Large Vocabulary Sign Language Recognition Based on Hierarchical Decision Trees

Gaolin Fang
Department of Computer Science and Engineering,
Harbin Institute of Technology
Harbin, 150001, China
glfang@jdl.ac.cn

Wen Gao
Institute of Computing Technology,
Chinese Academy of Sciences
Beijing, 100080, China
wgao@jdl.ac.cn

Debin Zhao
Department of Computer Science and Engineering,
Harbin Institute of Technology,
Harbin, 150001, China
dbzhao@jdl.ac.cn

ABSTRACT
The major difficulty for large vocabulary sign language or gesture recognition lies in the huge search space due to a variety of recognized classes. How to reduce the recognition time without loss of accuracy is a challenge issue. In this paper, a hierarchical decision tree is first presented for large vocabulary sign language recognition based on the divide-and-conquer principle. As each sign feature has the different importance to gestures, the corresponding classifiers are proposed for the hierarchical decision to gesture attributes. One- or two- handed classifier with little computational cost is first used to eliminate many impossible candidates. The subsequent hand shape classifier is performed on the possible candidate space. SOFM/HMM classifier is employed to get the final results at the last non-leaf nodes that only include few candidates. Experimental results on a large vocabulary of 5113-signs show that the proposed method drastically reduces the recognition time by 11 times and also improves the recognition rate about 0.95% over single SOFM/HMM.

Categories and Subject Descriptors
I.5.2 [**Pattern Recognition**]: Design Methodology – *Classifier design and evaluation*; H.1.2 [Models and Principles]: User/Machine Systems – *Human information processing.*

General Terms
Algorithms, Design, Experimentation, Human Factors, Languages.

Keywords
Sign language recognition, gesture recognition, hierarchical decision tree, Gaussian mixture model, finite state machine.

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 (SLR) 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, 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 of 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.

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. Attempts to automatically recognize sign language began to appear in the literature in the 90's. The recognition methods usually include rule-based matching, artificial neural networks, and hidden Markov models (HMM).

Kadous [1] demonstrated a system based on Powergloves to recognize a set of 95 isolated Australian sign language with 80% accuracy. Instance-based learning and decision-tree learning were adopted by the system to produce the rules of pattern. Matsuo et al. [2] 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 get sign language patterns.

Fels and Hinton [3] developed a system using a Dataglove with a Polhemus tracker as input devices. In their system, five neural networks were employed for classifying 203 gestures. Kim et al. [4] 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.

Grobel and Assan [5] 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. HMM was also employed by Hienz and Bauer [6] to recognize continuous German sign language with a single color video camera as input. Their research was an extension of the work by Grobel and Assan. An accuracy of 91.7% can be achieved in recognition of sign language sentences with 97 signs.

Liang and Ouhyoung [7] 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

ICMI'03, November 5–7, 2003, Vancouver, British Columbia, Canada.

recognition rate of 80.4% for 250 signs. In their system a Dataglove was used as input device and HMM was taken as recognition method.

Starner et al. [8] used a view-based approach for continuous American SLR. They used single camera to extract two-dimensional features and the extracted features are then taken as the input of HMM. The word accuracy of 92% or 98% was gotten when the camera was mounted on the desk or in a user's cap in recognizing the sentences with 40 different signs.

Vogler and Metaxas [9] 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. 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 [10], [11].

From the review above, we know that most researchers focus on small or medium vocabulary SLR in the signer-dependent domain. For large vocabulary sign recognition, the major difficulty lies in the huge search space due to a variety of recognized classes. How to reduce the recognition time without loss of accuracy is a challenging issue. In speech recognition, phoneme-based method was generally employed to tackle large vocabulary problem. However, there is no basic unit defined in the sign's lexical forms. The phonemes extracted manually or automatically were experimented on the small vocabulary. It is very difficult to extend these phonemes to act as the basic unit of whole sign language. Gao [12], [13] used Datagloves as input devices and HMM as recognition method. The system can recognize 5177 isolated signs with 94.8% accuracy in real time and recognize 200 sentences with 91.4% word accuracy in the signer-dependent domain. The state-tying HMM with one mixture component was employed to overcome the time-consuming problem due to the large vocabulary size. However, when this method was applied to signer-independent SLR, the recognition performance distinctly decreased.

