Signer-Independent Sign Language Recognition Based on SOFM/HMM

Gaolin Fang[1], Wen Gao[1, 2], Jiyong Ma[2]
[1]*Depatment of Computer Science and Engineering,*
Harbin Institute of Technology, Harbin, 150001, China
[2]*Institute of Computing Technology,*
Chinese Academy of Sciences, Beijing, 100080, China
fgl@vilab.hit.edu.cn wgao@ict.ac.cn

Abstract

The aim of sign language recognition is to provide an efficient and accurate mechanism to transcribe sign language into text or speech. State-of-the-art sign language recognition should be able to solve the signer-independent problem for practical application. In this paper, a hybrid SOFM/HMM system, which combines self-organizing feature maps (SOFMs) with hidden Markov models (HMMs), is presented for signer-independent Chinese Sign Language (CSL) recognition. We implement the SOFM/HMM sign recognition system. Meanwhile, results from the HMM-based system are provided as comparison. Experimental results show the SOFM/HMM system increases the recognition accuracy by 5% than HMM-based one. Furthermore, a self-adjusting recognition algorithm is also proposed for improving the SOFM/HMM discrimination. When it is applied to the SOFM/HMM system it can improve the recognition accuracy by 1.9%. All experiments are performed in real-time with the dictionary size 208.

1. Introduction

Sign language, as a kind of structured gesture, is one of the most natural means of exchanging information for deaf people. Sign language recognition has emerged as one of the most important research areas in the field of human-computer interaction. 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 becomes more convenient. Attempts to automatically recognize sign language began to appear in the literature in the 90's. Charaphayan and Marble [1] investigated a way using image processing to understand American Sign Language (ASL). Their system can correctly recognize 27 out of the 31 ASL symbols. Starner [2] used a view-based approach with a single camera to extract two-dimensional features as input to HMMs. The correct rate was 91% in recognizing the sentences comprised 40 signs. By imposing a strict grammar on this system, an accuracy of 97% was possible with real-time performance. Fels and Hinton [3] developed a system using a VPL DataGlove Mark II with a Polhemus tracker as input devices. In their system, the neural network method was employed for classifying hand gestures. Kadous [4] demonstrated a system based on Power Gloves to recognize a set of 95 isolated Auslan signs with 80% accuracy, with an emphasis on computationally inexpensive methods. Liang and Ouhyoung [5] used HMMs for continuous recognition of Tainwan Sign language with a vocabulary between 71 and 250 signs with Dataglove as input devices. However, their system required that gestures performed by the signer be slow to detect the word boundary. Grobel and Assan [6] used HMMs to recognize isolated signs with 91.3% accuracy out of a 262-sign vocabulary. They extracted the features from video recordings of signers wearing colored gloves. Vogler and Metaxas [7] used computer vision methods to extract the three-dimensional parameters of a signer's arm motions, coupled the computer vision methods and HMMs to recognize continuous American sign language sentences with a vocabulary of 53 signs. They modeled context-dependent HMMs to alleviate the effects of *movement epenthesis*. An accuracy of 89.9% was observed.

As the review of previous work showed, most researches on sign language recognition were done within the signer-dependent domain. For signer-independent sign language recognition, only Vamplew [8] in the literature reports a system based on CyberGlove to recognize a set of 52 signs independent of signers. The system employs a modular architecture consisting of multiple feature-recognition neural networks and a nearest-neighbour classifier to recognize isolated signs. But they use only a single glove to restrict the system to the recognition of one-handed signs.

Our previous system [9] can recognize 5177 isolated signer-dependent signs with 94.8% accuracy in real time and recognize 200 sentences with 91.4% word accuracy.

0-7695-1074-4/01 $10.00 © 2001 IEEE

It is necessary to investigate the signer-independent sign language recognition to improve the sign language recognition system robustness and practicability. However, different signers vary their hand shape size, body size, operation habit and so on, which bring about more difficulties in recognition. So recognition in signer-independent domain is more challenging than in signer-dependent one. The combination of powerful self-organizing performances of SOFMs with excellent temporal processing properties of HMMs within the novel scheme is investigated in order to improve the performance of HMM-based signer-independent sign language recognition systems. This investigation has also led to the development of a supervised learning method for updating SOFM weights. Furthermore, a self-adjusting recognition algorithm is proposed to improve the HMM discrimination after carefully analyzing the HMM probability density function (pdf). The SOFM/HMM system that employs this algorithm shows it is an effective means.

