Segmentation and Recognition of Motion Capture Data Stream by Classification

CHUANJUN LI, PUNIT R. KULKARNI and B. PRABHAKARAN
The University of Texas at Dallas

Three dimensional human motions recorded by motion capture and hand gestures recorded by using data gloves generate variable-length data streams. These data streams usually have dozens of attributes, and have different variations for similar motions. To segment and recognize motion streams, a classification-based approach is proposed in this paper. Classification feature vectors are extracted by utilizing singular value decompositions (SVD) of motion data. The extracted feature vectors capture the dominating geometric structures of motion data as revealed by SVD. Multi-class support vector machine (SVM) classifiers with class probability estimates are explored for classifying the feature vectors in order to segment and recognize motion streams. Experiments show that the proposed approach can find patterns in motion data streams with high accuracy.

Categories and Subject Descriptors: I.5.3 [**Multimedia**]: data mining

Keywords: Multimedia, classification, support vector machine, motion segmentation, gesture recognition, pattern analysis, singular value decomposition

1. INTRODUCTION

Recognition of human motion streams from 3D motion capture systems or data gloves can find wide applications in many situations, such as surveillance video systems, 3D animation and simulation-based training, gait analysis and rehabilitation and gesture recognition. Data streams generated in these applications have multiple attributes, each of which is for the angular values or positional coordinates of one sampling point or joint of a motion subject. Due to the multi-attribute property, many more challenges need to be addressed for pattern discovery (compared to mining uni-attribute time series data streams). Firstly, dozens of attributes need to be considered together to make motions meaningful. Secondly, similar motions can have different lengths, local shifting or scaling, and streams have transitions of variable lengths due to different ending and starting positions of neighboring motions, as shown in Figure 1. And finally, different attributes of both isolated motions and the motions in streams can have different variations. In addition, if global positional information is available, motions can be disguised by having very dissimilar data matrices. This is because similar motions carried out at different locations, following different trajectories, or in different orientations have different global positional coordinates at corresponding time instants.

Authors' addresses: C. Li, P. R. Kulkarni and B. Prabhakaran, Department of Computer Science, the University of Texas at Dallas, Richardson, TX 75080; email: {chuanjun,prk032000,praba}@utdallas.edu.

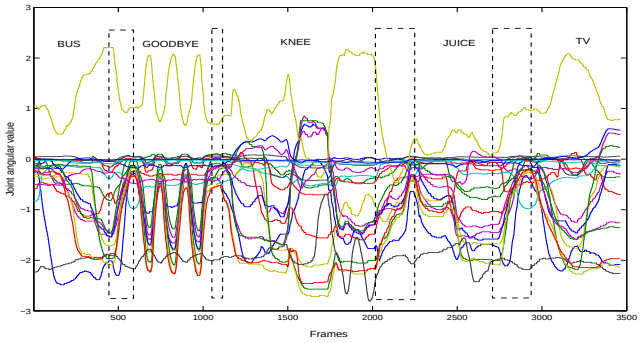

Fig. 1. Multi-attribute hand gestures for American Sign Language words BUS, GOODBYE, KNEE, JUICE and TV. Transitions between neighboring motions/signs are marked with the dotted boxes.

As shown in [Vapnik 1998] and [Li et al. 2005], when only isolated motions are considered and multiple examples for each motion are available, classification can have higher recognition rate than template matching with certain similarity or distance measure. An interesting question then comes into our mind: can we apply classification to segment multi-attribute streams for motion/pattern recognition and achieve high recognition rate?

Applying classification to segment human motion streams (integrated multiple time series) needs the uncertainty of classification so that differences in motion candidates can be compared. If the uncertainty can be at least qualitatively determined, it would be possible to segment and recognize human motion streams. The classification uncertainty can be captured if classifiers can output probabilities of classification in addition to class labels. For simplicity, we refer to complete motions, both training examples and complete test motions in streams, as patterns, and refer to other test motion candidates as sub-patterns.

