Semantic Quantization of 3D Human Motion Capture Data Through Spatial-Temporal Feature Extraction

Yohan Jin and B. Prabhakaran

Department of Computer Science
University of Texas at Dallas
Richardson, Texas 75083-0688, USA
{yohan,praba}@utdallas.edu

Abstract. 3D motion capture is a form of multimedia data that is widely used in animation and medical fields (such as physical medicine and rehabilitation where body joint analysis is needed). These applications typically create large repositories of motion capture data and need efficient and accurate content-based retrieval techniques. 3D motion capture data is in the form of multi-dimensional time series data. To reduce the dimensions of human motion data while maintaining semantically important features, we quantize human motion data by extracting Spatial-Temporal Features through SVD and translate them onto a 1-dimensional sequential representation through our proposed sGMMEM (semantic Gaussian Mixture Modeling with EM). Thus, we achieve good classification accuracies for primitive human motion categories (walking 92.85%,run 91.42%,jump 94.11%) and even for subtle categories (dance 89.47%,laugh 83.33%,basketball signal 85.71%,golf putting 80.00%).

1 Introduction

In terms of human perception, human can understand the "semantic" meaning of an image with small number of colors after color-quantization (that is, similar color shades are expressed with one representative color). Thus, the level of human understanding does not change after decreasing enormous amount of information for storage, which is quite useful for compression and recognition of multimedia data [7]. This is also true for human motion capture data. 3D human motion capture is widely used in various applications such as animation authoring and medical fields (such as physical medicine and rehabilitation). Here, sophisticated motion capture facilities aid in mapping the complex human motion in the three dimensional (3D) space. Each row of data matrix corresponds to a single frame that consists of information for 29 segments (corresponding to different parts of human body) depending on degree of freedom. The degree of freedom for each segment is the ability of the segment to rotate or translate along three axes, according to the hierarchical structure of the human segment. Altogether, these segments form a 62-dimensional data along time axis. Human motion is characterized by spatial relationships between each of these body

S. Satoh, F. Nack, and M. Etoh (Eds.): MMM 2008, LNCS 4903, pp. 318–328, 2007.

segments/joints over a period of time. Hence, we need to extract these spatial and temporal relationships from the multi-dimensional 3D motion capture data matrix and map them onto a lower-dimension (1-dimensional, in our approach) representation by the following steps;

- ***Spatial Feature Extraction:*** Human motions are primarily identified by movements of arms, legs, and the associated body or "torso". Hence, the human motion matrix is divided into three main body parts (or sub-matrices): 'torso','arms' and 'legs'. For extracting spatial relationships among body parts, we use Singular Value Decomposition (SVD)[12].
- ***Semantic Quantization Through Temporal Feature Extraction:*** We semantically quantize spatial 3-dimensional singular values by identify temporal distributions among body parts in a human by measuring the responsibility of each body component [torso, arms, and legs] and the associated combinations (8 combinations among the 3 components) to the observed frame values. For this, we map 3 dimensional singular values of frames into Gaussian Mixture Semantic Space. Then, we apply Expectation Maximization (EM)[11] to find the most probable semantic quantization value of each frame. We consider continuous singular spatial values as the "observation" values and discrete ("quantized") body components as the "latent (hidden)" values. Hence, multi-dimension 3D motion data frames can be reduced to 1-dimensional quantization value that can be used for classification and retrieval.

In this paper, we show that time series, multi-dimensional 3D human motion capture data can be semantically quantized through statistical modeling (Gaussian Mixture) based on temporal distribution of spatial features of human motion data. We proposed a novel approach which can extract human motion's semantic characteristics successfully during reducing dimensions so much and demonstrated its usefulness by matching the most closest motion with real subtle 3D human motion queries.

