Journal of Multidisciplinary Engineering Science and Technology (JMEST)
ISSN: 2458-9403
Vol. 4 Issue 10, October - 2017

Motion Recognition Based on Depth Feature and Dimensional Reduction

Jian Xiang,

School of Information and Electronic Engineering,

Zhejiang University of Science and Technology,
310023 Hangzhou, China

freenyspi@gmail.com

***Abstract*—Due to the emergence of many three-dimensional(3D) human motion databases, it has been a new challenge to effectively use motions to capture databases through analyzing and handling data about human motions. However, in-depth image has currently been a critical means to capture in-depth information about 3D human motions. Through this means, we can realize the retrieval and recognition of 3D human motions by carrying out data dimension reductions and feature fusion based on manifold-learning clustering methods on in-depth features of data about 3D human motions that we extracted**

Keywords—human motion, depth feature, dimensional reduction

I. INTRODUCTION

With the development as well as mature of motion capture and 3D scanning techniques, high-definition video technology has been improved constantly. Since 3D human motion data generated through various approaches in every year have been larger and larger, numerous 3D human motion databases have been established. Therefore, it has profound theoretical meanings as well as vital practical values to use these databases rationally and efficiently, to realize effective retrieval and recognition of 3D human motion databases, to use data that have been analyzed in recreation of 3D human motion data and to finally use them in fields including digital media[1][2].

Extracting in-depth features is currently an effective means in field of features extraction of 3D motions. In a scene, the distances between points and cameras can be expressed by in-depth images. In other words, the value of each pixel on in-depth images is the distance between corresponding point and camera lens in the scene. In motion capture techniques, in-depth information are very important information that can compensate information losses caused during the process from 3D spaces to 2D image mapping. It is their importance that forced researchers to explore numerous means for obtaining them. While in researches based on in-depth data, there emerges a great many of cheap in-depth cameras with higher precision, such as kinect produced by Microsoft, due to the reform and development of technologies. However, the precision rate of recognition is disguisedly promoted for in-depth data are major expressed by distance

II. BACKGROUND

Compared with traditional RGB images, in-depth images can illustrate 3D space information. When scenes in traditional images are converted from 3D to 2D, in-depth distance information are lost. However, by combining in-depth images with RGB images, 3D scenes can be recovered correctly, which can promote the precision of human motion recognition. In a in-depth image, every pixel point can stand for a real distance between objects in a scene. Then, how to apply these distance information in recognizing human motions has been a hot issue in the field of computer vision. However, as the release of relatively cheap sensors that can obtain in-depth image data, such as kinect produced by Microsoft, human motion recognition that based on in-depth images has been developed intensively.

As for features extraction on in-depth images, it can be classified into global features extraction and local features extraction. Recent years, researchers have achieved a lot findings in researches aiming at human motion recognition relevant to in-depth images[4][5][6].

In reality, before kinect, there had been plenty of researches on how to acquire in-depth information of a scene. In the working of Matyunin[3] et al., they acquire in-depth images through stereoscopic vision, while we collect RGB images through setting common cameras in certain angles. Since this method cannot correctly acquire in-depth information in a scene, we can apply multilevel collecting method in solving this problem and compared with traditional cameras, this idea has been a great progress.

There is a common idea to apply research achievements on RGB images in in-depth images. In the working of Tang et al[7]., they applied features of motion history images mentioned above in in-depth images.

In the working of Zhao et al[8]., they had once normalized in-depth images into gray scale images, after which features of spatio-temporal interest points mentioned above are extracted. However, this means does not take features of in-depth images into full play. Meanwhile, it cannot well use 3D space information

that it expresses. But, it is indeed a good idea for extracting features of spatio-temporal interest points on in-depth images, for it is a common idea to extract interest points before extracting features of every interest point. This idea is commonly used by domestic scholars.

The another common idea is to extract some new features based on features of in-depth images. In the working of Li et al[5]., they adapted an idea to project in-depth images. First of all, they extract the area that people located from in-depth images, after which projection will be carried out from three directions to extract pixel points from the edges of three 2D data obtained after projection. These pixel points can be regarded as current feature expressions of people.

