PERGAMON

Available at
www.ElsevierComputerScience.com
POWERED BY SCIENCE DIRECT

Pattern Recognition 37 (2004) 2389–2402

PATTERN RECOGNITION
THE JOURNAL OF THE PATTERN RECOGNITION SOCIETY

www.elsevier.com/locate/patcog

A Chinese sign language recognition system based on SOFM/SRN/HMM

Wen Gao[a,b], Gaolin Fang[a,*], Debin Zhao[a], Yiqiang Chen[b]

[a] *Department of Computer Science and Engineering, Harbin Institute of Technology, Harbin 150001, Republic of China*
[b] *Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100080, Republic of China*

Received 23 September 2003; accepted 27 April 2004

Abstract

In sign language recognition (SLR), the major challenges now are developing methods that solve signer-independent continuous sign problems. In this paper, SOFM/HMM is first presented for modeling signer-independent isolated signs. The proposed method uses the self-organizing feature maps (SOFM) as different signers' feature extractor for continuous hidden Markov models (HMM) so as to transform input signs into significant and low-dimensional representations that can be well modeled by the emission probabilities of HMM. Based on these isolated sign models, a SOFM/SRN/HMM model is then proposed for signer-independent continuous SLR. This model applies the improved simple recurrent network (SRN) to segment continuous sign language in terms of transformed SOFM representations, and the outputs of SRN are taken as the HMM states in which the lattice Viterbi algorithm is employed to search the best matched word sequence. Experimental results demonstrate that the proposed system has better performance compared with conventional HMM system and obtains a word recognition rate of 82.9% over a 5113-sign vocabulary and an accuracy of 86.3% for signer-independent continuous SLR.

Keywords: Sign language recognition; Chinese sign language; Hidden Markov model; Self-organizing feature map; Simple recurrent network

1. Introduction

Sign language as a kind of gestures is one of the most natural ways of exchanging information for most deaf people. The aim of sign language recognition is to provide an efficient and accurate mechanism to transcribe sign language into text or speech so that communication between deaf and hearing society can be more convenient. Sign language recognition (SLR), as one of the important research areas of human–computer interaction (HCI), has spawned more and more interest in HCI society. 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, the research on sign language and gesture recognition is likely to provide a shift paradigm from point-and-click user interface to a natural language dialogue-and-spoken command-based interface. In addition, it has many other applications, such as controlling the motion of a human avatar in a virtual environment via hand gesture recognition, learning demonstration for the robot, and multi-modal user interface in virtual reality system.

The major challenges that SLR faces now are developing methods that solve signer-independent continuous sign problems. Signer independence is highly desirable since it allows a system to be used straight out of the box and it allows the system to be built for the signer who is not known beforehand. The ability to recognize signer-independent and continuous sign language, without the introduction of artificial pause, has a profound influence on the naturalness of the human–computer interface. Therefore, their research is

* Corresponding author.
E-mail addresses: wgao@jdl.ac.cn (W. Gao), glfang@jdl.ac.cn (G. Fang), dbzhao@jdl.ac.cn (D. Zhao), yqchen@jdl.ac.cn (Y. Chen).

0031-3203/$30.00
doi:10.1016/j.patcog.2004.04.008

clearly an essential requirement for the widespread use of a SLR system.

For signer-independent SLR, there are two difficulties: (1) the model convergence difficulty caused by noticeable distinctions among different people signs. For a robust signer-independent recognition model, the training data must be collected from different signers. This makes the training data very massive. Since different people vary their hand shape size, body size, operation habit, rhythm, and so on, the noticeable distinctions between the data of the same sign due to different signers are almost larger than sign variations due to the change in the sign identity. (2) The lack of effective features extracted from different signers' data. Unlike speech recognition in which every speech feature has been profoundly explored, the research on the feature extraction of SLR is still in its infancy. How to effectively extract common features from different signers is a more challenging problem that needs to be solved.

For continuous SLR, the main issue is how to handle the movement epenthesis. The movement epentheses, i.e. transition movements between two signs, begin at the endpoint of the preceding sign and finish at the start of the following sign, which vary with 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 sign lexical forms, instead of merely affecting the performance of adjacent signs. In continuous speech recognition, context-dependent model such as biphone or triphone is generally employed for modeling the co-articulation. However, in continuous SLR no basic unit such as the phoneme of speech is defined in the sign dictionary yet. The number of subunits for the whole sign language extracted manually or automatically is so large that the training data become very sparse [1]. This leads to the impossibility to train the context-dependent models such as in Ref. [2] for overcoming the effect of movement epenthesis in large vocabulary SLR. Directly modeling the movement epenthesis [2] still has the same problem.

Two difficulties of signer-independent SLR lead to the fact that sign representations cannot be well modeled by conventional hidden Markov models (HMM). In this paper, SOFM/HMM is presented for modeling signer-independent isolated signs. The proposed method uses the self-organizing feature maps (SOFM) as a feature extractor for continuous HMM and its parameters are trained simultaneously in a global optimization criterion. SOFM transforms input signs into significant and low-dimensional representations that can be well modeled by the emission probabilities of HMM. Based on these isolated sign models, a SOFM/SRN/HMM model is proposed for signer-independent continuous SLR. This is because the segmentation method cannot only adapt well itself to scale with the increasing vocabulary size but also the solution strategy with the two-pass structure can effectively alleviate the effect of movement epenthesis. In the proposed method, the improved simple recurrent network (SRN) is used to segment continuous sign language in terms of transformed SOFM representations, and the outputs of SRN are taken as the HMM states in which the lattice Viterbi algorithm is employed to search the best matched word sequence. Experiments show that the proposed system has better performance than conventional HMM system.

The remainder of this paper is organized as follows. Section 2 reviews the related work. In Section 3, we give a SLR system overview. Section 4 presents SOFM/HMM for modeling signer-independent isolated signs. In Section 5, a SOFM/SRN/HMM model is proposed for signer-independent continuous SLR. Section 6 shows the experimental results and discussions. The conclusions and future work are given in the last section.

2. Related work

In this overview of related work, we focus on the previous work in sign language recognition. For details on gesture recognition, the survey in Refs. [3,4] gave a full review. Readers can also refer to Refs. [5–8] for more recent research.

Unlike general gestures, sign language is highly structured so that it provides an appealing test bed for new ideas and algorithms before they are applied to gesture recognition. Usually sign language recognition can be categorized into isolated SLR and continuous SLR and each can be further classified into signer-dependent and signer-independent according to the sensitivity to the signer.

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 sign language recognition focused on isolated signs. The recognition methods usually include rule-based matching, artificial neural networks, and hidden Markov models.

Kadous [9] demonstrated a system based on Powergloves to recognize a set of 95 isolated Australian sign languages with 80% accuracy. Instance-based learning and decision-tree learning were adopted by the system to produce the rules of pattern. Matsuo et al. [10] used the similar method to recognize 38 signs from Japanese sign language with a stereo camera for recording three-dimensional movements. Morphological analysis was used in their method to obtain sign language patterns.

Fels and Hinton [11] developed a system using a Dataglove with a Polhemus tracker as input devices. In their system, five neural networks were employed for classifying 203 signs. Kim et al. [12] used fuzzy min–max neural network and fuzzy logic approach to recognize 31 manual alphabets and 131 Korean signs based on Datagloves. An accuracy of 96.7% for manual alphabets and 94.3% for the sign words were reported. Waldron and Kim [13] also presented an expandable SLR system using the self-organizing maps to recognize a small set of isolated signs. They used Stokoe's transcription system to separate the hand shape, orientation and movement aspects of the signs.

Grobel and Assan [14] used HMM to recognize isolated signs with 91.3% accuracy out of a 262-sign vocabulary. They extracted 2D features from video recordings of signers wearing colored gloves. They also used HMM to recognize about 262 isolated signs from Netherlands sign language [15], with a single video camera as input device. Their system achieved recognition rates up to 94% for isolated signs.

For continuous SLR, as there is no clear pause between the individual signs, explicit segmentation of a continuous input stream into the individual signs becomes intractable. For this reason, together with the effect of movement epenthesis, work on isolated recognition often does not generalize easily to continuous sign recognition.

