International Journal of Multimedia and Ubiquitous Engineering
Vol.9, No.8 (2014), pp.211-218
http://dx.doi.org/10.14257/ijmue.2014.9.8.18

Motion Recognition based on Manifold Learning Spectral Clustering

Hongli Zhu and Jian Xiang

School of Information and Electronic Engineering, Zhejiang University City College, Hangzhou, China
School of Information and Electronic Engineering, Zhejiang University of Science and Technology, Hangzhou, China.
freexiang@gmail.com

Abstract

With the emergence of numerous 3D human motion capture databases, the effective analysis and handling of human motion data have become a major challenge so that the use of motion capture databases can be maximized. To reduce the high-dimensional complexity of data, a type of geometrical feature based on 2D geometrical space law is first extracted from human motion for the application of motion data into a low-dimensional subspace. With the aim of achieving a low-dimensional feature, identification and classification in different motions are then conducted through spectral clustering based on manifold learning to realize the automatic identification and retrieval of 3D human motion.

Keywords: *3D motion, spectral clustering, manifold learning, feature*

1. Introduction

With the development and gradual maturity of the motion capture and 3D scanning technique, as well as the advancement of high-definition video technology, large data volumes of 3D human motion data are produced every year through a variety of methods. At the same time, a vast number of 3D human motion databases have also been continuously developed. The reasonable and effective use of these databases has profound theoretical research significance and practical application value; the goals are to realize the effective retrieval of massive 3D human motion databases and utilize the data after analysis to achieve the re-creation of 3D human motion data, which are finally applied to fields, such as digital media [1-2].

Since the dimensionality of motion features extracted is very high, the distance between two motion data sequences is typically of similar order to the variation within each sequence due to the “curse of dimensionality”. High dimensional data also increases the computational complexity. Therefore, dimensionality reduction techniques are necessary to uncover intrinsic properties of the original data in a low-dimensionality subspace [3]. Tao [4] generalize Fisher’s linear discriminate analysis (FLDA) based on the geometric mean –based subspace selection to construct a low-dimensional feature subspace. High-order distance-based multiview stochastic learning (HD-MSL)[5] method for image classification is proposed, different features are integrated with the labeling information under a probabilistic framework to reduce dimensionality. A Semantic Preserving Distance Metric Learning (SP-DML) algorithm [6].

In recent years, a number of scholars have begun to focus on the identification of motion data mainly on the basis of similarity of two motions. The methods used can be categorized into two types from the perspective of similarity. Some methods perform identification according to the similarity of the motion numerical data. Meanwhile, other methods conduct identification according to the similarity of the motion data in logic.

ISSN: 1975-0080 IJMUE

Additionally, machine learning is widely utilized to classify motion data in the application of these methods. Li [7] used SVD to identify the motion feature and make the classified identification through SVM. Motion identification technology based on the descending dimension method is also proposed. For instance, some researchers continuously introduce the weight PCA [8], ISOMAP [9], and SDR [10] into the classified identification of motion data. Overall, the main purpose is to allow the original feature space to be projected to a specific subspace and hence make the classification on the low-dimension feature of the subspace through traditional methods, such as SVM and KNN. Although many domestic and foreign scholars have made significant efforts to classify and identify 3D human motion, they still confront many challenges. They fail to take advantage of the feature of 3D motion data and the relationship before and after the posture. Because of the high frame rate, motion data include numerous pieces of redundant information. Therefore, how can representative data be effectively obtained? The accuracy of identification has yet to be improved.

