MAHALANOBIS DISTANCE BASED POLYNOMIAL SEGMENT MODEL FOR CHINESE SIGN LANGUAGE RECOGNITON

Yu Zhou[1], Xilin Chen[2], Debin Zhao[1], Hongxun Yao[1], Wen Gao[3,1]

[1]School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China
[2]Key Lab of Intelligent Information Processing, Chinese Academy of Sciences (CAS)
Institute of Computing Technology, CAS, Beijing, China
[3]Institute of Digital Media, Peking University, Beijing, China
{yzhou, xlchen, dbzhao, hxyao, wgao }@jdl.ac.cn

ABSTRACT

Sign Language Recognition (SLR) systems are mostly based on Hidden Markov Model (HMM) and have achieved excellent results. However, the assumption of frame independence in HMM makes it inconsistent with the characteristic of strong temporal correlation in sign language signals. Polynomial Segment Model (PSM) explicitly represents the temporal evolution of sign language features as a Gaussian process with time-varying parameters. In this paper PSM is first introduced to SLR framework to solve the temporal correlation problem. Considering the correlation among the coefficients of polynomial trajectory's different orders, Mahalanobis distance is used as the classification criterion to evaluate the likelihood of test data. Experimental results show that our method outperform the conventional HMM methods by 6.81% in recognition accuracy.

Index Terms— Sign Language Recognition, Hidden Markov Model, Polynomial Segment Model, Mahalanobis distance

1. INTRODUCTION

SLR aims to transcribe sign language to text in order to make the communications between deaf and hearing society convenient. The first work on SLR appeared in the literature in 1990s. Since then SLR has become a hot research area around the world. Previous research works are mostly based on HMM.

Starner et al. [1] proposed a view-based approach for continuous American Sign Language recognition. They used a single camera to extract two-dimensional features as the input of HMM. While recognizing the sentences over a vocabulary of 40 signs, the word recognition accuracies of 92% and 98% were obtained when the camera was mounted on the desk and in the cap of the user, respectively. For continuous Taiwanese Sign Language recognition, Liang and Ouhyoung [2] employed the time-varying parameter threshold of the hand postures to determine the end points in a stream of gesture input. An average recognition rate of 80.4% was obtained over a vocabulary of 250 signs. In their system, HMM was employed, and a data glove was taken as input device. Vogler and Metaxas [3] used computer vision methods to extract the three-dimensional (3-D) parameters of motions from a signer's arm as the input to HMM, and recognized continuous ASL sentences over a vocabulary of 53 signs. Wang and Gao [4] designed a real-time system to recognize continuous Chinese Sign Language (CSL) sentences with a vocabulary of 4800 words. The data was collected from two CyberGloves and a 3-D tracker. They employed HMM with states tying, still frame detecting and search algorithm. Recognition rate of over 90% for continuous CSL recognition was achieved. McGuire, Hernandez-Rebollar et al. [5] realized a mobile one-way American Sign Language translator. They grounded their efforts in a particular scenario and described an initial recognition accuracy of 94% accuracy on a vocabulary of 141 words signed in phrases of fours signs using a one-handed glove-based system and HMM.

Although HMM has played an important role for a long time, it has two limitations: (1) weak duration modeling and (2) assumption of the conditional independence of observations belonging to the same state. Actually, the observations of consecutive frames are sometimes strong temporal correlated in SLR, which leads to poor performance of the HMM approach. In this paper, we adopt PSM to model CSL basic units so that we can solve the problem of temporal correlation. We argue that choosing Mahalanobis distance as the classification criterion is superior to using Euclidean distance in PSM evaluation.

The remaining part of this paper is organized as follows: Section 2 describes the characteristic of sign language trajectories to reveal how PSM models sign language trajectories more precisely than HMM; Then in Section 3 we give a brief introduction to PSM including its definition and parameter estimation; We propose the implementation of our method in Section 4; Experimental results are

reported in Section 5 and the paper is concluded in Section 6.

