A Similarity Measure for Motion Stream Segmentation and Recognition*

Chuanjun Li B. Prabhakaran
Department of Computer Science
The University of Texas at Dallas, Richardson, TX 75083
{chuanjun, praba}@utdallas.edu

ABSTRACT

Recognition of motion streams such as data streams generated by different sign languages or various captured human body motions requires a high performance similarity measure. The motion streams have multiple attributes, and motion patterns in the streams can have different lengths from those of isolated motion patterns and different attributes can have different temporal shifts and variations. To address these issues, this paper proposes a similarity measure based on singular value decomposition (SVD) of motion matrices. Eigenvector differences weighed by the corresponding eigenvalues are considered for the proposed similarity measure. Experiments with general hand gestures and human motion streams show that the proposed similarity measure gives good performance for recognizing motion patterns in the motion streams in real time.

Categories and Subject Descriptors: H.2.8 [Database Management]: Database Applications – Data Mining

General Terms: Algorithm

Keywords: Pattern recognition, gesture, data streams, segmentation, singular value decomposition.

1. INTRODUCTION

Motion streams can be generated by continuously performed sign language words [14] or captured human body motions such as various dances. Captured human motions can be applied to the movie and computer game industries by reconstructing various motions from video sequences [10] or images [15] or from motions captured by motion capture systems [4]. Recognizing motion patterns in the streams with unsupervised methods requires no training process, and is very convenient when new motions are expected to be added to the known pattern pools. A similarity measure with good performance is thus necessary for segmenting and recognizing the motion streams. Such a similarity measure needs to address some new challenges posed by real world motion streams: first, the motion patterns have dozens of attributes, and similar patterns can have different lengths due to different motion durations; second, different attributes of similar motions have different variations and different temporal shifts due to motion variations; and finally, motion streams are continuous, and there are no obvious "pauses" between neighboring motions in a stream. A good similarity measure not only needs to capture the similarity of complete motion patterns, but also needs to capture the differences between complete motion patterns and incomplete motion patterns or sub-patterns in order to segment a stream for motion recognition.

As the main contribution of this paper, we propose a similarity measure to address the above issues. The proposed similarity measure is defined based on singular value decomposition of the motion matrices. The first few eigenvectors are compared for capturing the similarity of two matrices, and the inner products of the eigenvectors are given different weights for their different contributions. We propose to use only the eigenvalues corresponding to the involved eigenvectors of the two motion matrices as weights. This simple and intuitive weighing strategy gives the same importance to eigenvalues of the two matrices. We also show that the 95% variance rule for choosing the number of eigenvectors [13] is not sufficient for recognizing both isolated patterns and motion streams. Our experiments demonstrate that at least the first 6 eigenvectors need to be considered for motion streams of either 22 attribute or 54 attributes, and the first 6 eigenvalues accounts for more than 99.5% of the total variance in the motion matrices.

2. RELATED WORK

Multi-attribute pattern similarity search, especially in continuous motion streams, has been widely studied for sign language recognition and for motion synthesis in computer animation. The recognition methods usually include template matching by distance measures and hidden Markov models (HMM).

Template matching by using similarity/distance measures has been employed for multi-attribute pattern recognition. Joint angles are extracted in [11] as features to represent different human body static poses for the Mahalanobis distance measure of two joint angle features. Similarly, momentum, kinetic energy and force are constructed in [2, 5] as activity measure and prediction of gesture boundaries for various segments of the human body, and the Mahalanobis distance function of two composite features are solved by dynamic programming.

*Work supported partially by the National Science Foundation under Grant No. 0237954 for the project CAREER: Animation Databases.