2 Related Work

To index 3D human motion data, there are approaches which use principle components of human motions. Li et al.[2] extracted geometric structure as exposed by SVD of matrices of human motion data and indexed using interval-tree based index structure and classified human motions with SVM on geometric extracted motion vectors [1]. Guodong et el.[6] selected small set of leading eigenvectors as principle features and tried to represent motion frames as simplified "cluster transition signature", which is conceptually similar to 1-dimensional quantization representation in this paper. Other approaches utilized hierarchical trees for indexing 3D human motions. Gaurav et al.[4] used hierarchical structure of the human body segments. Each level of index tree is associated with the weighted feature vectors of a body segment. Feng et al.[8] proposed content-based motion retrieval (CBMR) by building motion-index tree on hierarchial

motion description, which serves as a classifier to determine a sub-library that contains promising similar motions to the query example. For dealing with temporal invariance between similar motions, used "elastic match"[8], a combination of DTW (Dynamic Time Warping) and dynamic programming. To overcome the limit of DTW technique ('local scaling') for time-series data comparison, Keogh et al.[9] proposed a uniform scaling , which can scaling globally and showed that it can speed up indexing using bounding envelopes. Most recently, Muller et al.[5] contributed content-based human motion retrieval through "qualitative" geometric description for bridging the numerical and perceptual human motion similarity gap. However, a user has to select suitable features in order to obtain high-quality retrieval results. In this paper, we use SVD for extracting spatial characteristics of human motion frames while reducing dimensions similar to [2], and our approach don't need to specify the retrieval condition description. Furthermore, we also try to find temporal continuity with similar frames and represent each frames as semantic quantization values by applying statistical distribution modeling (Semantic Gaussian Mixture Space Modeling). Thus, this is the novel approach which exploit '*statistical distribution*' in 3D human motion database and '*quantize*' it for content-based 3D Human Motion Retrieval.

3 Spatial Feature Extraction

We consider one human motion is characterized by different combination of three main body parts: torso, arms and legs. Each body parts include several segments. For example, we divide 29 segments of human body into three sets, namely "torso," "arms," and "legs." The torso consists of 7 segments (with degree of freedom in parenthesis),namely root (6), lower back(3), upper back(3), thorax(3), lower neck(3), upper neck(3), and head(3) segments. The arms consists of 7 pairs of segments including left and right side, namely clavicle(2), humerus(3), radius(1), wrist(1), hand(2), fingers(1), and thumb(2). And, finally, legs consists of 4 pairs of segments including left and right side, namely femur(3), tibia(1), foot(2), and toes(1). For extracting spatial relationships and reducing dimensions (from 62 to 3 dimensions) among the 3 different body components, we separate a motion data matrix ($M_{f\times m}$) into three sub matrices ($M^{\alpha} = M_{f\times k}, M^{\beta} = M_{f\times j}, M^{\gamma} = M_{f\times r}$, where $m = k + j + r$) belonging to torso,arms and legs part respectively. From three sub matrices, SVD decomposes "singular" values (see Figure 1).

$$M^i = U\Sigma V^T, M^i v_1 = \sigma^i v_1, i \in \{\alpha, \beta, \gamma\} \tag{1}$$

Now, three "singular" values ($\sigma^{\alpha}, \sigma^{\beta}, \sigma^{\gamma}$) which represent torso,arms and legs parts are the coefficient of each frame as the spatial feature, then we have a reduced matrix $M_{f\times 3}$ for a single human motion clip. Singular values represent the periodic and continuous characteristics of human's motion. Increasing singular value of one body part indicates that part is used more intensively than other part for a particular motion (see Figure 2).