1.1 Proposed Approach

Since no classifiers take multi-attribute data sequences directly for classification, we propose to extract feature vectors by using singular value decomposition (SVD) to capture the major geometric structures of motion data matrices. Classification is then applied to the extracted feature vectors to recognize/reject the original test motion candidates. We choose support vector machine (SVM) classifiers for classification, considering that SVM is one of the most powerful classification techniques, and posterior probabilities can be successfully estimated [Vapnik 1998; Platt 2000]. Class probability estimates are used not only for recognizing class labels of complete motions or patterns, but also for rejecting incomplete motions or sub-patterns. For training, we propose to use complete pattern examples only, and no non-pattern/incomplete examples at all. For classifying the sequentially segmented motion candidates, temporal coherence is considered. The best class is chosen from the two classes which give the two highest probabilities. The class to which a larger number of motion candidates are classified is determined to be the best class and

the motion candidate which has the highest class probability is the best motion segment.

1.2 Our Contributions

In this paper, data streams of both one degree of freedom (DOF) hand gesture data and three DOFs motion capture data are studied. We segment multi-attribute streams and recognize stream patterns by exploring multi-class SVM classifiers with estimated class probabilities. We make the following main contributions:

(1) A framework is proposed for discovering patterns in multi-attribute motion streams. This framework considers SVD for feature extraction, SVM with class probabilities for classification and motion continuity for final segmentation and pattern recognition.
(2) We show that without including non-pattern/incomplete motions as negative examples for any class, classification can still be done by having only all the necessary complete patterns as positive class examples for training.
(3) Hand gesture streams can be recognized by classification with accuracy 82.43%, and human motion streams can be recognized by classification with accuracy 88.9%.

2. RELATED WORK

Multi-attribute data include positional coordinates and joint angular values of motion subjects and feature values can be extracted by using different approaches. In [Vlachos et al. 2004], 2D positional time series are transformed into angle/arc-length pair sequences that are translation, scale and rotation invariant for pattern recognition. Joint angles are extracted in [Qian et al. 2004] as features to represent human body static poses for a Mahalanobis distance measure. Similarly, momentum, kinetic energy and force are constructed in [Kahol et al. 2003; Dyaberi et al. 2004] as an activity measure and for prediction of gesture boundaries for various segments of the human body. In contrast, the approach developed in this paper is applicable to both angular data and positional data. If global positional data is available, positional coordinates are first transformed into local positional coordinates, and feature values are extracted from the transformed data.

Template matching by using similarity/distance measures has been employed for multi-attribute pattern recognition. Various similarity measures are defined for multi-attribute data in [Krzanowski 1979; Shahabi and Yan 2003; Yang and Shahabi 2004] based on principal component analysis. Inner products of principal components weighted by different strategies are considered in these papers. In [Li et al. 2004; Li et al. 2005], pattern recognition takes into account not only the dominating singular vectors and the associated singular values, but also the temporal orders of the patterns. As shown in [Li et al. 2004], if temporal orders are not considered, different patterns can have the identical corresponding singular vectors and associated singular values. [Li and Prabhakaran 2005] extends the work developed in [Li et al. 2004] by considering the angular similarities of several corresponding singular vectors weighted by the associated singular values.

Machine learning techniques have been utilized for pattern recognition [Shahabi et al. 2001; Starner et al. 1998]. [Shahabi et al. 2001] applies learning techniques

such as Decision Trees, Bayesian classifiers and Neural Networks to recognize static signs for a 10-sign vocabulary, and achieves 84.66% accuracy. Multi-attribute American Sign Language (ASL) motions are considered in [Starner et al. 1998], and five-word sentences are segmented at the word level and recognized by using hidden Markov models (HMMs) with 92-98% word accuracy. SVM combined with HMM have also been applied to visual speech [Gordan et al. 2002] and speech recognition [Ganapathiraju et al. 2004]. HMM models are used to model the temporal evolution of a stream via an underlying Markov process, while SVMs are used to post-process the data generated by a conventional HMM system. It is believed in [Gordan et al. 2002; Ganapathiraju et al. 2004] that SVMs are inherently static classifiers, while speech is a temporal process and cannot be modeled effectively by SVMs. If streams or motion sequences contain complex patterns from a large pattern set, the number of different state combinations can become astronomical, making a state machine approach (HMMs or Neural Networks) impractical. Although motion patterns are also temporal processes, we will show that motion streams can be segmented and recognized with high accuracy by using SVD, SVM with class probability estimates.

SVM classifiers with decision values, rather than class probability estimates, have been successfully applied to classify isolated patterns/multi-attribute motion data in [Li et al. 2005], where motion data are hand joint angles generated by using the data gloves called CyberGlove. In contrast, this paper uses the SVM classifiers with class probabilities to address the more challenging issue of segmenting motion streams.

