Study on the Technology of Real Time Motion Recognition and Based on the Retrieval Table

Hailong Jia
Center of modern education technology Xinxiang University, Xinxiang , China
Email: 47160216@qq.com

Pei Tang
Department of art design Xinxiang University, Xinxiang, China
Email:jhl_117@163.com

***Abstract*—This paper presents a real-time gesture recognition method based on motion retrieval table, motion data input in real time can be identified through calculation of the result of its parts in the corresponding retrieval table. The method is superior to the other popular ways in recognition accuracy and recognition speed.**

***Index Terms*—body movement, recognition, real-time, the retrieval table, algorithm research**

I. INTRODUCTION

In the modern society, highly overcrowded, complicated social relations, people face more and more unexpected and unusual events, almost all public places have deployed monitoring requirements; in scientific research or remote exploration, usually using intelligent robots to perform specific tasks, and moving target detection is one of the most basic technology; in traffic management system, especially in the monitoring system of key roads and highways, the motion detection technology is used for monitoring vehicle flow and emergencies , very convenient for traffic system control and management; in many other areas, such as animation, the game control, the sports analysis , it can also find an utilization[1].

It is a difficult point to understand and analyze the video pictures in real time by using the computer system. Mainly because the motion data are of high dimension, high specificity, high continuity and real-time performance. So it proposes higher demand for a very quick speed of the action recognition, which means the system is able to identify an action at the same time when the users perform it without waiting for the completion of this action. On the other hand, because the user's action is caught in the field, it requires that the identification method has the functions of fault tolerance, which means it can also detect or identify the illegal actions (i.e. actions which are not defined in the system database) made by the user.

The traditional methods of identification are difficult to simultaneously achieve the recognition speed and high recognition accuracy goals. For example: Alon et al., [1] use CDP to recognize gestures input in real-time, and improves the recognition accuracy and accelerate the speed of recognition through the sub pattern reasoning and independent template filtering strategy. However, based on dynamic programming, this method has great computational complexity, which is not suitable for many kinds of action. Chai et al. [2] divide the motion samples into groups according to their similarities and apply action prediction to each group by shaping a local PCA model. Li et al. [3] use SVD to construct the feature space and extract the most main component motion data, and based on this feature use SVM to achieve objective of real-time action recognition. Because the data dimension is reduced greatly, the identification speed of this method can be guaranteed, but they rarely consider the recognition problem of wrong action.

In this paper, through the analysis of large amounts of movement data, the author found the differences among similar action modes are mainly caused by the movement of local part of limb. Based on this, real-time gesture recognition method based on retrieval search, proposed in this paper, can be used to divide the motion data set into five subsets based on human limb structure (limbs and trunk, the five part) and process the data respectively in each subset, such as clustering, mapping, etc. For each type of action categories, the training data are gathered and integrated, and then get a construction of such a general model (GM). This model can concentrate main features of the action and can be used as a common template for them. Based on these GM, we can generate a motion index table for these five parts respectively. And the action data of real-time input can be identified respectively through calculation of the result of its parts in the corresponding retrieval table.

The first section of this paper presents the formation of GM and the construction method of retrieval table; section two introduces the method of using retrieval table to recognize the movement flow in real time; section three validates the algorithm on two publicly video database, and compares the experimental results with the existing algorithms; section four makes a summary of this paper. The following part is the body of this paper.

doi:10.4304/jcp.9.10.2414-2419

II. Generation of GM and the Structure of Retrieval Table

This paper applied content retrieval technology to action recognition problem. It divided the movements of the body into a plurality of independent parts and each of them can be processed in a clustering way. But compared with the existing methods, our method has the following two different aspects[3]. First, we construct a GM to capture the differences between the samples and to retain their principal features of each action class. Second, our retrieval table is based on the GM rather than directly on the action data characteristics.

Figure 1 shows a global view of the recognition method proposed in this paper. It is divided into two parts, the training part and recognition part. The former is for training the GM and structuring the retrieval table, while the latter is for action recognition by using the model or the retrieval table of the former one.