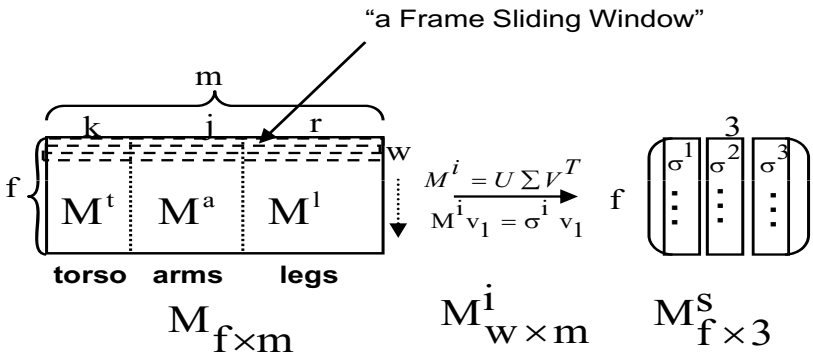

Fig. 1. SVD Spatial Feature Extraction using Frame Sliding Window

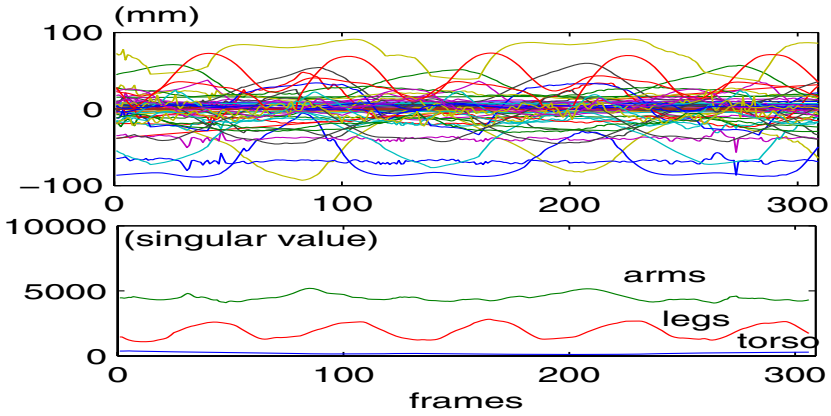

Fig. 2. SVD Spatial Feature Dimension Reduction ('walking' motion)

4 Semantic Gaussian-Mixture Quantization Through Temporal Feature Extraction

4.1 Gaussian Mixture Semantic Space

We map spatially extracted 3-dimensional singular values of each motion file into Gaussian Mixture Semantic Space [10] (see Figure 3), this space is to find "latent" semantic quantization components(A(rms), T(orso), and L(egs)) which is corresponding to the given observation (i^{th} frame of one motion file). It computes $\Re_{ki}$ of 3 body component, which is probability of "latent" component k's responsibility for observing i^{th} frame (O_i).

$$\Re_{ki} = P(k|O_i) = \frac{P(O_i|k)P(k)}{P(O_i)} = \frac{g(O_i;\mu_k)\Phi_k}{\sum_{k=1}^{K} P(O_i,k)} \tag{2}$$

Let $P(O_i|k)$ be the Gaussian function $g(O_i; \mu_k)$ of latent component k, and P(k) be the mixing parameter Φ_k of latent component k. $P(O_i)$ is the "prior" probability that we can get from marginalization of joint probability.

Human can express one action using more than one body component at the same time, so we need to extend Gaussian Mixture Semantic Space from three main body parts (Triangle) to a combination of the three main parts (Cube). We add three combined "latent" components corresponding to each edge of Triangle, which is 'TL', 'AT' and 'AL' respectively. Thus, overall number of "latent" component in Cube space is 8 including null (Φ) and all (TAL) components. Each quantizing component has its semantic meaning: for example, if one frame window has a mixture value close to 'L', it means that this frame window belongs to "legs intensive" actions. And if one frame window's mixture value is close to 'TAL', then it means that this frame has action using "legs" ,"arms" and "torso" actively.

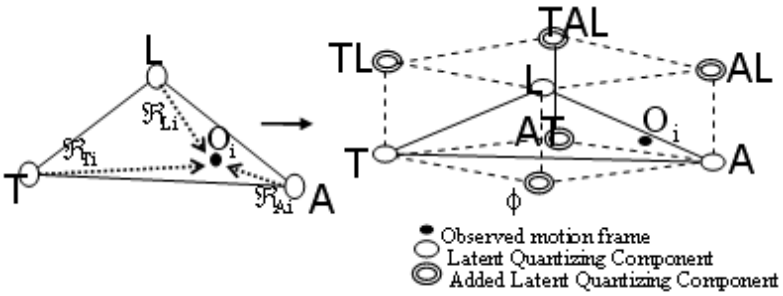

Fig. 3. Gaussian Mixture Semantic Space for finding "latent" component's responsibilities

4.2 Extension of GMM with EM for Semantic Quantization of Human Motion Data

We can extend GMM with EM (Expectation Maximization) for finding local maximal values based on the initial GMM values of human motions. After iteratively running with GMM and EM model, we get the locally maximized mixture value. Then, we assign each maximized value to the closest quantization value in order to get the quantized representation of each frame window from spatially extracted feature vectors through following steps.

1. Initialization Phase: After running the k-means cluster algorithm with all singular value coefficients matrices and 8 central points, we get the initial values (mixing Φ_k and initial memberships π_i) of GMM

$$\Phi_k = \frac{\sum_i (\pi_i; i = k)}{\sum_i \pi_i} \tag{3}$$

Every membership variable has the value of 0 or 1. We need to do this iteratively (from phase 2 to phase 5) until the values converge.

