ATLANTIS PRESS

Advances in Engineering, volume 100

International Conference on Manufacturing Engineering and Intelligent Materials (ICMEIM 2017)

Motion Segment based on Sparse Representation and 3D Features

Hongli Zhu [1], and Jian Xiang[2,*]

[1] School of Information and Electronic Engineering, Zhejiang University of Science and Technology, Hangzhou, China

[2] School of Information & Electronic Engineering, Zhejiang University City College, Hangzhou, China

* freexiang@gmail.com

Keywords: Motion Segment, Sparse Representation, 3D Features.

Abstract. With the emergence of a large number of 3D human motion capture database, which makes how to efficiently analyze and process the data of human body movement, and make use of the motion capture database become a new challenge. In order to reduce the high dimensional complexity of the data, firstly, a 3D dimensional feature based on 3D spatial and temporal characteristics is extracted from the motion of the human body, then, the motion data is re-expressed by the use of method for sparse representation, and different motion types are separated from long motion sequences, so that a motion database used for subsequent motion recognition and retrieval can be established.

Introduction

Currently, analysis of 3D human motion data still lacks complete and effective analysis and processing technology, which can not efficiently, rapidly, automatically and intelligently apply the large scale 3D human motion database to digital media field.

The 3D human motion data contains abundant semantic meanings of objects, events, behaviors, and scenes, and its characteristics of mass, non structure, high dimension and multi order bring a great challenge to the semantic understanding.

In recent years, compressed sensing and variable selection (when in the analysis of high dimensional data such as images, the variable selection is also called as feature selection in this application declaration) theories have been combined with methods, which are used for the formation of a more effective “sparse representation” on media data, becoming a hot issue in the fields of computer vision and machine learning, etc.. Compressed sensing makes the utilization of the future knowledge of “data being sparse and compressible” to achieve signal reconstruction, and in this aspect, some representative research works have been carried out by Donoho David and Emmanuel Candes of Stanford University, and Terence Tao (Tao Zhexuan) of University of California at Los Angeles, involving random matrix, signal recovery, sparsity measurement, etc. [1,2]

In view of the superiority of compressed sensing and variable selection in data processing, Wright and Ma Yi of the University of Illinois, Urbana-Champaign have introduced it into face recognition, and a new thought of carrying out feature selection by using the l1-- paradigm constraints model for recognizing human faces has been put forward[3]. Many features can be extracted from the media data, therefore, how to select effective sparse representation from high dimensional feature, and then study the theory and method of semantic understanding of media data on the basis of sparse expression has become a developing trend in the field of computer vision and pattern recognition, and it has been used in the visual word selection[4], image annotation[5], and image restoration[6] in succession. In the process of the identification of the real world, Urbana-Champaign have cooperated together to apply the sparse representation to visual object recognition, which has won the first prize in the PASCAL visual object recognition challenge (VOC2009) [7].

In order to realize this goal, aiming at 3D human motion data, we firstly make extraction of three-dimensional space-time characteristics, and then, the sparse representation of motion is given, and segmented.

Motion Data Features Extraction

The captured human motion data M is regarded as the human body posture sequence obtained by the discrete time point sampling, each sample point is a frame, and the posture of each frame is determined by 16 joint points. Therefore, at the arbitrary frame time i, the body posture is expressed as: $F_i = (p_i^{(1)}, r_i^{(1)}, r_i^{(2)}, ..., r_i^{(16)})$, where, $p_i^{(1)} \in P^3$ and $r_i^{(1)} \in R^3$ respectively represent the whereabouts and directions of Root joint points, that is, amount of translation and rotation; and $r_i^{(j)} \in R^3, j = 2...16$ represent the directions of non-Root joint points (rotation amount). According to the mutual relation of each joint point in the human skeleton, at any given time i, the location of non-Root joint point N_j in human skeleton can be gained through three-dimensional transformation Eq.1

$$\vec{p}_i^{(j)} = T_i^{(root)} R_i^{(root)} ... T_0^{(grandparent)} R_i^{(grandparent)}(t) T_0^{(parent)} R_i^{(parent)} \vec{p}_0^{(j)} \quad (1)$$

We calculate the world coordinates of each joint point by Eq.1, getting a fifty-one dimensional data, then, removing the root joint point, thus, the sixteen joints, a forty-eight dimensional data.

