MATEC Web of Conferences **44**, 02024 (2016)
DOI: 10.1051/matecconf/20164402024

A Somatosensory Interaction System based on Kinect

Xiu Bo Liang[1,a], Chao Wang[1] and Zhen Wang[2]

[1]*School of Software Technology, Zhejiang University, Ningbo, China*
[2]*College of Computer Science and Technology, Zhejiang University, Hangzhou, China*

Abstract. The somatosensory interaction technique is one form of the perceptual user interface which is used in video game and virtual reality more and more widly. In this paper, a somatosensory interaction system based on Kinect is presented. Firstly, the user performances his action in front of a kinect, the sensing data from kinect is preprocessed and the main features of the action are extracted. Secondly, the performaced action is recognized by the matching algorithm based on Dynamic Time Warping Hidden Markov Model. Finally, the recognized motion is employed to interact with the virtual human and virtual environment. A series of experiments have been done to test the availablity of our system. Results show that the recognition rate is high enough to be used in virtual reality applications.

1 Introduction

Somatosensory interaction is a kind of technology, referring that the user interacts with the machine through their body movements directly. This technology aims to build a more natural interaction environment, which simulates the user's scene to create a 3D virtual model. At the same time, it combines with the character recognition, gesture recognition technology, to identify the user's actions. The set of processes makes users seem to feel the real sense in the process of the use of somatosensory interaction. In recent years, with the rapid development of information technologies such as image processing and gesture recognition, somatosensory interaction gradually appears in the research and business field, which promotes the development of the somatosensory input devices. In this paper, we employ Kinect as the input device to develop a somatosensory interaction system. With our system, the user can interact with the virtual human by his body motion.

2 Related work

In recent years, new input device (e.g. Kinect) is used to implement the virtual human control interface based on the somatosensory interaction, which makes full use of the human perception ability to express the control intention in a natural way [1, 2]. Action recognition refers to the process that computers cognizing user's motion state by certain means over a period of time. The process of action recognition generally comprises two aspects: firstly extracting the features of motion, and secondly recognizing actions using the extracted motion features. After extraction to motion features, the next job is to identify corresponding actions which contains characteristics of these movements. In order to satisfy the visual rationality, the system can generate the action which conforms to the physical law [3-5]. Sequence of actions can be considered as trajectory in the model parameter space, and each different action category can be grouped into a subset of the model parameter space. Action recognition, is the process to classify the sequence of actions to be identified into a subset of the space. Common action recognition technologies are: action recognition based on template matching, action recognition based on statistics and so on. Action recognition based on template matching often uses the algorithm of dynamic time warping (DTW). Indexing uses lower-bounding functions to prune out the number of times DTW needs to be run for certain tasks such as clustering a set of time series or finding the time series that is most similar to a given time series [6][7].

Action recognition based on statistics usually applies statistical based approach to modelling the action sequence of common probability statistical models, and the common statistical probability models are known as Hidden Markov Model (HMM) [8], etc. HMM have been widely used in many fields, such as speech recognition, activity recognition from video, gene finding, gesture tracking [9]. There are two key reasons. First the models are very rich in mathematical structure and hence can form the theoretical basis for use in a wide range of applications. Second the models, when applied properly, work very well in practice for several important applications.

3 System overview

Kinect is a sensor based on the theory on optics, through the device, we can get a sequence of actions to the system. Kinect somatosensory input device can provide human skeleton data of the 20 joint point, the skeletal data including the space position about the joint points in the coordinate system of the Kinect. Thus, obtaining the skeletal data for a continuous period of time can reflect the movement trajectory of the joints in this time. Action recognition is based on these bone data, through a certain pretreatment process, the action data is normalized. The feature of these trajectories is extracted by using motion matching algorithm. After obtaining the characteristic, we need to match the real time data at first, and then analyse and process the acquired samples to generate the model and template which need to be used in the matching stage. Motion classification modeling is to model the action feature sample in order to provide the following feature matching process. The feature matching recognition

[a] Corresponding author: liangxb@cst.zju.edu.cn

requires the use of Dynamic Time Warping and Hidden Markov Model to identify these characteristics. Finally we get the recognition result which is used to interact with the virtual human.

