Journal of Physics: Conference Series

Human action recognition based on Kinect

To cite this article: Jiahui An *et al* 2020 *J. Phys.: Conf. Ser.* **1693** 012190

Human action recognition based on Kinect

Jiahui An[1, a], Xinrong Cheng[1*], Qing Wang[1], Hong Chen[1], Jiayue Li[1] and Shiji Li[1]

[1]Department of Computer Science and Technology, China Agricultural University, Beijing 100083, China

[*]cheng_xinrong@126.com

[a]s20183081286@cau.edu.cn

Abstract. With the emergence of Kinect, many research results have emerged in human action recognition based on skeleton information, which has promoted the development of human-computer interaction. In this paper, from the skeleton data obtained by Kinect, static features and dynamic features are extracted, and the two are merged; SVM classifier is used for action recognition. It is verified on the MSR Daily Activity 3D data set, and the experimental results show that the method in this paper improves the accuracy of action recognition.

1. Introduction

With the development of the information age, people hope that computers can become more and more intelligent, able to "see" and "listen" to the world like humans. Among them, computer vision technology can make computers "see" the world like humans. In the field of computer vision, action recognition technology has an important position. It can understand human actions and better interact with people. It has appeared in video surveillance, gaming, medical, and virtual reality fields.

In action recognition, action recognition based on 2D vision will be affected by factors such as illumination, occlusion, and shadow during image processing[1]. Recognizing human movements through wearable sensors has high accuracy and real-time performance. However, adding gyroscopes and accelerometers to collect human motion parameters will reduce the comfort of human body and destroy the naturalness of human-computer interaction [2].

The advent of the Microsoft Kinect depth camera has brought new ideas to action recognition. Using the skeleton information obtained from the Kinect depth camera for human action recognition can overcome the above-mentioned problems of illumination and noise. Therefore, the use of Kinect for action recognition is favored by more and more researchers. Many research results have been greatly improved in the accuracy and speed of human action recognition.

2. Related technology introduction

Action recognition using Kinect can be divided into the steps of skeleton data acquisition, feature extraction, and action recognition, as shown in the figure 1.

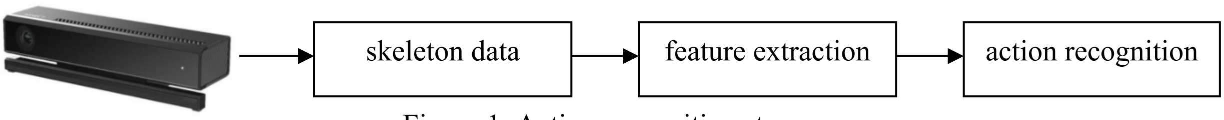

Figure 1. Action recognition steps.

2.1. Kinect gets skeleton data

The core of the somatosensory interaction device Kinect is an infrared device, through which Kinect can obtain information about the joint points of the human skeleton. In addition, Kinect also has the functions of dynamic tracking, image recognition, color imaging, and voice interaction.

Kinect obtains skeleton data from the depth image can be divided into 3 steps. First, the distance between the pixels in the depth image reflects the human body part. The human body foreground can be segmented based on this point, and then the human body contour image in the RGB image can be divided by edge detection. Split out. Secondly, the trained random forest model is used to obtain the probability distribution label, and the human body image is divided into 32 different parts. Finally, according to the obtained human body part tags, the local model method is used to merge each type of pixel points to form the three-dimensional coordinates of the human body joint points[3].

2.2. Research on action feature extraction

In human action recognition, if we can extract the motion features that effectively express the action, it is very important to the result of the action recognition, because it directly affects the final result of the action recognition. In order to complete action recognition, Carlsson et al. [4] performed shape matching between the key frames extracted from the action video and the saved action prototype.

