Large-Vocabulary Continuous Sign Language Recognition Based on Transition-Movement Models

Gaolin Fang, Wen Gao, *Member, IEEE*, and Debin Zhao

***Abstract*—The major challenges that sign language recognition (SLR) now faces are developing methods that solve large-vocabulary continuous sign problems. In this paper, transition-movement models (TMMs) are proposed to handle transition parts between two adjacent signs in large-vocabulary continuous SLR. For tackling mass transition movements arisen from a large vocabulary size, a temporal clustering algorithm improved from k-means by using dynamic time warping as its distance measure is proposed to dynamically cluster them; then, an iterative segmentation algorithm for automatically segmenting transition parts from continuous sentences and training these TMMs through a bootstrap process is presented. The clustered TMMs due to their excellent generalization are very suitable for large-vocabulary continuous SLR. Lastly, TMMs together with sign models are viewed as candidates of the Viterbi search algorithm for recognizing continuous sign language. Experiments demonstrate that continuous SLR based on TMMs has good performance over a large vocabulary of 5113 Chinese signs and obtains an average accuracy of 91.9%.**

***Index Terms*—Chinese sign language (CSL), dynamic time warping (DTW), hidden Markov model (HMM), sign language recognition (SLR), temporal clustering algorithm.**

I. Introduction

With the widespread use of computers in modern society, traditional human–computer interaction (HCI) technologies based on mouse and keyboard show their increasing limitations. Thus, research on multimodal HCI is becoming more and more important in real life. Sign language recognition (SLR), as one of the important research areas of HCI, has spawned more and more interest in HCI society. The goal of SLR is to provide an efficient and accurate mechanism to transcribe sign language into text or speech so that "dialog communication" between the deaf and hearing society can come true. From a user's point of view, the most natural way to interact with a computer would be through a speech and gesture interface. Thus, research on sign language and gesture recognition is likely to provide a shift paradigm from point-and-click user interface to a natural language dialog-and-spoken command-based interface. In addition, it has many other applications such as providing a "speaking aid" for deaf-mute people by integrating SLR and speech synthesis modules into a digital glove, controlling the motion of a human avatar in a virtual environment via hand gesture recognition, and having a learning demonstration for the robot.

SLR is used to deal with the recognition problem of the temporal pattern of multiple data streams. From the point of view of handling a time-series signal, SLR is very similar to speech recognition. However, some differences between them make SLR more difficult. Compared with traditional speech recognition, which only deals with one stream of speech signal data, SLR has to handle multiple data streams including hand shape, position, orientation, and movement. In speech recognition, the basic unit of recognition is phoneme, which has been defined in the speech lexicon. However, sign language has no basic unit of recognition yet, and how to find and define basic units is an open issue. If we extract basic units from each stream, the number of combinative units is too large: about $75 \times 75 \times 12 \times 12 \times 15 \times 15$ (denoting the number of left-hand shapes, right-hand shapes, left positions, right positions, left orientations, and right orientations, respectively). Thus, we must propose the corresponding algorithm according to the characteristic of SLR rather than the simple use of speech recognition methods.

The major challenges that SLR now faces are developing methods that will solve large-vocabulary continuous sign problems. Research on large-vocabulary continuous SLR has a profound influence on the naturalness of human–computer interfaces and is clearly an essential requirement for the widespread use of an SLR system. For continuous SLR, the main issue is how to handle the movement epenthesis. The movement epentheses, i.e., transition movements between two adjacent signs, begin at end of the preceding sign and finish at the start of the following sign, which vary with the sign contexts. The presence of movement epenthesis greatly complicates the recognition problem since it inserts a great variety of extra movements that are not present in the signs' lexical forms, instead of merely affecting the performance of adjacent signs.

In continuous speech recognition, context-dependent models such as biphone or triphone are generally employed for

Manuscript received December 13, 2004; revised April 19, 2005 and July 17, 2005. This work was supported in part by the Natural Science Foundation of China under Grant 60303018 and in part by the National High-Technology Development "863" Program of China under Grant 2001AA114160. This paper was recommended by Associate Editor D. Zhang.

G. Fang was with the Department of Computer Science, Harbin Institute of Technology, Harbin 150001, China. He is now with Fujitsu R&D Center, Beijing 100016, China (e-mail: glfang@cn.fujitsu.com).

W. Gao is with the Institute of Computing Technology, Beijing 100080, China (e-mail: wgao@ict.ac.cn).

D. Zhao is with the Department of Computer Science, Harbin Institute of Technology, Harbin 150001, China (e-mail: dbzhao@jdl.ac.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TSMCA.2006.886347

modeling the coarticulation. However, in continuous SLR, no basic unit such as the phoneme of speech is defined in the sign lexicon yet. The number of subunits for the whole sign language extracted manually or automatically is so large that the training data becomes very sparse [1]. This leads to the impossibility to train context-dependent models such as those in the literature [2] for large-vocabulary SLR. Direct models of the movement epenthesis between signs also have the same problem as context-dependent models.

However, transition movements are only related with the end of the preceding sign and the start of the following sign, so transition-movement models (TMMs) in terms of signs have many identical and very similar clusters. Thus, we can cluster transition movements so as to reduce their number and avoid the sparseness of training data. This also improves the generalization of transition movements, which is very suitable for large-vocabulary continuous SLR, with certain training samples of typical sentences. Nevertheless, transition movements are time-series vectors. The k-means clustering algorithm cannot handle temporal data because its distance measure builds between two spatial vectors. Volger and Metaxas [2] employed k-means to cluster the data frames of the start and end points of signs to produce less possible (end–start) combinative transition models. However, in a large vocabulary size, it is very difficult to obtain these data frames of start and end points. If we manually segment continuous signs into isolated signs, the huge workload and the introduction of man-made errors will make the task infeasible. Furthermore, it is also complicated to model the start and end points of movement epenthesis through isolated signs because there are distinct performance deviations between the isolated sign and the sign in a continuous sentence.