Multimedia data partitioning based on data clustering considers the whole picture as the entirety of data. According to the distance measurement method preliminarily defined, the similar relationship of the pixels is calculated to determine the different areas of the picture. In the last decade, people employed a variety of clustering algorithm methods, such as the K means clustering algorithm, in picture segmentation. Because the K means clustering algorithm is only suitable to the data of spherical space and numerous data structures are not of this shape in reality, the partitioning result of this algorithm is unsatisfactory. Recently, the spectral clustering algorithm [11] based on spectral graph theory has been a research hotspot. It demonstrates obvious advantages compared with traditional clustering algorithms. It can be implemented in the sampling space of different shapes and converge to the global optimum. This feature makes the algorithm widely applicable to various data. In addition, research on the spectral clustering algorithm has shown that to facilitate clustering, the Gaussian kernel should be used to calculate the similarity of any two points to form the similarity matrix. When the scale of the data set is large, the calculated amount of the eigenvector decomposition on the L matrix obtained from the similarity matrix and the memory capacity will be the bottlenecks. Meanwhile, the parameter ζ in the Gaussian kernel significantly affects the similarity of two points. However, the existing theory cannot effectively choose the most appropriate value automatically [12]. The ζ value can only be set according to people's experience. Different types of data sets will have different values, which result in unstable clustering results. From the preceding analysis, we can summarize the existing problems in the method into two main points: a) the spectral clustering is sensitive to the parameter, and b) the complexity of time and space is high.

To address the above problems, scholars have continuously conducted research and effectively integrated clustering learning methods and spectral clustering [13]; the goal is to improve the robustness of the clustering method and avoid the limitations of the parameter option. Significant research has also shown that the performance of the clustering ensemble is better than that of the single clustering algorithm in conducting clustering of data in any shape and scale [14].

This study considers the use of geometry feature extraction for 3D human motion. Then, it applies the spectral clustering method based on the manifold to conduct the identification and classification on 3D human motion. This process involves feature extraction to realize the purpose of rapidly and effectively identifying human motion.

2. The Module of Feature Selection and Extraction

Broadly speaking, the motion feature includes the time-order character and spatial feature of each joint. In our study, we mainly use the joint time and space feature to compare the similarity of motions in Figure 1. The joint feature can also be further categorized into the speed feature related to the time and angle feature relevant to the space. The former is utilized to describe the feature change of the motion. The latter is developed according to some priori knowledge (or assumption) of the described motion posture, which is closely related to the specific application.

A specific motion feature usually involves several different expression ways. Because of the difference in people's subjective consciousness, a so-called best expression for a specific feature does not exist. Actually, different expressions of motion features reflect some performances of the feature from different perspectives. For different data sets, the feature expressions used for the proper classification are also different. Additionally, choosing all extractable features to describe a motion is unreasonable because it will generate the so-called curse of dimensionality, which makes the original differentiable motion assembly become undistinguishable in practical treatment. Therefore, we should first select the proper feature class and the corresponding feature expression before extracting the features of all motions.

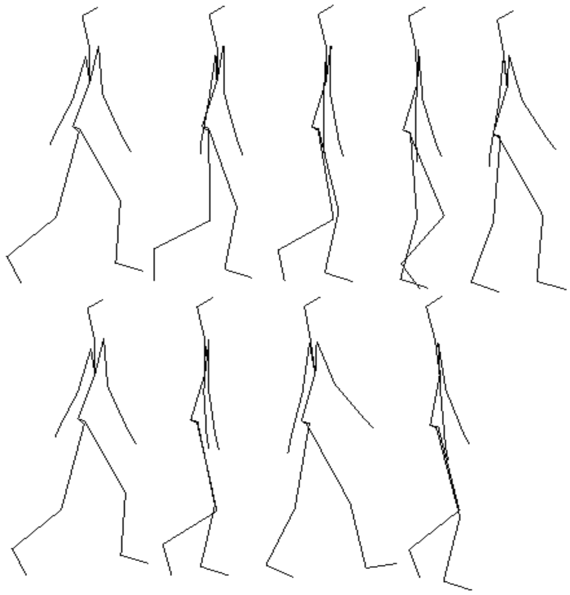


Figure 1. Two Motions (walk)

For the feature extraction of motion capture data, most methods start with the people's joint, including the position coordinates of the joint, speed, angular velocity, and relative angle, as well as the orientation information of the child node related to the joint. While each joint is moved, the angle relationship can be regarded as a type of "texture" feature of the motion to promote a sense of reality of the motion.