MDM/KDD 2005 Chicago, August 21, Chicago, Illinois, USA

Similarity measures are defined for multi-attribute data in [6, 12, 16] based on principal component analysis (PCA). Inner products or angular differences of principal components (PCs) are considered for similarity measure definitions, with different weighted strategies for different PCs. Equal weights are considered for different combinations of PCs in [6], giving different PCs equal contributions to the similarity measure. The similarity measure in [12] takes the minimum of two weighted sums of PC inner products, and the two sums are respectively weighted by different weights. A global weight vector is obtained by taking into account all available isolated motion patterns in [16], and this weight vector is used for specifying different contributions from different PC inner products to the similarity measure *Eros*. The dominating first PC and a normalized eigenvalue vector are considered in [7, 8] for pattern recognition. In contrast, this paper propose to consider the first few PCs, and the angular differences or inner products of different PCs are weighted by different weights which depends on the data variances along the corresponding PCs.

The HMM technique has been widely used for sign language recognition, and different recognition rates have been reported for different sign languages and different feature selection approaches. Starner et al. [14] achieved 92% and 98% word accuracy respectively for two systems, the first of the systems used a camera mounted on a desk and the second one used a camera in a user's cap for extracting features as the input of HMM. Similarly Liang and Ouhyoung [9] used HMM for postures, orientations and motion primitives as features extracted from continuous Taiwan sign language streams and an average 80.4% recognition rate was achieved. In contrast, the approach proposed in this paper is an unsupervised approach, and no training as required for HMM recognizers is needed.

3. SIMILARITY MEASURE FOR MOTION STREAM RECOGNITION

The joint positional coordinates or joint angular values of a subject in motion can be represented by a matrix: the columns or attributes of the matrix are for different joints, and the rows or frames of the matrix are for different time instants. Similarity of two motions is the similarity of the resulting motion matrices, which have the same number of attributes or columns, and yet can have different number of rows due to different motion durations. To capture the similarity of two matrices of different lengths, we propose to apply singular value decomposition (SVD) to the motion matrices in order to capture the similarity of the matrix geometric structures. Hence we briefly present SVD and its associated properties below before proposing the similarity measure based on SVD in this section.

3.1 Singular Value Decomposition

The geometric structure of a matrix can be revealed by the SVD of the matrix. As shown in [3], any real $m \times n$ matrix A can be decomposed into $A = U\Sigma V^T$, where $U = [u_1, u_2, \ldots, u_m] \in R^{m\times m}$ and $V = [v_1, v_2, \ldots, v_n] \in R^{n\times n}$ are two orthogonal matrices, and Σ is a diagonal matrix with diagonal entries being the singular values of A: $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_{\min(m,n)} \geq 0$. Column vectors u_i and v_i are the i^{th} left and right singular vectors of A, respectively.

It can be shown that the right singular vectors of the symmetric $n \times n$ matrix $M = A^T A$ are identical to the corresponding right singular vectors of A, referred to as eigenvectors of M. The singular values of M, or eigenvalues of M, are squares of the corresponding singular values of A. The eigenvector with the largest eigenvalue gives the first principal component. The eigenvector with the second largest eigenvalue is the second principal component and so on.

3.2 Similarity Measure

Since SVD exposes the geometric structure of a matrix, it can be used for capturing the similarity of two matrices. We can compute the SVD of $M = A^T A$ instead of computing the SVD of A to save computational time. The reasons are that the eigenvectors of M are identical to the corresponding right singular vectors of A, the eigenvalues of M are the squares of the corresponding singular values of A, and SVD takes $O(n^3)$ time for the $n \times n$ M and takes $O(mn^2)$ time with a large constant for the $m \times n$ A, and usually $m > n$.

Ideally, if two motions are similar, their corresponding eigenvectors should be parallel to each other, and their corresponding eigenvalues should also be proportional to each other. This is because the eigenvectors are the corresponding principal components, and the eigenvalues reflect the variances of the matrix data along the corresponding principal components. But due to motion variations, all corresponding eigenvectors cannot be parallel as shown in Figure 1. The parallelness or angular differences of two eigenvectors u and v can be described by the absolute value of their inner products: $|\cos\theta| = |u \cdot v|/(|u||v|) = |u \cdot v|$, where $|u| = |v| = 1$. We consider the absolute value of the inner products because eigenvectors can have different signs as shown in [8].