Starner et al. [16] used a view-based approach for continuous American SLR. They used a single camera to extract two-dimensional features as the input of HMM. The word accuracy of 92% or 98% was obtained when the camera was mounted on the desk or in an user's cap in recognizing the sentences with 40 different signs. HMM was also employed by Hienz and Bauer [17] to recognize continuous German sign language with a single color video camera as input. An accuracy of 91.7% can be achieved in recognition of sign language sentences with 97 signs. Furthermore, they developed the K-means clustering algorithm to get the subunits for continuous SLR [18]. The accuracy of 80.8% was achieved in the corpus of 12 different signs and 10 subunits. In large vocabulary sign language recognition, direct HMM is difficult to model a variety of movement epenthesis between signs.

Liang and Ouhyoung [19] employed the time-varying parameter threshold of hand posture to determine end-points in a stream of gesture input for continuous Taiwan SLR with the average recognition rate of 80.4% over 250 signs. In their system HMM was employed and a Dataglove was taken as input device. Sagawa and Takeuchi [20] used the changes of hand shape, orientation, and position to detect the borders of Japanese sign language words. They experimented 10 sentences and got 83.0% accuracy with top five choices. However, the fixed segmentation will result in the higher false recognition rate.

Vogler and Metaxas [2] used computer vision methods to extract the three-dimensional parameters of a signer's arm motions as the input of HMM, and recognized continuous American sign language sentences with a vocabulary of 53 signs. They, respectively, built context-dependent HMM and modeled transient movement to alleviate the effects of movement epenthesis. Experiments over 64 phonemes extracted from 53 signs showed that modeling the movement epenthesis has better performance than context-dependent HMM. The reported best accuracy is 95.83%. In addition, they used phonemes instead of whole signs as the basic units and achieved similar recognition rates to sign-based approaches over a vocabulary of 22 signs [21,22].

Gao et al. [23,24] used a dynamic programming method to obtain the context-dependent models for recognizing continuous Chinese sign language (CSL). Datagloves were used as input devices and state-tying HMM 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.

Previous research on sign language recognition focuses primarily on the signer-dependent field. There has been very little work reported on signer-independent sign language recognition. In fact, signer-independent SLR system has a promising perspective in practical applications as it can recognize a new signer's sign language without retraining the models. But the signer-independent recognition problem is very difficult to solve due to the great hurdles from a variety of sign variations among different signers.

To the best of our knowledge, there have been only two research works related to signer-independent isolated SLR. Vamplew and Adams [25] reported a signer-independent system based on a Cyberglove to recognize a set of 52 signs. Their system employed a modular architecture consisting of multiple feature-recognition neural networks and a nearest-neighbor classifier to recognize isolated signs. They got 94% recognition rate in the registered test set and 85% in the unregistered test set. We used the SOFM/HMM model to recognize signer-independent CSL over the 4368 samples from 7 signers with 208 isolated signs [26]. For signer-independent continuous SLR, even no research report was found in the literature except our early work in which the SRN/HMM model was applied for continuous Chinese SLR [27].

In this paper, SOFM/HMM is first presented for modeling signer-independent isolated signs. Based on these isolated sign models, a SOFM/SRN/HMM model is then proposed for signer-independent continuous SLR.

3. SLR system overview

3.1. System structure

The structure of SLR system based on SOFM/SRN/HMM is shown in Fig. 1. The sign/sentence samples collected by input devices are fed into the feature extraction module, and then, respectively, input into two related parts: SOFM/HMM based isolated SLR and SOFM/SRN/HMM based continuous SLR, where the SOFM/HMM training parameters in the isolated SLR are viewed as computing models of sign candidates in the lattice Viterbi of continuous SLR. In isolated SLR, it consists of SOFM/HMM training module and recognition module. In continuous SLR, the feature vector is first fed into SOFM, and then encoded to input into SRN. The outputs of SRN are recognized into the final sentences by HMM framework including lattice Viterbi and language models together with sign parameters of SOFM/HMM training.

3.2. Chinese sign language and its feature extraction

Chinese sign language (CSL) is the language of first-choice for 20.57 million deaf people in China. CSL consists

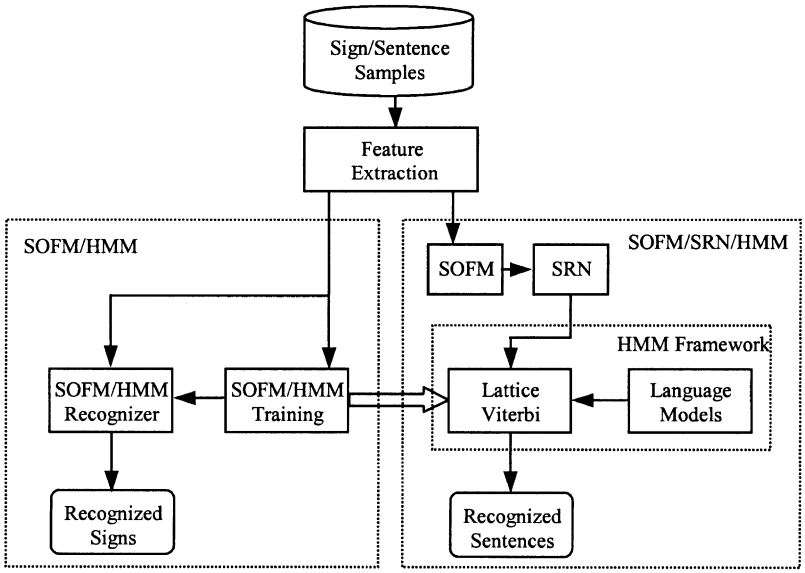

Fig. 1. The structure of sign language recognition system based on SOFM/SRN/HMM.

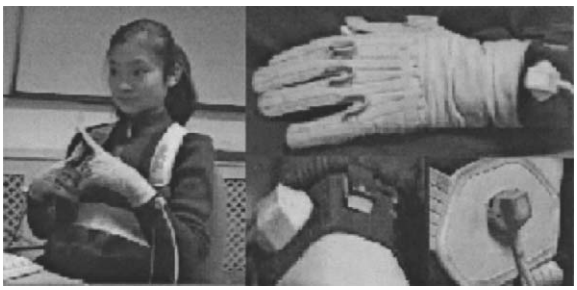

Fig. 2. The input device.

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 the semantic meaning as the main way of expression with the similar word order as the Chinese sentence, (2) Finger spelling and Chinese character-imitating gestures play a very important part of CSL. However, similar to Stokoe's analysis of American sign language [28], each Chinese sign can be broken into four parameters: hand shape, position, orientation and movement. These parameters are performed simultaneously and form multiple data streams which are the basis of sign language recognition.

Hand shapes are one of the primitive sign languages and reflect the information of hand configuration. For more accurately collecting 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 where the recognition system is moved from one place to another. In order to extract the invariant features to signer's position, the following method is proposed. First, two trackers are positioned on the wrist of each hand and another is mounted at signer's back, where the tracker at signer's back is chosen as the reference Cartesian coordinate system. And then, the position and orientation at 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 three-dimensional position vector and a three-dimensional 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 two hands. The data from different signers are calibrated by some fixed movements performed by each signer. In our experiments the 14 postures that can represent the min–max value ranges of the corresponding sensor and 75 basic hand shapes are defined. As each component in the vector has different dynamic range, its value is normalized to [0,1].

4. SOFM/HMM for modeling signer-independent isolated signs

In this section, we propose SOFM/HMM for modeling signer-independent isolated signs. The proposed model uses

SOFM as a feature extractor for continuous HMM and their parameters are trained simultaneously in a global optimization criterion. SOFM transforms input signs into significant and low-dimensional representations that can be well modeled by the emission probabilities of HMM. SOFM/HMM is represented in terms of SOFM/HMM architecture and SOFM/HMM based isolated SLR.

4.1. SOFM/HMM architecture

SOFM first introduced by Kohonen [29] has been successfully used in a variety of signal processing applications, especially in speech recognition. SOFM has shown significant potential for feature extraction in the situation where the nature of the feature of interest is not known before. The architecture of SOFM is a fully connected network with two layers, and each input is connected with every output by the adjustable weights. The outputs of SOFM in the form of two-dimensional lattices represent the corresponding eigenvector centroids. The weights are gradually adjusted while the training vectors are input into SOFM, so that the probability density of each centroid is becoming similar to that of the corresponding input vector.