However, these traditional 2D features cannot address the semantic gap between the logical similarity of the motion and mathematical similarity. As the image of two walking motions shows, the speed of the first image is five times faster than that of the second image. Nevertheless, perceptually, we still think that two motions are walking; they are similar. This concept is called logical similarity. However, they are not similar in the similarity matching of the 2D features of the two motions. This concept is called mathematical similarity. Through a series of experiments and the priori knowledge of humans, we can find that the motion in logical similarity does not always have mathematical similarity. Using space coordinates to obtain 2D similarity brings about unsatisfactory results. For instance, some motions that should have been similar are excluded from the retrieved results. Besides, the traditional 2D feature must use DTW technology to do the alignment in the time shaft. The calculation complexity of such a technology is very high and is therefore inappropriate in handling a large-scale data set.

We should instead take advantage of the geometrical relationship of each joint in motion postures to achieve the motion feature. In this way, the 2D geometrical feature achieved can efficiently solve the issue of the logical similarity of motion data.

To extract and express the two-dimensional geometrical feature, we introduce the concept of Boolean feature and use the Boolean function to express - $F : \mathrm{P}$? $\{0,1\}$. Obviously, any Boolean expression of the Boolean function is itself. For the vector composed of f Boolean functions, we can get a mixed function that is $F : \mathrm{P}$? $\{0,1\}^f$. As what we present above, we treat F as a feature function. The vector F(p) is the feature vector. Or simply speaking, it is a feature of the posture P. Through the combination $F \circ D$, the feature function is applied in a motion capture data.

Taking an example, a two-dimensional geometrical feature is utilized to express a fixed posture that the right toe is in the front of the plane composed of the left ankle, left hip and the root node. We define p_i 誇□ ^{3},1 i ? 4 as four 3D joints, of which $\langle p_1, p_2, p_3 \rangle$ stands for the base plane determined by the former three joints. The orientation determines the order of three joints. Hence, we make the definition as follows:

$$B(p_1, p_2, p_3; p_4) := \begin{cases} 1, \text{ if } p_4 \text{ lies in front of or on } \langle p_1, p_2, p_3 \rangle, \\ 0, \text{ if } p_4 \text{ lies behind } \langle p_1, p_2, p_3 \rangle \end{cases}$$

Through the above definition, for any four neighboring joints, we present the following feature function.

$$F_{plane}^{(j_1, j_2, j_3; j_4)}(P) := B(P^{j_1}, P^{j_2}, P^{j_3}; P^{j_4})$$

Among it, $F_{plane}^{(j_1, j_2, j_3; j_4)} : \mathrm{P}$? $\{0,1\}$.

We set that j$_1$='root', j$_1$='lankle', j$_3$='lhip' and j$_4$='rtoes'. Such two-dimensional feature can be expressed $F^r := F_{plane}^{(j_1, j_2, j_3; j_4)}$. The plane determined by j$_1$ j$_2$ j$_3$ is as what the picture shows. It is obvious to see that the value of the feature $F^r(P)$ is 1 for a natural standing posture of a person. For walking or running, if the right foot is moved to the back or the left foot is moved forward, the value of the feature is 0. While the position of the definition point of the feature F^r is changed and the orientation of the plane is overturned, we get another featureF^l. Now, we study the combination of feature function $F := F^r$? F^l . While the value of F is 1 and F^r, F^l are only 1, the left toe and right toe are in the front of the corresponding planes. Thus, it can be seen that the function F is most suitable for expression for the footstep circulating motion such as walking and running. If what a motion data $D : [1 : T]$ 췝 describes is such kind of motion, $F \circ D$ stand for two peaks of cycle of motion. Through peaks, it can be easy to achieve the speed of the motion. On the other hand, the feature F is invariable for the global position and orientation, the joint size, different local space deformation (such as sideway) and the vertical movement of legs.