Since eigenvalues are numerically related to the variances of the matrix data along the associated eigenvectors, the importance of the eigenvector parallelness can be described by the corresponding eigenvalues. Hence, eigenvalues are to be used to give different weights to different eigenvector pairs. Figure 2 shows that the first eigenvalues are the dominating components of all the eigenvalues, and other eigenvalues become smaller and smaller and approach zero. As the eigenvalues are close to zero, their corresponding eigenvectors can be very different even if two matrices are similar. Hence not all the eigenvectors need to be incorporated into the similarity measure.

Since two matrices have two eigenvalues for the corresponding eigenvector pair, these two eigenvalues should have equal contributions or weights to the eigenvector parallelness. In addition, the similarity measure of two matrices should be independent to other matrices, hence only eigenvectors and eigenvalues of the two matrices should be considered.

Based on the above discussions, we propose the following similarity measure for two matrices Q and P:

$$\Psi(Q, P) = \frac{1}{2}\sum_{i=1}^{k}((\sigma_i / \sum_{i=1}^{n} \sigma_i + \lambda_i / \sum_{i=1}^{n} \lambda_i)|u_i \cdot v_i|)$$

where σ_i and λ_i are the i^{th} eigenvalues corresponding to the i^{th} eigenvectors u_i and v_i of square matrices of Q and P, respectively, and $1 < k < n$. Integer k determines how many eigenvectors are considered and it depends on the number of attributes n of motion matrices. Experiments with hand gesture motions ($n = 22$) and human body motions ($n =$

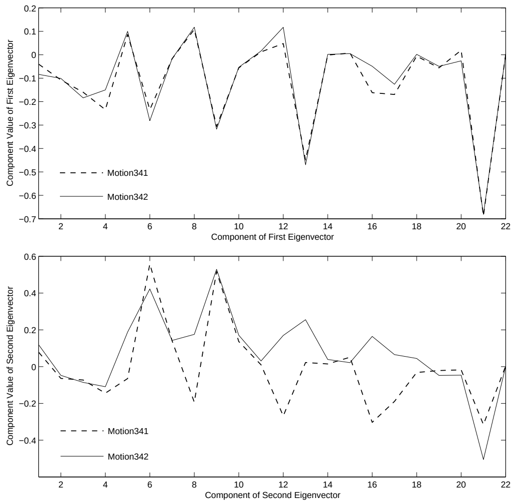


Figure 1: Eigenvectors of similar patterns. The first eigenvectors are similar to each other, while other eigenvectors, such as the second vectors shown in the bottom, can be quite different.

54) in Section 4 show that $k = 6$ is large enough without loss of pattern recognition accuracy in streams. We refer to this non-metric similarity measure as k Weighted Angular Similarity (kWAS) , which captures the angular similarities of the first k corresponding eigenvector pairs weighted by the corresponding eigenvalues.

It can be easily verified that the value of kWAS ranges over [0,1]. When all corresponding eigenvectors are normal to each other, the similarity measure will be zero, and when two matrices are identical, the similarity measure approaches the maximum value one if k approaches n.

3.3 Stream Segmentation Algorithm

In order to recognize motion streams, we assume one motion in a stream has a minimum length l and a maximum length $\mathcal{L}$. The following steps can be applied to incrementally segment a stream for motion recognition:

1. SVD is applied to all isolated motion patterns P to obtain their eigenvectors and eigenvalues. Let δ be the incremented stream length for segmentation, and let L be the location for segmentation. Initially $L = l$.

2. Starting from the beginning of the stream or the end of the previously recognized motion, segment the stream at location L. Compute the eigenvectors and eigenvalues of the motion segment Q.

3. Compute kWAS between Q and all motion patterns P. Update Ψ_{max} to be the highest similarity after the previous motion's recognition.

4. If $L+\delta < \mathcal{L}$, update $L = L+\delta$ and go to step 2. Otherwise, the segment corresponding to Ψ_{max} is recognized to be the motion pattern which gives the highest similarity Ψ_{max}, update $L = l$ starting from the end of the last recognized motion pattern and go to step 2.

