinguish these movements, we need to carry out a sequence like Pyramid. The N layer is divided into n non-overlapping sub-sequence. This paper uses 3 layers, and the final encoding is composed of the Fisher vectors of the (1 + 2 + 3) sub-sequence, that is, the eigenvectors of the 6 segments of the action sequence are calculated respectively. Finally, the corresponding Fisher vector is stitched together.

3.3 Action Classification

Support vector machine (SVM) is a classifier for solving two classification problems. It cannot be used to solve the multiple classification problem directly. In order to apply it to the multiple classification problem of action recognition, it needs to be popularized. This paper uses "one to one" strategy.

The "one to one" strategy refers to the assumption that the training samples have N categories. From which, two classes are taken as positive and negative samples of the two classification SVM, then a total of $N(N-1)/2$ two classification problems can be formed, and each of them is trained to get the SVM classifier. When testing, the test samples are entered into the classifiers to vote on the classification results. The category of the largest number of votes is the result of the identification.

4 Experiments and Analysis

The platform used in the paper is CPU/Intel i7-4790, Memory/8 GB, Kinect v2, Windows 8.1, Microsoft VS 2013, Windows SDK 2.0, LibSVM 3.22, OpenCV 2.4.10.

4.1 Public Dataset – MSR Action3D

The MSR Action3D data set is collected by a static generation of Kinect sensors at the Microsoft Institute, including twenty movements, such as hand waving, drawing, clapping and bending. The data set was performed by 10 experimenters. Each action was performed 2 $\sim$ 3 times, and finally there were 544 sets of experimental data.

The 20 actions are divided into three subsets. Each subset is composed of 8 actions, as shown in Table 1. AS1 and AS2 collect some similar actions, while AS3 sets some complex actions. In experiment, when calculating Fisher vector, the number K of Gauss elements in GMM is set to 128, and the normalized parameter of energy normalization $\alpha = 0.3$. The comparison of recognition rate in Tables 2 and 3 show that our method has a certain improving extent, which proves the effectiveness.

Table 1. Three subsets of MSR Actions3D.

Action Set 1 (AS1)	Action Set 2 (AS2)	Action Set 3 (AS3)
Horizontal arm wave	High arm wave	High throw
Hammer	Hand catch	Forward kick
Forward punch	Draw $\times$	Side kick
High throw	Draw $\sqrt{}$	Jogging
Hand clap	Draw O	Tennis swing
Bend	Two hand wave	Tennis serve
Tennis serve	Forward kick	Golf swing
Pickup & Throw	Side boxing	Pickup & Throw

Table 2. The comparison of recognition rate in MSR Actions3D with three subsets.

Method	AS1	AS2	AS3	Average
Histogram of 3D Joints [12]	87.98	85.48	63.46	78.97
EigenJoints [13]	74.50	76.10	96.40	82.33
Skeletal Quad [14]	88.39	86.61	94.59	89.86
Lie Group SE(3) [15]	95.29	83.87	98.22	92.46
Our method	93.81	90.09	96.30	**93.40**

Table 3. The comparison of recognition rate in MSR Actions3D with whole dataset.

Method	Recognition rate
Actionlet Ensemble [16]	88.20
Histogram of Norms in 4D [17]	88.89
Lie Group SE(3) [15]	89.48
The moving pose [11]	91.70
Our method	**92.08**

4.2 Own Dataset – PAR718

Based on the application scene of the home service robot, we build a data set called PAR718 in this paper. It collects 10 common movements in living rooms, including sitting, drinking water, staying, calling, which are performed by 14 experimenters (9 women and 5 men). Each performing is done two times for per movement, so the dataset has 280 action sequences in total. The length of the action is from 47 frames to 220 frames, with a total of 26633 frames.

Figure 4 is the depth map sequence of the key frame extracted from the "drink water" movement. The original sequence has 101 frames and 30 key frames are extracted.

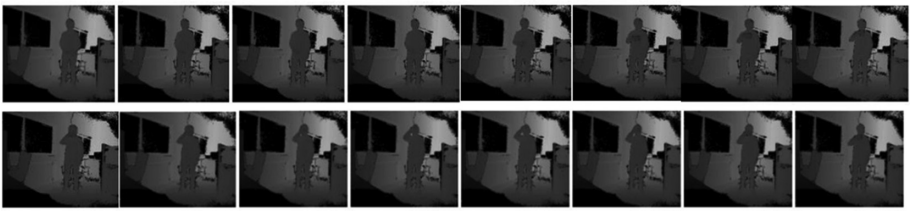

Fig. 4. The key frame sequences extracted from "drink water". From left to right, up to down, the numbers are 1st, 4th, 8th, 12th, 17th, 20th, 22th, 24th, 29th, 31th, 37th, 44th, 66th, 72th, 75th, 77th.

Figure 5 is the obfuscation matrix. In this experiment of feature extraction and action recognition classification, the K is 128 and $\alpha = 0.3$ when the energy is normalized. 280 action sequences are divided into training set and testing set. The accuracy is 98.57% (138/140). From Fig. 5, most of the actions are correctly identified, except for reading and writing, because they are too similar.