HMM has been proven to be one of the most successful statistical models in the area of speech recognition. But it has been pointed out in Ref. [30] that conventional HMM has some limitations. One of them is the assumption that the distributions of individual observation parameters can be well represented as a mixture of Gaussian or autoregressive densities is not always consistent with the fact. Another limitation is that HMM has a poorer discrimination than neural networks. In recent years, many researchers have employed HMM for sign language recognition and obtained inspiring results. However, these limitations still exist.

In fact, the method of combining artificial neural network (ANN) with HMM is an ideal alternative to overcome HMM's limitations, since the hybrid paradigm maintains an underlying HMM structure, capable of modeling long-term dependencies, with the integration of ANN, which provides probability estimation, discriminative training algorithms, and fewer parameters to estimate than those usually required in conventional HMM. Many ANN/HMM models (e.g. Refs. [31–33], the survey in Ref. [34] gave a full review), combining ANN with HMM at different levels, have been successfully applied to speech recognition. More recently, Corradini et al. [6] presented a hybrid classifier for gesture recognition, where SOFM is regarded as the quantizer of 32 defined subgestures for discrete HMM.

In this paper, we propose a SOFM/HMM model which is different from the ANN/HMM models used in speech recognition in terms of the architecture and the corresponding training algorithm. The proposed SOFM/HMM model is also different from Corradini's method because SOFM is presented in this paper as an implicit feature extractor of different signers for continuous HMM and their parameters are trained simultaneously in a global optimization criterion. The SOFM/HMM model uses SOFM as a feature extractor for continuous HMM, with the goal to transform input sign representations into compact, but significant, low-dimensional representations that can be well modeled by the emission probabilities of HMM. In this model, each eigenvector centroid of SOFM with different signers' transformation is regarded as one of the components in the state of HMM. The state probability density functions (pdf) of HMM are constructed in the form of the weighted sum of components. They can be computed through the Forward–Backward procedure or the Viterbi algorithm. The weights of SOFM are iteratively updated in the supervision of the computed state pdf.

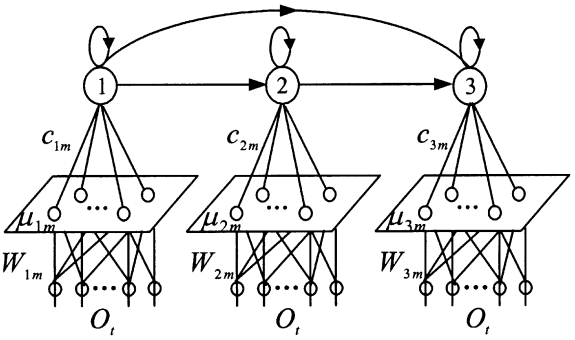

Fig. 3. The architecture of SOFM/HMM.

Let the vector of observation sequence $O_t = [o_{t1}, o_{t2}, \ldots, o_{tn}]$, $t = 1, \ldots, T$, where t is the time of observation sequence, and n is the dimensionality. The input vector O_t is linked with the SOFM/HMM neuron m by the weight vector W_{jm}, where j is the state variable and $W_{jm} = [w_{jm1}, w_{jm2}, \ldots, w_{jmn}]$. Denote M as the variable set of SOFM neurons and $|M|$ as the number of elements in the set.

Fig. 3 shows the architecture of a 3 state left–right model with skip. The input vector O_t at time t for all states is first transformed into the corresponding eigenvector centroid μ_{jm} by the linking weight W_{jm}, and then the neurons make up of the state pdf $b_j(O_t)$ of HMM in the form of the weighed sum.

The contribution probability to the state pdf of being the mth neuron in state j can be constructed as follows:

$$b_{jm}(O_t) = z \exp[-D(\mu_{jm}, O_t)], \tag{1}$$

where z is a constant. $b_{jm}(O_t)$ is the mth neuron's contribution to the state pdf, and it gradually decreases as the observation vector deviates from the corresponding eigenvector centroid. $D(\mu_{jm}, O_t)$ is the Euclidean distance between the observation O_t and the eigenvector centroid μ_{jm}. However, different signers' data vary their features, so the kth signer transformation $G^{(k)} = (R^{(k)}, \beta^{(k)})$ is imposed on the

weight vector of SOFM for the signer k data, and defined as $\mu_{jm} = R^{(k)} W_{jm} + \beta^{(k)}$.

Since the contribution to the state pdf $b_j(O_t)$ varies from different neurons, the coefficients that reflect the importance of the contribution are associated with $b_{jm}(O_t)$. Thus, $b_j(O_t)$ is defined as follows:

$$b_j(O_t) = \sum_{m=1}^{|M|} c_{jm} b_{jm}(O_t), \tag{2}$$

where $\sum_{m=1}^{|M|} c_{jm} = 1$. The weight vector W_{jm}, the transformation $G^{(k)}$ and the coefficient c_{jm} are calculated through the following re-estimation formulas.

Given the set of K observation sequences from different signers $O = [O^{(1)}, O^{(2)}, \ldots, O^{(K)}]$, where the kth signer data $O^{(k)} = [O_1^{(k)} O_2^{(k)} \cdots O_{T_k}^{(k)}]$, and the length T_k. Define the initial model parameters as λ, and the re-estimated model parameters as $\bar{\lambda}$. Q is a state sequence, and denoted as $Q = q_1, q_2, \ldots, q_{T_k}$, $q_i \in \{1, 2, \ldots, N\}$, where N is the number of states.

Assuming each observation sequence to be independent of every other observation sequence, a global optimization criterion is to adjust the parameters of the model λ to maximize K observation sequences, formulated as $\lambda^* = \arg\max_\lambda P(O \mid \lambda)$, where $P(O \mid \lambda) = \prod_{k=1}^{K} P(O^{(k)} \mid \lambda)$. For convenience, let $P_k = P(O^{(k)} \mid \lambda)$. The weight w_k is defined as $w_k = P(O \mid \lambda)/K P_k$, then $P(O \mid \lambda) = \sum_{k=1}^{K} w_k P_k$. Since $P(O \mid \lambda)$ depends on the hidden state variable S and the SOFM neurons variable M, it cannot be maximized directly. The maximization of $P(O \mid \lambda)$ is the problem of maximum likelihood estimation with missing values (i.e. hidden variables). The EM algorithm is a popular algorithm for maximum likelihood estimation given incomplete data samples. Similar to HMM parameter estimation, the auxiliary function $Q(\lambda, \bar{\lambda})$ is also introduced to facilitate the maximization of $P(O \mid \lambda)$, and constructed as

$$Q(\lambda, \bar{\lambda}) = \sum_{k=1}^{K} w_k Q_k(\lambda, \bar{\lambda}),$$

$$\text{where } Q_k(\lambda, \bar{\lambda}) = \sum_{Q} \sum_{M} \Big(P(O^{(k)}, Q, M \mid \lambda) \times \log P(O^{(k)}, Q, M \mid \bar{\lambda}) \Big). \tag{3}$$

It can be proved $Q(\lambda, \bar{\lambda}) \geqslant Q(\lambda, \lambda) \Rightarrow P(O \mid \bar{\lambda}) \geqslant P(O \mid \lambda)$ with the similar method in Ref. [35]. Thus, the maximization of $P(O \mid \lambda)$ is converted to get the critical point of $Q(\lambda, \bar{\lambda})$. The probability of being in state j at time t with the mth neuron accounting for $O_t^{(k)}$ is defined as: $\Phi_t^{(k)}(j, m) = P(q_t = j, m_t = m \mid O^{(k)}, \lambda)$.

Then we can get

$$\begin{aligned} Q(\lambda, \bar{\lambda}) = \sum_{k=1}^{K} \Bigg(& \sum_{i=1}^{N} \sum_{m=1}^{|M|} \Phi_1^{(k)}(i, m) \log \bar{\pi}_i \\ & + \sum_{i=1}^{N} \sum_{j=1}^{N} \sum_{t=1}^{T_k - 1} \sum_{M} P(q_t = i, q_{t+1} = j, \\ & M \mid O^{(k)}, \lambda) \log \bar{a}_{ij} \\ & + \sum_{j=1}^{N} \sum_{t=1}^{T_k} \sum_{m=1}^{|M|} \Phi_t^{(k)}(j, m) \log \bar{c}_{jm} \\ & + \sum_{j=1}^{N} \sum_{t=1}^{T_k} \sum_{m=1}^{|M|} \Phi_t^{(k)}(j, m) \log \bar{b}_{jm}(O_t^{(k)}) \Bigg) \\ & \times \frac{P(O \mid \lambda)}{K}. \end{aligned} \tag{4}$$

Through maximizing the individual terms 1–3 in $Q(\lambda, \bar{\lambda})$, we can obtain the re-estimation formulas for c_{jm}:

$$\bar{c}_{jm} = \sum_{k=1}^{K} \sum_{t=1}^{T_k} \Phi_t^{(k)}(j, m) \Bigg/ \sum_{k=1}^{K} \sum_{t=1}^{T_k} \sum_{m=1}^{|M|} \Phi_t^{(k)}(j, m). \tag{5}$$

The formulas for π_i and a_{ij} are the same as the conventional HMM [30].