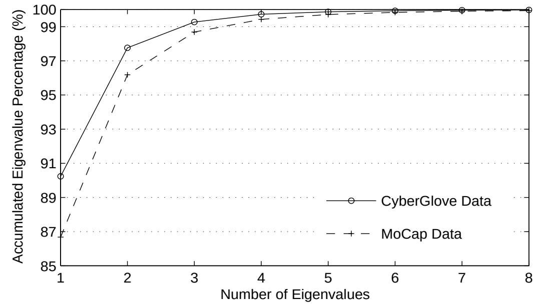


Figure 2: Accumulated eigenvalue percentages in total eigenvalues for CyberGlove data and captured human body motion data. There are 22 eigenvalues for the CyberGlove data and 54 eigenvalues for the captured motion data. The sum of the first 2 eigenvalues is more than 95% of the corresponding total eigenvalues, and the sum of the first 6 eigenvalues is almost 100% of the total eigenvalues.

4. PERFORMANCE EVALUATION

This section evaluates experimentally the performances of the similarity measure kWAS proposed in this paper. It has been shown in [16] that *Eros* [16] outperforms other similarity measures mentioned in Section 2 except MAS [8]. Hence in this section, we compare the performances of the proposed kWAS with *Eros* and MAS for recognizing similar isolated motion patterns and for segmenting and recognizing motion streams from hand gesture capturing CyberGlove and human body motion capture system.

4.1 Data Generation

A similarity measure should be able to be used not only for recognizing isolated patterns with high accuracy, but also for recognizing patterns in continuous motions or motion streams. Recognizing motion streams is more challenging than recognizing isolated patterns. This is because many very similar motion segments or sub-patterns needs to be compared in order to find appropriate segmentation locations, and a similarity measure should capture the difference between a complete motion or pattern and its sub-patterns. Hence, both isolated motion patterns and motion streams were generated for evaluating the performance of kWAS. Two data sources are considered for data generation: a CyberGlove for capturing hand gestures and a Vicon motion capture system for capturing human body motions.

4.1.1 CyberGlove Data

A CyberGlove is a fully instrumented data glove that provides 22 sensors for measuring hand joint angular values to capture motions of a hand, such as American Sign Language (ASL) words for hearing impaired. The data for a hand gesture contain 22 angular values for each time instant/frame, one value for a joint of one degree of freedom. The motion data are extracted at around 120 frames per second. Data matrices thus have 22 attributes for the CyberGlove motions.

One hundred and ten different isolated motions were generated as motion patterns, and each motion was repeated for three times, resulting in 330 isolated hand gesture motions. Some motions have semantic meanings. For example, the motion for BUS as shown in Table 1 is for the ASL sign "bus". Yet for segmentation and recognition, we only require that each individual motion be different from others, and thus some motions are general motions, and do not have any particular semantic meanings, such as the THUMBUP motion in Table 1.

The following 18 motions shown in Table 1 were used to generate continuous motions or streams. Twenty four different motion streams were generated for segmentation and recognition purpose. There are 5 to 10 motions in a stream and 150 motions in total in 24 streams, with 6.25 motions in a stream on average. It should be noted that variable-length *transitional noises* occur between successive motions in the generated streams.

Table 1: Individual motions used for streams

35 60 70 80 90 BUS GOODBYE
HALF IDIOM JAR JUICE KENNEL KNEE
MILK TV SCISSOR SPREAD THUMBUP

4.1.2 *Motion Capture Data*

The motion capture data come from various motions captured collectively by using 16 Vicon cameras and the Vicon iQ Workstation software. A dancer wears a suit of non-reflective material and 44 markers are attached to the body suit. After system calibration and subject calibration, global coordinates and rotation angles of 19 joints/segments can be obtained at about 120 frames per second for any motion. Similarity of patterns with global 3D positional data can be disguised by different locations, orientations or different paths of motion execution as illustrated in Figure 3(a). Since two patterns are similar to each other because of similar relative positions of corresponding body segments at corresponding time, and the relative positions of different segments are independent of locations or orientations of the body, we can transform the global position data into local position data as follows.