The organization of this paper is as follows. In section 2 we present the SOFM/HMM system architecture. In Section 3 we discuss the SOFM/HMM system training. In Section 4 we propose a self-adjusting recognition algorithm. In Section 5 we show the experimental results and their comparisons. The conclusion is given in the last section.

2. SOFM/HMM System Architecture

The SOFM firstly introduced by Kohonen [10] has found use in a variety of signal processing applications, especially in the speech recognition. The SOFM has shown significant potential for feature extraction in the situation where the nature of the feature of interest is not known in advance. The architecture of the SOFM is a fully connected network with two layers, and each input is connected with every output by the adjustable weight. The SOFM outputs in the form of two-dimensional lattices represented the corresponding vector centroids as the input vectors are fed into the SOFM and the weights are adjusted. This leads to that the probability density of each centroids is similar to that of the corresponding input vector.

HMMs have been proven to be one of the most successful statistical modeling methods in the area of speech recognition. It has been employed by more and more sign language recognition researchers in recent years and has produced good results. But there are some limitations of classical HMMs [11] for sign language recognition. Firstly, it is the assumption that the distributions of individual observation parameters can be well represented as a mixture of Gaussian or autoregressive densities. This assumption isn't always consistent with the fact. Secondly, HMMs have the poorer discrimination than neural networks. In the HMMs training, each word model is estimated separately using the corresponding labeled training observation sequences without considering the confused data (other models with similar behavior). However, the hybrid method combining SOFMs with HMMs is an ideal alternative. It has powerful self-organizing performance and needn't the predisposed pdf assumption and has excellent temporal processing properties.

Aiming at the first limitation of HMMs, this paper presents an alternative pdf scheme that each SOFM eigenveter centroid is regarded as one of the components in the state of HMMs. And this component forms the state pdf in term of the weighted sum. Then we can compute this state pdf by the Forward-Backward Procedure (or by the Viterbi algorithm). SOFM weights are iteratively updated in the supervision of computed state pdfs. In this way we combine the powerful self-organizing performances of SOFMs with excellent temporal processing properties of HMMs so that we improve the performance of HMMs-based sign language recognition systems. Aiming at the second limitation of HMMs, we present a novel self-adjusting recognition algorithm, which improves the SOFM/HMM discrimination by the posteriori probability modifying the current state pdfs. It will be discussed in Section 4.

Let the observation sequence, $O_t = [o_{t1}, o_{t2}, \cdots, o_{tn}]$, t=1,…,T, t is the time of observation sequence, n is the number of dimension. Each O_t is regarded as the input vector. O_t links the SOFM/HMM neuron m with the weight vector W_{jm}, where j is the state variable, $W_{jm} = [w_{jm1}, w_{jm2}, \cdots w_{jmn}]$.

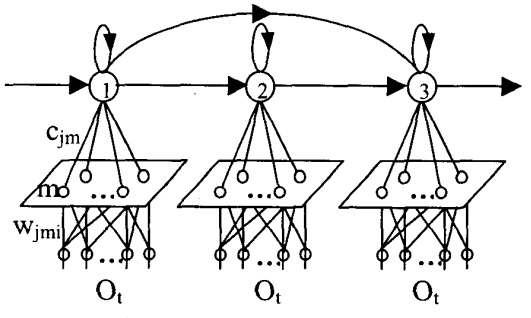

Figure 1. The architecture of SOFM/HMM

There is a 3 state left-right model with skip in Figure 1. And each state is respectively represented with 1, 2, 3. We can construct the contribution probability of being the mth neuron in state j to the state probability.

$$b_{jm}(O_t) = k \exp[-D(W_{jm}, O_t)] \quad (1)$$

Where k is a constant. The straightforward interpretation of $b_{jm}(O_t)$ is the mth neuron's contribution to the state probability, and it decreases as the observation vector deviates the corresponding neuron. $D(W_{jm}, O_t)$ is defined as the Euclidean distance between the observation O_t and the neuron m. $D(W_{jm}, O_t) = \sum_{i=1}^{n} (w_{jmi} - o_{ti})^2$.

