http://ansinet.com/itj

ITJ

ISSN 1812-5638

INFORMATION TECHNOLOGY JOURNAL

ANSInet
Asian Network for Scientific Information
308 Lasani Town, Sargodha Road, Faisalabad - Pakistan

Information Technology Journal 10 (10): 1957-1963, 2011
ISSN 1812-5638 / DOI: 10.3923/itj.2011.1957.1963

Underlying Semantic Annotation Method for Human Motion Capture Data

Lin Feng, Chang-You Xu, Bo Jin, Feng Chen and Zhi-Yuan Yin
School of Innovation Experiment, Dalian University of Technology,
Dalian 116024, Peoples' Republic of China

Abstract: On the issue of the representation model of the human motion 3D series, the most widely used methods were always based on numerical data. These methods could reduce the high dimensional 3D capture motion data and decrease the time complexity to a certain extent. However, the above mentioned traditional methods cannot extract the hidden useful domain physical knowledge, as well as meet the demands of current an intelligent computer processing on the numerical sequence. The present study proposed a new semantic annotation approach to obtain the linguistic tags on 3D motion data. Pre-processing on the human joint information should be implemented to appropriately achieve the spatio-temporal feature. The steps included: Constructing the human motion semantic category space, clustering the intermediate data through merging the kinematics knowledge and finally gaining the semantic annotations. At the end of the present study, the experiments showed that the proposed semantic approach could reasonably express the semantic information. In addition, there was also no absence of the essential domain knowledge of human motion data in the proposed method.

Key words: 3D human motion, semantic annotation, spatio-temporal features, motion semantic category space

INTRODUCTION

Human motion capture data records the variation information of the human joints in real world while humans perform daily actions and special sports. As a typical time series data, the human motion capture data and corresponding technology are extremely useful in fields (Hsu, 2011; Motlagh *et al.*, 2009), such as physical training, aided rehabilitative treatment, cartoon edition and virtual reality. Recently, many researchers have studied the above mentioned issue and achieved considerable knowledge. Under their study and promotion, human motion data analysis constitutes a hot spot of data mining research.

The concrete processes of human motion data analysis are as follows: Information annotation, motion recognition, motion retrieval and motion synthesis. Amongst them, motion information annotation is one of the most important representation methods based on follow up studying (Yang and Qin, 2010; Manivannan and Srivatsa, 2011). Constructing a potent motion annotation model is an indispensable and important precondition to detect the human motion domain-specific knowledge from the human motion capture data.

The current research on motion information annotation by Yang *et al.* (2008) employed spectral clustering for reprocessing and describing high-dimensional motion sequences in a simplified representation. Liang *et al.* (2010) coded the motion data to a character string, using text processing algorithms to pattern clustering and obtaining the corresponding string pattern annotation. Zhang *et al.* (2011) extracted the low-level numerical feature by Haar wavelet and Principal Component Analysis (PCA) and classified human activities by Hidden Markov Model (HMM). Muller and Roder (2005) proposed geometric features to quantize the relations between the specified body point of a pose and use the qualitative feature as a representative of the original human motion capture series data. In particular, the above mentioned methods classify and represent human motion in the temporal domain and provide a foundation for the movement sequence analysis, retrieval, identification and synthesis. However, these methods take the human movement series data as a character string or numeric text. Although, the within domain information of the movement sequence data is already shielded, it lacks the significance of a detailed description contained in the original human motion data. The reference filed semantic-based processing is still less developed. Muller and Roder (2006) and Liu *et al.* (2010) composed and encapsulated the lower level joints data of human motion by motion templates and labeled every motion template as a natural language parse. Jin and Prabhakaran (2011) extracted the spatio-temporal features through Singular

Corresponding Author: Bo Jin, School of Innovation Experiment, Dalian University of Technology, Dalian 116024, Peoples' Republic of China

Value Decomposition (SVD) and these features were formatted into an annotation sequential representation by their semantic Gaussian Mixture Modeling with EM (sGMMEM). The above mentioned methods do not analyze the human motion capture data at a profound semantic level. Thus, the movement data cannot achieve fine-grained knowledge and understanding.