2.3. Research on Action Recognition Method

At present, the commonly used action recognition methods include: template-based methods and probability statistics-based methods. The template-based method is intuitive and simple, and judges the motion category by comparing the similarity between the target to be detected and the template. Therefore, it lacks certain robustness. Ji et al. [5] used the dynamic time warping method to calculate the similarity between the action to be recognized and the action in the action library. The probabilistic statistical model represents the action as a continuous sequence of states, and the transition law between states can be represented by a time transfer function. Liu Fen [6]used Kinect sensor to generate human action depth map, establishes a three-dimensional human body model, uses the angle and modulus ratio of motion vectors as feature vectors, and uses SVM classifier for human body action classification and recognition.

3. Algorithm implementation

3.1. Action feature extraction

In each frame of the action sequence, there is a certain positional relationship between the joint points, so the spatial attributes of the action feature can be extracted. In addition, the same skeleton joint point has corresponding changes in each frame of the action sequence, based on which the time attribute of the action feature can be extracted. Therefore, when performing action feature extraction, we extract corresponding static and dynamic features based on the spatial and temporal attributes of the action feature.

Compared with Kinect V1, Kinect V2 has greatly improved its performance. From the original 20 joint points drawn in the visual range of the human body to 25 joint points, the neck, left hand tip, right hand tip, left thumb, and right thumb have been added. Coordinates, improve the recognition rate of gestures. However, the data set used in this article seldom involves hand movements, so the 20 joint point coordinates provided by Kinect V1 are used, as shown in the figure 2[7].

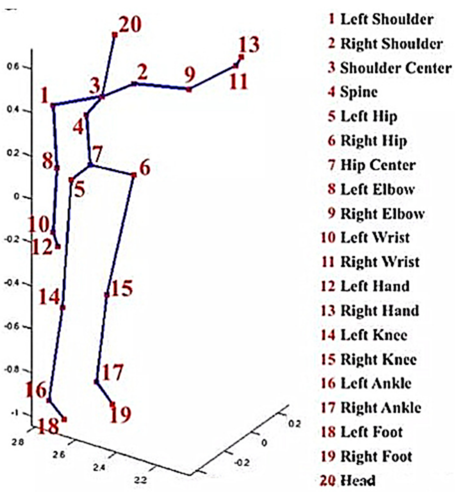

Figure 2. Human skeleton diagram.

The static feature is the position vector information between the skeleton joint points in each frame of the action sequence. The vector relationship between the 20 skeleton joint points is used to represent the static feature, that is, the vector form of the 19 segments of human limbs represented by the adjacent relationship of the joint points as shown in the table 1.

Suppose there is a human body action sequence X, the human body action can be expressed as formula (1).

$$X = (X^1, X^2, X^3, \cdots, X^t, \cdots, X^T) \tag{1}$$

Where T represents the number of frames of the action sequence, that is, the length of the action sequence. For a certain frame X^t of the action sequence, each contains 20 skeleton joint points, then $X^t=(V_1^t,V_2^t,\cdots,V_m^t,\cdots,V_{20}^t)$ .Where V_m^t represents the m-th joint point of the t-th frame of data. Then $V_m^t = (x_m^t, y_m^t, z_m^t)^T$. According to Figure 2, the relationship between human limbs as shown in the table 1. Static feature is shown in formula (2):

$$F_{static}=\begin{pmatrix} V_3^t\text{-}V_{20}^t,\ V_3^t\text{-}V_1^t,\ V_1^t\text{-}V_8^t,\ V_8^t\text{-}V_{10}^t,\ V_{10}^t\text{-}V_{12}^t,\ V_3^t\text{-}V_2^t,\ V_2^t\text{-}V_9^t,\ V_9^t\text{-}V_{11}^t,\cdots, \\ V_5^t\text{-}V_{14}^t,\ V_{14}^t\text{-}V_{16}^t,\ V_{16}^t\text{-}V_{18}^t,\ V_7^t\text{-}V_6^t,\ V_6^t\text{-}V_{15}^t,\ V_{15}^t\text{-}V_{17}^t,\ V_{17}^t\text{-}V_{19}^t \end{pmatrix} \tag{2}$$