2. Evaluation Phase: Let $\lambda(i,k)$ is the k quantizing component's Gaussian function value of the current motion observation (i^{th} frame window) with the recently updated Gaussian parameters from "Update Phase". At the first iteration, evaluate with initial parameters.

$$\lambda(i,k) = g(O_i;\mu_k) = \frac{1}{(2\pi\varepsilon)^{1/2}} exp(-\frac{\parallel O_i - \mu_k \parallel^2}{2\varepsilon}) \tag{4}$$

$$\Re_{ki} = \frac{\Phi_k \lambda(i,k)}{\sum_k \Phi_k \lambda(i,k)}, k \in \{\Phi, T, A, L, ..., TAL\} \tag{5}$$

3. Update Phase: Follows to the re-evaluated "responsible" parameters ($\Re_{ki}$) from previous phase, it updates Gaussian parameters ($\widetilde{\mu_k}$) and "mixing" parameters ($\widetilde{\Phi_k}$),

$$\widetilde{\mu_k} = \frac{1}{S_k} \sum_{i=1}^{I} \Re_{ki} O_i \tag{6}$$

$$\widetilde{\Phi_k} = \frac{S_k}{I}, S_k = \sum_{i=1}^{I} \Re_{ki} \tag{7}$$

4. Convergence Phase: In each iteration, we compute the maximum likelihood (ML) function with the updated Gaussian parameter $P(O|\mu)$ and mixing parameter $P(\Phi)$ from previous phase;

$$ML(O;\Phi,\mu) = \ln P(O|\Phi,\mu) = \ln P(O|\mu)P(\Phi) \tag{8}$$

Next, we compute the convergence of this maximum-likelihood (*ConvML*) function as;

$$ConvML = \ln \prod_{i=1}^{I} \sum_{k=1}^{K} (\widetilde{\Phi_k} - \Phi_k)(\lambda(i,k) - \lambda(i,\widetilde{k})) \tag{9}$$

where I is the number of observations and K is the number of quantizing components (K=8). $\widetilde{\Phi_k}$ and $\lambda(i,\widetilde{k})$ are updated mixing parameter and Gaussian function values with re-evaluated parameters. If the difference between current and previous iteration's is smaller than some convergence threshold ($ConvML < \varpi$), then the iterative process is complete and we proceed to the 'Quantization Phase'. Else, we go back to the 'Evaluation Phase' for the subsequent iteration.

5. Quantization Phase: After convergence of the maximum-likelihood function is reached, we generate the locally-maximized P matrix of one motion clip:

$$lmP(i,k) = \frac{P(i,k)\Phi_k^{conv}}{\sum_{k=1}^{K} P(i,k)\Phi_k^{conv}} \tag{10}$$

$lmP(i,k)$ is the maximized probability that component k is representative of the i^{th} motion frame. For each motion frame window, we can compute

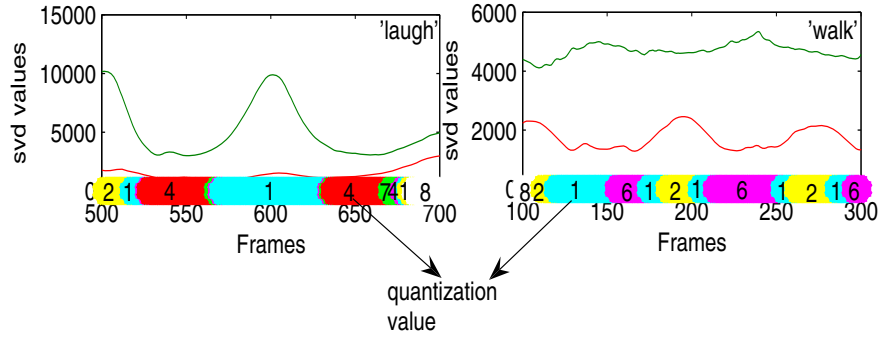

Fig. 4. Temporal Segmentation Effect From GMMEM Quantization Modeling ('laugh' & 'walk' examples)

the j^{th} quantizing component that has the maximum probability among all k components, and then assign the quantization value of this j^{th} component to that motion frame (QuantP(i)).

$$QuantP(i) = \arg\max_{k}(lmP(i,k)) \tag{11}$$

In Figure 4, we observe that quantization value representation from semantic GMMEM has been segmented with temporally similar frames [5][8]. Thus, finally extracted quantization value has spatial and temporal characteristics of a specific motion.