In this paper, TMMs are proposed to deal with transition parts between two adjacent signs. For tackling mass transition movements arisen from a large vocabulary size, a temporal clustering algorithm improved from k-means by using dynamic time warping (DTW) as its distance measure is proposed to dynamically cluster them; then, an iterative segmentation algorithm is presented for automatically segmenting transition parts from continuous sentences and training these TMMs through a bootstrap process. The estimated TMMs, together with sign models, are used for continuous SLR. Experiments demonstrate that continuous SLR based on TMMs has good performance on a large vocabulary.

The remainder of this paper is organized as follows. Section II reviews the related work. In Section III, we give the SLR system overview. In Section IV, the temporal clustering algorithm is proposed to dynamically cluster transition movements. Section V gives large-vocabulary SLR based on TMMs. Section VI shows the experimental results. The conclusions are drawn in the last section.

II. Related Work

Attempts to automatically recognize sign language began to appear in the literature in the 1990s. Following the similar path to early speech recognition, many previous attempts at SLR focused on finger spelling or isolated signs. Because finger spelling is a static gesture, the recognition algorithms are not very similar to isolated signs and continuous signs (belong to dynamic gestures). In this overview, we focus on previous related work on isolated and continuous SLR. The recognition methods of isolated SLR usually include rule-based matching [3], [4], fuzzy decision trees [5], artificial neural networks [6]–[9], and hidden Markov models (HMMs) [10], [11]. However, because there is no clear pause between the signs for continuous SLR, explicit segmentation of a continuous input stream into individual signs becomes intractable. For this reason, together with the effect of movement epenthesis, the research on isolated sign recognition often does not easily generalize to continuous SLR. For recognizing continuous sign language, several typical works here are performed.

Starner *et al.* [12] proposed a view-based approach for continuous American SLR. They used a single camera to extract two-dimensional features as input of the HMM. Word accuracies of 92% and 98% were obtained when the camera was mounted on the desk and in a user's cap, respectively, while recognizing the sentences over a vocabulary of 40 signs. An HMM was also employed by Bauer and Hienz [13] to recognize continuous German sign language with a single color video camera as input. An accuracy of 91.7% can be achieved during recognition of sign language sentences with 97 signs. Furthermore, they developed the k-means clustering algorithm to obtain the subunits for continuous SLR [14]. An accuracy of 80.8% was achieved in the corpus of 12 different signs with ten subunits. In large-vocabulary SLR, a direct HMM is difficult to use in modeling a variety of movement epentheses between signs arisen from a large vocabulary size.

Liang and Ouhyoung [15] employed the time-varying parameter threshold of hand postures to determine the end points in a stream of gesture input for continuous Taiwan SLR. An average recognition rate of 80.4% was obtained over a vocabulary of 250 signs. In their system, an HMM was employed, and a Dataglove was taken as input device. Sagawa and Takeuchi [16] used the changes of hand shape, orientation, and position to detect the borders of Japanese sign language words. They experimented ten sentences and obtained 83.0% accuracy with the top five choices. However, the fixed segmentation will result in a higher false recognition rate. Fang and Gao [17] proposed simple recurrent networks/HMMs (SRNs/HMMs) for signer-independent continuous SLR, where SRN is used as soft segmentation of continuous sign language. The system obtained 85% accuracy in recognizing 100 sentences from seven signers on a vocabulary of 208 signs. A critical issue in employing the segmentation strategy for continuous SLR in a large vocabulary size lies in finding an effective soft segmentation method and a whole framework to incorporate this method.

Vogler and Metaxas [2] used computer vision methods to extract the three-dimensional (3-D) parameters of a signer's arm's motions as input to the HMM and recognized continuous American sign language sentences over a vocabulary of 53 signs. They built a context-dependent HMM and then modeled transition movements to alleviate the effects of movement epenthesis. Experiments over the 64 phonemes extracted from the 53 signs showed that modeling movement epenthesis showed better performance than using context-dependent HMMs. The reported best accuracy is 95.8%. In addition, they

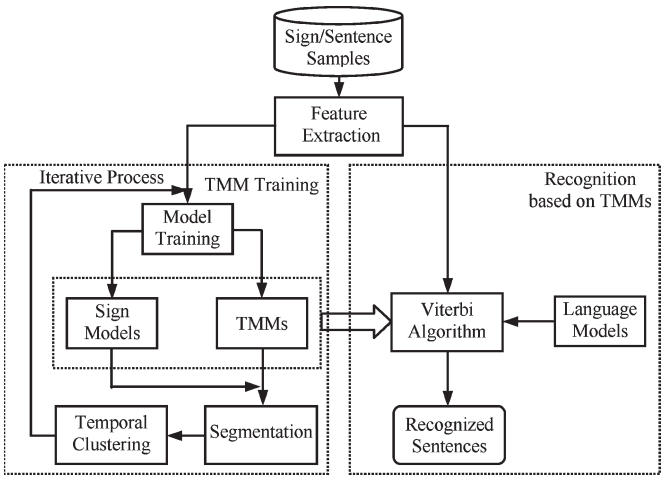


Fig. 1. Structure of a continuous SLR system based on TMMs.

used phonemes instead of whole signs as the basic units and achieved similar recognition rates as sign-based approaches over a vocabulary of 22 signs [18], [19].

Gao *et al.* [20], [21] used a dynamic programming method to obtain context-dependent models for recognizing continuous Chinese sign language (CSL). Datagloves were used as input devices, and a state-tying HMM was used as the recognition method. Their system can recognize 5177 CSL isolated signs with 94.8% accuracy in real time and recognize 200 sentences with 91.4% word accuracy. More recently, this system has been improved and expanded to signer-independent SLR [22]–[24].