3. FEATURE EXTRACTION

Multi-attribute motion data can be represented by matrices in which each column represents one attribute, and each row presents one recording of a motion stream. To classify motions of multiple attributes, feature vectors need to be extracted. This section describes how to extract features from motion data matrices for stream segmentation.

The geometric structure of a matrix can be revealed by the SVD of the matrix. As shown in [Golub and Loan 1996], any real $m \times n$ matrix A can be decomposed into $A = V\Sigma U^T$, where $V = [v_1, v_2, \ldots, v_m] \in R^{m\times m}$ and $U = [u_1, u_2, \ldots, u_n] \in R^{n\times n}$ 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 v_i and u_i are the i^{th} left and right singular vectors of A, respectively.

Given a unit 2-norm vector $x \in \Re^n$, applying $A \in \Re^{m\times n}$ to x results in an m dimensional vector Ax. The ith component of Ax is actually the projection of the ith row vector in A onto x. The ith largest singular value σ_i of A is actually the 2-norm or Euclidean length of the ith longest projected vector Ax which is orthogonal to all the $i-1$ longer orthogonal vectors. That is,

$$\sigma_i = \max_{U} \min_{x\in U, \|x\|_2=1} \| Ax \|_2$$

where the maximum is taken over all i-dimensional subspaces $U \subseteq \Re^n$ [Golub and Loan 1996]. Note that σ_1 is the largest 2-norm of A projections onto any x

directions:

$$\sigma_1 = \max_{\|x\|_2=1} \| Ax \|_2$$

Hence the right singular vectors are the corresponding projection directions, and the singular values account for the Euclidean lengths of different vectors Au_i projected by the row vectors in A onto different right singular vectors u_i.

When two motions are similar, the row vectors in the motion data matrices should cover similar trajectories in the n-dimensional space, hence the geometric structures of the motion data matrices are similar. Identically, for two similar motions, all corresponding singular vectors should be close to each other, and the corresponding singular values should also be proportional to each other. For realistic motions with variations, singular vectors associated with different singular values have different sensitivities to the motion variations. If a singular value is large and well separated from its neighbors, the associated singular vector would be relatively insensitive to small motion variations. On the other hand, if a singular value is among a poorly separated cluster, its associated singular vector would be highly sensitive to motion variations [G.W.Stewart 1973; Golub and Loan 1996].

Figure 2 shows the accumulative singular values for hand gestures and captured human subject motions. It shows that the first two singular values account for more than 95% of the sum of singular values, while the others might be very small. Accordingly, the corresponding first two singular vectors of similar motions, especially the first singular vectors would be close or parallel to each other as shown in Figure 3, while other singular vectors can be too sensitive to motion variations.

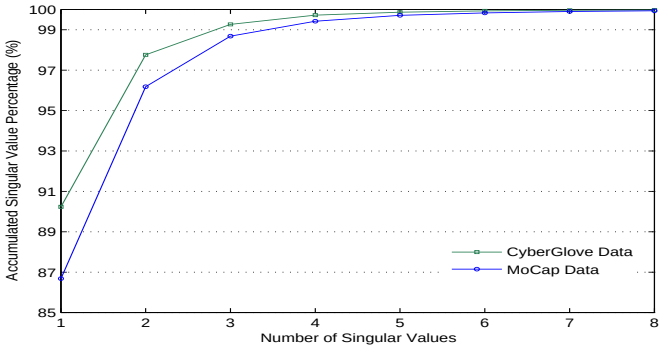

Fig. 2. Accumulated singular value percentages in total singular values for two data sources: Hand gesture data and captured human subject motion data. There are 22 singular values for one hand gesture motion, and 54 singular values for one motion capture motion.