In essence, the content-based human motion procession methods provide a lower dimensional representation. The above mentioned methods can solve the high-dimensional computational problems arising from the human motion sequences. However, these methods fail in supporting the follow-up stages of processing the human movement on a natural language level. Furthermore, in the field of time series data mining, Kovar and Gleicher (2004) have proved that a logic similar sequence is not always in the numerical similarity measures Zhu and Liang (2011), Shi *et al.* (2010) and Wei *et al.* (2009) used semantic tags and Ontology method to descript and organize raw data. To analyze the human motion capture data more accurately and conduct a deeper and intelligent analysis, it is necessary to propose a new representation model of human motion on a semantic level. Thus, the present study had proposed the semantic annotation approach of the human motion capture data, from the point of natural semantics and the reference semantic binary method of linguistics.

NOVEL SEMANTIC ANNOTATION METHOD

In this study, the analysis of human motion is done in vertical which signifies the physical knowledge in the numerical data. Finally, the semantic gap (Tapu *et al.*, 2009) problem between the logical similarity and the pure capture data is solved.

The outline of the main processes proposed in the present study is shown in Fig. 1. It is composed of semantic organization of the joints, spatio-temporal features extraction, motion category semantic mapping, motion cluster and semantic annotation.

Joints semantic organization: As the human body's experiences multiple degrees of freedom and uncertain appearances while moving, the human motion capture data has typical high-dimensional properties. Therefore, the essential stages to implement the proposed semantic annotation include preconditioning the original 3D human motion by locating the reference joints to the prescribed semantic body objects and acquiring the required data in form.

The present study uses the motion data from the Carnegie Mellon University's (CMU) human motion capture database and the human skeleton being acquainted, as shown in Fig. 2. The data from the CMU database provides a large number of joints contained in the human body. Thus, the CMU joints have been shortened to a suitable scale to fulfill the approach in the present study. Firstly, give an account of the joints used for human motion analysis. Next, define the JointsSet (JSet), Jset = {head, upperneck, lowerneck, thorax, lower back, root, lclavicle, lhumerus, lradius, rclavicle, rhumerus, lhipjoint, lfemur, ltibia, rhipjoint, rfermur, rtibia}, |Jset| = J and J, which is the amount of joints taken into account in the present study.

Motion feature extraction is a process to map the special movement characters information to the corresponding human joints. From the simplified joints set named JSet and the CMU skeleton (Fig. 2), the following steps were taken: Firstly, study the impact of human motion for inspecting the semantic motion type. Next, generate and sort out the bottom joints list which is a significant effect motion category (Table 1).

By reference to the physical training tutoring on the various body parts of human movement, the following table summarizes all the greater influence elements for the motion category. The underlying body objects and its contained joints are enumerated in Table 1.

To facilitate the above mentioned description, the subsequent section defines the following basic concepts:

Definition 1: Object (Ob) Underlying object of semantic movement as shown in Table 1. Considering a motion series file from the body parts, such as legs, arms and torso, based on the domain knowledge of the kinematics. Next we give the physical objects and their relationship between the corresponding human joints. Finally, we define Object as the meaning of the underlying body part, referred as Ob. For instance Object Ob_x and:

$$Ob_x = \{Joint_a, Joint_b, \ldots, Joint_i, \ldots\} \text{ and } Ob_x \subseteq Jset$$

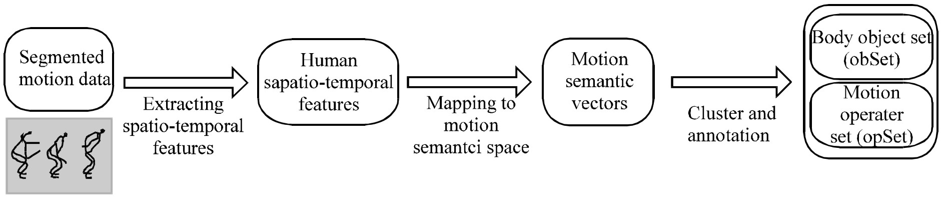
Segmented motion data
Extracting spatio-temporal features
Human sapatio-temporal features
Mapping to motion semantci space
Motion semantic vectors
Cluster and annotation
Body object set (obSet)
Motion operater set (opSet)

Fig. 1: Flowchart of semantic annotation for human motion capture data

Table 1: Body objects and corresponding joints

Body object	Bottom joints
head	head, upperneck
larm	lclavicle, lhumerus, lradius
rarm	rclavicle, rhumerns, rradius
torso	root, lowererback, thorax
lleg	lhipjoint, lfemur, ltibia
rleg	rhipjoint, rfemur, rtibia