Previous research on SLR primarily focuses on small- or medium-vocabulary continuous SLR. For large-vocabulary continuous SLR, to the best of our knowledge, no research report was found in the literature, except our early work in which TMMs were proposed for continuous Chinese SLR [25].

III. SLR System Overview

A. System Structure

The structure of a continuous SLR system based on TMMs is shown in Fig. 1. Sign/sentence samples collected by input devices are fed into the feature extraction module and then input into two related parts: TMM training and recognition based on TMMs. In the TMM training part, sign/sentence samples are trained into sign models and TMMs by the model training module (no TMMs in the first run). Then, these models are used to segment continuous sentence samples into sign parts and transition parts. Transition parts are clustered using the temporal clustering algorithm. We iterate this process until the convergence criterion is met. In the recognition part based on TMMs, the estimated TMMs and sign models obtained from the training part are viewed as candidates of the Viterbi search algorithm, together with language models (Bigram) for recognizing large-vocabulary continuous sign language.

B. CSL and Its Feature Extraction

Sign language, as a kind of gesture, is one of the most natural ways of exchanging information for most deaf people. CSL, as a kind of sign language, is the language of choice for 20.57 million deaf people in China. CSL consists of about 5500 conventional vocabularies, including postures and gestures. With the evolution of CSL, up-to-date CSL can express any meaning in natural spoken Chinese with the aid of finger spelling. CSL has the following unique features: 1) CSL is a kind of language using semantics as a main expressive way, and its sentence has a similar word order to Chinese written language and 2) finger spelling and Chinese character-imitating gestures play two very important parts of CSL. Similar to Stoke's analysis of American sign language [26], each Chinese sign can be broken into four parameters: hand shape, position, orientation, and movement. Hand shapes are one of the primitives of CSL and reflect information on the hand configuration. They are very stable and can be used to distinguish most signs. There are 75 basic hand shapes in CSL. The position of the hand is usually partitioned in terms of the signer's hand with respect to the defined three parts of his body: head, chest, and below chest. In each part, the position can be further subdivided into the body's left, right, and middle. In total, there are 12 positions defined in CSL. The orientation of the hand can be described in terms of two orthogonal directions: the facing of the palm and the direction to which the hand is pointing. There are 15 different orientations widely used in CSL. Movement differs from the other features because it is inherently temporal in nature. Its trajectory, which indicates the shape of an object, is described by time-series variation of hand position relative to the body part. These parameters are performed simultaneously and form multiple data streams, which are the basis of SLR.

Hand shapes are one of the primitives of sign language and reflect information on the hand configuration. To more accurately collect the variation information of hand shape and finger status, two Cybergloves are employed with the 18-dimensional data for each hand (see Fig. 2).

To collect the variation information of orientation and position, three Pohelmus 3SPACE-position trackers are used (see Fig. 2). However, the outputs of trackers cannot be directly used as sign language features because they vary with the position of the transmitter, especially in the situation when the recognition system is moved from one place to another. In order to extract invariant features to the signer's position, the following method is proposed. First, two trackers are positioned on the wrist of each hand, and another is mounted at the signer's back. The tracker at the signer's back is chosen as the reference Cartesian coordinate system. In addition, the position and orientation of each hand with respect to the reference system are calculated and can be taken as invariant features. By this transformation, the data consist of a relative 3-D position vector and a 3-D orientation vector for each hand, which do not change with the signer position and orientation.

In total, a 48-dimensional vector is formed, including the hand shape (36), position (6), and orientation vector (6) for the two hands. The data from different signers are calibrated by some fixed movements performed by each signer. In our experiment, the 14 postures that represent the min–max value ranges of the corresponding sensor and 75 basic hand shapes are defined. As each component in the vector has a different dynamic range, its value is normalized to [0, 1].

Fig. 2. Input devices.

IV. Temporal Clustering

Since the transition movements of continuous sign language are time-series vectors, the proposed clustering algorithm is required to handle not only spatial information but also temporal information. Furthermore, there is no criterion to describe how many clusters are reasonable, so the algorithm should be able to dynamically cluster transition movements according to their data distribution.

The k-means clustering algorithm cannot handle temporal data because its distance measure only builds between two spatial vectors. Wilpon and Rabiner [27] proposed a modified k-means algorithm (MKM) for producing robust matching templates for speaker-independent speech recognition. However, MKM cannot dynamically cluster the data. In this paper, a temporal clustering algorithm based on MKM is proposed to cluster time-series vectors. DTW is employed as the distance computation criterion because it can measure the distance between two temporal sequences by aligning different time signals and normalizing them to a warping function. In the algorithm, the corresponding skills are proposed to solve the issues of cluster splitting and combination. The proposed algorithm can automatically split and combine the centroids according to their data distribution to obtain a more reasonable cluster number and centers. The next subsection will discuss DTW-based distance computation and temporal clustering algorithm in detail.

A. DTW-Based Distance Computation

DTW is to search the best warping function using a dynamic programming technique so as to minimize the distance between two temporal sequences. Myers and Rabiner [28] have proposed several DTW algorithms for recognizing connected speech words and compared their performances. Let two temporal sequences be $X = (X_1, X_2, \ldots, X_{T_X})$ and $Y = (Y_1, Y_2, \ldots, Y_{T_Y})$, where X_i and Y_i are the 48-dimensional vectors. Define the warping function as $\phi = \{\phi(1), \phi(2), \ldots, \phi(N)\}$, $\phi(n) = (\phi_X(n), \phi_Y(n))$, where N is the "normal" duration of two sequences $\phi_X(n) \in \{1, \ldots, T_X\}$ and $\phi_Y(n) \in \{1, \ldots, T_Y\}$. The nth matching pair $\phi(n)$ consists of the $\phi_X(n)$ vector in X and the $\phi_Y(n)$ vector in Y.