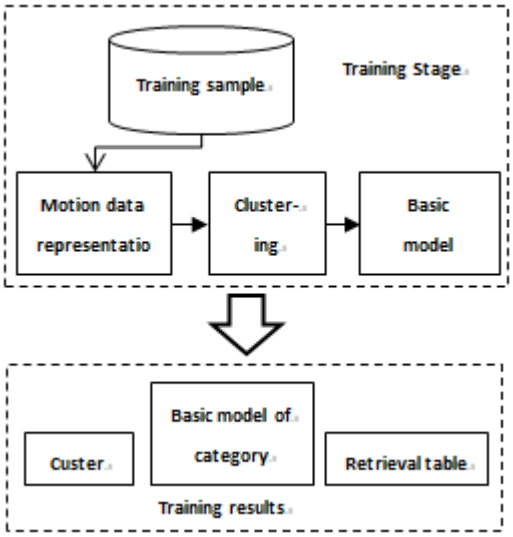

Figure 1. The basic model and the retrieval table training.

A. Motion Data Representation and Clustering

Although the action data can be intuitively represented in joint coordinate way, which is also conducive to the analysis of the corresponding changes between the joints, it still has a defect. Since different capturer may in different height and size and the same action they captured will have large difference, the following data processing (such as matching, identifying, etc.) will be in trouble. In order to facilitate unified treatment on the movement data of different capture acquisition, we will take 3D rotation joints (such as Euler angles) as the measurement method, among which the level of joint representation of the BVH (Biovision Hierarchy) format could be set as a data representation format. In the BVH data format, the angle value of each joint (except the root node "Root") is measured between it and its relative rotation of parent joint. Thus, as for an action M of m frames, it can be represented in the following formula (1):

$$M(t)=\{p(t),r\_0\ (t),r\_1\ (t),r\_2\ (t),\ldots,r\_|j|\ \ (t)\},$$

$$t=1,2,\ldots,m \qquad (1)$$

We use the BVH format, namely joint angle characteristics such as formula (1) described can stand for the action features, and we can remove the information of Root joint. As what is discussed earlier, in order to better capture the characteristics and difference of movement data, we will divide the human body into the trunk, left upper arm, right upper arm, left lower limb and right lower limb in five independent parts. Accordingly, all the action data is divided into five subsets based on the five parts and are respectively processed[4].

In order to set a further analysis for each subset data, we will Kmeans cluster each of them in a frame unit. In the process of clustering, the number of clusters is determined by the quantization error, including the definition of the quantization error is as shown in formula 2.

$$QE = \frac{\sum_{i=1}^{K}\sum_{x_j \in c_i} d(x_j, c_i)}{\sum_{i=1}^{K}|c_i|} \qquad (2)$$

Among them, K represents the number of generated cluster, C_i is the i[th] cluster and the number of its members and the cluster centers are respectively $|C_i|$ and c_i, while x_j refers to an arbitrary data point, and the function $d(\cdot)$ is as the distance function used in the clustering algorithm.

B. Training GM

In order to better model for every action category, we train a general model (GM) for each action category to capture the main feature of the category and to get a maximum concentration of differences between the samples. Assuming that each action category has a K training examples, through the cluster mapping, we use DTW to match the samples and integrate the results generated from GM[5].

As now each frame is a vector which consist of index numbers of five limbs movement, we need to define a similarity computing method for a new frame. For the two frames X and y, their similarity can be calculated by the formula (3) and (4), wherein P represents the index of feature value in a frame, x_p and y_p are respectively for the corresponding value of frame x and y.

$$s(x,y) = \begin{cases} h(x,y) & if\ \ h(x,y) > 2, \\ 0 & else \end{cases} \qquad (3)$$

$$h(x,y) = \sum_{p=1}^{5} h_p(x,y),\ \ h_p(x,y) = \begin{cases} 1 & if\ x_p = y_p, \\ 0 & else \end{cases} \qquad (4)$$

It can be seen from the above two formula that h (x, y) is the number of similar position in five sites between frame X and Y, and only when the number of similar position is for the most, which is more than or equal to 3, the global similarity of S (x, y) will be given positive matching scores, otherwise not counting the scores. In this way, those frames that only have 2 or less similar poses to each other are considered in big differences and will not participate in the similarity calculation, and effectively avoiding the negative impact of their calculation on the follow-up action similarity.