To overcome the difficulty from the large vocabulary size, a hierarchical decision tree is presented for sign language recognition in this paper based on the divide-and-conquer principle. As each sign feature has the different importance to gestures, the corresponding classifiers are proposed for the hierarchical decision to sign language attributes. One- or two-handed classifier with little computational cost is first used to eliminate many impossible candidates. The subsequent hand shape classifier is performed on the possible candidate space. SOFM/HMM classifier as a special component of hierarchical decision tree is employed to get the final results at the last non-leaf nodes that only include few candidates. To alleviate the effect of crisp classification errors, fuzzification is introduced in the decision tree, i.e., the classes that cannot be robustly classified will not be handled at this classifier, and they simultaneously enter next level for further decision. Experimental results show that the proposed method can drastically reduce the recognition time and also improve the recognition performance over single SOFM/HMM.

The remainder of this paper is organized as follows. In Section 2, we analyze sign language features. Section 3 proposes the feature classifiers in the hierarchical decision tree. In Section 4, the hierarchical decision tree for sign language recognition is presented. Section 5 shows experimental results. The conclusions are given in the last section.

2. SIGN LANGUAGE FEATURES

According to Stokoe's definition [14], each sign can be broken into four parameters: hand shape, orientation, position and motion. These parameters as four important features play an important role in sign language recognition. Furthermore, according to the number of participating hands in the sign performance, sign language can be divided into two categories: one-handed signs and two-handed signs. Thus, one- or two-handed is also one of the important features of sign language. Five features are respectively detailed as follows:

Hand shapes are one of the primitives of sign language and reflect the information of hand configuration. They are very stable and can be used to distinguish most signs. In the Chinese sign language dictionary, there are 75 basic hand shapes extracted by the sign language expert.

The orientation of the hand can be described in terms of two orthogonal directions - the facing of the palm, and the direction to which the hand is pointing. If we consider only six possible directions (up, down, left, right, towards the signer, away from the signer), then there are 15 different orientations used in Chinese sign language (CSL).

The position of the hand is usually partitioned in terms of the signer's hand relative to the defined three parts of his body: head, chest and below chest. In each part, the position can be further subdivided into body's left, right and middle. In total, there are 12 positions defined in CSL according to the hand with respect to the body part.

Motion differs from the other features in that it is inherently temporal in nature. It is difficult to enumerate the complete range of possible categories used within CSL, as many signs involve unique tracing motions which indicate the shape of an object. For this research only the 13 most commonly used motions were defined.

In the Chinese sign language dictionary, one-handed signs are always performed by right hand except for one sign "luo ma ni ya" by left hand. The difference between one-handed sign and two-handed sign is whether signer's left hand participates the action. In the one-handed sign performance, signer's left hand usually puts on the left knee and remains motionless. However, in the two-handed sign, left hand may either stay a fixed posture or perform a movement trajectory. The position and orientation information of left hand plays a dominant part in determining one- or two- handed signs.

3. FEATURE CLASSIFIERS

In this section, on the basis of the analysis of all the sign language features, Gaussian mixture model is first employed as one- or two-handed classifier, and then the finite state machine based method is proposed as hand shape classifier. At last, SOFM/HMM classifier as a special component of hierarchical decision tree is presented for tackling the signer-independent difficulties.