Before the singular vectors can be used to extract feature vectors, unique singular vectors should be obtained for each motion matrix and similar motion matrix should have singular vectors of similar directions or consistent signs. Let the SVD of a pattern matrix A be $A = W\Sigma U^T$, then $Au_1 = \sigma_1 v_1$. Since $\sigma_1 > 0$, singular vector u_1 can have opposite signs as long as the sign of v_1 is consistent. The following steps can be taken to have consistent signs for u_1 of similar patterns:

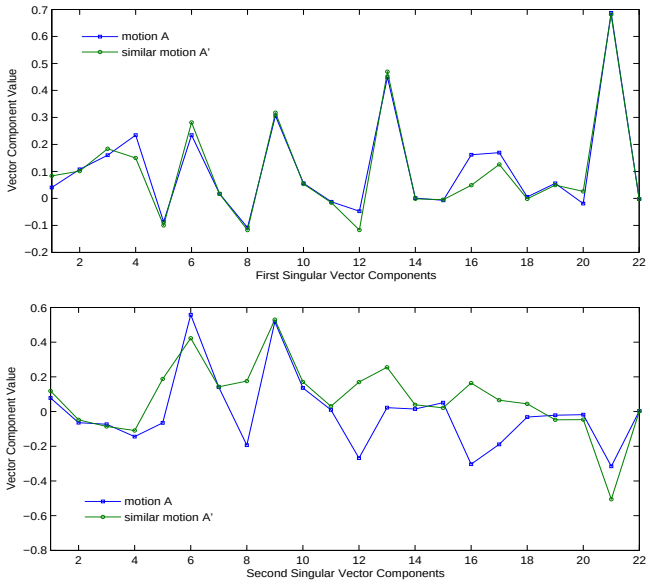

Fig. 3. Singular vectors of similar patterns. The first singular vectors are similar to each other, while other singular vectors, even the second vectors as shown in the bottom, can be quite different.

(1) Generate a matrix S with rows being the first right singular vectors u_1 of all known patterns.
(2) Subtract the elements of S by their corresponding column means, and let the resulting matrix with zero column means be S.
(3) Compute the SVD of S and let its first right singular vector be s_1.
(4) Project the first right singular vector u_1 of patterns (or pattern candidates) onto s_1 by computing $u_1 \cdot s_1$.
(5) Negate all components of any u_1 if the corresponding the inner product $u_1 \cdot s_1 < 0$, and let u_1 be the negated vector.

Step (3) computes the first right singular vector s_1, along which the first right singular vectors of all patterns have the largest variance. As $|u_1| = 1$ and $|s_1| = 1$, the inner product $u_1 \cdot s_1 = |u_1||s_1|\cos(\alpha)$ ranges over [-1, 1], where α is the angle between the two vectors. Since the projections of u_1 onto s_1 have the largest variances among projections on any unit vectors, we can expect that $u_1 \cdot s_1$ will not cluster around zero. Our experiments with hundreds of patterns of different sources show that no pattern has $|u_1 \cdot s_1| < 0.3$. Because similar patterns should have close projections $|u_1 \cdot s_1|$, reasonable variations in naturally performed similar patterns would not result in $u_1 \cdot s_1$ projections of opposite signs if their u_1 signs are the same. That is, only if the u_1 signs of similar motions are opposite can their $u_1 \cdot s_1$ projections have different signs. Hence, u_1 of similar motions would have the same sign by requesting $u_1 \cdot s_1 > 0$.

Similarly, the above steps can be repeated for u_2 with all u_1 replaced by u_2, resulting in consistent signs for the second singular vectors of similar motions.

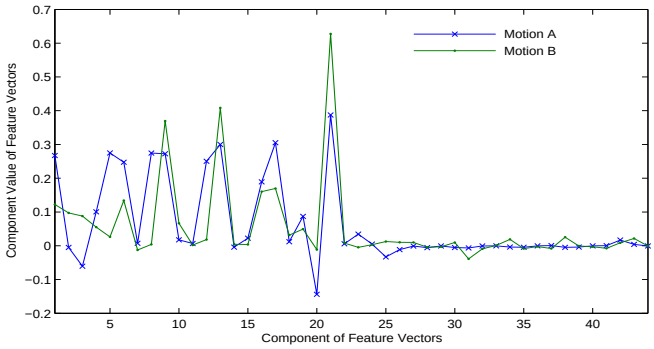

Fig. 4. Feature vectors of two different hand gesture motions A and B. The first half of a feature vector is w_1u_1, and the second half is w_2u_2.