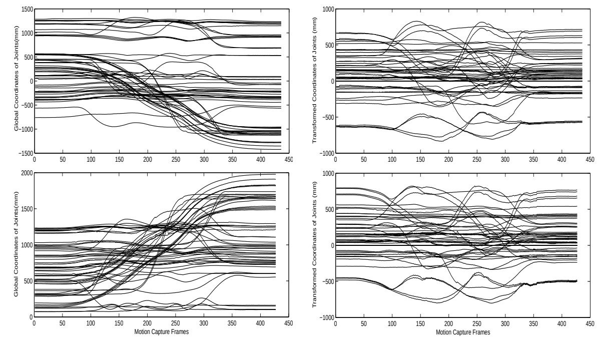
Let X_p, Y_p, Z_p be the global coordinates of one point on pelvis, the selected origin of the "moving" local coordinate system, and α, β, γ be the rotation angles of the pelvis segment relative to the global coordinate system axes, respectively. The translation matrix is T as follows:

$$T = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ -X_p & -Y_p & -Z_p & 1 \end{pmatrix}$$

The rotation matrix $R = R_x \times R_y \times R_z$, where

$$R_x = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & \cos\alpha & -\sin\alpha & 0 \\ 0 & \sin\alpha & \cos\alpha & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

$$R_y = \begin{pmatrix} \cos\beta & 0 & \sin\beta & 0 \\ 0 & 1 & 0 & 0 \\ -\sin\beta & 0 & \cos\beta & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$



(a) (b)

Figure 3: 3D motion capture data for similar motions executed at different locations and in different orientations: (a) before transformation; (b) after transformation.

$$R_z = \begin{pmatrix} \cos\gamma & -\sin\gamma & 0 & 0 \\ \sin\gamma & \cos\gamma & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

Let X, Y, Z be the global coordinates of one point on any segments, and x, y, z be the corresponding transformed local coordinates. x, y and z can be computed as follows:

$$[x \;\; y \;\; z \;\; 1] = [X \;\; Y \;\; Z \;\; 1] \times T \times R$$

The transformed data are positions of different segments relative to a moving coordinate system with the origin at some fixed point of the body, for example the pelvis. The moving coordinate system is not necessarily aligned with the global system, and it can rotate with the body. So data transformation includes both translation and rotation, and the transformed data would be translation and rotation invariant as shown in Figure 3(b). The coordinates of the origin pelvis are not included, thus the transformed matrices have 54 columns.

Sixty two isolated motions including Taiqi, Indian dances, and western dances were performed for generating motion capture data, and each motion was repeated 5 times, yielding 310 isolated human motions. Every repeated motion has a different location and different durations, and can face different orientations. Twenty three motion streams were generated for segmentation. There are 3 to 5 motions in a stream, and 93 motions in total in 23 streams, with 4.0 motions in a stream on average.

4.2 Performance of kWAS for Capturing Similarities and Segmenting Streams

We first apply kWAS to isolated motion patterns to show that the proposed similarity measure kWAS can capture the similarities of isolated motion patterns. Then kWAS is applied to motion streams for segmenting streams and recognizing motion patterns in the streams. We experimented with different k values in order to find out the smallest k without loss of good performance.

Figure 2 shows the accumulated eigenvalue percentages averaged on 330 hand gestures and 310 human motions, respectively. Although the first two eigenvalues account for

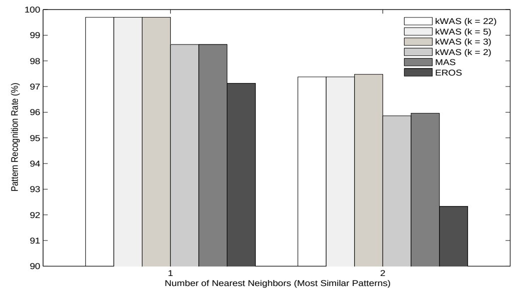


Figure 4: Recognition rate of similar CyberGlove motion patterns. When k is 3, kWAS can find the most similar motions for about 99.7% of 330 motions, and can find the second most similar motions for 97.5% of the them.