2. SIGN LANGUAGE TRAJECTORIES

In speech recognition, the number of phonemes is finite and each speech signal can be regarded as a sequence of homogeneous regions, where each homogeneous region corresponds to a phoneme. So the structure of HMM is consistent with the speech phonology. Nevertheless, in CSL recognition, there do not exist the basic units similar to phonemes in speech recognition. Moreover sign language trajectories are not always a sequence of homogeneous regions but a sequence of regions in which the observations are strong temporal correlated in consecutive frames.

Two etymon sequences with different time-space structure are plotted in Fig. 1. The sequence in Fig.-1-a is composed of 3 homogeneous regions. In each region the values of observations are very close, which can be viewed as independent and identically sampled from the same probability density function. This structure can be modeled well by HMM as in Fig.-1-c. Howbeit the sequence in Fig.-1-b is not composed of homogeneous regions, especially in the middle part. The observations in the middle part are time varying, thus they can't be modeled well by any HMM. We model this type of sequences with PSM, as shown in Fig.-1-d. Not only can PSM model the type of SLR trajectories as in Fig.-1-b, but also can model the type of SLR trajectories as in Fig.-1-a. We first introduce PSM in Section 3.

3. POLYNOMIAL SEGMENT MODEL

PSM was firstly described by Gish and Ng [6]. They used PSM to develop a secondary processing algorithm that rescored putative events hypothesized by a primary HMM word spotter trying to improve performance by discriminating true keywords from false alarms. Thereafter, many works showed that PSM can perform comparably well compared with HMM and are more powerful in simple tasks such as phoneme recognition. But PSM was not applied to large vocabulary speech recognition tasks because of the high computational complexity of its evaluation. Recently Li and Siu [7] proposed a new approach to evaluate the likelihood of a PSM segment by efficiently "accumulating" segment likelihood one frame at a time. This decreased the computational complexity of PSM evaluation and made PSM a model for speech recognition without the need of using other models for pre-segmentation. In [8] they showed that segment likelihood can be evaluated efficiently in an order of computational complexity similar to HMM, and they also introduced a fast PSM search algorithm that intelligently prunes the number of hypothesized segment boundaries. In this Section we only give an introduction to PSM. For more details about PSM, please refer to [6][7][8].

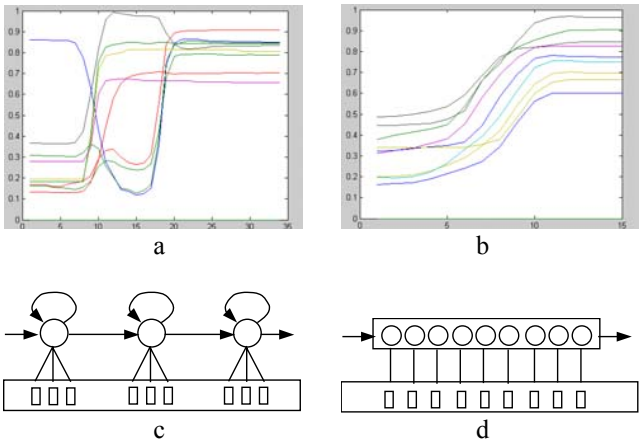

Figure. 1. The comparison between two different types of trajectories, the left is fit for HMM, the right is not fit, but both a and b can be modeled well by PSM.

Given a sign language segment O, the definition of its corresponding PSM is:

$$O = Z_N B + E \tag{1}$$

where O is a $N \times D$ observation matrix which contains N frames of D dimensional feature vectors. B is a $(R+1)\times D$ parameter matrix of a R^{th} order trajectory model and E is the residual error that is the same size as the feature vector O. Z_N is a $N \times (R+1)$ design matrix for an R^{th} order trajectory model that normalizes the segments of different frames to [0, 1].