After having unique singular vectors and singular values, we can construct feature vectors from the singular vectors and singular values. The first two singular vectors are the most dominating factors contributing to the similarity of two motions due to their associated large singular values. Other singular vectors are less reliable in capturing the similarities due to their associated singular values which might be small and approach zero. Since the first singular vectors are more reliable to reflect the geometric structures of motion data matrices than the second singular vectors, we can use their associated singular values as weights to reflect the reliability of the singular vectors. Hence, *feature vectors are constructed by concatenating the weighted first singular vectors* w_1u_1 *with the weighted second singular vectors* w_2u_2 as shown in Figure 4, where $w_i = \sigma_i / \sum_{k=1}^{n} \sigma_k$. These feature vectors are extracted by using only the prominent information from singular vectors and singular values.

4. STREAM SEGMENTATION BY CLASSIFICATION

This section discusses how to recognize patterns in multi-attribute streams by classifying the feature vectors extracted as above using SVM classifiers. The flow chart of the motion segmentation and recognition by classification is shown in Figure 5.

4.1 Classifier Selection

SVMs have demonstrated widespread successful uses in many pattern recognition problems [Burges 1998; Gordan et al. 2002; Ganapathiraju et al. 2004; Natsev et al. 2004; Li et al. 2005]. The good classification performance of a binary SVM is due to the optimal hyperplane which maximizes the margin, or the distance between separating hyperplane and the training examples nearest to the hyperplane.

Multi-class classifiers are commonly constructed from binary classifiers because two-class problems are much easier to solve than multi-class problems. There are two approaches to using binary classes for multi-class classification: one-versus-one and one-versus-rest. The former constructs $k(k-1)/2$ binary classifiers for k classes, and each binary classifier is trained only on the data of the two involved classes. The latter constructs k binary SVM classifiers, and each classifier is trained with data from one class as positive examples and data from all the other classes as

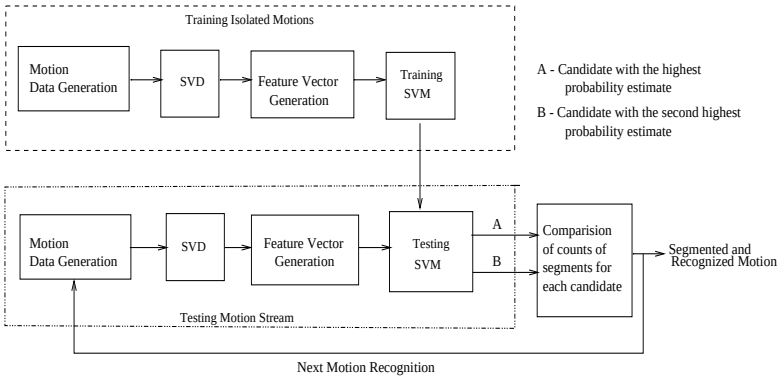

Fig. 5. Motion stream segmentation and recognition process.

negative examples. It has been shown by experiments on large-scale problems in [Hsu and Lin 2002] that in general the accuracy rate of one-versus-one multi-class SVM is higher than that of one-versus-rest, and the training time of one-versus-one multi-class SVM is less than that needed for one-versus-rest classifiers. Due to these reasons, we chose to use the one-versus-one multi-class SVM classifiers for classification. We propose to use multi-class SVM classifiers for distinguishing patterns from incomplete motions or sub-patterns segmented from a stream, and for classifying patterns. No sub-patterns or non-patterns are needed as negative examples, and all classes include only complete patterns.

Let the decision function resulting from the output of a standard binary SVM be

$$f(x) = \sum_{i=1}^{N} \alpha_i y_i K(x, x_i) + b$$

where N is the number of support vectors, $\{\alpha_i\}$ are non-negative Lagrange multipliers each of which corresponds to an example from the training data set, b is a bias term from training the SVM, K is a kernel function which allows a dot product in a higher dimensional feature space to be computed in the input space, and $\{y_i\}$ are the class assignments (-1 or +1) of training data vectors $\{x_i\}$.

The above decision function does not produce a probability. In many applications, a posterior probability, rather than an un-calibrated decision value sign(f), is needed for capturing the classification uncertainty. Efforts of mapping standard SVM outputs to posterior probabilities have been made in [Vapnik 1998; Platt 2000]. Platt [Platt 2000] uses a sigmoid function to estimate the binary class probability which is monotonic in the standard output f:

$$p(y = +1|f) = \frac{1}{1 + \exp(Af + B)}$$

The parameters A and B can be fitted by using maximum likelihood estimation.