III. In-depth Feature Extraction

Integrating efficiency and precision, we used the feature of Three Dimensional Motion History Images(3D-MHIs). This feature presents motions of objects in the way of brightness based on the template method of vision. When the motion gets closer to current time, its gray scale value will be higher, or its corresponding motion area in the picture will be brighter. When we got a video sequence, we first carry out background subtraction on it. Then, subtraction of foreground figures will be operated. 3D MHI is a expansion of traditional MHI. Supposing that M = [m1,m2, . . . mn] is a video sequence, of which mi is i-th pause and n is the amount of all pauses or the length of the video. Then, supposing that mt(x,y) is the pixel point on the x-th line and y-th row in the i-th frame, the formula used for computing features of MHI is shown as below:

$$H_t(x,y)=\begin{cases} t, & \text{当} \mathrm{m}_t(x,y)-m_{t-1}(x,y)>\alpha \\ Max(0,m_{t-1}(x,y)-1) & \text{else} \end{cases}$$

The effect of 3D-MHIs feature is much higher than other in-depth images features. However, the advantages in computing efficiency and storage spaces of 3D-MHIs feature are obvious: we neither need to keep previous image W and motion areas, nor do we need to carry out further operation on these data, as a result of which, it is high-efficiency, time-saving and space-saving. These are reasons why we use 3D-MHIs feature among methods we mentioned in this paper. What's the most noteworthy is that when we get in-depth images, we need to compute the size of bounding box around human based on in-depth videos. However, we cannot ensure that the height and profile of all performers are the same. Therefore, we need to operate normalization to them to make their pixel unified at 64×64. As a result, dimension of 3D-MHIs feature is 12288. This a huge figure. For this reason, we need to carry out dimension reduction on extracted in-depth data.

IV. Dimension Reduction and Fusion

We adapted Isomap in the fusion and dimension reduction of features above. ISOMAP is a means for manifold learning, while manifold is a notion in topology. Here, a manifold can be simply understood as a n-dimensional hook face in space $R^N(n\square N)$ determined by $f(x)=0, x\in R^n$ and it represents a topological space with Euclid spread everywhere in some part, while local Euclidean feature means that every point in the space has a neighborhood. However, Euclidean space is the most simple example for manifold, just like spherical surface like the earth is a slightly more complex one. General manifold can be adhered by some twisted straight pieces. ISOMAP is a typical spectral-analyzing manifold learning method. This dimension-reduction method features isometric mapping based on certain image space. Isometric mapping supposes that mappingφ from higher dimensional spaces to lower ones keeps geodesic distance remain unchanged between two points. In other words, as for two points m,m' on manifold as well as their projection on mapping $\varphi\theta,\theta'$, there is a formula $G(m,m')=\left|\theta-\theta'\right|$. In this formula, distance between two points on manifold $G(m,m')$ can be regarded as the minimum distance that a bug crawls from m to m' on manifold.

According to ISOMAP, it will be meaningful for two points on manifold when they have a relatively smaller Euclidean distance. Therefore, in this algorithm, adjacent image is first established with each point being connected only with its adjacent points. Manifold distance between two points in the same neighborhood approaches from their Euclidean distance, while manifold distance between two points in different neighborhood approaches from their minimum Dijkstra distance on adjacent image.

If we got N data points $\{x_i\}_{i=1}^N$, the procedures of clustering algorithm based on manifold learning will be listed as follows:

Major procedures of ISOMAP algorithm:

1. Establish a adjacency matrix G. There are two methods that can decide G's edges: ε neighborhood and K neighborhood. When point j is in point i's or point i is in point j's ε or K neighborhood, an edge with weight as $d_x(i,j)$ will be established between i and j.

2. Computing the minimum distance between points by Dijkstra or Flyod algorithm, we can get a matrix DG, in which all elements are minimum approaches of every pair of points in G:

$$D_G=\{d_G(i,j)\}$$

3. Carry out feature decomposition on L to get feature values of the top m $\{\lambda_i\}_{i=1}^m$ as well as their corresponding feature vector $\{v_i\}_{i=1}^m$.

So far, we get 3D human motion data based on in-depth features. We can apply these data in recognition and retrieval of human motion data, for follow-up and traditional framework of in-depth features as well as video features will realize the establishment and fusion of multimoding human databases.

V. Experimental Result

To show the effectiveness of this algorithm, we first verify it based on public data set MSR-ACTION 3D. There are 25 different motions with each contained multiple fragments in this data set. As a result, we get 781 in-depth videos whose resolution are 320×240. Then, we extract in-depth features and use manifold-learning dimension-reduction methods on this data set. After, that motion recognition and retrieval are carried out.