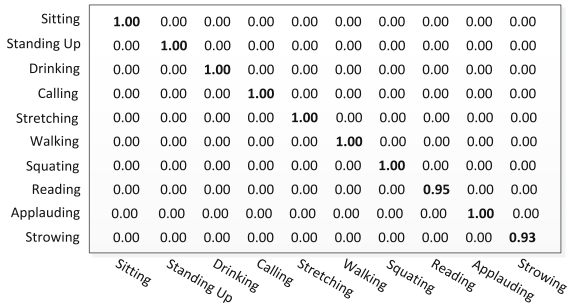

Fig. 5. The obfuscation matrix of own dataset.

5 Conclusion

Human action recognition based on vision has become a research hotspot nowadays. The difficulty of human action recognition based on deep data and human skeleton information is greatly reduced in target segmentation, and the accuracy is greatly

improved. This paper mainly studies the human action recognition method based on the 3D skeleton sequence obtained by Kinect, and puts forward a new framework of human body action recognition, which has achieved good recognition effect on the common dataset MSR Action3D and own dataset PAR718. Further effective identification of high similarity actions is the next step in this paper.

Acknowledgments. This paper is supported by the Guangxi Natural Science Foundation Project (2014GXNSFAA118395), the research project of Guangxi Colleges & Universities Key Laboratory of Intelligent Processing of Image and Graphics (GIIP201706), the National Natural Science Foundation Project (61763007), the key project of the Guangxi Natural Science Foundation (2017GXNSFDA198028).

References

1. Wang, Y., Li, Y., Ji, X.: Human action recognition based on global gist feature and local patch coding. Int. J. Signal Process. Image Process. Pattern Recognit. **8**(2), 235–246 (2015)
2. Kwak, N., Song, T.: Human action recognition using accumulated moving information. Int. J. Multimed. Ubiquitous Eng. **10**(10), 211–222 (2015)
3. Vinagre, M., Aranda, J., Casals, A.: A new relational geometric feature for human action recognition. In: Ferrier, J.-L., Gusikhin, O., Madani, K., Sasiadek, J. (eds.) Informatics in Control, Automation and Robotics. LNEE, vol. 325, pp. 263–278. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-10891-9_15
4. He, J.: Self-taught learning features for human action recognition. In: Proceedings of 2016 3rd International Conference on Information Science and Control Engineering, ICISCE 2016, pp. 611–615, 31 October (2016)
5. Dos Dawn, D., Shaikh, S.: A comprehensive survey of human action recognition with spatio-temporal interest point (STIP) detector. Visual Comput. **32**(3), 289–306 (2016)
6. Ijjina, E., Chalavadi, K.: Human action recognition using genetic algorithms and convolutional neural networks. Pattern Recognit. **59**, 199–212 (2016)
7. Wu, D., Sharma, N., Blumenstein, M.: Recent advances in video-based human action recognition using deep learning: a review. In: 2017 International Joint Conference on Neural Networks, IJCNN 2017–Proceedings, vol. 2017-May, pp. 2865–2872, 30 June (2017)
8. Sargano, A., Wang, X., Angelov, P., Habib, Z.: Human action recognition using transfer learning with deep representations. In: Proceedings of the International Joint Conference on Neural Networks, vol. 2017-May, pp. 463–469, 30 June (2017)
9. Ma, M., Marturi, N., Li, Y., Leonardis, A., Stolkin, R.: Region-sequence based six-stream CNN features for general and fine-grained human action recognition in videos. Pattern Recognit. **76**, 506–521 (2018)
10. Ofli, F., Chaudhry, R., Kurillo, G.: Sequence of the most informative joints (SMIJ): a new representation for human skeletal action recognition. J. Vis. Commun. Image Represent. **25** (1), 24–38 (2014)
11. Zanfir, M., Leordeanu, M., Sminchisescu, C.: The moving pose: an efficient 3D kinematics descriptor for low-latency action recognition and detection. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2752–2759 (2014)
12. Xia, L., Chen, C., Aggarwal, J.: View invariant human action recognition using histograms of 3D joints. In: 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 20–27 (2012)

13. Yang, X., Tian, Y.: Effective 3D action recognition using eigenjoints. J. Vis. Commun. Image Represent. **25**(1), 2–11 (2014)
14. Evangelidis, G., Singh, G., Horaud, R.: Skeletal quads: human action recognition using joint quadruples. In: 2014 22nd International Conference on Pattern Recognition (ICPR), pp. 4513–4518 (2014)
15. Vemulapalli, R., Arrate, F., Chellappa, R.: Human action recognition by representing 3D skeletons as points in a lie group. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 588–595 (2014)
16. Wang, J., Liu, Z., Wu, Y.: Mining actionlet ensemble for action recognition with depth cameras. 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1290–1297 (2014)
17. Tang, S., et al.: Histogram of oriented normal vectors for object recognition with a depth sensor. In: Lee, K.M., Matsushita, Y., Rehg, James M., Hu, Z. (eds.) ACCV 2012. LNCS, vol. 7725, pp. 525–538. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-37444-9_41