For multi-class classification, class probabilities can be estimated from binary class probabilities by pairwise coupling. Wu et al.[Wu et al. 2004] propose a multi-class probability approach which is more stable than other popular existing methods by using the following optimization:

Optimization:

$$\min_p \sum_{i=1}^{k} \sum_{j:j\neq i} (r_{ij}p_j - r_{ji}p_i)^2$$

under the constraints:

$$\sum_{i=1}^{k} p_i = 1, p_i \geq 0, \forall i$$

where r_{ij} are the binary class probability estimates of $\mu_{ij} \equiv P(y = i|y = i \text{ or } j, x)$ as obtained in [Platt 2000].

The above optimization problem can be solved using Gaussian elimination after some algebra as shown in [Wu et al. 2004]. After the probabilities are estimated for all classes, and the class with the largest posterior is chosen to be the winning class for a test vector: $\arg\max_i[p_i]$.

4.2 Stream Segmentation

We assume that a pattern in a stream has a minimum length l and a maximum length $\mathcal{L}$. Initially, a stream is segmented into multiple segments. Each segment starts at the beginning of the stream, and ends at $l + i \times \Delta$, where $\Delta > 0$ is the segment length difference, i is the number of segments ending between l and $\mathcal{L}$. The feature vector for each motion segment is constructed according to Section 3.

Ideally, a completely segmented pattern in a stream will have the highest probability of being classified to its class in the training set. When the optimal hyperplane divides the space between the correct and incorrect classes, this completely segmented pattern will lie on the side of the correct class. Similarly as the lengths of the motion candidates increase from l and the motion candidates approach a motion pattern, their feature vectors would become closer to the optimal hyperplane if they are on the side of any other classes, and move away from the optimal hyperplane if they are on its correct class side. The corresponding probability of the incomplete motion candidates being classified to the correct pattern class will eventually increase to the maximum as shown in Figure 6. Hence, the point with the highest probability should be the segmentation point for a complete pattern, and the class with the highest probability should be the right class for the segmented pattern.

In practice, the probability of certain motion candidate being classified to some class might be higher than the probability of a complete motion being classified to the right class as shown in Figure 7. This is because extracting feature vectors from motion data matrices inevitably loses some information, and this information loss, together with motion transitions and variations, can affect the classification probabilities when stream segmentation is considered. Nevertheless, we have observed that the class to which a completely segmented motion candidate is classified is almost always one of the two classes whose highest probabilities are the highest two among all highest class probabilities. We have also observed that if a completely segmented motion candidate is not classified to the class of the highest probability, the class to which it is classified has a much larger amount of motion candidates than the class of the highest probability as shown in Figure 7. This is because

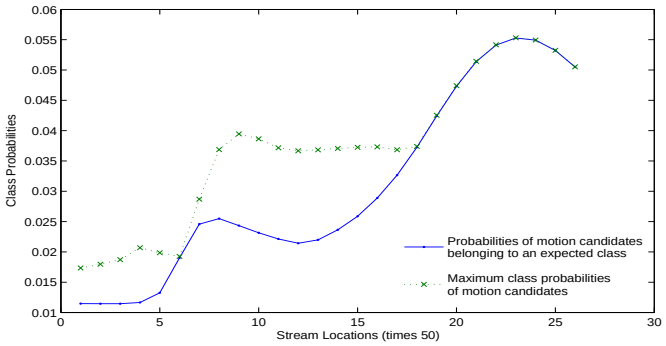

Fig. 6. Changes in estimated probabilities. The probability estimate reaches maximum as a pattern is completed, and decreases as the stream continues.

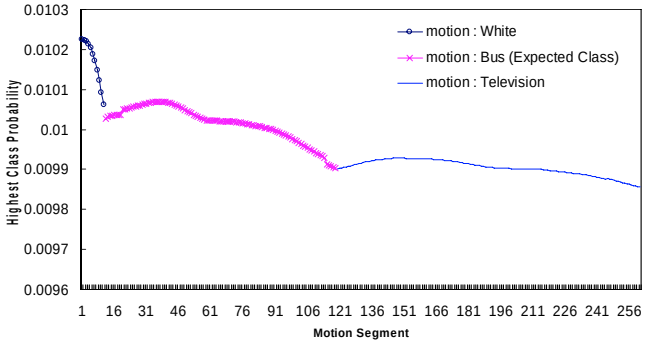

Fig. 7. The highest class probabilities of different motion segments in a stream beginning with motion Bus. Although the first motion in the stream is Bus, the first few segments have the highest probabilities of being classified to class White, while the second group of segments have the highest probabilities of being classified to class Bus, and others have the highest probabilities of being classified to class Television. Notice that the candidate count in the Bus class is much more than the candidate count in the White class.