As is shown in the figure below, recognition rate we acquired(DFDR method) is 89.1%. Compared with traditional methods for feature extraction, recognition rate of motion data extracted from in-depth features is higher than that from traditional methods. Comparative results are shown in the figure below.

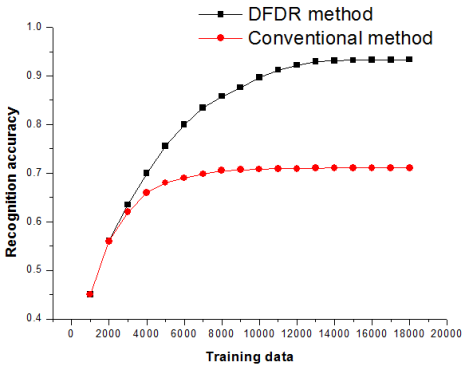

Fig.1 Comparison of the performance of conventional method with DFDR

Meanwhile, make a comparison on data sets of different motion data between this algorithm and others. As is shown in Table I, Table II, we found that the recognition rate of motion recognition of in-depth features aiming at complex motions is higher than other traditional means. This manifested that in-depth features can reflect the nature of human motions more precisely.

Comparisons on dimension-reduction algorithms, Table 3 is the comparison on retrieval time and precision of human motion retrieval system between classic PCA algorithms and dimension-reduction algorithm based on expanded ISOMAP. From it, we can find that some key information are lost in dimension reduction process when PCA dimension-reduction algorithm is used in some complex human motions, leading to the reduction of retrieval precision. As for ISOMAP, it conforms to the nonlinearity of human motions and it does not cause a large increase in retrieval time. Therefore, its adaptation is better than PCA dimension-reduction algorithm.

TABLE I RECOGNITION TIME

Motion clips	Recognition time(second)			
	Motion recognition by CIT		Motion recognition by DFDR	
	N=2000	N=10000	N=2000	N=10000
A(56)	2.7713s	14.3842s	1.7341s	2.9983s
B(113)	2.9113s	18.3823s	1.913s	3.3423s
C(172)	3.8193s	22.3094s	2.0923s	4.2444s
D(278)	5.1923s	36.4143s	2.5674s	5.3245s

TABLE II RECALL AND PRECISION

Motion clips	Recall		Precision	
	CIT	DFDR	CIT	DFDR
walk	0.75	0.968	0.91	0.97
run	0.691	0.959	0.84	0.985
jump	0.523	0.912	0.73	0.923
bunch	0.443	0.898	0.58	0.914

When it comes to a unknown motion, the first thing we need to do is to extract in-depth features of 3D human motion according to in-depth image. Then, through dimension-reduction algorithm, we can realized feature fusion and data reduction which will be brought to motion recognition and retrieval system before we get recognition and handle results.

References

[1]. [1] Pons-Moll G, Baak A, Helten T, Mueller M, Seidel H, Bodo Rosenhahn Multisensor-Fusion for 3D Full-Body Human Motion Capture, Proceedings of CVPR 2010
[2]. Shiratori T, Park H, Sigal L, Sheikh Y, Hodgins J, Motion Capture from Body-Mounted Cameras, Proceedings of ACM Siggraph 2011.
[3]. Sergey Matyunin, dmitriy Vatolin, Yury Berdnikov, Maxim Smirnov, Temporal filtering for depth maps generated by Kinect depth camera, 3DTV conference, 2011:1-4
[4]. Oreifej O, Liu z. Hon4d: Histogram of oriented 4D normal for activity recognition from depth sequences, IEEE Conference on Computer Vision and Pattern Recognition, 2013:716-723.
[5]. Li W, Zhang Z, Liu Z. Action recognition based on a bag of 3d points, CVPRW, 2010:9-14
[6]. Wang J, Liu z, Wu Y, et al. Mining actionlet ensemble for action recognition with depth cameras, CVPR, 2012:1290-1297
[7]. S. Tang, X. Wang, X. Lv, et al, Histogram of Oriented Normal Vectors for Object Recognition with a Depth Sensor. Computer Vision ACCV 2012:525-538
[8]. Y Zhao, Z Liu, L Yang, et al, Combining RGB and Depth map features for human activity recognition. Signal and Information Processing Association annual summit and conference. 2012.