5 Query Resolution

We can notice that similar motions have quite similar patterns of quantization values lie in time series since quantization values we chose is the implication

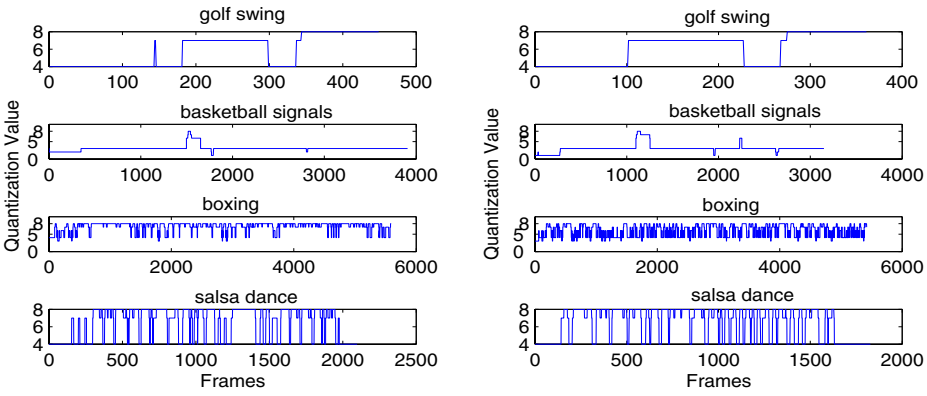

Fig. 5. Quantization Value Representation of Similar Motions

of body segment usages, quantization values are corresponding to semantic description of specific motion. For example, 'boxing' motion has repeated intensive action primitive and 'golf swing' has one big cycle of swing primitive action (see Figure 5). For similarity comparison, we transform quantization values of each motion into histogram values (see Figure 6). We randomly selected a query motion ('laugh','salsa dance' and so on) in the motion database [3] and search the database for the most closest match (using k=1) using kNN (K-Nearest Neighbor) classifier since we're interested in the best single matching motion as in [9].

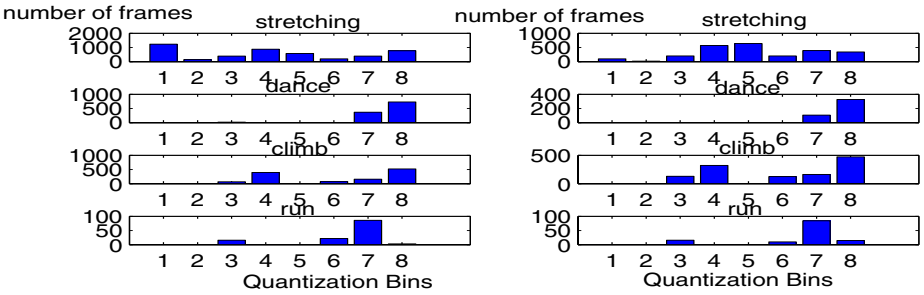

Fig. 6. Histogram Comparison of Quantized Similar Motions

6 Implementation and Performance Analysis

For experiments, we used the Pentium (R) D CPU 3.0 GHz and MATLAB 7.3.0 (R2006b) and the publicly used motion capture data files (CMU Motion Capture Database)[3] and chose 209 human motion clips (370,294 frames) in our experiment. There are 14 different semantic motion categories, which are 'dance','laugh', 'salsa dance','pantomime','siton & standsup','jump','golf swing', 'run','boxing','basketball signal','golf putting','walk','stretching','climb'. Classify and retrieve these subtle actions are quite challengeable problem because of following reasons;

- First, some action categories are semantically very similar to each other. For example, 'dance' vs. 'salsa dance', 'golf swing' vs. 'golf putting', and some other actions are syntactically similar, in that, their actions need to use similar body parts, but they have different meanings. For instance, 'laugh' vs. 'pantomime', 'pantomime' vs 'basketball signal'.
- Second, the lengths of each motion are quite different. For example, a motion may consists of only hundreds frames (it is only 1 second duration since frame rate of AMC motion capture data is 120 frames/sec.) and other one may have ten thousands frames (this motion has more than 125 seconds duration). Frame ranges in our experiment are from 124 frames to 15021 frames.