However, the contribution varies from different neurons. We introduce the weight to solve this problem.

$$b_j(O_t) = \sum_{m=1}^{|M|} c_{jm} b_{jm}(O_t) = \sum_{m=1}^{|M|} c_{jm} * k \exp[-D(W_{jm}, O_t)] \quad (2)$$

where $\sum_{m=1}^{|M|} c_{jm} = 1$.

The weight is computed by the reestimation formula. We will discuss the detail in next Section.

3. SOFM/HMM System Training

Let the set of K observation sequences for one sign as $O = [O^{(1)}, O^{(2)}, \cdots, O^{(K)}]$, where $O^{(k)} = [O_1^{(k)} O_2^{(k)} \cdots O_{T_k}^{(k)}]$ is the kth observation sequence, T_k is the time number of the kth observation sequence. The original model λ is defined as $\lambda = (\pi, A, B)$, the reestimated model $\bar{\lambda}$ is defined as $\bar{\lambda} = (\bar{\pi}, \bar{A}, \bar{B})$. We define S as a state sequence $S = q_1, q_2, \cdots, q_{T_k}$, denote the individual state set as $\{1,2,\cdots N\}$, where N is the number of states. We denote M as the variables set of SOFM neurons, $|M|$ as the number of elements in the set, where $|M|$ in every states is the same. We assume each observation sequence is independent of every other observation sequence, and our goal is to adjust the parameters of the model λ to maximize $P(O \mid \lambda) = \prod_{k=1}^{K} P(O^{(k)} \mid \lambda)$. Since $P(O \mid \lambda)$ depends on the hidden state variables and SOFM neurons variables M it cannot be maximized directly. The MLE optimization is then solved by introducing the auxiliary function $Q(\lambda, \bar{\lambda})$, and iterating the following two steps for i =1, 2, ···:

Expectation: $Q(\lambda, \bar{\lambda}) = E[\log \prod_{k=1}^{K} P(O^{(k)} \mid \bar{\lambda}) \mid O, \lambda]$

Maximization: Updates the parameters as

$\bar{\lambda} \leftarrow \arg\max_{\hat{\lambda}} Q(\lambda, \hat{\lambda})$

$Q(\lambda, \bar{\lambda})$ can be expressed as:

$$Q(\lambda,\bar{\lambda}) = \sum_{k=1}^{K} \sum_{S} \sum_{M} \frac{P(O^{(k)}, S, M \mid \lambda)}{P(O^{(k)} \mid \lambda)} \log P(O^{(k)}, S, M \mid \bar{\lambda}) \quad (3)$$

where

$$\log P(O^{(k)}, S, M \mid \bar{\lambda}) = \log(\prod_{t=1}^{T_k} \bar{a}_{q_{t-1}q_t} \bar{b}_{q_t m_t}(O_t^{(k)}) \bar{c}_{q_t m_t})$$

$$= \log \bar{\pi}_{q_1} + \sum_{t=1}^{T_k - 1} \log \bar{a}_{q_t q_{t+1}} + \sum_{t=1}^{T_k} \log \bar{b}_{q_t m_t}(O_t^{(k)}) + \sum_{t=1}^{T_k} \log \bar{c}_{q_t m_t}$$

$$m_t \in M,\ q_t \in \{1,2,\cdots,N\}$$

Given the model λ and the kth observation sequence, we define

$$\alpha_t^{(k)}(i) = P(O_1^{(k)} O_2^{(k)} \cdots O_t^{(k)}, q_t = i \mid \lambda)$$

$$\beta_t^{(k)}(i) = P(O_{t+1}^{(k)} O_{t+2}^{(k)} \cdots O_{T_k}^{(k)} \mid q_t = i, \lambda)$$

as the forward variable and the backward variable respectively. We can compute them by Forward-Backward Procedure [11].