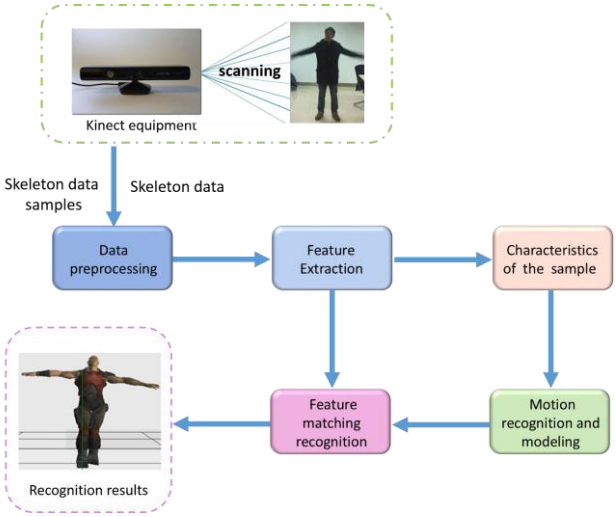

Fig 1. Overview of the somatosensory interaction system.

4 Data pre-processing and feature extraction

4.1 Data pre-processing

Action data is collected in accordance with the number of frames, and therefore we can not guarantee to collect these data at the time of the action started, and stop when the operation is completed. In this situation, we need to pre-process the data to ensure that the action data is correct.

There are two common situations when we collect the motion data: user has not started action in the starting frames of data; users has completed the action before the end of the motion data. The start and end of the action of the motion data is useless, therefore, it can be removed.

After removal of the useless part of motion data, the length of the sequence has changed. In order to ensure the consistency of the input data, it should be interpolated and sampled according to the specified length. By interpolation, we can obtain a continuous curve with all the known points, and the continuous curve is the fitting function of these data points.

4.2 Feature extraction

Feature to be replaced the action sequences. Features include the direction vector of the motion sequences, the distance, the speed of movement and so on. The process of extracting motion feature comprises calculating characteristic values from the original data, then the characteristic value de-noising processing.

4.2.1 Extraction speed feature

The data extracted from the Kinect device is coordinate data about the key points of the spatial position. The data is relative to each person's range of motion, not universal. It cannot represent a sequence of actions accurately. Therefore we need to extract location-independent features from these position sequences to represent this set of actions. Kinect samples data in a fixed time interval (30 frames per second), so the difference between each two frames can represent the speed of this action. That can be considered $v_t = P_{t+1} - P_t$. Applying the above formula, it can be considered to obtain the instantaneous velocity action. The characteristics is not relative with spatial location, so you can use the feature value in the subsequent process.

4.2.2 Extraction direction feature

In order to further simplify the input data, we can consider the motion direction of the corresponding motion from the instantaneous velocity of motion. For a set of coordinate data sequences, from the first data, the vector is calculated by the coordinates of the front and rear two points. We need to define an array of pre-set directions, multiply the vector with the direction of the various directions, take the maximum value, and put the value into the corresponding position of the array. Finally, the new array is added to the feature array. There are also some noise in the direction of the extracted features, the result is not good enough. In order to deal with this situation, we can choose the direction of the characteristic sequence of linear filter. The method is to scan the directional feature sequence, which will be all less than three consecutive equal points, and then replace it with the data in front of these points.

4.2.3 Feature extraction based on principal component analysis

Analysing principal component is a multivariate statistical method for representing the original variable sequence by using the linear transformation to select several major variables.

The process of principal component analysis [10] of data pre-processing is to subtract the mean number from each dimension and then divided by its standard deviation to achieve standardization, that is, set X is the $n \times m$ sample matrix, the result matrix is Y, each line represents a sample, and each column is a dimension of the sample. Set i as the line number, j for the column number. Then the result matrix of each Y is given by the formula (1):

$$Y_{ij} = \frac{X_{ij} - \overline{X}_j}{S_j} \tag{1}$$

Where $\overline{X}_j$ is the mean value of the j dimension, and S_j^2 is the variance of the j dimension. Next, the covariance matrix is solved by the standard matrix, and the covariance matrix method can be given by formula (2):

$$C = \frac{Y^T Y}{(n-1)} \tag{2}$$

After solving the covariance matrix, we can decompose the eigenvalues of c, get the feature vector matrix and orthogonal, then take the dimension of the corresponding dimension of the k eigenvalues, and the new feature vector p which is the projection matrix. In the end, the sample can be obtained by the dimension reduction matrix, which is projected by the projection matrix p.