Through experiments and priori knowledge, we also define a series of other types' two dimensional geometrical features. The difference from the definition of the plane with three joints is that it can obtain the base plane with the normal vector given by two joints. Taking the plane with the orthogonal vector from the joint "chest" to the joint "neck" as an example, through this plane, it is convenient to judge whether the head is above or under the neck.

The other type's geometrical feature is to check two joints and two body parts or whether a joint and the body part are in a near area. In other words, whether are they close enough? Here, we need to consider the mutual touch of two hands or the situation that a hand touches the head and leg. Therefore, we are required to determine a large enough threshold value for the near area, to try to avoid the errors. Two-dimensional feature is also used to judge the specific part of the human body, such as arms, legs or the bending or the stretch of the body. Such geometrical feature is expressed by utilizing the proper body angle, including the angle between the thigh and shank, the angle between the upper arm and the forearm or the angle between the chine and the left and right thighs. The experiment result indicates that 120°is a good threshold value which can be used to tell whether different parts of the body is bending ($a < 120$) or stretch (a ³ 120). Finally, some non-geometrical features are applied to make up for the whole feature class, such as the absolute speed and the relative speed of some joints.

In the motion retrieval, users choose a certain geometrical feature set in accordance with their own needs. Also, according to a vast number of different types of motions, it is required to use different geometrical features to accomplish the retrieval. Thus, the selection of the geometrical features should meet these needs. Besides, the features chosen should have the mutual orthogonality. That is, there are no several features used to describe the same specific motion. Currently, two-dimensional feature can only be manually chosen through the interaction with users. It cannot still be automatically extracted.

3. Spectral Clustering

The traditional clustering algorithm such as K-means clustering algorithm and the expectation maximization algorithm is established based on the sampling space of the convex spherical shape. However, while the sampling space is not convex, the algorithm will be stuck into the local optimum. The spectral clustering is a kind of clustering methods on the basis of the graph theory. It uses the partition idea of the graph theory to improve the K-means clustering algorithm, to make it do the clustering in the sampling space of any shape. Meanwhile, it contributes to let the clustering converge to global optimum.

The spectral clustering algorithm is founded on the basis of the spectral graph theory. Compared with the traditional clustering algorithm, it is equipped with advantages that it can do clustering in the sampling space of any shape and converge to the global optimum.

This algorithm firstly defines an affinity matrix which describes the similarity of the data points in pairs according to the sampling dataset given and calculates the feature value and vector of the matrix. Secondly, it chooses the proper feature vector to cluster different data points. The spectral clustering algorithm is firstly used to the fields such as the computer vision and VLSI design. Recently, it begins to be applied in the machine study, which rapidly becomes the research hotspot in the international machine learning field.

It is built on the spectral graph theory. So, its nature is to transfer the clustering problem into the optimal partition problem of the graph. It is a kind of point pair clustering algorithms, which has a good application prospect for the data clustering.

3.1. Manifold Learning Spectral Clustering (MLSC)

ISOMAP is a kind of way of manifold learning. The manifold is the concept of topology. The manifold can be simply understood as n dimensional surface in space determined by. It can be expressed that each local is the Euclid topological space. The local Euclid feature means that any point in the space has a neighborhood. In fact, the Euclidean space is the simplest case of the manifold. The sphere like the surface of the earth is quite a complicated case. The general manifold can be formed through bending lots of flat slices and adhering them.