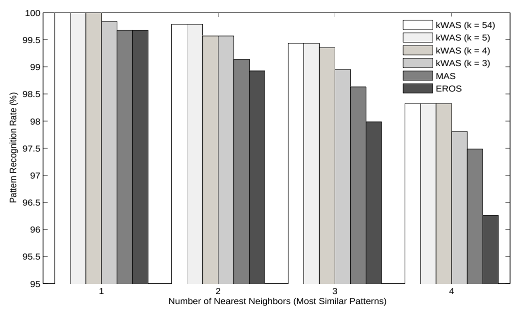


Figure 5: Recognition rate of similar captured motion patterns. When k is 5, by using kWAS, the most similar motions of all 310 motions can be found, and the second most similar motions of 99.8% of the 310 motions can also be found.

more than 95% of the respective sums of all eigenvalues, considering only the first two eigenvectors for kWAS is not sufficient as shown in Figure 4 and Figure 5. For CyberGlove data with 22 attributes, kWAS with $k = 3$ gives the same performance as kWAS with $k = 22$, and for motion capture data with 54 attributes, kWAS with $k = 5$ gives the same performance as kWAS with $k = 54$. Figure 4 and Figure 5 illustrate that kWAS can be used for finding similar motion patterns and outperforms MAS and *Eros* for both hand gesture and human body motion data.

The steps in Section 3.3 are used for segmenting streams and recognizing motions in streams. The recognition accuracy as defined in [14] is used for motion stream recognition. The motion recognition accuracies are shown in Table 2. For both CyberGlove motion and captured motion data, $k = 6$ is used for kWAS, which gives the same accuracy as $k = 22$ for CyberGlove data and $k = 54$ for motion capture data, respectively.

Figure 6 shows the time taken for updating the candidate segment, including updating the matrix, computing the SVD of the updated matrix, and computing the similarities of the segment and all motion patterns. The code implemented in C++ was run on one 2.70 GHz Intel processor of a GenuineIntel Linux box. There are 22 attributes for the CyberGlove streams, and 54 attributes for the captured

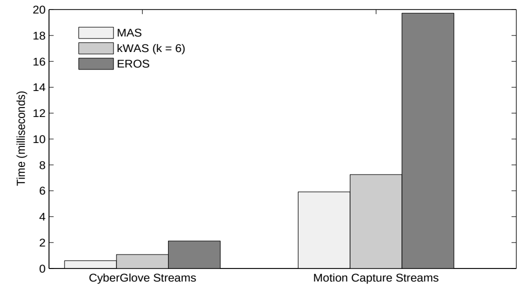


Figure 6: Computation time for stream segment update and similarity computation.

Table 2: Stream Pattern Recognition Accuracy (%)

Similarity Measures	CyberGlove Streams	Motion Capture Streams
Eros	68.7	78.5
MAS	93.3	78.5
kWAS (k=6)	94.0	94.6

motion streams. Hence updating captured motion segments takes longer than updating CyberGlove motion segments as shown in Figure 6. The time required by kWAS is close to the time required by MAS, and is less than half of the time taken by using *Eros*.

4.3 Discussions

kWAS captures the similarity of square matrices of two matrices P and Q, yet the temporal order of pattern execution is not revealed in the square matrices. As shown in [7], two matrices with the identical row vectors in different orders have identical eigenvectors and identical eigenvalues. If different temporal orders of pattern execution yield patterns with different semantic meanings, we need to further consider the temporal execution order, which is not reflected in the eigenvectors and eigenvalues and has not been considered previously in [6, 12, 16].

Since the first eigenvectors are close or parallel for similar patterns, we can project pattern A onto its first eigenvector u_1 by Au_1. Then similar patterns would have similar projections (called projection vectors hereafter), showing similar temporal execution orders while the projection variations for each pattern can be maximized. The pattern projection vectors can be compared by computing their dynamic time warping (DTW) distances, for DTW can align sequences of different lengths and can be solved easily by dynamic programming [1]. Incorporating temporal order information into the similarity measure can be done as for MAS in [7] if motion temporal execution orders cause motion pattern ambiguity to kWAS.