For a set of segments $\{O_1, O_2, \ldots, O_M\}$, the maximum likelihood estimation for the PSM parameter matrix $\hat{B}$ and the residual covariance $\hat{\Sigma}$ are given by

$$\hat{B} = \left[\sum_{m=1}^{M} Z_{N_m}^T Z_{N_m}\right]^{-1}\left[\sum_{m=1}^{M} Z_{N_m}^T O_m\right] \tag{2}$$

and

$$\hat{\Sigma} = \frac{\sum_{m=1}^{M}\left(O_m - Z_{N_m}{}^T B_{N_m}\right)^T\left(O_m - Z_{N_m}{}^T B_{N_m}\right)}{\sum_{m=1}^{M} N_m} \tag{3}$$

4. MAHALANOBIS DISTANCE BASED PSM

For SLR tasks, if we consider PSM as a set of Gaussians with a time-varying mean, given a specified PSM, the log likelihood can be computed using equation (4):

$$\begin{aligned} L_N\left(O \mid \hat{B}, \hat{\Sigma}\right) = & -\frac{N}{2}\left[D\log(2\pi) + \log\left|\hat{\Sigma}\right|\right] \\ & -\frac{1}{2} tr\left[\left(O - Z_N\hat{B}\right)\hat{\Sigma}^{-1}\left(O - Z_N\hat{B}\right)^T\right] \end{aligned} \tag{4}$$

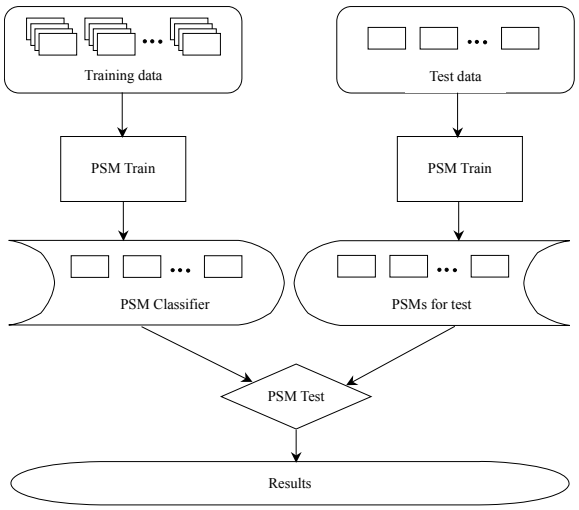

Figure. 2. Our sign language recognition framework.

The triplet, $\{B, \Sigma, N\}$, are the sufficient statistics of O, we can compute $L_N\left(O \mid \hat{B}_\lambda, \hat{\Sigma}_\lambda\right)$ as below:

$$\begin{aligned} & L_N\left(O \mid \hat{B}, \hat{\Sigma}\right) \\ & = L_N\left(B, \Sigma \mid \hat{B}, \hat{\Sigma}\right) \\ & = -\frac{N}{2}\left[D\log(2\pi) + \log\left|\hat{\Sigma}\right|\right] - \frac{N}{2} tr\left[\hat{\Sigma}^{-1}\Sigma\right] \\ & \quad -\frac{1}{2} tr\left[Z_N\left(B-\hat{B}\right)\hat{\Sigma}^{-1}\left(B-\hat{B}\right)^T Z_N{}^T\right] \end{aligned} \tag{5}$$

In our work we compute the likelihood of O given a specified PSM by the means of comparing the PSM representing O with this specified PSM. In equation (5) we have shown that the likelihood of O given a specified PSM is equal to the likelihood of PSM representing O given a specified PSM, that's $L_N\left(O \mid \hat{B}, \hat{\Sigma}\right) = L_N\left(B, \Sigma \mid \hat{B}, \hat{\Sigma}\right)$. In speech recognition, there are enough sample segments to train one PSM so the computation for the residual error matrix Σ is straightforward. However, in large vocabulary CSL recognition, only several samples (in our work, 3) are available for a basic unit (in our work, etymon), which results in the singularity of the residual error matrix $\hat{\Sigma}$. In that case we can't compute the log likelihood directly.