The dynamic feature refers to the vector feature of the direction change between the skeleton joint points in the current frame and the previous frame in the action sequence. For the data X^t in X, dynamic feature is shown in formula (3) or (4).

$$F_{dynamic} = \{V_m^t - V_m^{t-1} | m = 1,2,\cdots,20\} \tag{3}$$

$$F_{dynamic} = \{x_m^t - x_m^{t-1},\ y_m^t - y_m^{t-1},\ z_m^t - z_m^{t-1} | m = 1,2,\cdots,20\} \tag{4}$$

Human actions can be regarded as composed of continuous static postures, that is, an action sequence contains multiple frames of information, and the correlation between the front and back information between frames is relatively large, so it is difficult to accurately describe a single action feature by extracting only a single action feature. Human action, considering that the information contained in human action is not only related to the spatial position between joints, but also has a certain relationship in time, so this article uses the corresponding static and dynamic hybrid characteristics to describe human actions.

Table 1. Limb relationship.

Joint 1	3	3	1	8	10	3	2	9	11	7	4	7	5	14	16	7	6	15	17
Joint 2	20	1	8	10	12	2	9	11	13	4	3	5	14	16	18	6	15	17	19

3.2. SVM action recognition method

The basic idea of support vector machine (SVM) is to solve the separation hyperplane that can correctly divide the training data set and have the largest geometric interval. Taking two dimensions as an example, the black dots and white dots in Figure 3 below belong to two different categories. The purpose of SVM is to ask for a line to "best" distinguish these two types of points. The sum of the distance between a certain line and the closest points on both sides of it is the margin. For example, the band formed by the two dashed lines in Figure 4 is the margin. Therefore, SVM is to find the hyperplane that can distinguish the two categories and maximize the margin[8].

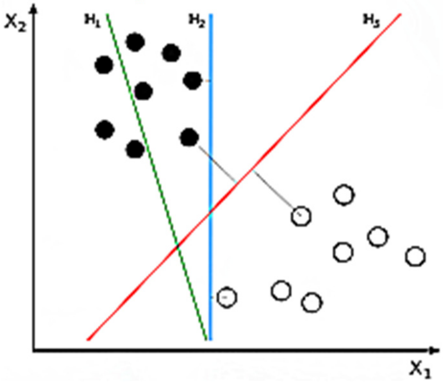

Figure 3.SVM diagram.

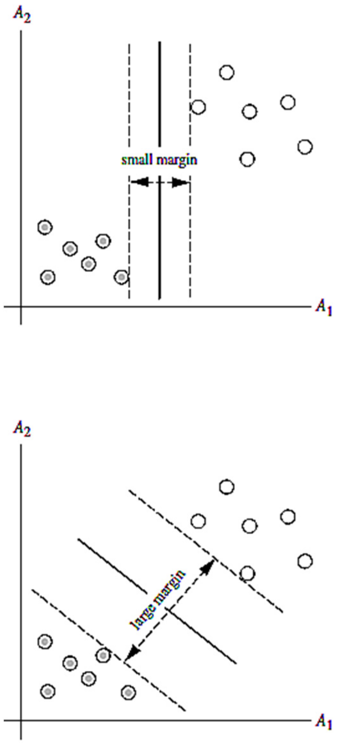

Figure 4.Margin graph.

For the condition of linearly separable data, suppose a given training data $T = \{(x_1, y_1), (x_2, y_2), \cdots, (x_N, y_N)\}$ in a feature space, where $x_i \in R^n$, x_i Is the i-th feature vector. $y_i \in \{-1, +1\}$ is the category label, $i = 1,2, \cdots, N$.