C. Structuring the Retrieval Table

Content retrieval technique was demonstrated to be very effective in the similarity search problem of large-scale data and fast in processing speed. Based on the GM trained above, we can construct a content retrieval table for each part of the body (namely the left arm, left leg, right arm, right lower extremity and trunk) respectively. In the process of constructing, each GM is ergodic and its frame records will be converted to the contents of retrieval table. The size of each retrieval table is determined by the corresponding parts of the cluster size. For example, for the trunk, if the corresponding motion data cluster to the formation of K clusters, the generating torso motion retrieval table also contains the K records[6].

Each record consists of two attributes, namely the record label and record content. Wherein the record label, corresponding with family index, can be used as a unique identifier for recording and the record content includes the location information of frames Indexed by the label in all the GM.

D. Identification Algorithm Based on GM and the Retrieval Table

In this paper, we introduced the action recognition algorithm based on GM and the retrieval table, which is used to solve different recognition problems and includes split action pattern recognition and real-time stream of action recognition, as is shown in figure 2.

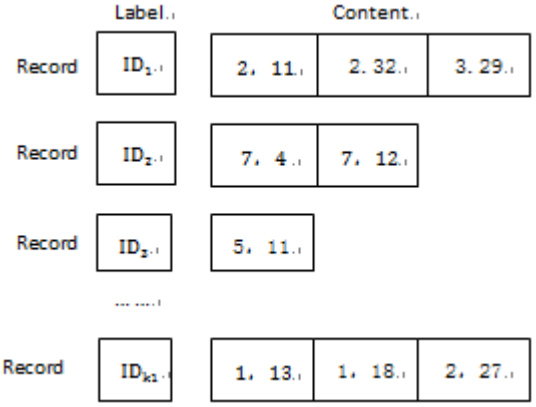

Figure 2. The retrieval table example

In the process of recognition, the general template representing each category (GM) can be used as a template pattern to directly match with the test action and for further recognition. In the process of recognition, each test sample is to map all the frames in accordance with the definition of body parts to the corresponding cluster and remove redundant index sequences in, and then matched with the GM. Because the GM of each frame is composed of five groups of index values, and each frame of test sample is a five value vector, so the former similarity function s (x, y) (see formula (3)) cannot be directly used in this case. Assuming that X and Y respectively represents the test sample and some certain frame of GM, in order to facilitate the comparison between them, we will redefine $h_{y_p}(x, y)$ in formula (4) as $h'(x, y)$ (as is shown in formula (5) below), and thus support the similarity calculation of s(x, y).

$$h'_{y_p}(x,y) = \begin{cases} 1 & if\ x_p \in y_{p'} \\ 0 & else \end{cases} \quad (5)$$

In the process of recognition, each frame of input actions is accordingly divided into eight parts, and every part will be mapped to trained clusters respectively. If the frame is the key frame (i.e. the mapping result of this frame is different from that of the previous one), the various parts of the frame will be indexed in the corresponding retrieval results table. The score of these results after accumulation can be used as a basis for identification, and inputting the end of action mode and whether it is legal action can also be matched to judge. The specific recognition process is as shown in figure 3.