The probability of being in state j at time t with the mth neuron accounting for $O_t^{(k)}$ is defined as

$$\Phi_t^{(k)}(j,m) = P(q_t = j, m_t = m \mid O^{(k)}, \lambda)$$

$$= \frac{\alpha_t^{(k)}(j)\beta_t^{(k)}(j)}{\sum_{j=1}^{N} \alpha_t^{(k)}(j)\beta_t^{(k)}(j)} \cdot \frac{c_{jm} b_{jm}(O_t^{(k)})}{b_j(O_t^{(k)})} \quad (4)$$

We can maximize every item in $Q(\lambda, \bar{\lambda})$. Let $\frac{\partial Q(\lambda, \bar{\lambda})}{\partial \bar{c}_{jm}} = 0$, $\nabla_{\bar{W}_{jm}} Q(\lambda, \bar{\lambda}) = 0$, and we can get the reestimation formulas for c_{jm}, W_{jm}.

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

$$\bar{W}_{jm} = \sum_{k=1}^{K} \sum_{t=1}^{T_k} \Phi_t^{(k)}(j,m) O_t^{(k)} / \sum_{k=1}^{K} \sum_{t=1}^{T_k} \Phi_t^{(k)}(j,m) \quad (6)$$

The reestimation formulas for π_i, a_{ij} are the same as the classical HMMs. Scaling is employed in the computation of Forward-Backward variables to avoid the underflow in the programming. The detail refers to [11]. An important aspect of the reestimation procedure is that the stochastic constraints of c_{jm}, namely $\sum_{m=1}^{|M|} c_{jm} = 1$.

The training procedure of SOFM/HMM is as follows:

1. Initialize the parameter of $\pi_i, a_{ij}, c_{jm}, W_{jm}$.
2. Reestimate the parameters with the reestimation formulas and all observation sequences for the corresponding word.
3. Terminate the procedure, if the convergence criterion is met, and the parameters are the model of this word; otherwise replace the old parameters with the new ones, and return 2.

4. Self-adjusting recognition algorithm

In the HMM-based recognition, we employ the Bayesian decision based on the minimum error rate to get the recognition result. The Bayesian decision is composed of the rules that can minimize $P(e)$ (the mean error rate). This paper presents a self-adjusting recognition algorithm, which modifies a class original pdf with the posteriori probability of this class in the whole set. For simplicity we illustrate that this algorithm can reduce the error rate in the one-dimensional and two-class condition. We can extend this result to the multi-dimensional and multi-class condition as the same rationale.

$P(e)$ is defined as: $P(e) = \int_{-\infty}^{\infty} P(e \mid x)p(x)dx$,

where $p(x)$ is the pdf.

As to the two-class problem:

$$P(e \mid x) = \begin{cases} P(w_1 \mid x), & if\ P(w_2 \mid x) > P(w_1 \mid x) \\ P(w_2 \mid x), & if\ P(w_1 \mid x) > P(w_2 \mid x) \end{cases} \quad (7)$$

Let t is the interface between two classes. When vector x is one dimension, t is a point in the axis x.

$$P(e) = \int_{-\infty}^{t} P(w_2 \mid x)p(x)dx + \int_{t}^{\infty} P(w_1 \mid x)p(x)dx$$
$$= \int_{-\infty}^{t} p(x \mid w_2)P(w_2)dx + \int_{t}^{\infty} p(x \mid w_1)P(w_1)dx \quad (8)$$

In Figure 2, $P(e)$ is the intersectant area of $p(x \mid w_1)P(w_1)$, $p(x \mid w_2)P(w_2)$. Now we consider how to reduce this area?