The maximization of the fourth term in $Q(\lambda, \bar{\lambda})$ can be carried on through the following two steps. First, we maximize the Q-function with respect to the signer transformation while keeping the value of W_{jm} fixed. The detailed description of derivation can be found in Ref. [36]. After some simplification, the re-estimated formulas can be expressed as follows:

$$\begin{aligned} \bar{r}_{in} = \sum_{j=1}^{N} \sum_{m=1}^{|M|} \sum_{t=1}^{T_k} \Phi_t^{(k)}(j, m) (w_{jmn} \bar{w}_{mi} & \\ - \sum_{s \neq n} r_{is} w_{jms} - b_i) \Bigg/ & \\ \sum_{j=1}^{N} \sum_{m=1}^{|M|} \sum_{t=1}^{T_k} \Phi_t^{(k)}(j, m) w_{jmn}^2, & \end{aligned} \tag{6}$$

$$\bar{b}_i = \sum_{j=1}^{N} \sum_{m=1}^{|M|} \sum_{t=1}^{T_k} \Phi_t^{(k)}(j, m) \left(\bar{w}_{mi} - \sum_{s} r_{is} w_{jms} \right) \Bigg/ \sum_{j=1}^{N} \sum_{m=1}^{|M|} \sum_{t=1}^{T_k} \Phi_t^{(k)}(j, m), \tag{7}$$

where $\bar{w}_{mi} = \sum_{j=1}^{N} \sum_{t=1}^{T_k} \Phi_t^{(k)}(j, m) o_{ti}^{(k)} / \sum_{j=1}^{N} \sum_{t=1}^{T_k} \Phi_t^{(k)}(j, m)$, r_{in} and b_i are the elements of $R^{(k)}$ and $\beta^{(k)}$.

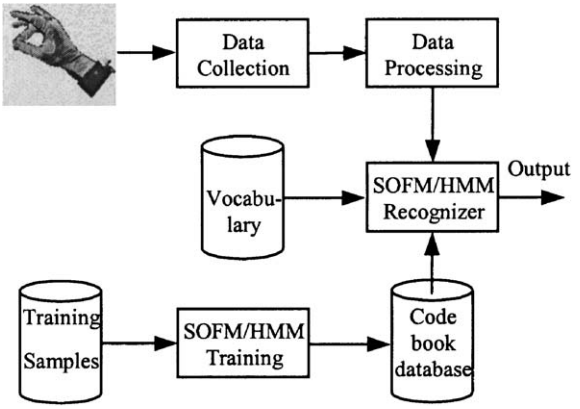

Fig. 4. The isolated sign language recognition system.

Second, we maximize the Q-function with respect to the weight vector of SOFM W_{jm} using the fixed signer transformation $G^{(k)}$, and obtain the re-estimation formula for W_{jm}:

$$\bar{W}_{jm} = \sum_{k=1}^{K}\sum_{t=1}^{T_k} \Phi_t^{(k)}(j,m)\cdot \bar{R}^{(k)^T}\cdot(O_t^{(k)} - \bar{\beta}^{(k)}) \Big/ \sum_{k=1}^{K}\sum_{t=1}^{T_k} \Phi_t^{(k)}(j,m)\bar{R}^{(k)^T}\bar{R}^{(k)}. \quad (8)$$

Scaling is usually employed in the implementation of forward–backward procedure to avoid the underflow. For more details, refer to Ref. [30].

4.2. SOFM/HMM based isolated SLR

The SOFM/HMM based isolated SLR is illustrated in Fig. 4. It has two modules: training module and recognition module. The training module includes three parts: training samples, SOFM/HMM training and codebook database. The recognition module consists of four parts: data collection, data processing, vocabulary and SOFM/HMM recognizer. The model parameters of SOFM/HMM recognizer are obtained from the codebook database.

Training: Before collecting word data, every signer performed a series of samples which contain 75 basic hand shapes and 14 postures of reflecting the min–max value of the finger-bending sensor. These data are referred to as calibrated samples. Then, several samples for every word in the vocabulary set are collected from different signers. The vocabulary and the corresponding samples form training samples. One SOFM/HMM model is built for each word in the vocabulary through the following training procedure:

(1) Using calibrated samples to calculate the signer transformation $G^{(k)}$ through Eqs. (5)–(8) with the similar steps of maximum likelihood linear regression [36].
(2) For training samples of each sign, initialize the parameters of c_{jm}, W_{jm}.
(3) Re-estimate the parameters by Eqs. (5) and (8) based on the calculated signer transformation $G^{(k)}$.
(4) If the convergence criterion is met, the parameter is saved and continues to the next; otherwise replace old parameters with new ones and return to (3).

Recognition: The procedure of recognition is to select the model from the codebook database that can well represent the observation sequence. Given the observation sequence $O = O_1 O_2 \cdots O_T$, the probability $P(O \mid \lambda_v)$ is computed for each codebook λ_v in the codebook database $\lambda = \{\lambda_1, \lambda_2, \ldots, \lambda_V\}$. $P(O \mid \lambda_v)$ is approximated as $P(O \mid \lambda_v) = P(O, Q^* \mid \lambda_v)$, where Q^* is the best state sequence among the state spaces for the given O, which can be obtained through the Viterbi algorithm search [30]. The recognized class can be obtained through the following formula:

$$v^* = \arg\max_{1 \leqslant v \leqslant V} P(O \mid \lambda_v). \quad (9)$$

5. SOFM/SRN/HMM models for continuous SLR

In this section, a SOFM/SRN/HMM model is presented for signer-independent continuous SLR based on the previous isolated sign models. The improved SRN is used to segment continuous Chinese sign language in terms of transformed SOFM representations. The outputs of SRN are taken as the HMM states in which the lattice Viterbi algorithm is employed to search the best word sequence.

The SOFM/SRN/HMM model is described by SRN-based segmentation and HMM framework. SRN-based segmentation consists of improved SRN and segmentation of continuous sign language. HMM framework is represented by lattice Viterbi algorithm and language models.

5.1. SRN-based segmentation

5.1.1. Improved SRN

Jordan [37] first described recurrent networks in 1986. Along this line, Elman [38] developed a simple recurrent network in 1990. In the past few years, recurrent networks due to their good dynamic memory performances have been successfully applied to speech recognition [39], handwriting recognition [40], and gesture recognition [41,42].

Fig. 5 shows the structure of simple recurrent network. The input unit I_t receives the first input at time t. The context unit C_t is initially set to 0.5. Both the input unit and context unit activate the hidden unit H_t; and then the hidden unit feeds forward to activate the output unit O_t. The hidden

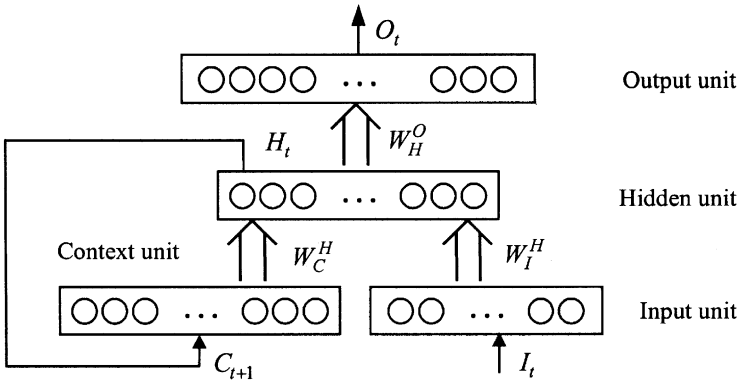

Fig. 5. The structure of simple recurrent network.

unit also feeds back to activate the context unit C_{t+1} which constitutes the forward activation at time $t+1$. In the network the context unit is connected one-to-one corresponding to the hidden unit through a unit of time-delay, i.e.

$$C_t = \begin{cases} H_{t-1}, & t \geqslant 2, \\ \overrightarrow{0.5}, & t = 1. \end{cases} \tag{10}$$

W_I^H, W_C^H and W_H^O are defined as the weight matrices of input unit to hidden unit, context unit to hidden unit and hidden unit to output unit. Φ and ψ denote the bias of hidden unit and output unit. Then H_t and O_t can be expressed as follows:

$$H_t = f(C_t \cdot W_C^H + I_t \cdot W_I^H - \Phi), \tag{11}$$

$$O_t = f(H_t \cdot W_H^O - \psi), \tag{12}$$

where $f(\cdot)$ is the standard sigmoid activation function: $f(x) = (1 + e^{-x})^{-1}$.