5 Action recognition

5.1 Action recognition based on DTW

Template matching is a common method in pattern recognition. In the process of template matching, we calculate the similarity of the two sequences. And the similarity is generally measured by the distance between the feature vectors in feature space. Several commonly used distance, Euclidean distance, square and distance, absolute distance, weighted distance, etc.
Two n-dimensional feature vector X_1 and X_2,

$$X_1 = [x_{11}, x_{12}, ..., x_{1n}]^T, X_2 = [x_{21}, x_{22}, ..., x_{2n}]^T.$$

The Euclidean distance of the two vectors is defined as:

$$\|X\| = \|X_1 - X_2\| = \sqrt{(x_{11}-x_{21})^2 + (x_{12}-x_{22})^2 + \cdots + (x_{1n}-x_{2n})^2} \tag{3}$$

Square and distance defined as:

$$\|X\|^2 = (x_{11}-x_{21})^2 + (x_{12}-x_{22})^2 + \cdots + (x_{1n}-x_{2n})^2 \tag{4}$$

Absolute value distance is defined as:

$$|X| = |x_{11}-x_{21}| + |x_{12}-x_{22}| + \cdots + |x_{1n}-x_{2n}| \tag{5}$$

Weighted distance is defined as:

$$D = \alpha_1 |x_{11}-x_{21}| + \alpha_2 |x_{12}-x_{22}| + \cdots + \alpha_n |x_{1n}-x_{2n}| \tag{6}$$

Which $\alpha_1, \alpha_2, ..., \alpha_n$ are the corresponding eigenvalues of the weighted parameters. According to the characteristics of data preprocessing process, we can have the motion rate, directional vector characteristic value and the characteristic value of PCA dimension. When using the speed of motion for identification, the distance between the two motion sequences can be measured by Euclidean distance.

It can't achieve a high recognition rate that comparing and matching the action sequence and template directly. Because even for the same action, on the one hand, in the different time, made by different people, its duration can't often be ensure, on the other hand, with one action of each fragment, everyone duration also tend to be long or short, and the template cannot keep consistent. If the sequence of the template is used to match the template using Euclidean distance directly, it often leads to a low recognition rate. DTW is a nonlinear time alignment algorithm, which will be extended or shortened to a better matching template to calculate the correlation between the two time series. DTW to match the pattern as shown in figure 2.

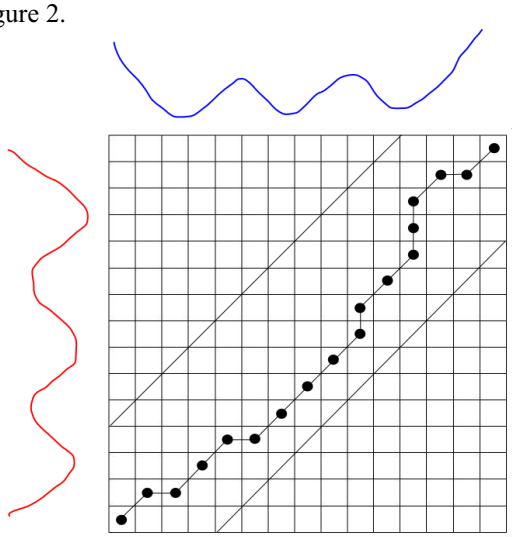

Fig 2. Dynamic time warping to match two sequences

After calculating the similarity between the sequence to be detected and the template using DTM, it can determine whether the sequence belongs to the same category with the test sequence according to the pre-set threshold. So the choice of the template has a great influence on the recognition. If the template cannot reflect all the characteristics of a type of action well, the recognition rate of the action will be reduced. In order to reduce the impact of template selection on the recognition rate, the same category of multiple templates are selected to compare when we decide the category of the test sequence at last, and the nearest neighbour algorithm[10] is introduced to classify the test sequences.

5.2 Action recognition based on HMM

Hidden Markov model is a recognition model used in the filed involved speech signal recognition based on statistics widely. When analysing the time-varying non-stationary sequence, the model can yield good results. As the action sequence is also the time variant signal, so the use of HMM to identify the action sequence, the sample size is large enough, can also achieve a high recognition rate. HMM practical application of the process need to solve the following three issues: [11]: assessment of the problem, decoding and learning problems. Forward-backward algorithm can calculate the probability of a given observation sequence in a particular model. So it can be used to solve the problem of HMM's assessment. By dynamic programming, the Viterbi algorithm can be used to determine the global optimal path, which can be used to solve the decoding problem. HMM model parameters selection and optimization problems are usually solved by Baum-Welch algorithm, Baum-Welch

algorithm is an iterative algorithm, the initial parameters of the model are given by the user, the algorithm through continuous iteration, these parameters will gradually tend to be more reasonable value.

6 Experimental Results

In order to verify the validity of the system, we carried out some experiments. The process of action matching is to classify the motion data calling the action matching module. First, the real-time motion data is pre-processed to obtain the feature vector. Then the feature vector is used as the input to classify the action matching module. The final action matching module will give the classification results. In the experiment, we draw the digital 1-9 as a test case using the right hand gesture, and test the dynamic time warping and the action matching module of Hidden Markov model. Test case of the action form, as shown in figure 3.