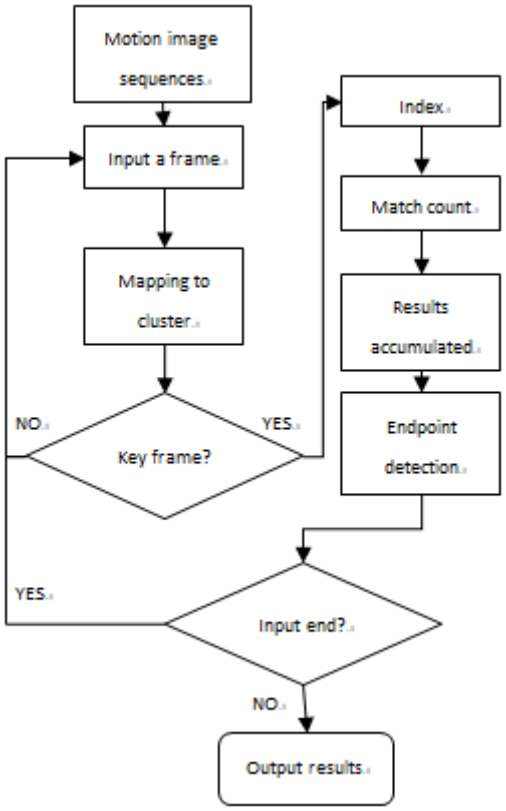

Figure 3. Action recognition process based on the retrieval table

III. Real-time Recognition Of Stream Of Action Based On The Retrieval Table

In the process of recognition, each frame of the input motion needs to follow the previous introduction successively, dividing the body into five parts first and then mapping each part to the trained cluster respectively. If the frame is the key frame, then the various parts of it will be retrieved in the corresponding retrieval table. The scores which are obtained after an accumulated result can be used as a basis for identification, and the end of input action mode and whether it is a legal action can also be judged by the match situation[7]. The specific recognition process is shown in figure 3, including three steps: match counting, scores accumulation and endpoint monitoring.

A. The Match Counting

Key frame is represented by a five dimensional vector, where each dimension represents the mapping of a part of the frame posture. We retrieve the match of each dimension in the corresponding table. For example, if the value of the current mapping frame in the trunk parts is d, then all content recorded in d in the torso retrieval table will be search results of the dimension. Thus, for these five dimensions, five sets of search results will be returned[8]. So we need to make further analysis of these results so that the matching results to the frame can be obtained.

Assuming that TO, LU, LL, RU and RL represent the set of search results which come from the retrieval table of trunk, left upper limb, left lower limb, right upper limb and right lower limb, we defined the following two kinds of strategy counting the match results.

Definition 1: [Whole-body matching (WBM)]. If a (GM) exists in the above five sets at the same time, that means the key frame is quite similar to the posture contained in the j_{th} frame in GM, so we can define the Whole-body Matching (WBM) of the key frame and the j_{th} frame in GM. Each key frame may not exist or may have multiple WBM in these collections, and WBM can be obtained by calculating the intersection of these five sets out, as is shown in formula (6):

$$WBM=TO\cap LU\cap LL\cap RU\cap RL \quad (6)$$

Definition 2: [Left side or Right side Body Matching (LRBM)]. Because of a big difference between actions, we found that although the two movements are similar, when they are in the matching, many key frames may not always return WBM results. Therefore, in order to avoid the negative effect brought by the rigid matching, we defined the left side or right side body matching (LRBM), i.e. if GM appears in the sets of trunk, left arm and left leg at the same time, we believe that the current key frame matches with the $Frame\#_{th}$ frame in GM approximately. The results of LRBM can be calculated by the formula (7):

$$LRBM=(TO\cap LU\cap LL)\cup(TO\cap RU\cap RL) \quad (7)$$

It's easy to learn, from the above two kinds of definitions, that the WBM results of the same key frames must be a subset of its LRBM results. Because if the five parts are all matched, the left side or the right side will also be similar with each other, and vice versa.

B. Results Accumulation

Based on the retrieval results of last step, a simple way of matching results accumulation is to select the GM from GM (including WBM and LRBM) which is matched the highest number as a result of recognition. However, this approach does not take into account the timing issue of frames in action. For example, the two different actions, such as "squat" and "stand up", can easily be misidentified, which due to their similar frame but different frame timing. Therefore, we need to consider the consistency in timing sequence of the input frame with the matching frame in GM[9].