With the introduction of context units, the outputs of network depend not only on the external inputs but also on the previous internal states that rely on the results of all the preceding external inputs. Thus SRN can memorize and utilize a relatively larger preceding context [38].

Though SRN can memorize the preceding context, it cannot utilize the following context information. Two approaches are presented to modify SRN for efficiently utilizing the following context information. One approach is to take the following context vector as a part of the input vector. The input vector can be redefined as $I_t = [I_t \ I_{t+1}]$, and the rest of calculations are the same as the standard SRN. So the following context information can be exploited. Another approach is to input the training samples, in turn and in reverse turn, into the SRN with the same architecture. Then one forward SRN and one backward SRN are trained, and the forward SRN can memorize the preceding context while the backward SRN memorizes the following context. Thus, the context information can be assimilated. Experiments show that the former approach has better performance than the latter for segmentation of continuous sign language. Thus, the former improved SRN is adopted in our SOFM/SRN/HMM model.

5.1.2. Segmentation of continuous sign language

In the segmentation of continuous sign language, we need to guarantee that segmentation has a high recall of detecting all actual segments. For the case of detecting several segments for an actual word, we will solve it later through the lattice Viterbi search algorithm in the following HMM framework.

After sign data input and processing, we can obtain the 48-dimensional data. If the 48-dimensional data are taken as the inputs of SRN directly, due to mass training data, it is difficult to train a converging SRN with a good performance. Thus, SOFM is employed as the feature extraction network to reduce the training data dimension. The extracted features after being encoded are fed into SRN.

Our experiments show that continuous sign language after being transformed by SOFM has stronger segment properties, i.e. distinct fluctuation in the movement epenthesis between signs and stabilization within one sign. An example is shown in Fig. 6, where *x*-axis represents the frame number of the sign language sentence, and *y*-axis denotes the transformed SOFM outputs with one associated quantization number for each frame. In the figure, Chinese sign word “我们” (we) has two segments and the signs of “什么” (what), “时候” (time), and “走” (go) only have one segment, and the fluctuating parts between the two signs are movement epentheses. One segment in continuous sign language can be considered as a potential phoneme. Phoneme, as the basic unit of sign language, is defined as a dynamic continuous sign data of the changes of hand shape, position and orientation being very stable.

Distinct fluctuation in the movement epenthesis between signs and stabilization within one sign constitute the movement epenthesis characteristic (see Fig. 6). This

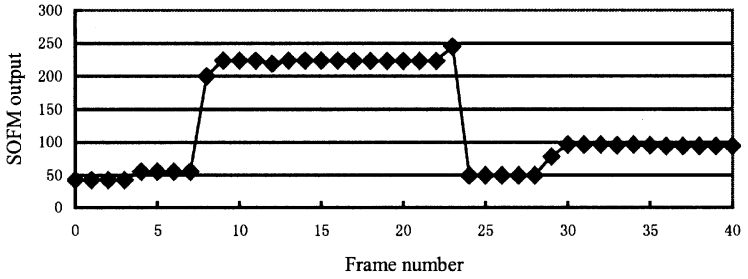

Fig. 6. The segment property of sign language "我们什么时候走" (when will we leave).

characteristic results in the fact that SRN-based segmentation may have good performance. Furthermore, we can discard severe fluctuation transition frames judged by the SOFM outputs near the SRN segmentation point so that partial movement epentheses are removed in the implementation. After the feature extraction of SOFM, the outputs of SOFM are fed into SRN. 256 output units and 48 input units are selected for SOFM. 16 input units, 15 hidden units and 3 output units are chosen for SRN. The following paragraphs will explain SRN-based segmentation in detail.

Input: SRN has 16 input units, where 8 units are for the encoding of 256 output units of SOFM, and the other 8 units are for the following contexts. The values of the input are formulated as $I_t^i \in \{0, 1\},\ i = 1, 2, \ldots, 16$.

Output: The 3 output units are defined as: the left boundary of segments 1, the right boundary of segments 2, and the interior of segments 3. The corresponding units are represented by $o_t^1,\ o_t^2$ and o_t^3. $O_t = \{o_t^1, o_t^2, o_t^3\}$ is defined as follows:

$$O_t = \begin{cases} [1 \quad 0 \quad 0] & \text{Output is 1} \\ [0 \quad 1 \quad 0] & \text{Output is 2} \\ [0 \quad 0 \quad 1] & \text{Output is 3} \end{cases} \tag{13}$$

Training: The target segments of training cannot be obtained straightforwardly because sign language is continuous and there is no mark between signs. Thus, automatic segmentation approach is employed to find the target segments. Let the sample sentence $W = w_1 w_2 \cdots w_k$ in the training set and the corresponding frame sequence $T = t_1 t_2 \cdots t_l$. If $t_i \in w_m,\ t_{i+1} \in w_{m+1}$, then frame t_i is the right boundary of segments and frame t_{i+1} is the left boundary of segments. Each state probability of frame t_i belonging to word $w_m\ (m = 1, \ldots, k)$ is calculated through the SOFM/HMM model with the isolated sign language model parameters. In the calculated state probability space, the constrained Viterbi algorithm following the word sequence $w_1 w_2 \cdots w_k$ is used to search the best segment sequence. The segment results are regarded as the target outputs of SRN training.

Back-propagation through time [43] is introduced as the SRN learning algorithm. 800 samples over 400 different continuous CSL sentences in the training set are transformed by SOFM, and the SOFM outputs together with the following contexts are fed into the SRN. The errors between the SRN outputs and the targets are propagated back using back-propagation and then the network weights are adjusted. At the beginning of learning, the weight matrices, the bias of hidden units, and the output units are initialized to the random value $(-1, +1)$. The feedback units are initialized to activations of 0.5.

Recognition: Continuous signs in the test set are fed into the SOFM at first. Then the SOFM outputs together with the following contexts are fed into the SRN. The segmentation result of SRN is $i^* = \arg\max_i (o_t^i)$ at time t. The adjacency property of the left and the right boundary of segments is used as the constraint in the segmentation.

5.2. HMM framework

The segmentation results of SRN are fed into the HMM framework. One sign candidate, which consists of one or several segments, is viewed as a state of HMM. The probability of the sign candidate as an isolated word can be regarded as the state emission probability of HMM. The transition relations are built among words relevant to sign candidates. Since each sign may include several segments (usually 2–4 segments), we should search the best path in these segments. This can be shown in the following two aspects: the selection of the recombined segment sequence and the selection of the best word sequence from the recombined segment sequence. In our framework, the lattice Viterbi algorithm is proposed to solve the best search problem. Compared with the standard Viterbi algorithm that searches the frames one by one, the lattice Viterbi algorithm can span one or more segments to search, so the recombined segment sequence accompanying the corresponding word sequence can be obtained simultaneously.

5.2.1. Lattice Viterbi algorithm

One sign candidate is defined as a triple $\langle t, t', w\rangle$, starting at segment t, ending at segment t' and representing word $w,\ 0 \leqslant t < T,\ t < t' \leqslant T$. All triple sets are defined as $L = \{x \mid x = \langle t, t', w\rangle\}$. We introduce the accumulator

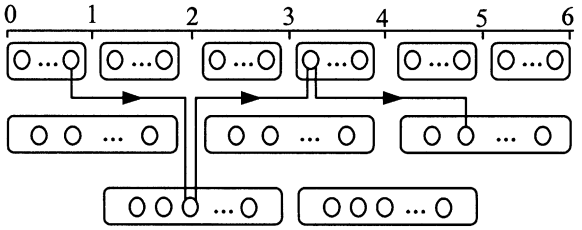

Fig. 7. An example of the lattice Viterbi algorithm.