The measure $d(\phi_X(n), \phi_Y(n))$ is defined as the Euclidean distance. The goal of DTW is to search the minimal accumulating distance and the associated warping path, i.e.,

$$D(X, Y) = \min_{\phi} \sum_{n=1}^{N} d\left(\phi_X(n), \phi_Y(n)\right). \tag{1}$$

The warping function in our experiment satisfies end-point constraint, monotony constraint, and one-step local continuity constraint. The one-step local constraint refers to the case in which, when the current warping function pair is (i, j), its last step has only three choices: $(i-1, j)$, $(i-1, j-1)$, and $(i, j-1)$. Unlike in speech recognition, we do not put any region constraint to the DTW search so as to get the best path among all the possible candidates.

The minimum partial accumulated distortion along a path from (1, 1) to (i_X, i_Y) is defined as

$$D(i_X, i_Y) = \min_{\phi T'} \sum_{n=1}^{T'} d\left(\phi_X(n), \phi_Y(n)\right) \tag{2}$$

where $\phi_X(T') = i_X$ and $\phi_Y(T') = i_Y$.

The auxiliary parameter $\psi(i_X, i_Y)$ is defined to record a point before point (i_X, i_Y) in the local optimal path. The recursive relations according to the constraints are given as follows:

$$D(i_X, i_Y) = \min_{(i'_X, i'_Y)} \left[D(i'_X, i'_Y) + d(i_X, i_Y)\right] \tag{3}$$

$$\psi(i_X, i_Y) = \arg\min_{(i'_X, i'_Y)} \left[D(i'_X, i'_Y) + d(i_X, i_Y)\right] \tag{4}$$

where $(i'_X, i'_Y) \in \{(i_X - 1, i_Y), (i_X - 1, i_Y - 1), (i_X, i_Y - 1)\}$.

Through the dynamic programming search, the minimal distance $D(X, Y)$ between two temporal sequences and the associated warping function pair ϕ are simultaneously obtained.

B. Temporal Clustering Algorithm

Let $\Pi = \{O_1, O_2, \ldots, O_V\}$ be a data set with V temporal sequences to be clustered. The temporal clustering algorithm is used to dynamically cluster these data into c centers, i.e.,

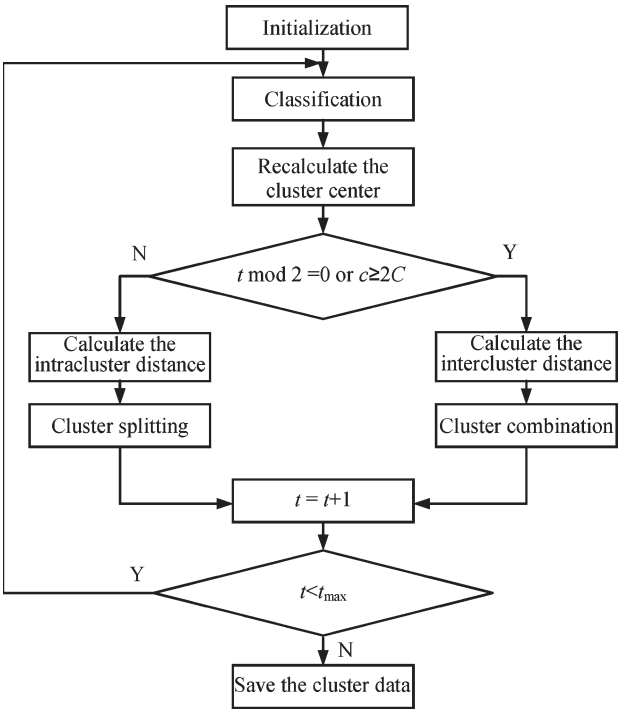


Fig. 3. Flowchart of the temporal clustering algorithm.

$\Pi = \cup_{j=1}^{c} \Gamma_j$. The flowchart of the temporal clustering algorithm is shown in Fig. 3.

The detailed algorithm is described as follows.

1) *Initialization:* Calculate all distances $d(O_i, O_j)$ using DTW. Set the initial parameters; c is the number of clusters, C is the expected number of clusters, θ_N is the minimum number of samples in each cluster, θ_C is the threshold of the intercluster distance that determines whether to combine or not, t is the number of iterations, and $t_{\max}$ is the maximum iterations. The method described in [27] is employed to set the initial cluster centers. It splits the clusters from one to the expected number by a one-to-two method step by step.
2) *Classification:* According to the minimum DTW distance rule, each sample is classified into the corresponding center.

   For each cluster, if its sample number is less than θ_N, this cluster is discarded. Set $c = c - 1$, and then reclassify the samples in this cluster.
3) *Recalculate the Cluster Center:* The recalculation is described by the following two steps.
   (a) First, find the pseudoaverage center O'. A particular element in the cluster has the largest population of elements (subset of the cluster) whose distance to the particular sample falls within the threshold. If several patterns have the same largest count of samples with distances below the threshold, then the element that has the smallest average distance to all samples in the subcluster is chosen as the pseudoaverage center.
   (b) Second, all samples in Γ_j are warped to the pseudoaverage center O'. We then group the samples according to their individual warping paths with respect to O'. The vectors that are aligned to the same index i are then averaged to produce an average vector for the new cluster. The resultant sequence with vectors indexed from 1 to $T_{O'}$ (duration for O') is the average cluster center $m(\Gamma_j)$.
4) If $t \bmod 2 = 0$ or $c \geq 2C$, then GO TO step 6); else, GO TO step 5).
5) *Cluster Splitting:* Calculate the intracluster distance λ_j for each cluster j, i.e.,

$$\lambda_j = \frac{1}{\|\Gamma_j\|} \sum_{O \in \Gamma_j} d\left(m(\Gamma_j), O\right), \qquad j = 1, 2, \ldots, c. \tag{5}$$

