DTW/ISODATA algorithm and Multilayer architecture in Sign Language Recognition with large vocabulary

Feng Jiang, Hongxun Yao, Guilin Yao
Department of Computer Science, Harbin Institute of Technology, China
86-451-6416485
fjiang@hit.edu.cn

***Abstract*-Up to now analytical or statistical methods have been used in sign language recognition with large vocabulary. Analytical methods such as Dynamic Time Wrapping (DTW) or Euclidian distance have been used for isolated word recognition, but the performance is not satisfactory enough because it is easily interfered by noise. Statistical methods, especially hidden Markov Models are commonly used, for both continuous sign language and isolated words and with the expansion of vocabulary the processing time becomes increasingly unacceptable. Therefore, a multilayer architecture of sign language recognition for large vocabulary is proposed in this paper for the purpose of speeding up the recognition process. In this method the gesture sequence to be recognized is first located at a set of words that are easy to be confused (confusion set) through a global cursory search and then the gesture is recognized through a more aborative local search and the generation of confusion set is realized by DTW/ISODATA algorithm. Experiment results indicate that it is an effective algorithm for Chinese sign language recognition.**

Keywords: Sign language recognition, DTW/ISODATA, Multilayer architecture.

I. INTRODUCTION

Sign language is composed of a number of basic gestures. The deaf people use different combinations of hand configurations and hand motion for their speechless communication. The aim of sign language recognition is to provide an accurate and effective mechanism to transcribe sign language into text or speech. Recently, there have been strong efforts in developing multi-functional perception and natural interfaces between users and systems, these systems are based on gesture recognition [1-5].

Two approaches for modeling sequence of gesture are prevalent in sign language recognition: the template-matching approach and the statistical approach. The template-matching approach is to design one or more templates for each word and then seek the optimal alignment between the incoming pattern and each of the templates. The alignment process is carried out through the DTW algorithm [6]. We will henceforth refer to each template together with the DTW algorithm as a DTW model. The statistical approach to sign language recognition lays a more theoretically sound foundation to modeling by attributing a statistical model, usually a hidden Markov model, to each word. With this approach, we formally consider the recognition problem as a statistical classification task. In the real time gesture recognition system, one of the crucial problems is the speed of recognition. The probability computation step is generally performed with the Viterbi algorithm and requires on the order of $V \cdot N^2 \cdot T$ computations, where V is the size of vocabulary, N is the hidden states number and T is the length of gesture sequence. The computation complexity for recognition increases with the expansion of the vocabulary size. Increased processing power can improve performance, but it does not improve the performance/cost ration. A solution to this problem is to design a signer-independent recognizer which is not based on direct comparisons of the feature vectors sequence. Having classified the vocabulary space into confusion sets of sign vocabulary, we locate the incoming feature vector sequences at one of the confusion sets first and then the further recognition task focuses only on the identified set. The corresponding Multilayer architecture fusing DTW and HMM is presented in the later section. The structure of this paper is as follows: section 2 introduces the data collection and the Multilayer architecture; section 3 discusses the DTW/ISODATA algorithm; the experimental results and their comparisons are shown in section 4 and the conclusion is given in the last section.

II. SYSTEM ARCHITECTURE

A . Data Collection

Currently there are two major approaches to collect gesture data: the visual approach, obtaining data from a video camera, and the instrumental approach, involving special measuring devices. We choose the latter because it provides a concise representation of hand shapes and does not require sophisticated preprocessing. By the instrumental approach, data is collected from two Cyber Gloves and three 3SPACE-position trackers which function as input devices. Two of the three position trackers are fixed on each of the two wrists and another on the back of human body as the reference tracker. The Cyber Gloves collect the variations and measure angles at 18 major joints while the position trackers collect the variations of orientation, position, movement trajectory. The data from position trackers can be converted in the following way: the Cartesian coordinate system of the trackers at each hand together with the reference is calculated as invariant features. Through this transformation, the data can be presented as a relative three-dimensional position vector and a

three-dimensional orientation vector of each hand. The range of each component is different and should be normalized to $0-1$.