5. CONCLUSIONS

This paper has proposed a similarity measure kWAS for motion stream segmentation and motion pattern recognition. kWAS considers the first few k eigenvectors and computes their angular similarities/differences, and weighs contributions of different eigenvector pairs by their corresponding eigenvalues. Eigenvalues from two motion matrices are given equal importance to the weights. Experiments with CyberGlove hand gesture streams and captured human body motions such as Taiqi and dances show that kWAS can recognize 100% most similar isolated patterns and can recognize 94% motion patterns in continuous motion streams.

6. REFERENCES

[1] D. Berndt and J. Clifford. Using dynamic time warping to find patterns in time series. In *AAAI-94 Workshop on Knowledge Discovery in Databases*, pages 229–248, 1994.

[2] V. M. Dyaberi, H. Sundaram, J. James, and G. Qian. Phrase structure detection in dance. In *Proceedings of the ACM Multimedia Conference 2004*, pages 332–335, Oct. 2004.

[3] G. H. Golub and C. F. V. Loan. *Matrix Computations*. The Johns Hopkins University Press, Baltimore,Maryland, 1996.

[4] L. Ikemoto and D. A. Forsyth. Enriching a motion collection by transplanting limbs. In *Proceedings of the 2004 ACM SIGGRAPH/Eurographics symposium on Computer animation*, pages 99 – 108, 2004.

[5] K. Kahol, P. Tripathi, S. Panchanathan, and T. Rikakis. Gesture segmentation in complex motion sequences. In *Proceedings of IEEE International Conference on Image Processing*, pages II – 105–108, Sept. 2003.

[6] W. Krzanowski. Between-groups comparison of principal components. *J. Amer. Stat. Assoc.*, 74(367):703–707, 1979.

[7] C. Li, B. Prabhakaran, and S. Zheng. Similarity measure for multi-attribute data. In *Proceedings of the 2005 IEEE International Conference on Acoustics, Speach, and Signal Processing (ICASSP)*, Mar. 2005.

[8] C. Li, P. Zhai, S.-Q. Zheng, and B. Prabhakaran. Segmentation and recognition of multi-attribute motion sequences. In *Proceedings of the ACM Multimedia Conference 2004*, pages 836–843, Oct. 2004.

[9] R. H. Liang and M. Ouhyoung. A real-time continuous gesture recognition system for sign language. In *Proceedings of the 3rd. International Conference on Face and Gesture Recognition*, pages 558–565, 1998.

[10] K. Pullen and C. Bregler. Motion capture assisted animation: texturing and synthesis. In *SIGGRAPH*, pages 501–508, 2002.

[11] G. Qian, F. Guo, T. Ingalls, L. Olson, J. James, and T. Rikakis. A gesture-driven multimodal interactive dance system. In *Proceedings of IEEE International Conference on Multimedia and Expo*, June 2004.

[12] C. Shahabi and D. Yan. Real-time pattern isolation and recognition over immersive sensor data streams. In *Proceedings of the 9th International Conference on Multi-Media Modeling*, pages 93–113, Jan 2003.

[13] A. Singhal and D. E. Seborg. Clustering of multivariate time-series data. In *Proceedings of the American Control Conference*, pages 3931–3936, 2002.

[14] T. Starner, J. Weaver, and A. Pentland. Real-time american sign language recognition using desk and wearable computer based video. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(12):1371–1375, 1998.

[15] C. J. Taylor. Reconstruction of articulated objects from point correspondences in a single image. *Computer Vision and Image Understanding*, 80(3):349–363, 2000.

[16] K. Yang and C. Shahabi. A PCA-based similarity measure for multivariate time series. In *Proceedings of the Second ACM International Workshop on Multimedia Databases*, pages 65–74, Nov. 2004.