3.1 One- or Two- Handed Classifier

Gaussian mixture model (GMM) is in essence one of the multivariate probability density functions. It has been successfully used as a classifier in a variety of applications. According to the estimation, one-handed sign and two-handed sign probability distributions can be approximately described by the GMM. The estimation process is designed as follows: in the training set, one- and two- handed signs are manually split, and then we calculate the frequency distributions of one- and two- handed signs at each point for every dimension of the vector. Figure 1 shows the frequency distributions of one- and two- handed signs in the x-component value of the left hand position vector, where the value of x-axis is the result of the x-component value [0, 1] multiplied by 25, and the value of y-axis is the result of the number of the estimated signs with this x-axis value divided by the total number of one- or two- handed signs. The similar distributions are observed in other components of the left hand position and orientation vectors. From the curves of statistical results, we can speculate that one-handed sign and two-handed sign probability density can be approximated by the respective GMM. Therefore, in this paper, GMM is employed to determine whether a gesture is represented by one hand or two hands.

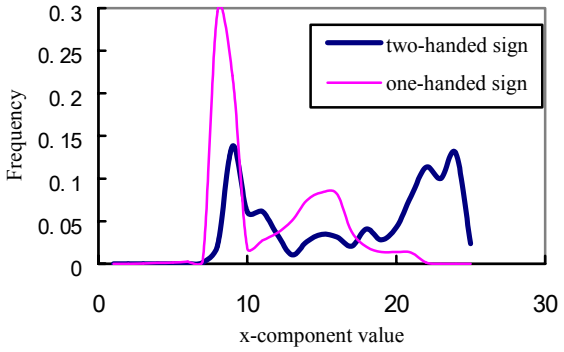

Figure 1. The frequency distributions of one- and two- handed signs in the x-component value of the left hand position vector

GMM can be described by the mixture parameter, the mean vector and the covariance matrice, and formulated as $\lambda = \{\pi, \mu, \Sigma\}$. The probability density function of an observation x is represented by the linear combination of Gaussian density: $p(x \mid \lambda) = \sum_{i=1}^{M} \pi_i p(x \mid i)$, where x is the observed vector of d dimensions, M is the number of mixture term, and π_i is the mixing parameter and satisfies the constraint $\sum_{i=1}^{M} \pi_i = 1$.

$p(x \mid i)$ is the Gaussian probability density function of d dimensions: $p(x \mid i) = \dfrac{\exp\{-\frac{1}{2}(x-u_i)^T \Sigma_i^{-1}(x-u_i)\}}{(2\pi)^{d/2} \mid \Sigma_i \mid^{1/2}}$, where u_i is the mean vector, and Σ_i is the covariance matrice.

The parameter λ can be trained using the Expectation-Maximization (EM) algorithm. There are two key issues in the model training: initialization and the selection of training samples.

Initialization. In the EM algorithm, the parameter λ need to be initialized. However, no general theoretical framework but empirical or experimental appoach is employed to solve this problem. Here, k-means clustering is used to get the mean and convariance of centriods as the initialization values. π_i^0 is initialized to $1/M$. The mixture term M is determined according to the distribution of training data and the classification accuracy of one- or two- handed signs. In our classification experiments, M is set to the values from 5 to 30. From the experiments, the classification performance grows with the mixture term M. When the value is greater than 15, the classification performance doesn't improve or even slightly decline but the classification time increases. Thus, M in the one-handed GMM and two-handed GMM are both set to 15, where the mixture terms are set to the same for the comparability of their probabilities.

The selection of training samples. How to select typical samples for training one-handed and two-handed GMM is a difficult issue. Different training data will produce different results, that is, too many data will make the model training difficult to converge, and not enough data will train the model that cann't be well generalized. Through the experiments, the following strategy is taken. For a one-handed sign, left hand stays motionless and its data are very stable, so the stablest frame is extracted from all the frames of this word. For a two-handed sign, if left hand is in motion, then all the data frames are extracted as the training data. If left hand is motionless, the extraction method is the same as one-handed signs. In all those training data, only the position and orientation information of left hand is used for classification.

Classification: Given the frame sequence for one sign $O = o_1 o_2 \cdots o_T$, the probabilities of belonging to one-handed and two-handed signs for each frame are calculated with the trained GMM. The probabilities can be expressed as $P_i(o_j)$, $i = 1,2$, where 1 denotes the one-handed sign and 2 for the two-handed sign. The classes can be gotten through the following formula: $i^* = \arg\max_{i \in \{1,2\}} (\sum_{j=1}^{T} \delta_i(o_j))$, $\delta_k(o_j) = \begin{cases} 1 & if \ \ k = \arg\max_{i \in \{1,2\}} P_i(o_j) \\ 0 & otherwise \end{cases}$.