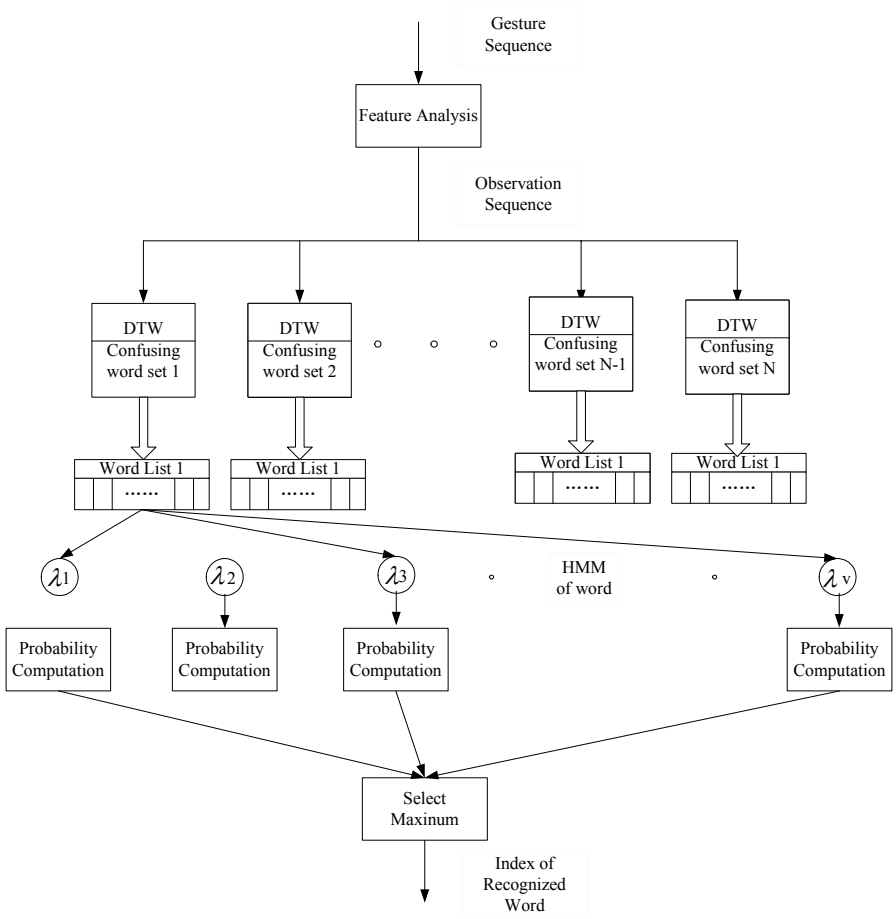

Figure 1. Multilayer architecture with two-stage hierarchy

B. Multilayer Architecture In Sign Language Recognition

As the most popular and effective method in sign language recognition, DTW and HMM are intrinsically related with each other, and on the other hand, they also have their own features. Links between DTW and HMM have been suggested in the literature. Juang[7]presented a unified view of the two models and observed that DTW searches for the best alignment path while the HMMs like hood function sums the density along all possible alignment paths. Bridle [8] independently observed similar links between DTW and HMMs and noted that the Baum-Welch estimation algorithm for HMMs can be viewed as a generalization of the segmental k-means estimation method widely used in DTW. Unlike the HMM, DTW has no concrete mathematical formulation. Generally speaking, DTW can provide a higher level of granularity in the movement path comparing with HMM [9].When choosing a suitable recognition engine, the performance depends on several factors including an acceptable processing time, and a degree of noise interference. Much research done before have the result that HMM has better accuracy than DTW and the performance of DTW is not satisfactory enough because of the noise. According to the analysis above, form the view of multi-scale observation, DTW is effective in wide scale observation and HMM is suitable for solving the analysis of the detail. Hence it is feasible to devise a recognition engine which combines these two methods.