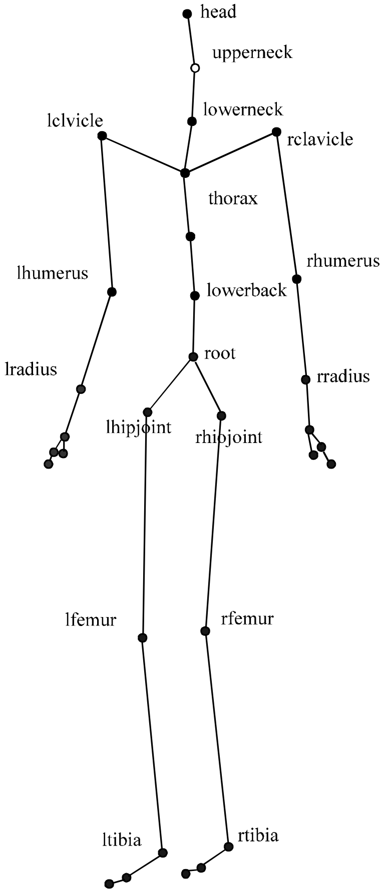

Fig. 2: Brief CMU human skeleton (L: Left, R: Right)

Thus, we give the definition of the Object set as ObjectSet (ObSet), by formulating Table 1, ObSet = {head, larm, rarm, torso, lleg, rleg}, |ObSet| = M and obviously M = 5.

Spatial-temporal features extraction: Considering an integrated human movement accomplished by the whole joints, every action of the joint mainly consists of shifting and rotation in the real 3D world. Consequently, we imply sub-space analysis to extract the spatio-temporal features from the capture data. At the same time, we try to maintain the original series characters as much as possible. For this purpose, He and Niyogi (2003) proposed Locality Preserving Projections (LPP). LPP can map the original high-dimensional data to lower-dimensional sub-space and extract complete essential information through preserving the neighborhood structure and the optimally chosen best weights. Based on the study of sub-space technology, we employ LPP to manipulate the 3D motion capture data to obtain the spatio-temporal features.

LPP aims to transform the space toward seeking for a low-dimensional Y which corresponds to X, its internal performance through the target function to find a certain linear transformation matrix W and extract feature information Y by the equation:

$$Y = W^T X$$

With the known existing l training samples, $X = [x_1, x_2, ..., x_i, x_l] \in R^{m\times l}$, a certain linear transformation matrix W can be gained by minimizing the following objective function:

$$W = \operatorname{argmin}\left(\sum_{i,j}\left(W^T x_i - W^T x_j\right)^2 S_{ij}\right) \tag{1}$$

The weight matrix S can be constructed as a definition:

$$S_{ij} = \begin{cases} e^{-\frac{\|x_i - x_j\|^2}{t}}, x_j \text{ is one of k neigborhood nodes of } x_i \\ 0, \text{other} \end{cases} \tag{2}$$

where, t is a constant greater than 0.

From the target Eq. 1, we can find that after reduction, the feature information can maximize to preserve the original construction character. With algebra on Eq. 1, the final result is:

$$\frac{1}{2}\sum_{i,j}\left(W^T x_i - W^T x_j\right)^2 S_{ij} = W^T X(D - S)X^T W \tag{3}$$

where, D is a l×l diagonal matrix and the elements of D are described as:

$$D_{ii} = \sum_j S_{ij}$$

We can make the transformation on Eq. 1 once more. The problem can be conversed to supplicate the minimizing Eigen values and Eigen vectors of the transformation matrix W. The calculation process is uncomplicated:

$$XLX^T w_j = \lambda XDX^T w_j \tag{4}$$

where, L = D-S, while $W = [w_1, w_2, ..., w_j, ..., w_d]$

Finally, we can make use of the transformation equation $y_j = W^T x_j$, j = 1, 2, ..., l to acquire the feature information matrix Y of the original matrix X.

In the present study, we can construct the initial motion capture matrix seq which refers to the active joints. Next, we can extract the lower dimensional information matrix seqMap by the LPP algorithm and receive the spatio-temporal features as the foundation for next vectoring and clustering.