For a given data set T and hyper plane $\omega . x + b = 0$, define the geometric interval of the hyperplane about the sample points (x_i, y_i) as $\gamma_i = y_i \left(\frac{\omega}{\|w\|} . x_i + \frac{b}{\|w\|}\right)$, the minimum value of the geometric interval of all sample points in the hyperplane is $\gamma = \min_{i=1,2,\cdots N,} \gamma_i$.

The problem of solving the maximum split hyper plane of the SVM model can be expressed as the formula (5) and formula (6).

$$\max_{\omega,b} \gamma \tag{5}$$

$$s.t. \quad y_i\left(\frac{\omega}{\|\omega\|}.x_i + \frac{b}{\|\omega\|}\right) \geq \gamma, i = 1,2,\cdots,N \tag{6}$$

After sorting out, the maximum segmentation hyper plane problem of the SVM model can be expressed as the formula (7) and formula (8).

$$\min_{\omega,b} \frac{1}{2}\|\omega\|^2 \tag{7}$$

$$\mathrm{s.t.}\, y_i(\omega.x_i + b) \geq 1, i = 1,2,\cdots,N \tag{8}$$

This is a convex quadratic programming problem with inequality constraints, and its dual problem can be obtained by using the Lagrange multiplier method as the formula (9), formula (10) and formula (11).

$$\min_{\alpha} \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\,\alpha_j y_i y_j\left(x_i.x_j\right) - \sum_{i=1}^{N}\alpha_i \tag{9}$$

$$\mathrm{s.t.}\sum_{i=1}^{N}\alpha_i y_i = 0 \tag{10}$$

$$\alpha_i \geq 0, i = 1,2,\cdots,N \tag{11}$$

However, in actual situations, there is almost no completely linearly separable data. In order to solve this problem, the concept of "soft interval" is introduced, which allows certain points to not satisfy the constraint $y_j(\omega.x_j + b) \geq 1$. In order to measure how soft this interval is, a slack variable ξ_i is introduced for each sample, so that $\xi_i \geq 0$, and $1 - y_i(\omega \cdot x_i + b) - \xi_i \leq 0$.

After increasing the soft interval, our optimization goal becomes the formula (12) and formula (13).

$$\min_{\omega} \frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{m}\xi_i \tag{12}$$

$$\mathrm{s.t.}\, y_i(\omega.x_i + b) \geq 1 - \xi_i, \xi_i \geq 0, i = 1,2,\cdots,N \tag{13}$$

Where ξ_i is the "slack variable", which represents the degree to which the sample does not meet the constraints. C is called the penalty parameter. The greater the value of C, the greater the penalty for classification.

For samples that are linearly inseparable in a finite-dimensional vector space, map them to a higher-dimensional vector space, and then learn to obtain a support vector machine by maximizing the interval, which is a nonlinear SVM. Use x to represent the original sample point, and Φ(x) to represent the new vector after x is mapped to the new feature space. Then the split hyper plane can be expressed as $\mathrm{f(x)} = \omega.\Phi(\mathrm{x}) + \mathrm{b}$.

After the low-dimensional space is mapped to the high-dimensional space, the dimensionality will be very large, which will cause trouble in calculating the inner product. The kernel function can be used, because the inner product of the kernel function and the nonlinear mapping function is the same, $\mathrm{K}(X_1, X_2) = \phi(X_1).\phi(X_2)$, at this time the amount of calculation is much less than the inner product. Common kernel functions are polynomial kernel $\mathrm{K}(X_1, X_2) = (X_1^T X_2)^n$; radial basis function kernel (Gaussian kernel) $\mathrm{K}(X_1, X_2) = \exp\left(-\frac{\|X_1 - X_2\|^2}{2\sigma^2}\right)$; Laplacian kernel $\mathrm{K}(X_1, X_2) = \exp\left(-\frac{\|X_1 - X_2\|}{\sigma}\right)$; Sigmoid kernel $\mathrm{K}(X_1, X_2) = \tanh(a(X_1^T X_2) - b), a, b > 0$.