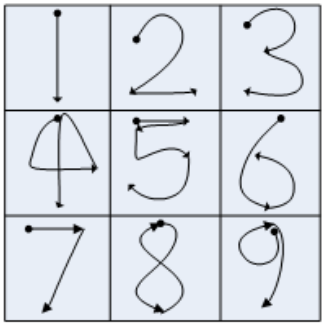

Fig 3. Test action pattern.

The test operation is carried by the user waving the right hand in the space, and the trajectory is shown in Figure 4. The solid points represent the position of the action at the beginning of the movement, and the movement of the arrow and line represents the trajectory of the movement. Each of the 20 sets of tests are used for testing. HMM and DTW were used to identify these tests, and the results are shown in figure 4:

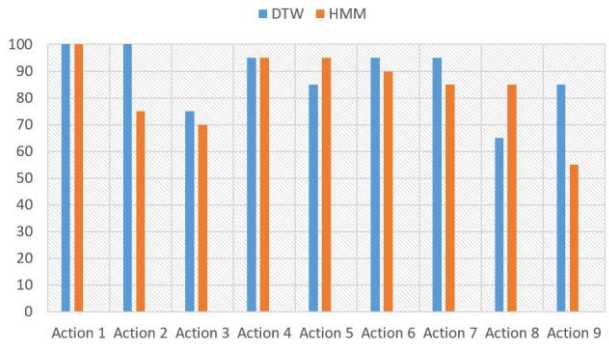

Fig 4. The test results.

Using cross validation test method, all the collected data samples are divided into test set and training set, the test set contains 20 samples, the training set contains 80 samples. From the results, we know whether the use of HMM for recognition or the use of DTW recognition, most of the action can achieve more than 75% of the recognition rate. If there are some context constraints, such as the type of action that may appear in a restricted scene, increase the morphological differences between the actions, or restrict the action of the front and back, you can also continue to improve the recognition rate of these gestures.

7 Conclusion

In this paper, we develop a somatosensory interaction system using Kinect. The captured motion should be pre-processed and feature-extracted before using for recognition. DTW and HMM are selected as the motion matching algorithm. Experimental results show that the recognition rate is high enough to be used in virtual reality applications. However, the system still exists several problems. The way of interaction is simple and rough, and it does not present an exclusive interface of somatosensory interaction well. In the feature, we will try to overcome these shortcomings.

Acknowledgement:
This work was partly supported by Ningbo Natural Science Foundation (Grant no: 2013A610064) and intelligence industry talent base of Ningbo.

References:

1. A. SHIRATORI T, HODGINS J K. Accelerometer-based user interfaces for the control of a physically simulated character [J]. ACM Transactions on Graphics, 2008, 27(5): 1-9.
2. H. SHUM, E. HO, Y. JIANG, et al. Real-time posture reconstruction for Microsoft Kinect[J]. IEEE Transactions on Cybernetics, 2013, 43(5): 1357–1369.
3. Luigi T.De Luca, *Propulsion physics* (EDP Sciences, Les Ulis, 2009) .
4. W. ZHAO, J. ZHANG, J. MIN, et al. Robust realtime physics-based motion control for human grasping[J]. ACM Transactions on Graphics, 2013, 32(6): 1–12.
5. F. HAHN, S. MARTIN, B. THOMASZEWSKI, et al. Rig-space physics [J]. ACM Transactions on Graphics, 2012, 31(4): 1–8.
6. P. HÄMÄLÄINEN, S. ERIKSSON, E. TANSKANEN et al. Online Motion Synthesis Using Sequential Monte Carlo [J]. ACM Transactions on Graphics, 2014, 33(4).
7. E. Keogh, Exact Indexing of Dynamic Time Warping. In VLDB, pp. 406-417. Hong Kong, China, 2002.
8. S. Kim, S. Park & W. Chu. An Index-based Approach for Similarity Search Supporting Time Warping in Large Sequence Databases. In Proc. 17th Intl. Conf. on Data Engineering, pp. 607-614. Heidelberg, Germany, 2001.
9. Elliott R J, L. Aggoun,Moore J B.Hidden Markov Models[M].Springer,1995.
10. T. Cover, P. Hart, Nearest neighbor pattern classification[J]. Information Theory, IEEE Transactions on,1967,13(1):21-27.
11. I. Jolliffe, Principal component analysis[M].John Wiley & Sons, Ltd,2005