Motion is expressed as:

$$M_s = (F_1, F_2, ..., F_i, ..., F_n), F_i = (p_{i1}, p_{i2}, ..., p_{ij}, ..., p_{i16}), p_{ij} = (x, y, z) \quad (2)$$

Several space division rules are defined as followings:

Defining the spatial transformation of motion B=($b_1, b_2, ..., b_n)'$, $b_i = (s_{i1}, s_{i2}, ..., s_{i16})$, where b_i is the space transformation of joint i in the motion, s_{ij} means the space transformation of joint i at the frame of j . Supposing s_{aj} means the space transformation of a joint point a on the upper half of the body.

Sparse Representation of Motion

In the analysis and processing of motion data, we use one n-dimensional vector m∈Rn to represent a motion sequence, the vector here can be obtained from arranging the all joint point data in order in the graph, and can also be a certain feature vector of motion. Thus, supposing we have y sequences of different roles of a certain sequence, such as m1,m2,⋯,my ∈ Rn, for a new motion sequence m ∈ Rn, there should be:

$$\text{m}= \sum_{i=1}^{y} \beta_i m_i \quad (3)$$

Where, β_i means linear representation, its matrix form is:

$$m_{y\times 1} = T_{y\times m} x_{m\times 1} \quad (4)$$

Where, x is coefficient vector, next, on the basis of such a compact representation, we will give the sparse representation of motion, for motions to be recognized, we give a linear representation of the different roles of the same motion. This representation is compact. In the actual situation, the database is stored with multi role data of all motions. In this section, we will give a global

representation of the sequence of motion to be identified in the entire database, supposing that there are k motions in the database, the i motion has yi different characters of motion data. For the i motion, it will extract yi n-dimensional feature vectors, marked as m1,m2,⋯,my ∈ Rn. Where, the first subscript i represents the first i motion, the second subscript j represents the j motion. Again, we use a matrix to represent these vectors: Ti = [m1,m2,⋯,my] ∈ $R^{n\times y_i}$, then, for each motion sequence, it is corresponding to a matrix Ti, a total of k motion sequences, there are k such matrices T1; T2;⋯ ; Tk, these matrices are concatenated to obtain a large matrix that represents all the motions in the entire database:

$$T = [T1; T2; \cdot \cdot \cdot ; Tk] \in R^{n\times y} \tag{5}$$

Now, we consider a representation of the motion sequence to be recognized globally. Supposing that the motion to be recognized is from i motion sequence, its feature is f, then this motion sequence is put into the overall situation by following equation:

$$f = \sum_{i=1}^{y_j} \beta_i m_{j,i} = \beta_1 m_{j,1} + ... + \beta_{y_j} m_{j,y_j} \tag{6}$$

The contribution of a same motion to it is shown in the above formula, the contribution of different people to it is 0, so there is

$$F = 0 \cdot m_{1,1} + ... + 0 \cdot m_{j-1,y_i-1} + \beta_1 m_{j,1} + \beta_2 m_{j,2} + ... + \beta_{y_j} m_{j,y_j} + 0 \cdot m_{j+1,1} + ... + 0 \cdot m_{k,y_k} \tag{7}$$

We point out that when the number of motion types in the database is more, that is, the larger k, the linear representation of the Eq.8 is sparse. Because, there are only Yi non-zero elements in y-dimension vector x, and yi/y ≈ 1/k □ 1, namely ni □ n. That is, non-zero elements in the vector x only account for a very small part. At this point, we give a sparse representation of the motion to be recognized.

For the input motion, when it is represented with global motion linearity, only different sequences of the same motion have a greater contribution, and the contribution of the rest of motion is close to 0 (not exactly 0 for there is an error).

After the completion of the sparse identification of motion, in this paper, a simple and effective algorithm is designed to segment the long moving sequence. As follows:

Firstly, normally, the time consuming of the human body to finish a complete action is not less than 1 second. Therefore, for an unknown motion, the appropriate threshold can be set up according to its frame rate, namely the minimum continuous frame of each full motion.

Categories of all pose in the motion sequence can be calculated through sparse representation, and the continuous divided same kind of attitude will be merged into a collection according to the time sequence of motion.