– Third, variety of frame sizes and human actors (subjects) in the same action category. For example, 'pantomime' category includes 4 different subjects and sizes are from 3508 to 15021, 'climb' has 4 subjects and their size is from 1198 to 6956 frames and so on.

In Figure 7, for showing how much dimensional reduction from spatial feature reduction (62 to 3 dimensions) affects the 3D motion recognition. We compare kNN classification accuracies with original (62) data dimension and reduced (3) dimension that have been quantized using K-means technique. It demonstrates that dimensional reduction by extracting spatial features doesn't deteriorate motion recognition rate, but it increase the accuracies for some motion categories.

We compare with LVQ (Learning Vector Quantization) technique[13] and our proposed sGMMEM(semantic GMMEM) quantization method (in Figure 8), LVQ and sGMMEM quantized spatially extracted data, which got from the first stage of our approach. We can observe sGMMEM can improve kNN classification accuracies so much than LVQ does with 3 dimensional human motion data (see Figure 8). Finally, in Table 1,we can see overall precision and recall values for 14 different motion categories. sGMMEM quantization shows quite good performance in most of motion categories. Especially, for such subtle motion categories as dance (89.47%),laugh (83.33%),basketball signal (85.71%),golf putting(80.00%), golf swing(81.81%) show good recall values. In other hands, LVQ demonstrates much lower accuracies about those subtle motion categories (dance(31.57%),laugh(33.33 %),golf swing(36.36%) and so on). About primitive motion categories, sGMMEM achieves more than 90% precision accuracies (walk (92.85%), run (%91.42),jump (94.11%)) -see Table 1.

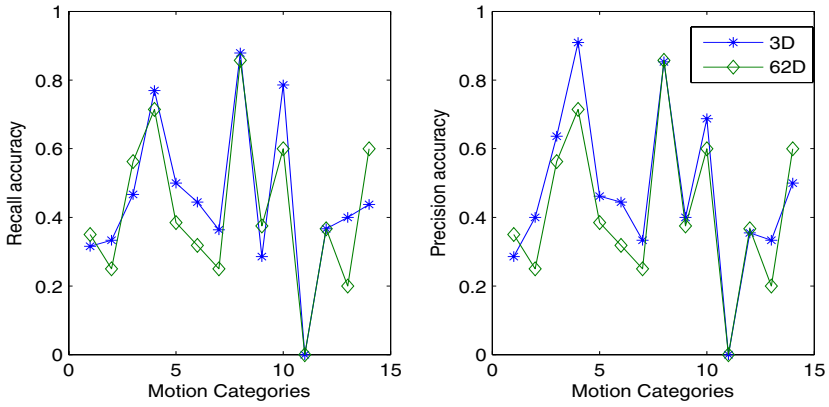

Fig. 7. Comparison Accuracies with 3-dimensional and 62-dimensional Motion Representations

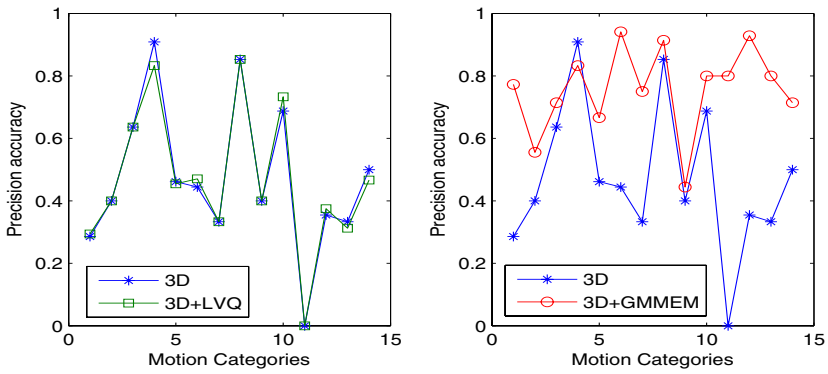

Fig. 8. Comparison Accuracy Improvements between LVQ and Semantic GMMEM Quantization with 3-dimensional Motion Representations