Motion semantic category space: The semantic knowledge can be discovered from the human motion capture data, to bridge the semantic gap between the natural language logical similarity as human assessment and the numerical 3D series data. Thus, the present study proposes the concept of motion semantic space. The following section exposits the details of the concept. Kolli and Boufaida (2011) and Song *et al*. (2011) proposed method to descript conception which based on features of knowledge and its measurement.

Inspecting one of the body objects from ObSet, first distribute the object with a common undisputed motion description. Next, construct the motion semantic space T with physical significance which regards the spatio-temporal feature matrixes of each standard motion description as the basis of vector space. Specifically speaking, first stipulate several proverbial and obvious natural language distinctive perceptive physical motion series as the standard category movement sequences:

$$STD = (std_1, std_2, ..., std_i, ..., std_{SK})$$

Then, import the idea of space and use STD which involves the human semantic perception as a basis to construct a motion semantic space. In the above mentioned semantic space which has cognition of the meaning of human action, quantify the unknown category motion capture data to a vector. Subsequently, calculate the vector to observe the special physical action character in a unified operation. Assume that an unknown motion capture data seq uses the above mentioned definition. We can get a vector $(\|seq\text{-}std_1\|, \|seq\text{-}std_2\|,...,\|seq\text{-}std_{SK}\|)$ in the semantic space T. As each basis of contains different domain special knowledge of the physical category, every note in corresponds to a semantic sense. Thus, we can bridge the semantic and numerical data. Theoretically, the basis of vector space in the present study may not be orthogonal. However, the oblique space also has the ability to describe the homogeneous objects. As an example, take the object torso and choose the stoop and straight belt as the two standard motions to construct torso motion semantic space.

The semantic space has a preferable ability to decrepit the physical category of the torso, as shown in Fig. 3. The choused spatio-temporal feature of the standard capture data stdl has the physical mean of stoop while std2 is another feature matrix that contains the straight belt preservation. In the semantic space construct by these two bases, the unknown category motion seq can be represented as a distance vector $(\|seq\text{-}std_1\|, \|seq\text{-}std_2\|)$. The torso needs to make bending movements during high jumping, such that there are similar characters on the bending down action of the waist. Finally, we can find that the motion data of a high jump out performs at the axis of the stoop. A similar motion, such as wandering which can finish without bending the torso, can be located on the top right corner of the space. Thus, based on using the reasonable quantitative method, the motion category semantic space mentioned in the present study could provide efficient input data for clustering.

Motion semantic annotation: First, quantize the spatio-temporal feature matrix, which is mentioned, to a vector in the motion semantic space and cluster these vectors into several rough categories. Then, artificially analyze and dress the clustering results to a perceptual consistent physical behavior classification. As a result of this approach, the present study obtains all the underlying objects and the corresponding semantic operations annotations.

Initially, utilize the k-NN model to cluster the feature matrix of every body object and gain the relevant rough physical action classifications. Next, by means of the AMCViewer (CMU Graphics Lab, 2007), examine all the rough classifications where by the delicate differences between the rough clustering results emerge. Finally, annotate the domain special descriptions. All the body objects have a set of corresponding operation semantic

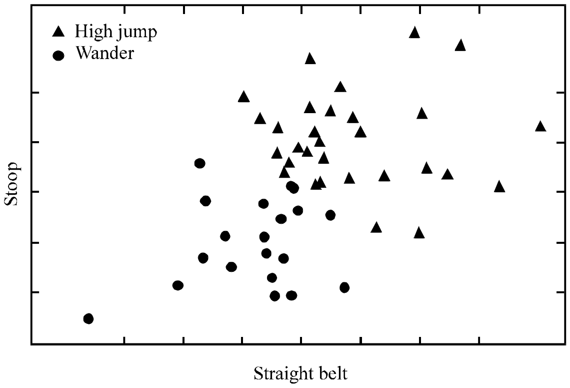

Fig. 3: Torso motion semantic space

domain knowledge annotations. Clustering the spatio-temporal feature matrix of the right arm and manually reorganizing. Then the semantic annotation actions of the right arm can be sorted over as the following: Arm wave, arm swing, elbow, arm raise and put up hands.

The natural language description through the semantic annotation can be formalized as follows:

Definition 2: Operate (Op) Operate is the natural language of the associative movements corresponding to a named body object. Certain objects which perform actions, such as a lift swing may get the corresponding semantic annotation as Op_x = swing

Thus, the present study provides the definition of the Operate set as OperateSet (OpSet). The OperateSet contains all the physical semantic annotations of every object in ObSet, |OpSet| = N where N is the account of the operation annotations set.