$\delta(t, t', w)$ that collects the maximum probability of covering the triple $\langle t, t', w\rangle$. To keep track of the best path, the auxiliary argument $\psi(t, t', w)$ is defined as the previous triple pointer of the local maximum $\delta(t, t', w)$. We denote $b(t, t', w)$ as the emission probability of word w covering segments from position t to t'. $P(w \mid w')$ is defined as the transition probability from word w' to w, which is estimated by language models. The lattice Viterbi algorithm is described as follows:

(1) *Initialization*: Let $\delta(0, t, w) = b(0, t, w)$ and $\psi(0, t, w) = NULL$.
(2) *Recursion*: $\delta(t, t', w) = \max_{\langle t'', t, w'\rangle \in L} \delta(t'', t, w')\ P(w \mid w')b(t, t', w)$ and $\psi(t, t', w) = \arg\max_{\langle t'', t, w'\rangle \in L} \delta(t'', t, w')\ P(w \mid w')$.
(3) *Termination*: $P^* = \max_{\langle t, T, w\rangle \in L} \delta(t, T, w)$ and $\langle t_1^*, T, w_1^*\rangle = \arg\max_{\langle t, T, w\rangle \in L} \delta(t, T, w)$.
(4) *Path backtracking*: Let $T = t_0^*$, $\langle t_{i+1}^*, t_i^*, w_{i+1}^*\rangle = \psi(t_i^*, t_{i-1}^*, w_i^*)$ is iterated until $\langle t_{k+1}^*, t_k^*, w_{k+1}^*\rangle = NULL$, and the generated word sequence $w_k^* \cdots w_1^*$ is the best path.

The calculation of the sign candidate probability $b(t, t', w)$ is similar to that of isolated sign language recognition. So the SOFM/HMM model used for isolated SLR is also employed for the calculation of sign candidate probability. The probabilities of all possible candidates are regarded as the state emission probability of the HMM framework. In the SOFM/HMM model, SOFM is used as an implicit feature extractor of different signers for continuous HMM and it transforms input signs into significant and low-dimensional representations that can be well modeled by the emission probabilities of HMM. Thus, it can alleviate the effect of signer-independent problem in continuous SLR.

Fig. 7 illustrates the search results of the lattice Viterbi algorithm with a continuous sign language sentence of six segments. 1, 2, 3, 4 and 5 are the boundaries of segments. Each rectangle denotes one sign candidate segment in which the circles represent candidate words. The search result is four recombined segments: $\langle 0\,1\rangle\langle 1\,3\rangle\langle 3\,4\rangle\langle 4\,6\rangle$, and the corresponding words in each segment construct the best word sequence.

5.2.2. Language models

Language models are an attempt to capture regularities of natural language by a large amount of training data for improving the performance of sign language recognition. It plays the role of the prior and guarantees the recognized sentence, which is well interpreted from the grammar point of view, can be selected with the maximum probability. A simple but effective way is to use an n-gram model in which the probability of appearance of w_i is assumed to depend only on the preceding $n-1$ words, that is

$$P(S) \stackrel{\text{def}}{=} P(w_1, w_2, \ldots, w_N) = \prod_{i=1}^{N} P(w_i \mid w_{i-n+1}^{i-1}). \quad (14)$$

In our SOFM/SRN/HMM model, bigram is adopted for continuous SLR. If the training corpus is not large enough, many bigrams will not appear in the training data and many others will only appear once or twice. So the Katz smoothing technique [44] is employed to make the estimated probability robust for unseen data in our language model. The corpus used to estimate the bigram probabilities consists of 200Mbytes from China Daily between the years 1993–1995 and the Family Collection Book. 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 corpus.

6. Experiments and discussions

In our experiments, two Cybergloves and three Pohelmus 3SPACE-position trackers are used as data input devices. Two systematic experiments are carried out: the first is to test the ability of SOFM/HMM for modeling signer-independent signs that is evaluated by an accuracy of isolated SLR, and the second is to evaluate the SOFM/SRN/HMM model for signer-independent continuous SLR.

6.1. Isolated sign language recognition evaluation

The proposed SOFM/HMM model for signer-independent isolated SLR is to evaluate on a large-vocabulary with 5113

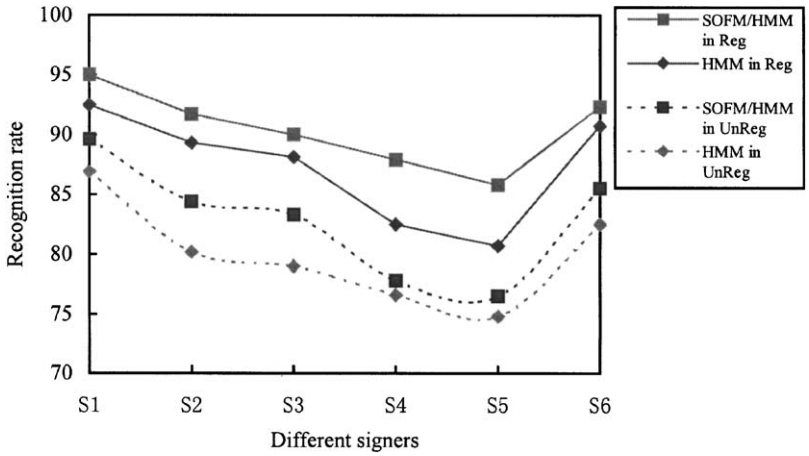

Fig. 8. The isolated sign recognition performance on a large vocabulary of 5113 signs.

signs. Experimental data consist of 61,356 samples over 5113 isolated signs from 6 signers with each performing signs twice. The vocabulary is taken from the Chinese sign language dictionary excluding the synonymous words with the same gestures. One group data from 6 signers are referred to as the registered test set (Reg) and the other 11 group data are used as the training samples. Using the cross validation method, 10 group data samples from 5 signers are used as the training samples and the other signer data are referred to as the unregistered test set (Unreg).

Fig. 8 shows the test results of HMM and SOFM/HMM, in which HMM has 3 states and 5 mixture components and SOFM/HMM has 3 states and 5 initial SOFM neurons. The average recognition rates of 90.5% for SOFM/HMM and 87.3% for HMM are observed for the registered test set. For the unregistered test set, the average recognition rate of 82.9% and 80.0% are obtained. Experiments show that SOFM/HMM increases the recognition accuracy by 3% than HMM on the registered and unregistered test sets.

From the experiments above, we know that SOFM/HMM has better performance than HMM. The possible reasons are as follows. First, SOFM is trained as a feature extractor for continuous HMM in a global optimization criterion to transform signer-independent input signs into compact, but significant and low-dimensional representations that can be well modeled by the emission probabilities of HMM. Second, the combination of powerful self-organizing performances of SOFM and excellent temporal processing properties of HMM in a novel scheme may compensate each other to obtain better results. Third, in SOFM/HMM, less parameters need to be re-estimated than in conventional HMM, thus SOFM/HMM is more inclined to convergence in the model training. From the training procedure of HMM, we find that the changes for some sign models don't occur at the means of multi-mixture of HMM, i.e. the means remain unaltered, but occur at the corresponding covariances. For recognition the means play a more important role in scoring one sign than covariance. Clearly, this phenomenon is due to the model converging to the local optimum.

6.2. Continuous sign language recognition evaluation

In this section, the SOFM/SRN/HMM model for signer-independent continuous SLR is evaluated. The 61,356 samples over 5113 isolated signs are collected from 6 signers with each performing signs twice. The 10 group samples from 5 signers with each having two group data are regarded as the isolated sign training set. The isolated sign models are trained through the SOFM/HMM method. We select 2 from 5 signers in the isolated sign training set, represented by A, B, and the signer in the test set represented by C. Continuous sign language database consists of 2400 samples from these 3 signers with 400 different continuous CSL sentences. The sentences are chosen from the 200 Mbytes corpus of China Daily and the Family Collection Book and consists of 3–14 words with the average 6.55 words per sentence. There are 6 group data marked with $A_1, A_2, B_1, B_2, C_1, C_2$. In the SOFM/SRN/HMM model, A_1 and B_1 are chosen as the training set for SOFM, SRN and embedded training. A_2, B_2 are referred to as the registered test set, and C_1, C_2 as the unregistered test set.