Table 1. Accuracy Comparison for all Motion Categories

Categories	LVQ		Semantic GMMEM	
	precision	recall	precision	recall
1.dance	0.2857	0.3157	0.7727	0.8947
2.laugh	0.4000	0.3333	0.5555	0.8333
3.salsa	0.6363	0.4666	0.7142	0.6666
4.pantomime	0.9090	0.7692	0.8333	0.7692
5.siton standup	0.4615	0.5000	0.6666	0.6666
6.jump	0.4444	0.4444	0.9411	0.8888
7.golf swing	0.3333	0.3636	0.7500	0.8181
8.run	0.8529	0.8787	0.9142	0.9696
9.boxing	0.4000	0.2857	0.4444	0.5714
10.baskeball signal	0.6875	0.7857	0.8000	0.8571
11.golf putting	0.0000	0.0000	0.8000	0.8000
12.walk	0.3548	0.3666	0.9285	0.8666
13.stretching	0.3333	0.4000	0.8000	0.4000
14.climb	0.5000	0.4375	0.7142	0.6250
average	0.4713	0.4533	0.7596	0.7591

7 Conclusions and Future Work

In this paper, we extract spatial (using SVD) and temporal features (using statistical distribution of spatial feature values) of 3D motion clips and map into Gaussian Mixture Semantic Space. The mapped values are then quantized into one of the 8 motion components (corresponding to body movements involving Arms, Torso, and Legs) using Expectation Maximization (EM). In performance analysis, although multi-dimensional(62) human motion capture data are reduced to one-dimensional expression, we demonstrated our approach can keep the semantic characteristics of each human motion clip. As the sequentially

quantized compact representation imply human motion's semantics (body components relationship in the statistical distribution), the quantized representation could bridge the gap between 3D human motion matrix data and human understandable languages with possible extensions with information retrieval research area such as motion vocabulary and motion classification with LSI (Latent Semantic Indexing) through translating 1-dimensional quantization values into repeated rules.

References

1. Li, C., Kulkarni, P.R., Prabhakaran, B.: Motion Stream Segmentation and Recognition by Classification. International Journal of Multimedia Tools and Applications (MTAP) by Springer-Verlag 35(1) (2007)
2. Li, C., Pradhan, G., Zheng, S.Q., Prabhakaran, B.: Indexing of Variable Length Multi-attribute Motion data. In: Proc. of the Second ACM International Workshop on Multimedia, Washington D.C., USA, pp. 75–84 (November 2004)
3. CMU Motion Capture Library, `http://mocap.cs.cmu.edu/`
4. Pradhan, G.N., Li, C., Prabhakaran, B.: Hierarchical Indexing Structure for 3D Human Motion. In: Int'l Proceedings of ACM Multimedia Modeling Conference (MMM) 2007, Singapore, January 9-12 (2007)
5. Muller, M., Roder, T., Clausen, M.: Efficient content based retrieval of motion capture data. ACM Transactions on Graphics (TOG) 24, 677–685 (2005)
6. Liu, G., Zhang, J., Wang, W., McMillan, L.: A system for analyzing and indexing human-motion databases. In: Proc. 2005 ACM SIGMOD International conference on Management of data (2005)
7. Ketterer, J., Puzicha, J., Held, M.: On Spatial Quantization of Color Images. IEEE Transactions on Image Processing 9, 666–682 (2000)
8. Liu, F., Zhuang, Y., Wu, F., Pan, Y.: 3D motion retrieval with motion index tree. Computer Vision and Image Understanding 92, 265–284 (2003)
9. Keogh, E., Palpanas, T., Zordan, V.B., Gunopulos, D., Cardle, M.: Indexing large human-motion databases. In: Proc. 30th VLDB Conference, Toronto, Canada, pp. 780–791 (2004)
10. Bishop, C.M.: Pattern Recognition and Machine Learning. Springer, Heidelberg (2006)
11. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm. Journal of Royal Statistical Society B39, 1–38
12. Golub, G.H., Loan, C.F.: Matrix Computations. The Johns Hopkins University Press, Baltimore, Maryland (1996)
13. Kohonen, T., Kangas, J., Laaksonen, J., Torkkola, K.: A program package for the correct application of Learning Vector Quantization algorithms. In: Proceedings of the International Joint Conference on Neural Networks, Baltimore, pp. 725–730 (June 1992)