All the objects of the human body and the corresponding motion semantic annotations can characterize the whole human action in a good condition. At the same time, the collection of these two sets, such as ObSet and OpSet can fulfill to represent the human motion in the natural language. To facilitate the representation and calculation, the definition of the human motion elements named Sememe Library (SL) is defined as follows: The objects and semantic annotations are equivalent to the term elements in the engineering web and biological hierarchical structure. The present study proposes the human motion semantic approach to achieve the underlying elements of the set to the bottom-up parsed approach of the high level kinesiology knowledge and SL = ObSet∪OpSet. SL is a natural language complete domain semantic elementary database, which transforms the computable numerical similarity measures to semantic logical similarity.

THE EXPERIMENTS

In order to validate the effectiveness of the proposed semantic annotation method, we implement the present method on the publicly available human motion capture data from CMU. After the experiment, we find that the different categories of the human movement own limited-scale operations. Reorganizing the rough clustering results can achieve the essential body objects and relevant semantic annotation. Here, we detail the particular method used in the experiment. Then, the obtained results are analyzed.

The 20.16 MB .asf and .amc files concerning the human joints downloaded from CMU contain each frame of data which includes 31 joints of the body and 55 degrees of freedom. After transforming the .amc joints information to the 3D space coordinate sequence by inverse kinematics, we finally obtain 32.9 MB data and 135 motions with 27000 frames.

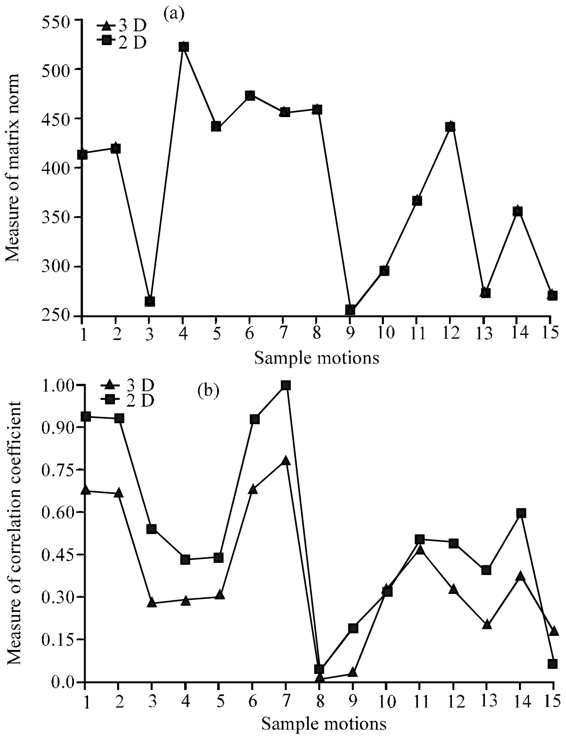

Fig. 4 (a-b): Different measures for the spatio-temporal feature matrix, (a) Measure of matrix norm and (b) Measure of correlation coefficient

The 3D data of human motion, as a high-dimensional series data, contains each of the joints of the moving skeleton. In order to reasonably and effectively deal with the human motion, the experiment extracts the spatio-temporal feature matrix by the above mentioned LPP algorithm and the structure vector of the feature matrix in the motion semantic space. Essentially, the work of structuring is a measurement process of the matrix. We look for a preferable measurement mechanism between the matrixes.

Using the matrix norm can get a better measure for the same classification movement, which is chosen from the CMU open database. In the present experiment motion series data 5, 6, 7 and 8 are in the same physical category, while 12, 13 are entirely different actions. According to the principle, similar motions have a fine contrast. At the same time, the measure between the different motions is comparatively large. In Fig. 4a, whether the dimension is reduced to 3 or 2, the measure of the matrix norm on the feature matrixes of motion 5, 6, 7 and 8 are roughly of the same value. However, the measure of 12, 13 are on an entirely different level. The measure of correlation coefficient showed in Fig. 4b does not have the ability to distinguish the difference between the different motions 3, 4, 5. Thus, it cannot issue the correct decision on the same categories 5, 6, 7, 8. Consequently, in the present study we determine the matrix norm to measure the feature matrixes.