In order to avoid this problem and its possible negative effects, we propose the following two principles in the search and matching of key frame of the input data:

Firstly, try to minimize the timing difference between matched frames. As for the input frame, it may have more than one match in the same GM. In this case, if the match of its previous frame is (GM, f) in GM, select the nearest one whose index is greater than f as the matching from the candidate frames. On the other hand, in order to avoid interval excessive situation between two adjacent matched frames in GM, we will restrain the candidate matching on the interval. For example, (GM, fi) is the candidate matching, that is $fi>f$, but we require that the interval between fi and f must be in a certain range, for example, we must ensure that $|fi-f|< \delta$ (δ is a constant), or the matching is a failure.

Secondly, LRBM must be the first and WBM second. Based on the first principle, the input key frame may have the LRBM and WBM results to meet the conditions at the same time. In order to avoid repeated computation of LRBM and WBM, simultaneously to enhance the elastic matching of input action, we search its LRBM results according to the first principle in search the match of the key frame. If the LRBM does not exist, then end the matching search of the current key frame. If the LRBM exists, then judge that the match is also WBM according to definition 2, and if the judgment is true, the matching performance of the corresponding GM will increase Δ ($\Delta\rangle$ 1), or increase only 1.

The above two principles respectively consider the issue of time sequence and space difference in search of matching of key frame. Principle one makes the input frame keep a temporal consistency with the matching frame in GM, and effectively avoids the local optimal problem in finding the best match for the possible key frame through the introduction of interval limits. Principle two ensures the matching priority through LRBM, which makes the input frame always find its match in the largest possibility so as to accurately classify the similar actions. Finally, the recognition of input action depends on the highest scored GM.

C. Endpoint Detection

For real-time input action flow, in most cases it comes from online capture in continuous manner, thus the input action may contain a number of different categories of action mode, and may also be incorrect action (i.e. undefined action) [10].Therefore, when we are in the real-time segmentation and recognition for these legal action patterns, we need to determine if the current input action frame is the legal action. In this step, our solution is based on the following two discoveries: (1) although the action modes are different in length, the legal action pattern length has a lower limit. If the length of the current input action is less than the lower limit, we will take it as incomplete action or illegal action. (2) it is very difficult for the key frame to continue matching with its corresponding GM frame after the current input key frame end frame movement pattern, which is difficult to make the GM results continue to grow. Therefore, we require that if the following three conditions are true at

the same time, we will believe that the end of the current input action has been detected, and the input pattern is recognized as legitimate and consistent with the GM_k class.

(1) GM_k obtains the highest cumulative score for all in GM.

(2) There currently has no key frames k consecutive matches GM_k.

(3) The current length of the input operation is GM_k matching frame is greater than the minimum limit.

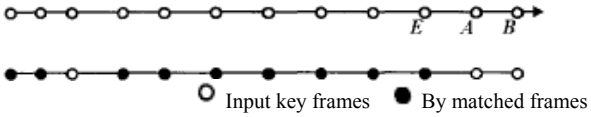

Figure 4. An endpoint detection example

In the process of recognition, if one GM at a time meets the above conditions, we are back to the matching process, find out the matching position of the key frame in GM, and take the position as the endpoint of current input action mode and continue doing the segmentation. Figure 4 shows a similar example, in which the position E is considered as the end of the current input action. On the other hand, if it only meets the first two conditions but not the conditions (3), then the current input action is identified as illegal action, and split it out. After any of the above two kinds of circumstances occur, we reinitialize the recognition program to process the input action later[11].

IV. Design And Analysis Of Experiments

We apply recognition method based on GM and the retrieval table to action recognition experiment to test their performance. This experiment is implemented in Matlab environment of a personal computer whose CPU is Intel Pentium 3.0GHz and internal memory is 2G.