In our experiments, embedded training and un-embedded training are defined. In embedded training, the sentences are automatically segmented into words and the corresponding data by automatic segmentation approach (see Section 5.1.2), and the isolated sign models together with the models trained using the segmented word data are taken as the candidate models in the calculation of segment emission probability. In un-embedded training, only the isolated signs models are used.

Table 1
The continuous sign recognition rates of un-embedded training on the unregistered test set

Method	Recognition rate (%)
HMM	66.5 ($S=423$, $I=276$, $D=178$)
SOFM/SRN/HMM	72.4 ($S=395$, $I=117$, $D=210$)

Table 2
The continuous sign recognition rates of embedded training on the unregistered test set

Method	Recognition rate (%)
HMM	82.9 ($S=208$, $I=162$, $D=78$)
SOFM/SRN/HMM	86.3 ($S=173$, $I=50$, $D=135$)

Table 3
The continuous sign recognition rates of embedded training on the registered test set

Method	Recognition rate (%)
HMM	89.2 ($S=117$, $I=105$, $D=62$)
SOFM/SRN/HMM	91.3 ($S=91$, $I=35$, $D=102$)

We test the performances of SOFM/SRN/HMM for signer-independent continuous SLR in the following three experiments: one experiment is to use un-embedded training on the unregistered test set, a second experiment is to use embedded training on the unregistered test set, and a third experiment is to use embedded training on the registered test set. The experimental results are listed in Tables 1 , 2 , 3 , respectively.

All experiments are performed with the bigram language model. S, I and D denote the error numbers of substitution, insertion and deletion, respectively. The number of signs in the test set is 2620. Table 1 shows that the recognition rate of 72.4% for SOFM/SRN/HMM is obtained when using un-embedded training on the unregistered test set. SOFM/SRN/HMM outperforms HMM by 5.9% recognition rate. It can be easily seen that because no knowledge of continuous sign language is applied, the recognition rate is not satisfied. In Tables 2 and 3, embedded training is employed to utilize the sentence information. The recognition rate of 86.3% for SOFM/SRN/HMM is shown in Table 2 on the unregistered test set. Table 3 illustrates the recognition rate of 91.3% for SOFM/SRN/HMM on the registered test set. Experiments show that SOFM/SRN/HMM with embedded training increases the recognition rate by 3.4% than HMM on the unregistered test set and 2.1% on the registered test set. In our three experiments of continuous SLR, the SRN segment recalls (i.e. Number of correct segments/Number of all actual segments) are respectively 93.8%, 94.5% and 97.1%. However, soft-segmentation instead of fixed-segmentation is employed in SRN segmentation and the sign boundary is decided in the lattice Viterbi algorithm. The recognition results are also improved through the product scores of language models (LM) and sign candidate probability, where LM can guarantee to select those sign candidates whose emission probabilities are not top-1 but well interpreted from the grammar point of view.

Compared with conventional HMM, SOFM/SRN/HMM has higher recognition rates. This may be due to the following reasons:

(1) HMM uses the continuous sign Viterbi algorithm which is liable to be influenced by movement epenthesis, whereas SOFM/SRN/HMM alleviates the effects of movement epenthesis by discarding the transition frames between signs that can be judged by the SOFM outputs near the SRN segmentation point.
(2) Unlike HMM which searches the best state sequence, SOFM/SRN/HMM gets the best word sequence that is more suitable for the language model.
(3) The isolated sign Viterbi algorithm (ISVA) employed by SOFM/SRN/HMM can get the higher accuracy than the continuous sign Viterbi algorithm (CSVA) used by conventional HMM. In recognition, ISVA only searches all states of the word in each frame, while CSVA searches not only all states of this word but also the states of other words, so ISVA is more accurate and less time-consuming than CSVA for the same sign data recognition.

7. Conclusions and future work

In this paper, a sign language recognition system is developed both for isolated signs and continuous signs in the signer-independent field. In this system, SOFM/HMM is first presented for modeling signer-independent isolated signs. The proposed method uses SOFM as a feature extractor for continuous HMM so as to transform input signs into significant and low-dimensional representations that can be well modeled by the emission probabilities of HMM. Based on these isolated sign models, the SOFM/SRN/HMM model is proposed for signer-independent continuous SLR. The improved SRN is used to segment continuous sign language in terms of transformed SOFM representations, and the outputs of SRN are taken as the HMM states in which the lattice Viterbi algorithm is employed to search the best matched word sequence. Experiments on the vocabulary of 5113 signs show that SOFM/HMM reaches a word accuracy of 90.5% in the registered test set and 82.9% in the unregistered test set, respectively. SOFM/HMM increases the recognition rates by 3% than HMM. For signer-independent continuous SLR, experimental results demonstrate that the proposed SOFM/SRN/HMM model has an accuracy of 91.3%, 86.3% and 72.4%, respectively, in embedded training and registered, embedded training and

unregistered, and un-embedded training and unregistered test sets. The experiments also show that this model has higher accuracy than conventional HMM.

Though we have researched into signer-independent sign language recognition, there are still many issues to be further investigated: (1) Effective feature extraction from different signers: Can explicit effective features be extracted through the transformation to frequency domain such as in speech recognition? It is a challenging issue that deserves further study. (2) Compact training sentences for a general model: Different from spoken language, sign language has no large amount of continuous sign language corpus. Its corpus is collected from the expert teachers. How to use compact and representative continuous sign language sentences to train a general recognition model is a very promising issue. (3) Minimum unit definition in SLR and its extraction: Due to no lexical defined basic units, how to tackle the distinct definition and effective extraction of minimum units is a barrier to prevent them as basis units for large vocabulary SLR? (4) The use of statistical sign language models: Chinese sign language is a kind of language mainly using the semantic meaning as ways of expression and it has many synonymous words with the same gestures. So sign language unit to natural language is a one-to-many map. Using statistical sign language models, the vocabulary size can be enlarged through the vocabulary maps in the post-processing sentence generation, which is very useful for SLR. (5) The utilization of non-manual parameters in sign language: Non-manual parameters in sign language include gaze, facial expression, mouth movement, position and motion of the trunk and head. Incorporating the understanding of non-manual parameters into sign language recognition is a further direction.

Acknowledgements

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

This work was supported in part by Natural Science Foundation of China (Grant No. 60303018), National Key-Basic Research Initialize (Grant No. 2001cca03300) and National High-Technology Development '863' Program of China (Grant No. 2001AA114160).

References

[1] C.L. Wang, W. Gao, S.G. Shan, An approach based on phonemes to large vocabulary Chinese sign language recognition, in: Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition,2002, pp. 411–416.

[2] C. Vogler, D. Metaxas, Adapting hidden Markov models for ASL recognition by using three-dimensional computer vision methods, in: Proceedings of the IEEE International Conference on Systems, Man and Cybernetics,1997, pp. 156–161.

[3] V.I. Pavlovic, R. Sharma, T.S. Huang, Visual interpretation of hand gestures for human–computer interaction: a review, IEEE Trans. Pattern Anal. Mach. Intell. 19(7) (1997) 677–695.

[4] Y. Wu, T.S. Huang, Vision-based gesture recognition: a review, in: Proceedings of the International Gesture Workshop,1999, pp. 103–115.

[5] H.K. Lee, J.H. Kim, An HMM-based threshold model approach for gesture recognition, IEEE Trans. Pattern Anal. Mach. Intell. 21(10) (1999) 961–973.

[6] A. Corradini, H.J. Böhme, H.M. Gross, A hybrid stochastic-connectionist approach to gesture recognition, Int. J. Artif. Intell. Tools 9(2) (2000) 177–204.

[7] H.S. Yoon, J. Soh, Y.J. Bae, H.S. Yang, Hand gesture recognition using combined features of location, angle and velocity, Pattern Recognition 34(7) (2001) 1491–1501.

[8] J. Triesch, C. Malsburg, A system for person-independent hand posture recognition against complex backgrounds, IEEE Trans. Pattern Anal. Mach. Intell. 23(12) (2001) 1449–1453.