a large number of the candidates which are close to the best candidate also have the highest probabilities of being classified to the expected class. Hence, we obtain two classes with the highest probabilities instead of only one class with the highest probability. To choose the final class for a sequence of motion segments, we consider further the number of motion candidates being classified to each of the two classes. The class *with higher candidate count* is chosen to be the best one, and the segment, which has the highest probability among all the candidates being classified to the same class, is chosen to be the best motion candidate.

The next pattern recognition starts from the end of the last recognized pattern in the stream, and the same process repeats until the remaining stream has length less than the minimum stream length l as illustrated in Figure 5. In the next section, we will experiment with the above segmentation approach by using both one DOF hand gesture data and three DOFs data from the Vicon motion capturing system.

5. EXPERIMENT EVALUATION

5.1 Data Generation

Hand gesture streams and human motion streams are used for performance evaluation. Hand gestures were generated by using a data glove called CyberGlove, and human motions were captured by using 16 Vicon cameras and the Vicon iQ Workstation software.

CyberGlove Data: A CyberGlove is a fully instrumented 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 DOF. The motion data are extracted at 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 3 times. That is, each of the 110 classes has 3 examples. Twelve different motion streams were generated for segmentation and recognition purpose. A gesture stream contains 5 to 10 gestures concatenated by brief transitions.

Motion Capture Data: If we have global 3D positional information such as 3D motion capture data, patterns might be disguised by the differences in motion execution, such as different locations, different orientations or different paths. Since two patterns are similar to each other because of similar relative positions of corresponding subject joints/segments at corresponding time, and the relative positions of different joints are independent of locations or orientations of the subject, we can transform the global position data into local position data. The transformed data are the position coordinates of different joints relative to a moving coordinate system with the origin at some fixed point of the subject, for example the pelvis, the moving coordinate system is not necessarily aligned with the global system, and it can rotate with the subject. So data transformation includes both translation and rotation, and the transformed data would be translation and rotation invariant.

motion data matrices have 54 columns for coordinates of 18 joints. One hundred isolated motions including Taiqi, Indian dances, and western dances were performed for generating captured motions, and each motion was repeated for 5 times. Every motion repetition has a different location and can face different orientations. Hence we have 100 classes of motion patterns, and each of the classes has 5 examples for SVM training. Twelve motion streams were also generated for stream segmentation. The motion streams include 3 to 5 different length motion patterns each and the patterns in the motion streams have various-length transitions.

5.2 Performance of Classification

k-fold cross validations are used for training of SVMs, where k is 3 for hand gestures and 5 for 3D captured motions. The average cross validation accuracy is 96.7% for the isolated hand gestures, and is 97.7% for the isolated captured motions.

We use the recognition accuracy Acc as defined in [Starner et al. 1998]:

$$Acc = \frac{N - (I + D + S)}{N}$$

where N is the total number of motions/patterns in the test motion streams, I is

the number of insertions, D is the number of deletions, and S is the number of substitutions.

For the hand gesture data streams, there are 74 patterns in the 12 streams. Recognizing 74 patterns results in 11 insertions, 1 deletion and 1 substitution. The accuracy is 82.43%. For the motion capture data streams, there are 45 patterns in the 12 streams. The accuracy is 88.9% with 5 deletions. For comparison, the most related similarity measure Eros as proposed in [Yang and Shahabi 2004] is applied to the hand gesture data streams and the motion capture data streams. The results, as listed in Table I, show that the proposed classification approach to segmenting and recognizing multi-attribute streams gives high accuracy and performs better than Eros.