   Find the cluster $\Gamma_{j\max}$ with the maximum intracluster distance. If $\|\Gamma_{j\max}\| \geq 2\theta_N$ or $c \leq C/2$, then split $\Gamma_{j\max}$ as follows: Search for two temporal sequences O_{p1} and O_{p2} that satisfy $d(O_{p1}, O_{p2}) \geq d(O_{p3}, O_{p4})$ for any other pairs O_{p3}, O_{p4} in $\Gamma_{j\max}$. Two sequences O_{p1} and O_{p2} are used as new cluster centers to replace the original cluster. Set $c = c + 1$, and then GO TO step 7).
6) *Cluster Combination:* For all the cluster centers, calculate the intercluster pairwise distances $d(m(\Gamma_i), m(\Gamma_j))$. Find the pair with the minimum interclass distance $d(m(\Gamma_p), m(\Gamma_q))$, if $d(m(\Gamma_p), m(\Gamma_q)) < \theta_C$. Then, combine Γ_p and Γ_q. Using DTW, the optimal path between the sequences Γ_p and Γ_q is obtained. Let T be the warping path length for ϕ, and the new cluster $\overline{m}$ is calculated as follows:

$$\overline{m}_k = \frac{1}{2}\left(m(\Gamma_p)_{\phi_X(k)} + m(\Gamma_q)_{\phi_Y(k)}\right), \qquad k = 1, 2, \ldots, T. \tag{6}$$

   Replace these two clusters with the new cluster $\overline{m}$, and set $c = c - 1$.
7) $t = t + 1$. If $t < t_{\max}$, then return to step 2). Otherwise, save the cluster data, and exit.

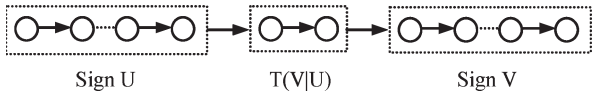


Fig. 4. TMM between two signs.

V. Large-Vocabulary SLR Based on TMMs

For continuous SLR, the main issue is how to handle the movement epenthesis. In fact, the method of modeling movement epenthesis can alleviate the effect of movement epenthesis. However, the number of all the possible combinations between signs is so large, especially in a large vocabulary size, that a large amount of continuous sentences is required to train these combinative models. There are no such large corpora in the SLR field at present. Furthermore, due to the lack of lexical definition in the sign lexicon for the movement epenthesis, it is difficult to model these movement epentheses using separately collected sign data. However, movement epentheses are usually related with the end of the preceding sign and the start of the following sign. In Fig. 4, Sign U and Sign V denote any two adjacent signs in continuous sign language, and T(V|U)

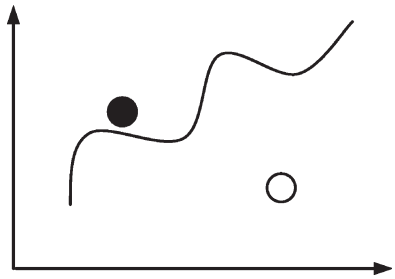


Fig. 5. Trajectory of a sign in the sign space.

represents the TMM from Sign U to Sign V. Different transition movements between two adjacent signs have identical and very similar (end–start) sequences. Thus, we can reasonably cluster them into one class. This will not only reduce the number of transition movements to avoid the sparseness of training data but also improve their generalization. These clustered models are very suitable for large-vocabulary SLR.

In continuous sign language, the start and end points of the corresponding signs cannot be known, so it is infeasible to segment transition movements through manual annotation. In this paper, the training algorithm of TMMs is proposed to automatically segment transition movements from continuous sentences and simultaneously estimate their model parameters.

The iterative segmentation algorithm of TMM training is described in detail as follows.

1) With initial isolated HMM models, continuous sign sentences with recorded text are segmented into the corresponding isolated sign sequence using automatic segmental clustering in HMM [29].
2) Set the transition parts from the last state of the preceding sign to the first state of the following sign as the initial values of current transition movements.
3) Cluster transition movements through the temporal clustering algorithm. Then, train the TMMs with the clustered transition data, and train the sign models with the segmented sign data and isolated sign data.
4) Use the newly trained models (TMMs and sign models) to segment continuous sentences into signs and the corresponding transition movements. Judge whether the number of transition frames has changed compared with the last segmentation. If it has changed, then return to step 3); otherwise, save the trained models, and exit.

After the training, the TMMs and new sign models are built. They are combined into the candidate models of the Viterbi search algorithm [29] for recognizing large-vocabulary continuous sign language. However, because the number of candidate models is so big, the pruning operation has to be employed to improve system performance.

Each sign has its trajectory in sign space. If an observation vector is close to this trajectory, then the sign may be active at that time; else, the sign will be inactive. Fig. 5 demonstrates a trajectory of a sign in the sign space, where a black dot represents an active sign and a white dot represents an inactive sign. For each observation vector, judging whether a sign is active is very important to speed up the recognition process. If only a small fraction of signs is active at the current frame, the most likely active signs are those which are active at the previous frame due to the continuous property of gesture movement trajectory. Only these active signs need to be further searched at the next frame; thus, a large amount of computation cost can be saved.

According to the preceding analyses, the rules of adding candidate words and removing candidate words are made during the search process. The details are described as follows.