In the concrete implementation of the semantic annotation experiment there is a requirement to determine the target dimension, account of clustering and clustering centers. To calculate the average clustering accuracy base on the matrix norm with the different parameters on the sample motions and the effect of the experiment, first select the parameters and corresponding cluster centers as the base conditions for the follow-up operations. Next, after estimating the intrinsic dimension on the 3D motion capture data through a manifold learning algorithm, the target reduce dimension d is resolved as d∈{1, 2, 3}. Finally, take object lleg which represents the left leg. The average accuracy of clustering is shown in Fig. 5.

We take the body object lleg as the experimental subject. Next, we carefully select wander, soldier match, high jump and forward jump as the standard distinctive descriptions of the left leg as a basis of vector space, whereby the bending range and pattern of leg is quite different. Comparing to the manufactory reference category list, we can draw out that the reduced original 3D motion data to 3 achieves higher average accuracy. At the same time, a reduction to 2 may get stable accuracy. As a result of randomly chosen initial cluster centers in the k-NN algorithm, the experimental accuracy is not higher than the traditional approach. However, as an important advantage of semantic annotation, the cluster result we obtain through LPP and the semantic space contain the domain special knowledge of human action. These results do a great favor to collate the natural language description of human action instead of collecting the semantic information directly from the abundant raw 3D motion data. Making decisions according to the experimental result in Fig. 5, we designate the number of cluster k = 5 and target reduce dimension d = 3 and tread the corresponding cluster centers as the default input for annotating of object lleg. This is done through artificially ordering and annotating with the integrated physical domain knowledge. We calculate the final semantic results as Fig. 6.

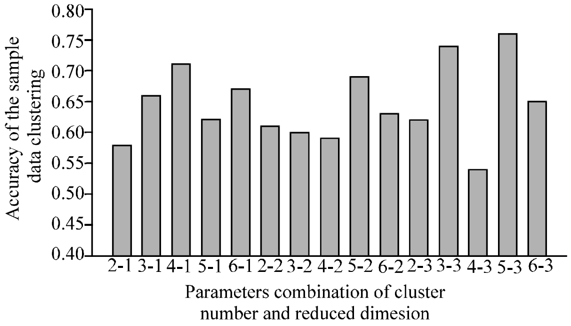

Fig. 5: Accuracy of clustering under different parameters

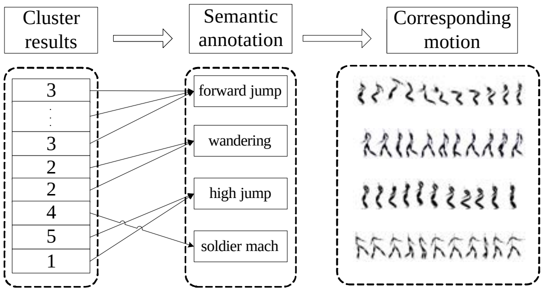

Fig. 6: Semantic annotation of object lleg

Under the parameters k = 5 and d = 3 clustering on object lleg we achieve the final result, as shown in Fig. 6. The k-NN algorithm divides the motion in one category with slight differences into two different motions, as shown in Fig. 6. At the same time, the present experiment also reflects the effectiveness of the spatio-temporal feature. Therefore, both the high jump and forward jump are bent leg aggressive. However, the high jump acts in the vertical direction while the forward jump acts both in the vertical movement and horizontal displacement. Nevertheless we can distinguish the two similar movements as the extracting features. Consequently, the proposed approach achieves two categories of motions and discriminates them on a semantic level. The rough cluster results assist to analyze and make known the categories. We achieve the final natural semantic descriptions of object Ob = lleg list as Op = {standing jump, large span bending, march swing, normal wander}. The different body objects applying the above mentioned same method can complement the set of elemental SememeLibrary which contains the operator semantic annotations and corresponds to all the objects in ObSet.