The action data used in this experiment is divided into two parts, one is from the motion capture Lab of Beijing Institute of Technology, the other comes from the opened HDM05 motion data set[12].

We applied recognition methods based on GM in split action mode set of Polytechnic U and HDM05 to test its performance. For each data set, we have conducted action pattern recognition experiments of 3 groups of the user independently, for each experiment, we choose the samples which were collected by three captors as the training set, while the other two captor's samples as test sample. For the first group, the first three captors' data can be used as the training data, and then the latter two captors' samples can be used for testing. Moreover, for each type of action, we just choose the first two samples for training, that is to say each kind of action has only 6 training samples[13].

Table 1
The Recognition Accuracy Rate of DTW、LCS、SW and PDTW Based on GM in A and B Data Sets Respectively

		Group 1	Group 2	Group 3	Average
Polytech-nic U Data set	GM+DTW	0.9474	0.9474	0.9385	0.9443
	GM+LCS	0.8772	0.8684	0.9210	0.8887
	GM+SW	0.8246	0.8772	0.8508	0.8507
	GM+PDTW	0.9386	0.9386	0.9474	0.9414
HDM05 Data set	GM+DTW	0.9138	0.8501	0.9166	0.8934
	GM+LCS	0.9482	0.8834	0.9666	0.9327
	GM+SW	0.9655	0.9100	0.9832	0.9496
	GM+PDTW	0.9828	0.8835	0.9668	0.9442

Table 1 shows the recognition accuracy of the DTW, LCS, SW and PDTW based on GM in the segmentation action mode set of Polytechnic U and HDM05. The result showed that, among all the methods, GM+DTW performed best in Polytechnic U data set while got the lowest cognition accuracy rate in HMD05 data set, and on the other hand, GM+SW is on the contrary with GM+DTW in the two data sets[14]. This is because the HDM05 data set has more significant differences in the time domain than Polytechnic U data set, so in the HDM05 data set, SW can effectively deal with the matching of unequal-length samples in same category by punishment mechanism, but not like DTW, which can easily bring the over fitting problem for matching the frame in the longer samples repeatedly. However, in the Polytechnic U data set, SW cannot produce its advantage because this time domain difference is small, and the punishment mechanism may even have a negative impact, so its performance is not as good as DTW. Because PDTW inherits the advantages of DTW and SW, it shows better stability in the two data sets, and the recognition accuracy in the two data sets is close to their optimal results[15].

In order to verify the action division according to the structure of human body and consider its superiority respectively, we analyze the movements of the body as a whole process in two data sets--- the science and technology U and HDM05. When dealing with the motion data as a whole process, we choose θ=45 as the quantization error of the Kmeans clustering, and after clustering there generates an average of 220 and 140 groups in data sets of the science and technology U and HDM05 after three sets of experiments. Figure 5 compares the recognition results of GM+PDTW under two schemes based on the body division and the body whole. As you can see, the classification performance of the body dividing scheme in the two data sets is better than that of the body whole scheme, especially in the data set HDMOS, this advantage is more obvious. This is because the data set HDM05 covers the basic daily movements, such as running, jumping, etc. and these

actions can bring about greater intra-class differences. For example, as for the two "walking" movements, one is a "walking and swinging left arm" action, while the other is the "walking but not shaking left arm" action. From people's sense of speaking, they belong to the same category, but due to the different movements of arm, it may cause the difference value of the movement data. Therefore, when the action is calculated and recognized as a whole, the recognition accuracy rate will be affected seriously, but our method based on the body division can avoid the disadvantages.