[9] M.W. Kadous, Machine recognition of Auslan signs using PowerGloves: towards large-lexicon recognition of sign language, in: Proceedings of the Workshop on the Integration of Gesture in Language and Speech,1996, pp. 165–174.

[10] H. Matsuo, S. Igi, S. Lu, Y. Nagashima, Y. Takata, T. Teshima, The recognition algorithm with non-contact for Japanese sign language using morphological analysis, in: Proceedings of the International Gesture Workshop,1997, pp. 273–284.

[11] S.S. Fels, G.E. Hinton, Glove-talk: a neural network interface between a data-glove and a speech synthesizer, IEEE Trans. Neural Networks 4(1) (1993) 2–8.

[12] J.S. Kim, W. Jang, Z. Bien, A dynamic gesture recognition system for the Korean sign language (KSL), IEEE Trans. Syst. Man Cybernet. 26(2) (1996) 354–359.

[13] M.B. Waldron, S. Kim, Isolated ASL sign recognition system for deaf persons, IEEE Trans. Rehabilitation Eng. 3(3) (1995) 261–271.

[14] K. Grobel, M. Assan, Isolated sign language recognition using hidden Markov models, in: Proceedings of the International Conference on System, Man and Cybernetics,1997, pp. 162–167.

[15] M. Assan, K. Grobel, Video-based sign language recognition using hidden Markov models, in: Proceedings of the International Gesture Workshop,1997, pp. 97–109.

[16] T. Starner, J. Weaver, A. Pentland, Real-time American sign language recognition using desk and wearable computer based video, IEEE Trans. Pattern Anal. Mach. Intell. 20(12) (1998) 1371–1375.

[17] B. Bauer, H. Hienz, Relevant features for video-based continuous sign language recognition, in: Proceedings of the Fourth International Conference on Automatic Face and Gesture Recognition,2000, pp. 440–445.

[18] B. Bauer, K.F. Kraiss, Towards an automatic sign language recognition system using subunits, in: Proceedings of the International Gesture Workshop,2001, pp. 64–75.

[19] R.H. Liang, M. Ouhyoung, A real-time continuous gesture recognition system for sign language, in: Proceedings of the Third International Conference on Automatic Face and Gesture Recognition,1998, pp. 558–565.

[20] H. Sagawa, M. Takeuchi, A method for recognizing a sequence of sign language words represented in a Japanese sign language sentence, in: Proceedings of the Fourth International Conference on Automatic Face and Gesture Recognition,2000, pp. 434–439.
[21] C. Vogler, D. Metaxas, Toward scalability in ASL recognition: breaking down signs into phonemes, in: Proceedings of the International Gesture Workshop,1999, pp. 400–404.
[22] C. Vogler, D. Metaxas, A framework for recognizing the simultaneous aspects of American sign language, Comput. Vis. Image Und. 81(3) (2001) 358–384.
[23] W. Gao, J.Y. Ma, J.Q. Wu, C.L. Wang, Sign language recognition based on HMM/ANN/DP, Int. J. Pattern Recognition Artif. Intell. 14(5) (2000) 587–602.
[24] W. Gao, J.Y. Ma, X.L. Chen et al., HandTalker: A multimodal dialog system using sign language and 3-D virtual human, in: Proceedings of the Third International Conference on Multimodal Interface,2000, pp. 564–571.
[25] P. Vamplew, A. Adams, Recognition of sign language gestures using neural networks, Austral. J. Intell. Inform. Process. Syst. 5(2) (1998) 94–102.
[26] G.L. Fang, W. Gao, J.Y. Ma, Signer-independent sign language recognition based on SOFM/HMM, in: Proceedings of the IEEE ICCV Workshop Recognition, Analysis and Tracking of Faces and Gestures in Real-time Systems,2001, pp. 90–95.
[27] G.L. Fang, W. Gao, A SRN/HMM system for signer-independent continuous sign language recognition, in: Proceedings of the Fifth International Conference on Automatic Face and Gesture Recognition,2002, pp. 312–317.
[28] 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), Linstok Press, University of Buffalo, 1960.
[29] T. Kohonen, The self-organizing maps, Proc. IEEE 78(9) (1990) 1464–1480.
[30] R. Rabiner, A tutorial on hidden Markov models and selected applications in speech recognition, Proc. IEEE 77(2) (1989) 257–285.
[31] Y. Bengio, R.D. Mori, G. Flammia, R. Kompe, Global optimization of a neural network-hidden Markov model hybrid, IEEE Trans. Neural Networks 3(2) (1992) 252–259.
[32] H. Bourlard, N. Morgan, Continuous speech recognition by connectionist statistical methods, IEEE Trans. Neural Networks 4(6) (1993) 893–909.
[33] D. Albesano, R. Gemello, F. Mana, Hybrid HMM-NN modeling of stationary-transitional units for continuous speech recognition, Inf. Sci. 123(1) (2000) 3–11.
[34] E. Trentin, M. Gori, A survey of hybrid ANN/HMM models for automatic speech recognition, Neurocomputing 37(1) (2000) 91–126.
[35] X.L. Li, M. Parizeau, R. Plamondon, Training hidden Markov models with multiple observations—a combinatorial method, IEEE Trans. Pattern Anal. Mach. Intell. 22(4) (2000) 371–377.
[36] C. Leggetter, P. Woodland, Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models, Comput. Speech Lang. 9(2) (1995) 171–186.
[37] M.I. Jordan, Serial order: a parallel distributed processing approach, ICS Report 8604, Institute for Cognitive Science, University of California, San Diego, 1986.
[38] J.L. Elman, Finding structure in time, Cognitive Sci. 14(2) (1990) 179–211.
[39] T. Robinson, An application of recurrent nets to phone probability estimation, IEEE Trans. Neural Networks 5(2) (1994) 298–305.
[40] A. Senior, A.J. Robinson, Forward–backward retraining of recurrent neural networks, Adv. Neural Inf. Process. Syst. 8 (1996) 743–749.
[41] K. Murakami, H. Taguchi, Gesture recognition using recurrent neural networks, in: Proceedings of the CHI'91 Human Factors in Computing Systems,1991, pp. 237–242.
[42] A. Corradini, Real-time gesture recognition by means of hybrid recognizers, in: Proceedings of the International Gesture Workshop,2001, pp. 34–46.
[43] P.J. Werbos, Backpropagation through time: what it does and how to do it, Proc. IEEE 78(10) (1990) 1550–1560.
[44] S.X. Kate, Estimation of probabilities from sparse data for the language model component of a speech recognizer, IEEE Trans. Acoust. Speech Signal Process. 35(3) (1987) 400–401.

About the Author —WEN GAO received his Ph.D. degree in Computer Science, Harbin Institute of Technology, China, 1988 and Ph.D. in Electronics Engineering, University of Tokyo, Japan, 1991. He was a Research Fellow at Institute of Medical Electronics Engineering, the University of Tokyo, in 1992; a Visiting Professor at Robotics Institute, Carnegia Mellon University, in 1993; a Visiting Professor at MIT AI Lab, from May 1994 to December 1995. Now he is a professor of Institute of Computing Technology, Chinese Academy of Sciences. His research interests include pattern recognition and artificial intelligence, image understanding, data compression, hand gesture recognition, multimodal interface, and computer vision. He has published 7 books and over 260 scientific papers.

About the Author —GAOLIN FANG is a Ph.D. candidate at Harbin Institute of Technology. He received his MS degree in computer science from Harbin Institute of Technology, China, 2000. His research interests include: pattern recognition, multimodal human computer interaction, machine learning, and statistical language models. He has published 15 scientific papers.

About the Author —DEBIN ZHAO received his Ph.D. and MS degree in Computer Science from Harbin Institute of Technology, China, in 1998 and 1988. He was a Research Fellow with Department of Computer Science, City University of Hong Kong. And now he is a professor with Department of Computer Science, Harbin Institute of Technology. His research interests include multimedia data compression, image processing, and multimodal human-machine interface. He has published 2 books and over 50 scientific papers.

About the Author —YIQIANG CHEN received the B.S. degree and the M.S. degree in computer science from XiangTan University, China, in 1996 and 1999, respectively, and the Ph.D. degree in computer science from the Institute of Computing Technology, Chinese Academy of Science, Beijing, China, in 2003. He is currently an Assistant Researcher with the Digital Laboratory, Institute of Computing Technology. His research interests include machine learning, multimodal interfaces, and bioinformatics.