ISOMAP is a kind of typical manifold learning method of the spectral analysis. This kind of descending dimension method is based on the specific image space, equipped with the isometric mapping feature. The isometric mapping feature assumes that the geodesic distance also refers to the distance kept between two points in the map φ from the high dimensional space to the low dimensional space. In other words, for two points in the manifold as well as their projections in the map φ, there is. The distance in the manifold can be imagined as the shortest distance for an insect moving from m to. ISOMAP algorithm thinks that in the manifold, the Euclidean distance between two points has the significance only while it is quite small. Thus, the algorithm firstly constructs an adjacent map. Each point is only connected with its neighboring point. The geodesic distance between two neighboring points approximates with their Euclidean distance. For the geodesic distance of two points in different adjacent maps approximates with their shortest Dijkstra distance in the adjacent maps.

If we get N data points, the spectral clustering algorithm with the manifold structure has the following steps.

3.2. The Main Steps of ISOMAP Algorithm

1. The adjacent matrix G is built. There are two methods determining the side of the adjacent matrix that are εneighborhood and k neighborhood. While the point j is on the point I or I is located onεneighborhood and k neighborhood of j, the side with the weight should be set between i and j.

2. It uses Dijkstra or Flyod algorithm to calculate the shortest distance between point pairs so as to get the matrix. All elements are the shortest paths of each point pairs in G.

3. It makes the feature decomposition on L, to obtain the feature value before m and the corresponding feature vector.

4. M feature vectors are lined into a N x m matrix, regarding each line as a point in the m dimensional space. And then, K-means clustering algorithm is applied. The type of each line in the clustering result is the category that the corresponding node in the original space belongs to.

Compared with K-means clustering, the spectral clustering has the following advantages.

The spectral clustering only needs the similarity matrix among data while K-mean requires that the data must be the vector in the original Euclidean space.

The computation complexity is lower than that of K-mean, especially in the operation in data with high dimension such as the text data and image data.

It utilizes the partition idea of the graph theory to improve the K-means clustering algorithm, to make it able to do clustering in the sampling space of any shape and converge to the global optimum.

4. The Experimental Result and Analysis

The experimental data of this paper is obtained from CMU MOCAP database. The experimental dataset involves 300 different motions taken by different people including 9 sports commonly seen that are walking, running, jumping, kicking, boxing, tumbling, picking, swing and standing. Their numbers in CMU database can be seen in Table 1. The reason why we choose the data from multi numbers is to increase the complexity of the motion to check the effectivity of the method. Each motion takes 100 frames, without the repetition. Among it, there are 48 joints in each posture. The length of the skeleton has been uniform so as to weaken the influence brought about by different bodies.

Through the experimental research, it puts forward the performance of the clustering method while it deals with the dataset and conducts the comparison among the experimental results with various clustering methods including L2, SVM, PCA and K-mean.

The experiment adopts two evaluation standards: 1) run time; 2) clustering accuracy.

In the experiment, it makes 10 times of clustering on 6 sports commonly seen with our method clustering method (MLSC) and chooses the optimal values of the time and the accuracy. In Table 1 We can see that the run time and accuracy of the classification are closely relevant to the complexity of the sport types. Currently, the clustering effect and efficiency for the sport types commonly seen are quite good. However, while it refers to the relatively complicated or non-single sport, the classification effect and efficiency are quite poor.

Table 1. Performance of Classification

Motion Data	Sequencings classification accuracy	Motion recognition accuracy %
Walk	97.1	99.3
Run	96.3	99.0
kick	93.2	97.7
jump	92.4	97.2
dance	90.1	89.2
Bunch	89.2	87.8

Table 2. The Performance by Different Method

method	Sequencings classification accuracy %	Motion recognition accuracy %	Average Recognition time (200 motion sequcences)
MLSC	90.0	89.2	2.4
L2	76.3	67.1	2.3
SVM	81.8	80.2	1.2
PCA	75.9	80.7	1.7

Meanwhile, we compare the experimental results between MLCS method and various clustering methods such as L2, SVM, PCA and K-mean. The result shows that, aiming to different datasets, our methods reflect the good classification accuracy. Yet, in the classification time, it is a bit slower than that of PCA and K-means clustering. Thus it can be seen that the spectral clustering still remains to be improved in the efficiency while it enhances the accuracy.