Then, the number of gestures in each collection is calculated, if the number of the attitude collection is less than threshold, then it is called as illegal collection, after excluding these illegal collections, the remaining can be called as a candidate collection.

Finally, all attitude categories of the whole motion are summed up, the gesture of the label indicates the transition or unknown gesture. Meanwhile, we can also get the category of each segment.

Experimental Results and Analysis

In order to verify the motion segmentation method proposed in this paper, a total of 20 motion sequences are extracted directly from the motion database, containing 32013 frames, which are not

modified. Each motion sequence contains at least two different types of motion segments and they are naturally transitional. 10 types of common human motion can be found in this sequence, including walking, running, jumping, kicking, boxing, picking, dancing, squatting, leapfrog, etc..

The correct rate of segmentation on this motion sequence of our method is 96.7%, and the check rates are shown in the below table. Compared with the traditional method of motion segmentation, the results are shown in below table:

Table 1. comparison of performance with several methods

Methods	recall	precision
Conventional methods	68.2%	69.2%
Manifold	86.9%	87.1%
Our method	96.7%	94.9%

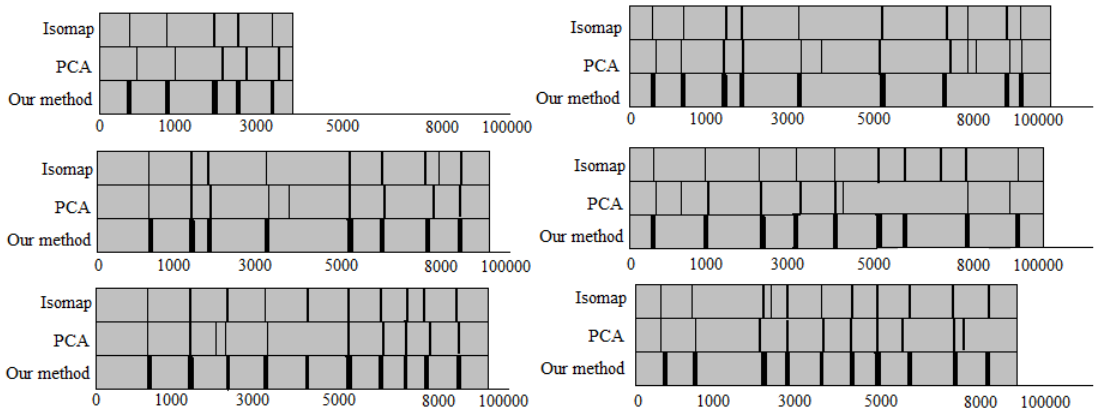

Figure 1. segmentation by some methods

Conclusion

The results show that the method in this paper can effectively segment the motion sequence.

But, we also shall notice that there is a shortage of this method, which is it can only make segmentation on the data with similar motions in the training data. For example, if there are some types of motion are unknown in a long sequence of motion, then, the types can't be identified or transitional motion types, which cannot be directly and correctly classified.

References

[1]Donoho, D., Compressed Sensing, IEEE Transactions on Information Theory, 52(4):1289-1306, 2006

[2]Cades, E., Tao, T., Reflections on compressed sensing, IEEE Information Theory Society Newsletter, 58(4), 20-23, 2008

[3]Wright, J., Yang, A., Ganesh, A., Sastry, S., Ma, Y., Robust face recognition via sparse representation, IEEE Transactions on Pattern Analysis and Machine intelligence, 31(2):210-227,2009

[4]g, J., Yu, K., Gong, Y., Huang, T., Linear spatial pyramid matching using sparse coding for image classification, Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2009

[5], D., Kakade. S., Langford, J., Zhang, T., Multi-label prediction via compressed sensing, Proceedings of Advances in Neural Information Processing Systems (NIPS), 2009

[6]ral, J, Elad, M., Sapiro, G., Sparse representation for color image restoration, IEEE Transactions on Image Processing, 17(1):53-69, 2008

[7]g, Y., Huang, T., Lv, F., Wang, J., Wu, C.,, Wu, W., Yang,J., Yu, K., Zhang, T.,, Zhou, X., Image classification using Gaussian mixture and local coordinate coding，The PASCAL Visual Object Classes Challenge Workshop, 2009