4. Experimental results

This article chooses the LIBSVM toolbox to implement the recognition algorithm, because it is simple and provides many default SVM parameters[9]. We set the SVM type to C_SVC, the kernel function type is RBF, and at the same time, the data is scaled, and the scale range is [-1,1]. The correct rate was evaluated using cross-validation. This experiment selects the MSR Daily Activity 3D data set for verification, which is collected by the Kinect device and contains 16 action types, including RGB, depth, and skeleton data. 10 people, each performed each action twice, a total of 320 action samples, a total of 3*320 files[10].

In the experiment, we used the combination of dynamic features and static features to achieve 78.88% accuracy through support vector machines. Thi Lanle [11] used dynamic time warping for action recognition, and the correct rate is 54% on the MSR DailyActivity3D data set. Shi Xiangbin [12] used the K-means clustering algorithm to extract the key frames in the human action video sequence.

Taking the position of the joint points in the key frame and the angle of the skeleton as features, using the SVM classifier for classification, the accuracy rate on the MSR Daily Activity 3D data set is 62%. It can be seen that the method in this paper has significantly improved the action recognition rate.

5. Conclusion

In this paper, from the skeleton data obtained by Kinect, static features and dynamic features are extracted to represent actions; SVM classifier is used for action recognition. It is verified on the MSR Daily Activity 3D data set, and the experimental results show that the method in this paper improves the accuracy of action recognition.

Acknowledgments

Thanks to my mentor Cheng Xinrong, Wang Qing and Chen Hong. Thanks for their valuable suggestions for my article.

References

[1] Chen Bin, Shi Yan. Research and development of somatosensory virtual mouse based on Kinect [J]. Software, 2016, 37(02): 46-49 (in Chinese).

[2] Mao Yijie. Research on continuous action recognition of human body based on Kinect [D].Sichuan: University of Electronic Science and Technology of China, 2017:3-28(in Chinese).

[3] Shotton, J., Fitzgibbon, A., Sharp, T., et al. (2011) Real-time human poses recognition in parts from a single depth image. In: Proceedings of the IEEE Conference on Recognition. Colorado Springs. pp. 1297-1304.

[4] Carlsson, C., Carlsson, S., Sullivan, S. (2001) Action recognition by shape matching to key frames. In: Proceedings of the Workshop on Models versus Exemplars in Computer Vision. Colorado. pp. 1-8.

[5] Ji, R., Yao, H., Sun, X. (2011) Actor-independent action search using spatial temporal vocabulary with appearance hashing. Pattern Recognition, 44 (3): 624-638.

[6] Liu Fen, Wu Zhipan. A Kind of Human Action Recognition Algorithm Based on Kinect and SVM [J]. Research and development, 2019, 18(011): 1007-1423 (in Chinese).

[7] Yang Ming, Liu Yang, Li Shaowei. A calculation method based on Kinect skeletal joint angle [J]. Science and Technology Information, 2018, 16(18): 102-103 (in Chinese).

[8] Wu Hongfeng. Research on human action recognition based on kinect human skeleton model [D]. Shanghai: Shanghai Normal University, 2016:6-38(in Chinese).

[9] Chang, C.C., Lin, C. J. (2011) LIBSVM: a library for support vector machines. ACM Transactions on Intelligent Systems and Technology, 2(1):2-27.

[10] Wang J, Liu Z C, Wu Y, Yuan J S. (2012) Mining actionlet ensemble for action recognition with depth cameras. In: Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). USA. pp.1290-1297.

[11] Thi Lanle, Minhquoc Nguyen. (2013) Human posture recognition using human skeleton provided by Kinect. In: Proceedings of the 2013 International Conference. USA.pp.340-345.

[12] Shi Xiangbin, Liu Shuanpeng, et al. Human action recognition method based on key frame [J]. Journal of System Simulation, 2015, 27(10): 2401-2408 (in Chinese).