1) **Adding candidate words:** Calculate the first state probability of all the words, excluding the candidates of the last frame. If the probability of a word is greater than a certain threshold, this word will become a candidate of the current frame. At the same time, the other state probabilities of this word need not be further calculated at the current frame.
2) **Removing candidate words:** For all the candidates of the last frame, if all the state path scores of a word are less than a certain threshold, then this word is removed from the current candidates, and the paths with this frame as the tail will not be further expanded.

VI. Experiments and Discussions

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

Experimental data consist of 51130 sign samples over 5113 isolated signs from two signers, with each sign having ten samples. The vocabulary is taken from the CSL dictionary. Eight samples are used as the training set, and the rest of the samples of each signer are used as the isolated sign test set. The continuous sign language database consists of the 3000 sentence samples with 750 different sentences over a vocabulary of 5113 signs. These data are collected from two signers represented by S_1 and S_2, with each performing the sentences twice. The sentences are extracted from the 200-MB corpora, which are composed of China Daily between the years 1993 and 1995 and the Family Collection Book. The 200-MB corpora are also used to estimate Bigram probabilities, where Bigram is adopted as language models in our continuous SLR framework. As sign language is somewhat different from natural language, e.g., the function words are always omitted and sometimes the subject and the predicate are hyperbatic, some adaptations to these linguistic characteristics are imposed on the training corpora of the language models.

The first experiment validates that the proposed temporal clustering algorithm can effectively cluster similar or the same sequences into one class. The database consists of 1268 samples from 317 signs, which are selected among 5113 signs at random, and each sign has four samples. Because the corresponding classes are known beforehand, we can judge whether the clustering results are reasonable. The expected cluster center is initially set to 317. After the processing of the temporal

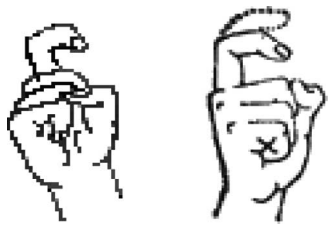


Fig. 6. Description of the signs J and Ninety. (Left) J. (Right) Ninety.

clustering algorithm, the 309 cluster centers can be obtained. The 301 centers are the same as the sign data, i.e., each has four samples. The remaining eight centers are the results of the sample combinations of the two signs.

In the eight centers, they can be classified into three categories. One is that the two signs have the same action, such as zhu-ren (director) and zhu-chi (preside). The second is that the two signs have the same posture but only a small difference in position, such as zhong-zu (race or tribe) and zhong-lei (category). The third is that the two signs have the very similar posture, where one has slight movement and the other has not. For example: "J" and "jiu-shi" ("ninety"), where the sign J is static and the sign ninety has slight movement of first finger. Fig. 6 shows the description of the signs J and ninety.

From the preceding experiments, we know that the temporal clustering algorithm can effectively cluster these segments with high similarity into one class.

The second experiment is to analyze the factors influencing the isolated sign accuracy. This is because the parameters of isolated sign models, as a basis of continuous SLR, have direct influence on the continuous SLR performance. Only when these signs are correctly modeled is it possible for them to be recognized in the continuous sentence. There are two factors that directly influence the recognition accuracy: the number of states N and the number of mixture components M in the HMM. In the HMM training, these two factors need to be manually set through experiments, and other parameters of models can be automatically trained through reestimation. N depends on the number of potential phonemes of the sign, where phoneme is defined as a segment dynamic continuous sign data where the variability of hand shape, position, and orientation is very stable. The value of M is determined by the distribution of sign data. It reflects the differences of training data for one sign, and if the data difference is larger, the value of M is becoming bigger. To obtain the best parameters for HMM, we perform different experiments when N is set to 2, 3, 4, and 5 and M is set to 1, 2, 3, 4, 5, and 6, respectively.

As shown in Fig. 7, the best accuracy of 95.4% can be obtained on 5113 isolated signs when $M = 3$ and $N = 3$. When M grows from 1 to 3, the recognition performance is also improved. However, if M increases from 3 to 6, the recognition rate remains similar or even slightly decreases. Thus, $M = 3$ is viewed as the best number of mixture components. Though $N = 5$ and $N = 3$ have the comparative accuracy from Fig. 7, $N = 3$ is chosen because of its less computational complexity in recognition.

The third experiment is used to evaluate the performance of TMMs for continuous SLR. The 3000 samples are divided into

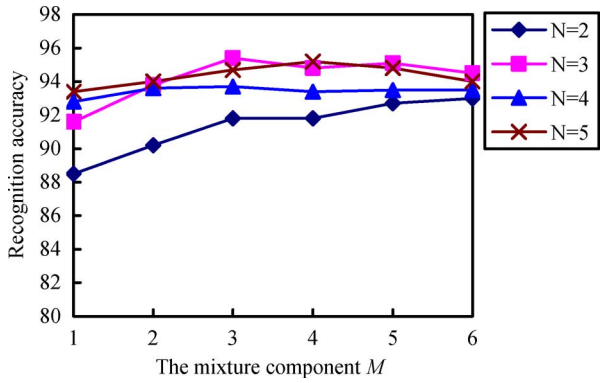


Fig. 7. Relations between the model parameters M and N and the isolated sign recognition accuracy.

TABLE I
PERFORMANCES OF CONTINUOUS SLR BASED ON TMMs

Signer	Accuracy	Recognition time(s/w)
S_1	90.8% (S=279, I=53, D=127, N=4994)	1.29
S_2	93.0%(S=234, I=30, D=85, N=4994)	1.25
Average	91.9%	1.27

two groups, with one group per signer (S_1, S_2). Among the 1500 sentence samples of each group, 750 sentences are used as training, and another 750 samples are used as the test set. These sentences consist of the words from 3 to 15, with an average of 6.6 words for each sentence.