Table I. Stream Pattern Recognition Accuracy (%)

Data Source	Classification	Eros
Hand Gesture	82.43 (N=74, I=11, D=1, S=1)	71.6 (N=74, I=8, D=8, S=5)
Motion Capture	88.9 (N=45, D=5)	80.0 (N=45, D=6, S=3)

5.3 Discussions

Segmenting and recognizing motion streams is more challenging as indicated by the different recognition accuracies of isolated motions and motion streams. Motions can have different durations, and our extracted feature vectors are not affected by the differences in motion data lengths. On the other hand, long transitions between motions in a stream and large motion variations might make the feature vectors less discriminative. All the streams we experimented with have reasonably long transitions and small motion variations. Future work can be done for motions with large variations and long transitions. For example, a different motion candidate generation algorithm might be needed for streams with long transitions so that transitions can be explicitly identified. As the results in Table I show, insertions and deletions cause most of the problems for stream segmentation and recognition, while few substitutions occurs. This suggests that our extracted feature vectors are very discriminative, yet long motion transitions might be misrecognized to be some motions, and very short motions might be deleted, or be combined with neighboring motions, due to motion variations.

For motions following similar trajectories in different directions, vectors indicative of motion directions can be constructed as in [Li et al. 2005] and dynamic time warping distances of the motion direction vectors can be used for motion direction recognition.

6. CONCLUSIONS

In this paper, multi-class SVMs with probability estimates are proposed for segmenting multi-attribute human motion streams and recognizing patterns in them. SVD is applied to extract feature vectors for the multi-attribute motions. The extracted feature vectors are constructed by concatenating the first singular vectors

with the corresponding second singular vectors weighted by the associated singular values. Two classes with the highest probabilities are chosen to be the class candidates for motion candidates of one motion pattern, and the class with higher candidate count is the winning class. The candidate which has the highest probability of being classified to the best class is the best segmented motion candidate. Experiments with hand gesture data and 3D motion capture data show that SVMs with probability estimates combined with SVD can segment and recognize motion patterns in multi-attribute motion streams with high accuracy.

REFERENCES

Burges, C. J. 1998. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery* 2, 121–167.

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

Ganapathiraju, A., Hamaker, J. E., and Picone, J. 2004. Application of support vector machines to speech recognition. *IEEE Transactions on Signal Processing 52,* 8, 2348–2355.

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

Gordan, M., Kotropoulos, C., and Pitas, I. 2002. Application of support vector machines classifiers to visual speech recognition. In *Proceedings of the International Conference on Image Processing*. 24–28.

G.W.Stewart. 1973. Error and perturbation bounds for subspace associated with certain eigenvalue problems. *SIAM Review 15,* 4, 727–764.

Hsu, C.-W. and Lin, C.-J. 2002. A comparison of methods for multi-class support vector machines. *IEEE Transactions on Neural Networks* 13, 415–425.

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

Krzanowski, W. 1979. Between-groups comparison of principal components. *J. Amer. Stat. Assoc. 74,* 367, 703–707.

Li, C., Khan, L., and Prabhakaran, B. 2005. Real-time classification of variable length multi-attribute motion data. *International Journal of Knowledge and Information Systems (KAIS)*.

Li, C. and Prabhakaran, B. 2005. A similarity measure for motion stream segmentation and recognition. In *Proceedings of the Sixth International Workshop on Multimedia Data Mining*.

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

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

Natsev, A., Naphade, M. R., and Smith, J. R. 2004. Semantic representation, search and mining of multimedia content. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge discovery and Data Mining*. 641–646.

Platt, J. C. 2000. Probabilistic outputs for support vector machines and comparison to regularized likelihood methods. In *Advances in Large Margin Classifiers*, A. J. Smola, P. L. Bartlett, B. Scholkopf, and D. Schuurmans, Eds. MIT Press, Cambridge, MA. URL citeseer.nj.nec.com/platt99probabilistic.html.

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

Shahabi, C., Kaghazian, L., Mehta, S., Ghoting, A., Shanbhag, G., and McLaughlin, M. 2001. Analysis of haptic data for sign language recognition. In *Proceedings of the 9th International Conference on Human Computer Interaction*. 441 – 445.

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

Starner, T., Weaver, J., and Pentland, A. 1998. 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.

Vapnik, V. N. 1998. *Statistical Learning theory*. Wiley, New York.

Vlachos, M., Gunopulos, D., and Das, G. 2004. Rotation invariant distance measures for trajectories. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge discovery and Data Mining*. 707–712.

Wu, T.-F., Liu, C.-J., and Weng, R. C. 2004. Probability estimates for multi-class classification by pairwise coupling. *Journal of Machine Learning Research* 5, 975–1005.

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