We compute the similarity between two PSMs by computing the distance between two parameter matrixes $B_{PSM-Test}$ and $B_{PSM-Templet}$. $B_{PSM-Test}$ is the parameter matrix of PSM which is estimated by test sample, and $B_{PSM-Templet}$ is the parameter matrix of PSM which is estimated by training samples. The framework of our method is shown in Fig. 2.

In our method, given each CSL etymon we compute two PSMs corresponding to training data and test data separately. Then at the recognition stage for each PSM trained from test data we compute the distance between this PSM and every PSM trained from training data. The classification result is the PSM whose distance is the nearest from the PSM trained from test data.

To measure the distance between two parameter matrix B and $\hat{B}$, we adopt Euclidean distance firstly:

$$ED\left(B, \hat{B}\right) = \sum_{i=1}^{D} \sqrt{\sum_{j=0}^{R}\left(b_{ji} - \hat{b}_{ji}\right)^2} \tag{6}$$

where b_{ji} and $\hat{b}_{ji}$ corresponding to the jth row and ith column element of B and $\hat{B}$ respectively.

Euclidean distance supposes that the coefficients of polynomial trajectory's different orders contribute equally to the evaluation likelihood and are not correlated, which is not consistent with CSL recognition. For this reason Mahalanobis distance [9] was used in our work:

$$MD(B, \hat{B}) = \sum_{i=1}^{D}\left(b_i - \hat{b}_i\right)^T \Sigma^{-1}\left(b_i - \hat{b}_i\right) \tag{7}$$

The computation of equation (7) consists of much redundancy, so we reformulate equation (7) as:

$$MD(B, \hat{B}) = sum\left(\left(\left(B-\hat{B}\right)\left(B-\hat{B}\right)^T\right) \circ \Sigma^{-1}\right) \tag{8}$$

where the matrix $X \circ Y$ represents the Hadamard product of X and Y, and $sum(X)$ represents the sum of all elements of matrix X. By this reformulation we can decrease the computation complexity greatly.

5. EXPERIMENTAL RESULTS

In our experiments, two Cybergloves and three Polhemus 3SPACE-position trackers are used as input devices. Two trackers are positioned on the wrist of each hand and another is fixed at the back of the signer (as the reference tracker). The Cybergloves collect the variation information of hand shape with the 18-dimentional data each hand, and the position trackers collect the variation information of orientation, position and movement trajectory.

In order to extract the invariant features to signer's position, the tracker at the back of the signer is chosen as the reference Cartesian coordinate system, and the position and orientation at each hand with respect to the reference system are calculated and can be taken as invariant features. After this transformation, the data consists of a relative three-dimensional position vector and a three-dimensional orientation vector for each hand, which do not change with the position and orientation of the signer. In the case of two hands, a 48-dimensional vector is formed, including the hand shape, position and orientation. The data from different signers are calibrated by some fixed postures performed by each signer. In our experiments 14 postures that can represent the min-max value ranges of the corresponding sensor are defined. As each component in the

vector has different dynamic range, its value is normalized to [0, 1].

Previous work [10] has shown that the recognition rate based on etyma is comparable to that based on words, so we use etyma as basic recognition units. All the experiments are carried on the large vocabulary with 2438 etyma. Experimental data consists of 59512 samples over 2438 signs from 6 signers, and each signer performs each sign four times. The vocabulary is taken from the CSL dictionary. One group data from each signer are referred to as test set and the other 3 groups data are used as the training set.

We compared the recognition performance of HMM, Euclidean distance based PSM and Mahalanobis distance based PSM. The results are shown in Fig. 3. Compared with HMM, the accuracy is improved by 1.13% with Euclidean distance based PSM, and when Mahalanobis distance based PSM is used the accuracy is improved by 6.81%.

The reason of the accuracy improvements lies in that PSM considers the frame correlation which is not included by HMM. The reason of that Mahalanobis distance based PSM outperform Euclidean distance based PSM is that Mahalanobis distance considers the correlation among coefficients of different orders whereas Euclidean distance does not consider it.