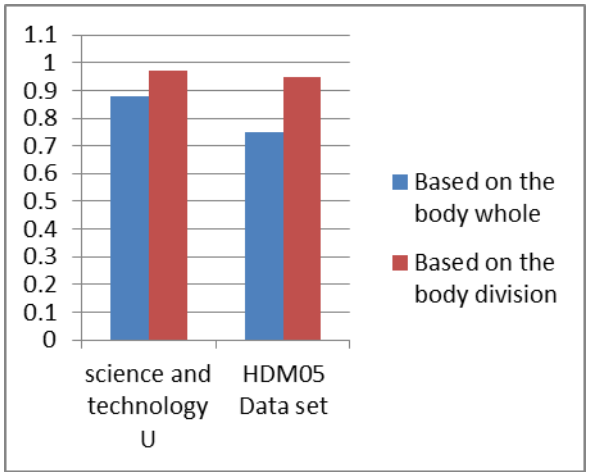

Figure 5 The recognition results based on the methods of the body whole and the body division

V. Conclusion

Through abundant analysis and study of action data, this paper comes to some conclusions. First of all, when it is in the recognition of simple static background action, a basic model for each action category can be used to capture the difference between the samples and to retain their major features. Secondly, the paper presents a retrieval table recognition method based on the basic model of limb. This method uses content-based retrieval technology, overcomes the defects of inconvenient similar action recognition, and improves the speed and accuracy of human action recognition.

References

[1] Alon J, Athitsos V, Yuan Q. et al. A unified framework for gesture recognition and spatiotemporal gesture segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2009,31 (9): 1685-1690.

[2] Clai J, Hodgins J. Performance animation from low-dimensional control signals. ACM Transactions on Graphics (TOG), 2005. 21(3):686-690.

[3] Li C. Khan L, Prabhakaran B. Real-time classification of variable length multi-attribute motions. Knowl. Inf. Syst.. 2008,10::163-183.

[4] Yi Gao． On Preliminary Research into Correlativity between Mechanical Characteristics of Sole of Foot and Upper Limb while Walking［J］.Journal of Computers,2011,10：2068-2075.

[5] Poppe R． A survey on vision-based human action recognition，Image and Vision Computin［J］，2010，28: 976-990.

[6] Qin L J，Zhu F.A new method for pose estimation from line correspondences[J].Acta Automatic Sinica，2008，34（2）：130-134.

[7] Duan F Q，Wu F C，Hu Z Y.Pose determination and plane measurement using a trapezium[J].Pattern Recognit Lett，2008，29（3）：223-231.

[8] Davrondzhon Gafurov, Kirsi Helkala． Biometric Gait Authentication Using Accelerometer Sensor［J］. Journal of Computers，2006，10：51-59.

[9] Kahol K. Tripalhi K. Panchanathan S. Documenting motion sequences with a. personalized annoLation system. IEEE Multimedia, 2006. 13(1):37-45.

[10] Raptis M, Kirovski D, Hoppe H. Real-time classification of dance gestures from skeleton animation. Proceedings of Proceedings of the 2011 ACM SIGGRAPII/Enrographics Symposimn on Computer Animation. ACM. 2011. 147-156.

[11] Lee J, Chai J, Reitsma P, et al. Interactive control of avatars aniiuated with human motion data. ACM Transactions on Graphics, 2002, 21 (3):491-500.

[12] Chan J, Leung H. Poizner H. CoiTelalion among joint motions allows classification of Parkinsonian versus normal 3-D reaching. IEEE Transactions on Neural Systems and Rehabilitation Engineering, 2010, 18(2):142-149.

[13] Li C. Khan L, Prabhakaran B. Real-time classification of variable length multi-attribute motions. Knowl. Inf. Syst.. 2008.10:163-183.

Hailong Jia. He was born in 1982.3, he incepted his bachelor degree in Air Force Engineering University in 2004, he obtained his master degree in Beijing Industry University in 2010, and his major is computer application technology. Now he is working in Center of modern Education Technology of Xinxiang University as a lecture. His research directions are image recognition, information and communication engineering.

Pei Tang. He was born in 1975.1, he incepted his bachelor degree in Zhengzhou Liberation Army Electronic Institute of Information Technology in 1998, he obtained his master degree in Yunnan University in 2006, and his major is computer application technology. Now he is working in Department of Art Design of Xinxiang University as a lecture. His research directions are data mining, network application.