Acknowledgements

This work is based upon work funded by Natural Science Foundation of P.R. China under Grant No 51246002 and Zhejiang Provincial Natural Science Foundation of China under Grant No. Y1080366, Y1100219.

References

[1] G. Pons-Moll, A. Baak, T. Helten, M. Mueller, H. Seidel, B. Rosenhahn, "Multisensor-Fusion for 3D Full-Body Human Motion Capture", Proceedings of CVPR, (**2010**).

[2] L. Zhongxiang, Z. Yueting, L. Feng and P. Yunhen, "Video Based Human Body Animation", Journal of Computer Research and Development, vol. 2, no. 40, (**2003**).

[3] J. Yu, L. Feng, S. Hock-Soon, L. Cuihua and L. Ziyu, "Image classification by multimodal subspace learning", Pattern Recognition Letters, vol. 9, no. 33, (**2012**).

[4] D. Tao, X. Li and X. Wu, "Geometric Mean for Subspace Selection", IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, vol. 2, no. 31, (**2009**).

[5] J. Yu, Y. Rui, Y. Y. Tang and D. Tao, "High-order Distance based Multiview Stochastic Learning in Image Classification", IEEE TRANSACTIONS ON CYBERNETICS, vol. 99, (**2014**).

[6] J. Yu, D. Tao, J. Li and J. Cheng, "Semantic preserving distance metric learning and applications", Information Science, (**2014**).

[7] B. Prabhakaran and C.-j. Li, "Indexing of motion capture data for efficient and fast similarity search", Journal of Computers, vol. 3, no. 1, (**2006**).

[8] E. Fiume and K. Forbves, "An efficient search algorithm for motion data using weighted pca [C]", Proceedings of ACM SIGGRAPH / Eurograohics Symposium on Computer Animation, (**2005).**

[9] Y.-t. Zhuang, F. Wu, J. Xiang and J.-g. Weng, "Ensemble learning hmm for motion recognition and retrieval by isomap dimension reduction", JOURNAL OF ZHEJIANG UNIVERSITY SCIENCE A., vol. 12, no. 7, (**2006**).

[10] A. Shyr, R. Urtasun and M. I. Jorda, "Sufficient dimension reduction for visual sequence classification", Proceedings of CVPR, (**2010).**

[11] C. Ding and H. Simon, "A min-max cut algorithm for graph partitioning and data clustering", In Proceedings of IEEE Int. Conf. Data Mining, (**2001**).

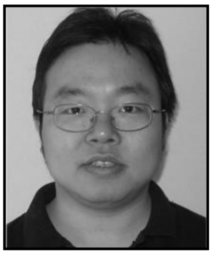
[12] U. von Luxburg, "A tutorial on spectral clustering", Statistics and Computing, vol. 14, no. 17, **(2007).**

[13] X. R. Zhang, "Spectral Clustering Ensemble Applied to SAR Image Segmentation", IEEE Geoscience and Remote Sensing Society, vol. 7, no. 46, **(2008).**

[14] A. Strehl and J. Ghosh, "Cluster Ensembles—A knowledge reuse framework for combining multiple".

Authors

Hongli Zhu, she received his Master's Degree in Communication Engineering from ZheJiang University in 2007. He is currently an associate professor at the School of Information and Electronic Engineering, ZheJiang University of Science and Technology, HangZhou, China. His current .research interests include multimedia analysis, static learning and computer animation.



Jian Xiang, he received his Ph.D. degree from the College of Computer Science and Technology, Zhejiang University. He has been associate professor at Zhejiang University of Science and Technology. He is a member of the China Computer Federation. His research interests include multimedia analysis and retrieval, computer animation and statistical learning, etc. E-mail:freexiang@gmail.com