All experiments are performed with Bigram language models on the PIV1600 (512-MB memory) personal computer. S, I, and D denote the number of substitution, insertion, and deletion errors, respectively. The number of signs in the whole test set is 4994, and the number of transition movements without clustering is 3945. The expected cluster center is initially set to 800. After the processing of the temporal clustering algorithm, the 546 cluster centers can be obtained. The candidates for recognition consist of 546 clustered TMMs and 5113 sign models, where their models are three states and three mixture components. Table I shows that an average accuracy of 91.9% for TMMs is obtained on the test set. In the two test sets, the insertion errors are $I = 53$ and $I = 30$, which is the smallest proportion of all errors, respectively. Thus, we know that the proposed method can effectively alleviate the effect of movement epenthesis. Experiments also demonstrate that when signers perform the sign sentence with natural speed, continuous SLR based on TMMs can be performed in real time without clear delay, which is about 1.27 s per word (s/w).

Table II demonstrates the comparison of different methods in continuous SLR. Compared with context-dependent models, our system has a better recognition rate of 94.3% on the test set of 200 different sentences, which then increases by about 2.9%. On the test set over a large vocabulary of 5113 signs, the experiments with TMM and without TMM are performed, where the model without TMM is identical to direct HMM. The accuracies of HMM and TMM are 78.2% and 91.9%, respectively, on the test set of 750 different sentences with a large vocabulary of 5113 signs. From their comparison, we

TABLE II
COMPARISON OF DIFFERENT METHODS IN CONTINUOUS SLR

Methods	Accuracy	The number of different sentences
Context-dependent models	91.4%	200
TMM	94.3%	200
HMM	78.2%	750
TMM	91.9%	750

know that continuous SLR based on TMMs is better than direct HMM. The reason is that HMM by segmenting transition parts into two adjacent signs is difficult to use in modeling various transition movements with great variations arisen from a large vocabulary size and TMM by clustering transition models can effectively solve this issue. Through TMM, we can solve the transition movement issue in a large vocabulary size, which direct HMM and context-dependent model cannot deal with, while retaining comparable time complexity. Because the clustered models improve TMM generalization, the proposed model can scale well with the vocabulary size for recognizing a larger vocabulary sign language.

VII. CONCLUSION

In this paper, continuous SLR based on TMMs is first implemented on a large vocabulary of 5113 signs. For tackling mass transition movements that have arisen from a large vocabulary size, a temporal clustering algorithm is proposed to dynamically cluster them; then, an iterative segmentation algorithm for automatically segmenting transition parts from continuous sentences and training these TMMs through a bootstrap process is presented. The clustered models can improve the TMM generalization and is very suitable for large-vocabulary continuous SLR with certain training samples of typical sentences. Experimental results demonstrate that continuous SLR has an average accuracy of 91.9% on 1500 test sentence samples over a large vocabulary of 5113 signs. Furthermore, the temporal clustering algorithm can be further extended to extract the basic units from CSL and automatically seek the anonymous gestures.

ACKNOWLEDGMENT

The authors would like to thank the editors and the anonymous reviewers, whose invaluable comments and suggestions led to a greatly improved manuscript.

REFERENCES

[1] C. L. Wang, W. Gao, and S. G. Shan, "An approach based on phonemes to large vocabulary Chinese sign language recognition," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit.*, 2002, pp. 411–416.

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

[3] M. W. Kadous, "Machine recognition of Auslan signs using Powergloves: Towards large-lexicon recognition of sign language," in *Proc. Workshop Integration Gesture Language Speech*, 1996, pp. 165–174.

[4] H. Matsuo, S. Igi, S. Lu, Y. Nagashima, Y. Takata, and T. Teshima, "The recognition algorithm with non-contact for Japanese sign language using morphological analysis," in *Proc. Int. Gesture Workshop*, 1997, pp. 273–284.

[5] G. L. Fang, W. Gao, and D. B. Zhao, "Large vocabulary sign language recognition based on fuzzy decision trees," *IEEE Trans. Syst., Man, Cybern. A, Syst., Humans*, vol. 34, no. 3, pp. 305–314, May 2004.

[6] S. S. Fels and G. E. Hinton, "Glove-talk: A neural network interface between a data-glove and a speech synthesizer," *IEEE Trans. Neural Netw.*, vol. 4, no. 1, pp. 2–8, Jan. 1993.

[7] J. S. Kim, W. Jang, and Z. Bien, "A dynamic gesture recognition system for the Korean sign language (KSL)," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 26, no. 2, pp. 354–359, Apr. 1996.

[8] P. Vamplew and A. Adams, "Recognition of sign language gestures using neural networks," *Aust. J. Intell. Inf. Process. Syst.*, vol. 5, no. 2, pp. 94–102, 1998.

[9] M. B. Waldron and S. Kim, "Isolated ASL sign recognition system for deaf persons," *IEEE Trans. Rehabil. Eng.*, vol. 3, no. 3, pp. 261–271, Sep. 1995.

[10] K. Grobel and M. Assan, "Isolated sign language recognition using hidden Markov models," in *Proc. Int. Conf. Syst., Man and Cybern.*, 1997, pp. 162–167.

[11] M. Assan and K. Grobel, "Video-based sign language recognition using hidden Markov models," in *Proc. Int. Gesture Workshop*, 1997, pp. 97–109.

[12] 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, no. 12, pp. 1371–1375, Dec. 1998.

[13] B. Bauer and H. Hienz, "Relevant features for video-based continuous sign language recognition," in *Proc. 4th Int. Conf. Autom. Face Gesture Recognit.*, 2000, pp. 440–445.

[14] B. Bauer and K. F. Kraiss, "Towards an automatic sign language recognition system using subunits," in *Proc. Int. Gesture Workshop*, 2001, pp. 64–75.

[15] R. H. Liang and M. Ouhyoung, "A real-time continuous gesture recognition system for sign language," in *Proc. 3rd Int. Conf. Autom. Face Gesture Recognit.*, 1998, pp. 558–565.