We construct a function $\eta(w,x) = \dfrac{p(x \mid w)}{\sum_{i \in W} p(x \mid i)}$,

$\{x \mid p(x \mid w) > \delta\}$, where W the set of all classes, δ is the very small critical value and the values that are smaller than it are ignored. The straightforward interpretation of $\eta(w,x)$ is that it is a posteriori probability of the w class in the whole set. We can transfer the original pdf into the new pdf by scaling $\eta(w,x)$. The new pdf is defined as $\hat{p}(x \mid w)$

$$\hat{p}(x \mid w) = \eta(w,x)p(x \mid w) \quad (9)$$

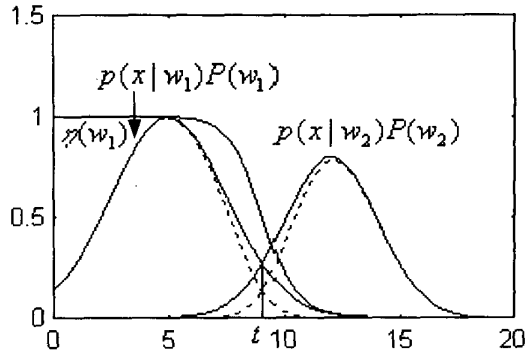

Figure 2. The comparison of error rate

Out of the intersectant area of pdfs, $\eta(w,x) = 1$, the pdf will retain the same; in the intersectant area, $\eta(w,x) < 1$, the pdf will converge to the respective centroid. Thus the intersectant area will shrink and the mean error rate will decrease. $\hat{p}(x \mid w_1)P(w_1)$, $\hat{p}(x \mid w_2)P(w_2)$ are respectively represented with the dashed in Figure 2. The new mean error rate $\hat{P}(e)$ is the intersectant area of two dashed. Compared with $P(e)$, $\hat{P}(e)$ is clearly reduced.

Let $p(x \mid w) = b_{wj}(O_t)$, $\eta(w,O_t) = \dfrac{b_{wj}(O_t)}{\sum_{i \in W} b_{ij}(O_t)}$

$\{O_t \mid b_{wj}(O_t) > \delta\}$, where each word is regarded as one class. $b_{wj}(O_t)$ is the pdf of being in wth words model in state j. From the analysis above we can reduce the mean error rate by scaling $\eta(w,O_t)$. The formula is as follows:

$$\hat{b}_{wj}(O_t) = \eta(w,O_t)b_{wj}(O_t) \quad (10)$$

The recognition procedure with the self-adjusting algorithm is as follows:

1. For the observation sequence $O = O_1O_2 \ldots O_T$, we compute $b_{wj}(O_t)$ in all words.
2. Compute $\hat{b}_{wj}(O_t)$ by the $b_{wj}(O_t)$.
3. Decode with Viterbi algorithm in term of $\hat{b}_{wj}(O_t)$.
4. The result is the word that has the maximum probability of decoding among all words.

5. Experiments and Comparisons

Input: We use two CyberGloves and three Pohelmus 3SPACE-position trackers as input devices. Two trackers

are positioned on the wrist of each hand and another is fixed at back (the reference tracker). The CyberGloves collect the variation information of hand shapes with the 18-dimensional data at each hand, and the position trackers collect the variation information of orientation, position, movement trajectory.

Data processing: Data from position trackers can be converted as follows. The reference Cartesian coordinate system of the trackers at back is chosen, and then the position and orientation at each hand with respect to the reference Cartesian coordinate system are calculated as invariant features. Through this transformation, the data are composed of a relative three-dimensional position vector and a three-dimensional orientation vector for each hand. Furthermore, we calibrate the data of different signers by some fixed postures because everyone varies his hand shape size, body size, and operation habit. For two hands, they formed a 48-dimensional vector in total. However, the dynamic range of each component is different. Each component value is normalized to ensure its dynamic range is 0-1.

Experiments: The data are collected from 7 signers with each performing 208 isolated signs 3 times. The vocabulary is the words from the elementary textbooks of 1-2 grades for Chinese deaf pupil. We select 5 from 7 signers, which are regarded as the registered signers. The rest two are referred to as the unregistered signers. Each in the registered signers contributes to two group data as training samples (in total 10 groups). Five group data from the rest one in registered signers are referred to as the registered test set (Reg.). The samples from the unregistered signers are referred to as the unregistered test set (Unreg.). Test samples are performed in real time, that is, collection and recognition are parallel without distinct delay.