CONCLUSIONS

The present study has demonstrated that it is possible to combine the semantics analysis and domain-special knowledge into the semantic annotation of the human motion capture data. Through the obtained experimental results, it is obviously that our method is a reasonable and effective scheme to organize and achieve the elemental semantic description of the 3D motion data. This is performed by extracting the spatio-temporal feature matrix and the corresponding physical elemental human body unit objects through applying the LPP algorithm and kinemics knowledge. However, there is one important difference between the traditional methods and the proposed method. The latter method provides the necessary basis for an intelligent computer processing of the motion data by organizing the information in the semantics way. The barriers between the underlying numerical data and the semantic approach are shielded as much as possible. Thus, semantic annotation provides a new way for users to search the motion data and synthesize the movement on-demand. This is as convenient as using the natural language in a friendly interaction. Similarly, as the semantic approach in fields, such as engineer semantic web and gene semantic search, establishing a perfect set of elemental Sememe Library needs an investment of a larger long-term supplement for the additions and amendments. Moreover, the semantic method of human motion application is not yet mature. Thus, future studies can focus on how to construct a well balanced semantic web and a more high-layered description on the whole body movement.

ACKNOWLEDGMENTS

This study is supported by National Natural Science Foundation of P.R. China (Grant No. 61173163, No. 51105052) and Program for New Century Excellent Talents in University (Grant No. NCET-09-0251).

REFERENCES

CMU Graphics Lab, 2007. CMU Graphics Lab Motion Capture Database. http://mocap.cs.cmu.edu/

He, X.F. and P. Niyogi, 2003. Locality preserving projections. Proceeding of the Advances in Neural Information Processing Systems Conference 16, (NIPS'03), Vancouver, Canada, pp: 153-160.

Hsu, K.S., 2011. Application of a virtual reality entertainment system with human-machine sensor device. J. Applied Sci., 11: 2145-2153.

Jin, Y. and B. Prabhakaran, 2011. Knowledge discovery from 3D human motion streams through semantic dimensional reduction. ACM Trans. Multimedia Comput. Commun. Appl., (In press).

Kolli, M. and Z. Boufaida, 2011. Composing semantic relations among ontologies with a description logics. Inform. Technol. J., 10: 1106-1112.

Kovar, L. and M. Gleicher, 2004. Automated extraction and parameterization of motions in large data sets. ACM Trans. Graphics, 23: 559-568.

Liang, P., Y.Z. Yang and Y. Cai, 2010. Pattern mining from saccadic motion data. Proceedings of the 10th International Conference on Computational Science, May 31, Amsterdam, Netherlands, pp: 2529-2538.

Liu, W., X. Liu, W. Xing and B. Yuan, 2010. Study on semantic control in motion synthesis. Proceedings of 10th International Conference on Signal Processing (ICSP), Oct. 24-28, Beijing, China, pp: 1182-1185.

Manivannan, R. and S.K. Srivatsa, 2011. Semi automatic method for string matching. Inform. Technol. J., 10: 195-200.

Motlagh, O., S.H. Tang, N. Ismail and A.R. Ramli, 2009. A review on positioning techniques and technologies: A novel AI approach. J. Applied Sci., 9: 1601-1614.

Muller, M. and T. Roder, 2005. Efficient content-based retrieval of motion capture data. ACM Trans. Graph., 24: 677-685.

Muller, M. and T. Roder, 2006. Motion templates for automatic classification and retrieval of motion capture data. Proceedings of the 2006 ACM SIGGRAPH/Eurographics Symposium on Computer Animation, Aire-la-Ville, Switzerland.

Shi, H., G. Zhou and P. Qian, 2010. An attribute-based sentiment analysis system. Inform. Technol. J., 9: 1607-1614.

Song, S., Z. Guo and P. Chen, 2011. Fuzzy document clustering using weighted conceptual model. Inform. Technol. J., 10: 1178-1185.

Tapu, R., E. Tapu, B. Mocanu and E. Dragulanescu, 2009. A survey of the low-level descriptors used for content based multimedia retrieval. Metalurgia Int., 14: 12-15.

Wei, S., M. Qin-Yi and G. Tian-Yi, 2009. An ontology-based manufacturing design system. Inform. Technol. J., 8: 643-656.

Yang, B. and Z. Qin, 2010. Composing semantic web services with PDDL. Inform. Technol. J., 9: 48-54.

Yang, Y.D., Wang L.L. and A.M. Hao, 2008. Motion string: A motion capture data representation for behavior segmentation. J. Comput. Res. Dev., 45: 527-534.

Zhang, H., Z. Liu and H. Zhao, 2011. Human activities for classification via feature points. Inform. Technol. J., 10: 974-982.

Zhu, S. and Z. Liang, 2011. Semantic scene segmentation for advanced story retrieval. Inform. Technol. J., 10: 98-105.