III. DTW/ISODATA ALOGRITHM

One important step in the Multilayer architecture is to build the confusing vocabulary set. The DTW/ISODATA algorithm is presented as the solution to this problem.

We turn to use the Dynamic Time Warping algorithm with local constraints on path to compute the distance between two gesture sequences which are allowed to have different lengths. The DTW performs a time alignment and normalization by computing a temporal transformation function allowing two sequences to be matched. Given two sequences to compare, if we consider a table having the signals in the first row and column respectively, the temporal function can be seen as a path in the table. The global path cost (locally accumulated over the time) represents the dissimilarity between the two sequences and the template signal with the least path cost is the closest among the inputs. Finally the best match distance is compared with a threshold distance value to determine whether the identity claim should be accepted or rejected.

Having got the distances of every two samples in the vocabulary space, we adopt an unsupervised classification using ISODATA algorithm to get the confusion set[10]. The clustering procedure which provides a set of rules for splitting and combining existing clusters to be used to obtain a final

clustering is agglomerative and hierarchical. These clusters are represented by prototypes which have the minimum sum of distance to the other samples in the same cluster. In our work the index of prototype is defined as follows:

$$P(l) = \arg\max_j \sum_{k=1}^{n_l} d\left(C_l^j, C_l^K\right) \quad (1)$$

Where $d(\bullet,\bullet)$ is a DTW-based distance, n_l is the samples in the lth cluster, C_l^j is the j th sample in the lth cluster. The ISODATA algorithm has some further refinements by splitting and merging of clusters (JENSEN, 1996). Clusters are merged either when the number of members (gesture sequence) in a cluster is less than a certain threshold or when the centers of two clusters are closer than a certain threshold. Clusters are split into two different clusters if the cluster standard deviation exceeds a predefined value and the number of members (gesture sequencexs) is twice as much as the threshold for the minimum number of members.

We define T as the threshold on the number of samples in clustering and N_c as the approximate number of clusters, and denote S^2 as maximum spread for parameter splitting. D_m is the maximum distance separation for merging and $N_{\max}$ is the maximum number of clusters that can be merged at each step. The details of the algorithm are now presented as DTW/ISODATA algorithm:

1. Compute the pair wise dissimilarity in the training set based on DTW algorithm.

2. Choose an initial set of N_c cluster centers and partition the gesture sequences into clusters using the current cluster centers with the minimum distance assignment procedure. If any cluster has less than T members, decrease N_c and recluster. Continue this process until all clusters have T members.

3. Split Clusters. If $N_c<2N_D$ and iteration is odd then split any cluster whose samples form sufficiently disjoint groups according to the following rule:

Criterion for splitting: Compute the average distance d_k for samples in each cluster to their cluster and a measure of spread defined as follows

$$d_k = \frac{1}{N_k}\sum_{l=1}^{N_k} d(C_k^l, C_k^{p(k)}) \quad (2)$$

$$\sigma_k^2 = \frac{1}{N_k}\sum_{l=1}^{N_k} d^2\left(C_k^l, C_k^{p(l)}\right) \quad (3)$$

Let d_{avg} be the weighted average distance given by

$d_{avg} = \sum_{k=1}^{N_c} N_k d_k$. If for any $\sigma_k^2 > S^2$ then the cluster is split provided either $d_k > d$ And $N_k > 2(T+1)$ or $N_c < N_D/2$ The original cluster center is replaced by two new centers displaced slightly along the axis of largest variance.

Criterion for Merging: Compute the pair wise distances d_{ij} between cluster centers $d_{ij} = d(\mu_i, \mu_j)$. If any pairs of distances correspond to a distance less than the threshold D_m , then the pairs are merged. The two clusters P_i and P_j represented by smallest d_{ij} are combined eliminating P_i and P_j from further merging. The next smallest distance between two clusters, neither of which is cluster P_i nor P_j will also cause a merger of the two respective clusters. This process is continued until the number of merges equal $N_{\max}$ or there are no more to merge whichever occurs first.