Table 1. The comparison of different results

Signer		HMMs	SOFM/ HMM	Self-adjusting
Reg.	A	95.2%	98.6%	99.5%
	B	88.0%	95.2%	97.1%
	C	95.7%	96.7%	97.6%
	D	89.4%	91.3%	93.3%
	E	85.1%	94.7%	95.7%
	Mean	90.7%	95.3%	96.6%
Unreg.	F	82.2%	88.5%	90.4%
	G	84.1%	88.0%	89.9%
	Mean	83.2%	88.2%	90.1%

Table 1 reports respectively test results of HMMs, SOFM/HMM and SOFM/HMM with the self-adjusting recognition algorithm (Self-adjusting), where HMMs have 3 states and 5 mixture components, and SOFM/HMM has 3 states and 5 initial SOFM neurons. 90.7%, 95.3%, 96.6% of mean recognition rates are respectively observed in Reg., and 83.2%, 88.2%, 90.1% in Unreg.

Experimental results show that SOFM/HMM increases the recognition accuracy by 4.6% than HMMs in the registered test, 5% in the unregistered test. When the self-adjusting algorithm is applied to the SOFM/HMM system, the results show it increases the recognition accuracy by 1.3% in the registered test, 1.9% in the unregistered test.

6. Conclusions

This paper presents the SOFM/HMM, which classical HMMs and SOFMs are combined within novel scheme, for the signer-independent CSL recognition. In SOFM/HMM architecture we introduce the training and recognition procedure. We implement the signer-independent sign recognition system with the SOFM/HMM, which has 96.6% recognition rates in Reg. and 90.1% in Unreg. The experiments show the SOFM/HMM system increases the recognition accuracy by 5% than HMM-based one. Furthermore, a self-adjusted recognition algorithm is illustrated how to improve the SOFM/HMM discrimination in the two-class condition. The experiments show this algorithm can improve the recognition accuracy by 1.9% when it is applied to the SOFM/HMM signer-independent CSL recognition system.

Acknowledgment

This work has been supported by National Science Foundation of China (contract number 69789301), National Hi-Tech Development Program of China (contract number 863-306-ZD03-01-2), and 100 Outstanding Scientist foundation of Chinese Academy of Sciences.

References

[1] C. Charayaphan and A. Marble, "Image processing system for interpreting motion in American Sign Language", *Journal of Biomedical Engineering*, 1992, 14, pp. 419-425.

[2] T. Starner and A. Pentland, "Visual Recognition of American Sign Language Using Hidden Markov Models", *International Workshop on Automatic Face and Gesture Recognition*, Zurich, Switzerland, 1995, pp. 189-194.

[3] S. S. Fels and G. Hinton, "GloveTalk: A neural network interface between a DataDlove and a speech synthesizer", *IEEE Transactions on Neural Networks*, 1993, Vol. 4, pp. 2-8.

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

[5] R. H. Liang and M. Ouhyoung, "A real-time continuous gesture recognition system for sign language", *In Proceeding of*

the Third International Conference on Automatic Face and Gesture Recognition, Nara, Japan, 1998, pp.558-565.
[6] K. Grobel and M. Assan, "Isolated sign language recognition using hidden Markov models", *In Proceedings of the International Conference of System, Man and Cybernetics*, 1996, pp. 162-167.
[7] C. Vogler and 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*, Orlando, 1997, pp.156-161.
[8] P. Vamplew, "Recognition of sign language gestures using neural networks", *The 1st European Conference on Disability, Virtual Reality and Associated Technologies*. 1996.
[9] W. Gao, J. M. Ma, et al., "HandTalker: A Multimodal Dialog System Using Sign Language and 3-D Virtual Human", *Advances in Multimodal Interfaces-ICMI 2000*, pp 564-571.
[10] T. Kohonen, "The Self-Organizing Maps", *Proceedings of the IEEE*, 1990, vol. 78, no. 9, pp 1464-80.
[11] R. Rabiner, (1989), A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition, *Proceedings of the IEEE*, 1989,Vol. *77*, No. 2, pp.257-285.