[16] H. Sagawa and M. Takeuchi, "A method for recognizing a sequence of sign language words represented in a Japanese sign language sentence," in *Proc. 4th Int. Conf. Autom. Face Gesture Recognit.*, 2000, pp. 434–439.

[17] G. L. Fang and W. Gao, "A SRN/HMM system for signer-independent continuous sign language recognition," in *Proc. 5th IEEE Int. Conf. Autom. Face Gesture Recognit.*, 2002, pp. 312–317.

[18] C. Vogler and D. Metaxas, "Toward scalability in ASL recognition: Breaking down signs into phonemes," in *Proc. Int. Gesture Workshop*, 1999, pp. 400–404.

[19] ——, "A framework for recognizing the simultaneous aspects of American sign language," *Comput. Vis. Image Underst.*, vol. 81, no. 3, pp. 358–384, Mar. 2001.

[20] W. Gao, J. Y. Ma, J. Q. Wu, and C. L. Wang, "Sign language recognition based on HMM/ANN/DP," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 14, no. 5, pp. 587–602, 2000.

[21] W. Gao, J. Y. Ma, X. L. Chen *et al.*, "HandTalker: A multimodal dialog system using sign language and 3-D virtual human," in *Proc. 3rd Int. Conf. Multimodal Interface*, 2000, pp. 564–571.

[22] G. L. Fang, W. Gao, and D. B. Zhao, "Large vocabulary sign language recognition based on hierarchical decision trees," in *Proc. 5th Int. Conf. Multimodal Interface*, 2003, pp. 125–131.

[23] Y. Q. Chen, W. Gao, G. L. Fang, and Z. Q. Wang, "CSLDS: Chinese sign language dialog system," in *Proc. IEEE ICCV Int. Workshop Anal. Modeling Faces Gestures*, 2003, pp. 236–237.

[24] W. Gao, G. L. Fang, D. B. Zhao, and Y. Q. Chen, "A Chinese sign language recognition system based on SOFM/HMM/SRN," *Pattern Recognit.*, vol. 37, no. 12, pp. 2389–2402, Dec. 2004.

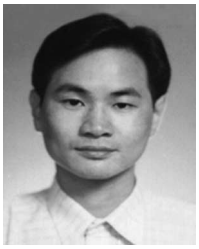
[25] ——, "Transition movement models for large vocabulary continuous sign language recognition," in *Proc. IEEE 6th Int. Conf. Autom. Face Gesture Recognit.*, 2004, pp. 553–558.

[26] W. C. Stokoe, *Sign Language Structure: An Outline of the Visual Communication System of the American Deaf. Studies in Linguistics: Occasional Papers 8 (Revised 1978).* Buffalo, NY: Linstok, 1960.

[27] J. G. Wilpon and L. R. Rabiner, "A modified K-means clustering algorithm for use in isolated word recognition," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 33, no. 3, pp. 587–594, Jun. 1985.

[28] C. S. Myers and L. R. Rabiner, "A comparative study of several dynamic time-warping algorithms for connected word recognition," *Bell Syst. Tech. J.*, vol. 60, no. 7, pp. 1389–1409, 1981.

[29] R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, no. 2, pp. 257–285, Feb. 1989.



Gaolin Fang received the M.S. and Ph.D. degrees in computer science from the Harbin Institute of Technology, Harbin, China, in 2000 and 2004, respectively.

From 2000 to 2004, he was a Research Assistant with the Joint R&D Lab, Chinese Academy of Sciences, Beijing, China. In 2003, he was a Visiting Research Assistant at Microsoft Research Asia, Beijing. Since 2004, he has been with the Fujitsu R&D Center, Beijing. He has published more than 20 scientific papers. His research interests include intelligent human–machine interaction, statistical language models, pattern recognition, and machine learning.

Wen Gao (S'88–M'99) received the M.S. and Ph.D. degrees in computer science from the Harbin Institute of Technology, Harbin, China, in 1985 and 1988, respectively, and the Ph.D. degree in electronics engineering from the University of Tokyo, Tokyo, Japan, in 1991.

In 1992, he was a Research Fellow with the Institute of Medical Electronics Engineering, University of Tokyo. In 1993, he was a Visiting Professor at the Robotics Institute, Carnegie Mellon University, Pittsburgh, PA. From 1994 to 1995, he was a Visiting Professor at the AI Laboratory, Massachussetts Institute of Technology, Cambridge. He is currently with the Institute of Computing Technology, Beijing, China. He is also the Vice President of the University of Science and Technology of China, Hefei; the Deputy President of the Graduate School of the Chinese Academy of Sciences, Beijing; a Professor of computer science with the Harbin Institute of Technology; and an Honor Professor of computer science with the City University of Hong Kong, Hong Kong. He is also the Head of the Chinese National Delegation to the MPEG Working Group (ISO/SC29/WG11). He has published seven books and more than 200 scientific papers. His research interests include signal processing, image and video communication, computer vision, and artificial intelligence. He is the Editor-in-Chief of the *Chinese Journal of Computers*.

Dr. Gao was the General Cochair of the IEEE International Conference on Multimodel Interface in 2002.

Debin Zhao received the B.S., M.S., and Ph.D. degrees in computer science from the Harbin Institute of Technology, Harbin, China, in 1985, 1988, and 1998, respectively.

He has been an Associate Professor with the Department of Computer Science, Harbin Institute of Technology and a Research Fellow with the Department of Computer Science, City University of Hong Kong, Kowloon, from 1989 to 1993. He is currently a Professor with the Department of Computer Science, Harbin Institute of Technology. He has authored or coauthored more than 50 publications. His research interests include data compression, image processing, and human–machine interface.