4. Go to Loop.

IV. RESULT AND DISCUSSION

We use two Cyber Gloves and three3SPACE-position tracker as input devices. Two trackers are positioned on the wrist of each hand and another is fixed at back. The Cyber Gloves collect the variation information of hand shapes with the 18-dimensional data at each hand, and the position trackers collect the variation information of orientation, position, movement trajectory.

The data are collected from 7 signers with each signer performing 4942 isolated words twice. The words of vocabulary are chosen from sign language dictionary of china. We select 4 from 7 signers as the registered signers. The rest are referred to as the unregistered signers. Each of the registered signers contributes two groups of data as training samples; the samples from the unregistered signers are referred to as the unregistered test set.

Table 1. The comparison of different results

signer	Recognition rate in %		Recognition speed in second	
	HMM	Multilayer architecture	HMM	Multilayer architecture
ljh	89.48	93.83	2.369	0.137
llq	86.26	91.23	2.366	0.121
mwh	90.45	95.11	2.358	0.152
Average	88.73	93.39	2.364	0.137

Table 1 reports respectively test results of HMMs and method of multilayer architecture, where HMMs have 3 states and 5 mixture components. 88.73% and 93.39% of mean recognition rates are respectively observed A exciting performance of processing time can be seen in the table, the average recognition speed have increased distinctly under the multilayer architecture.

V. CONCLUSION

This paper presents the Multilayer architecture in Sign Language Recognition for the signer-independent CSL recognition, where classical DTW and HMM is combined

within an initiative scheme. In the two-stage hierarchy we define the confusion sets and introduce DTW/ISODATA algorithm as the solution to build confusion sets in the vocabulary space. The experiments show that the Multilayer architecture in Sign Language Recognition increases the average recognition speed by 94.2% and the recognition accuracy 4.66% more than the HMM-based recognition method.

REFERENCES

[1] S. S. Fels and G. Hinton, "Glove Talk: A neural network interface between a DataGlove and a speech synthesizaer", IEEE Transactions on Neural Networks, 1993, Vol. 4, pp.2-8.

[2] M. W. Kadous, "Machine recognition of Auslan signs using PowerGlove: Towards large-lexicon recognition of sign language", proceeding of workshop on the Integration of Gesture in Language and Speech, Wilmington, DE, 1996, pp.165-174.

[3] C. Vogler, D. Metaxas, "Toward Scalability in ASL Recognition: Breaking Down Signs into Phonemes", In Proceedings of Gesture Workshop, Gif-sur-Yvette, France,1999, pp. 400-404.

[4] R.H. Liang, M. Ouhyoung, "A Real-time Continuous Gesture Recognition System for Sign Language", In Proceeding of the Third International Conference on Automatic Face and Gesture Recognition, Nara, Japan, 1998, pp. 558-565.

[5] G.L. Fang, W. Gao, "A SOFM/HMM System for Person-Independent Isolated Sign Language Recognition", INTERACT2001 Eight IFIP TC.13 Conference on Human-Computer Interaction, Tokyo, Japan, 2001, pp.731-732.

[6] H. Sakoe and S. Chiba, "A dynamic programming approach to continous speech recognition," in *Proc. Int. Congress on Acoustics*, Budapest, Hungary, 1971, 20 C-13.

[7] B.-H. Juang, "On the hidden Markov model and dynamic time warping for speech recognition – A unified view," AT&T Bell labs. Tech. J. , vol. 63, no. 7, pp. 1213-1243, sept. 1984.

[8] J.S.Bridle, "Stochastic models and template matching: Some important relationships between two apparently different techniques for automatic speech recognition," in proc. Inst. Acout., vol 6, 1984, pp. 452a-452h.

[9] R. Stapert, "A Segmental Mixture Model," Ph.D., Univ.Wales, Swansea,U.K., 2000.

[10] D. Hall and G. Ball. Isodata: a novel method of data analysis and pattern classification. Technical report, Stanford Research Institute, 1965.