6. CONCLUSIONS AND FUTURE WORK

In this paper, to solve the problem of strong temporal correlation in consecutive frames, we introduce PSM to CSL recognition for the first time. At recognition stage Euclidean distance and Mahalanobis distance are separately used as the classification criteria. Both Euclidean distance based PSM and Mahalanobis distance based PSM are argued to be superior to HMM. Furthermore, compared with Euclidean distance based PSM, Mahalanobis distance based PSM can achieve considerable improvements. The improvements may arise from the fact that Mahalanobis distance criterion considers correlations among the coefficients of the polynomial trajectories' different orders. Finally by reformulating the equation for computing Mahalanobis distance, we can decrease the computational complexity greatly.

In the future, we will focus on the signer independent recognition and signer adaptation using PSM. Since PSM describes signal segments directly, we can also extract CSL phonemes using PSM.

ACKNOWLEDGMENT

This research is partially sponsored by Natural Science Foundation of China under contract No.60533030 and No.60603023, and also by Natural Science Foundation of Beijing Municipal under contract No.4061001.

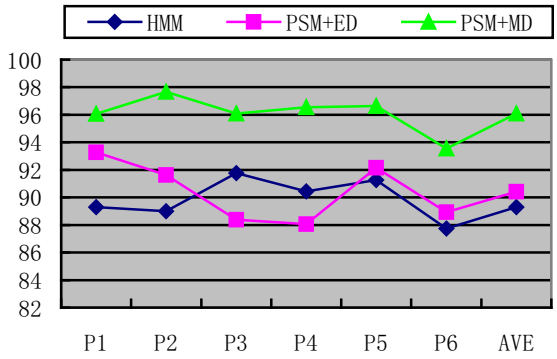

Figure. 3. Comparison among recognition rates of HMM, PSM+ED and PSM+MD.

7. REFERENCES

[1] T. Starner, J. Weaver, and A. Pentland, "Real-time American sign language recognition using desk and wearable computer based video," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, pp. 1371-1375, December, 1998.

[2] R. H. Liang and M. Ouhyoung, "A real-time continuous gesture recognition system for sign language," *Proc. Int. Conf. Autom. Face Gesture Recognit*, pp. 558-565, 1998.

[3] C. Vogler and D. Metaxas, "Adapting hidden Markov models for ASL recognition by using three-dimensional computer vision methods," *IEEE Int. Conf. Syst., Man Cybern*, pp. 156-161, 1997.

[4] C. L. Wang, W. Gao and Z. G. Xuan, "A Real-Time Large Vocabulary Continuous Recognition System for Chinese Sign Language," *Pacific Rim Conference on Multimedia*, pp. 150-157, 2001.

[5] R. M. McGuire, J. Hernandez-Rebollar, T. Starner, V. Henderson, H. Brashear, D. S. Ross, "Towards a one-way American sign language translator," *IEEE International Conference on Automatic Face and Gesture Recognition* , pp. 620-625, 2004.

[6] H. Gish and K. Ng, "A segmental speech model with applications to word spotting," *Proc. of ICASSP*, pp. 447–450, 1993.

[7] C. F. Li and M. Siu, "An efficient incremental likelihood evaluation for polynomial trajectory model using with application to model training and recognition," *Proc. of ICASSP*, pp. I-756- I-759, 2003.

[8] C. F. Li, M. Siu and S-K. Au-Yeung, "Recursive likelihood evaluation and fast search algorithm for polynomial segment model with application to speech recognition," *IEEE Transactions on Audio Speech and Language Processing*, vol. 14, pp. 1704-1718, September, 2006.

[9] R.D. Maesschalck, D. Jouan-Rimbaud, D.L. Massart, "The Mahalanobis distance," *Chemometrics and Intelligent Laboratory Systems*, vol. 50, pp. 1-18, January, 2000.

[10] C. L. Wang, X. L. Chen and W. Gao, "A Comparison Between Etymon- and Word-Based Chinese Sign Language Recognition Systems," *Gesture Workshop,* pp. 84-87, 2005.