After all the training samples are classified using above method, the candidate words associated with one handed sign and two handed sign are generated, which will be used by the following hand shape classifier.

3.2 Hand Shape Classifier

Lee et al. [15] used finite state machine (FSM) to segment the motion of Korean sign language. Hong et al. [16] also used FSM to recognize gestures, where each state is modeled as a multivariate Gaussian function. A gesture can be described as an ordered sequence of hand shape states in spatial-temporal space and well modeled by a finite state machine, whose states consist of 75 basic hand shapes. The structure of an FSM is like that of an HMM, where each state can jump to either itself or its next state. FSM has the advantages of easy interpretation and faster classification rate, and can well solve the different frame alignment for the same sign. Furthermore, hand shape is very

stable in all the features of sign language and it plays a very important role in distinguishing most signs due to its distinct feature discrimination, so it is feasible to use it as a classifier of the hierarchical decision tree. Thus, in this paper, FSM-based method is proposed for hand shape classifier, which is regarded as part of the hierarchical decision tree.

The training algorithm of FSM-based hand shape classifier is described as follows.

1) Clustering. Basic hand shapes extracted by the experts from the dictionary of sign language are regarded as initial centroids, and then the k-means clustering algorithm is employed to get new centroids in the training set.
2) Fuzzy vector quantization (FVQ). Fuzzy N-best results are outputted at each frame with new centroids. For utilizing the context information of sign frames to supervise the quantization, the Viterbi algorithm is employed to get the best vector quantization sequence on the fuzzy N-best results.
3) Pattern extraction and pruning. The word patterns are extracted from the quantization results. Pattern to word is a many-to-many map, where the patterns are regarded as the classification criterion. For the better generalization ability, simple pruning is operated on the extracted patterns.
4) Candidate word generation. Training samples are classified using FSM, and each branch in the classifier denotes one pattern. After all the samples in the training set are handled, the candidate word set associated with each pattern is generated, which will be used by SOFM/HMM classifier.

To eliminate the effect of noise and make the extracted patterns have better generalization ability, two key techniques are employed in this algorithm: fuzzy vector quantization, and pattern extraction and pruning.

3.2.1 Fuzzy vector quantization

In the vector quantization, every frame data is independent of each other, so the noise will have the direct influence on the quantization results. However, a gesture consists of several basic hand shapes. The changes between hand shapes are very stable, that is, slow changes from one series of hand shapes to another series of hand shapes. This context information can be utilized to supervise the quantization through FVQ so that the algorithm can reduce the effect of noise and get the well-generalized patterns.

Given the frame sequence for one sign $O = o_1 o_2 \cdots o_T$, where T denotes the frame number. Define $V = v_1 v_2 \cdots v_T$, $v_t \in \{1,2,\cdots,75\}$ as one of the quantization sequence. In FVQ, the vector o_t is quantized as not only the top scoring output but the N top scoring outputs. The corresponding probability is associated with the N top outputs, where $N = 3$. The Viterbi algorithm is employed to get the best vector quantization sequence among all the results, and formulated as: $V^* = \arg\max_V b_{v_1}(o_1) \prod_{t=2}^{T} a_{v_{t-1}v_t} b_{v_t}(o_t)$.

Transition probability between the quantization results v_{t-1}, v_t is defined as follows: $a_{v_{t-1}v_t} = \begin{cases} 1 & v_{t-1} = v_t \\ 0.1 & v_{t-1} \neq v_t \end{cases}$, where the values of 1 and 0.1 are manually set through the experiments.

Emission probability of the frame o_t being quantized as v_t is defined: $b_{v_t}(o_t) = \frac{\exp(-d_{v_t}(o_t))}{\sum_v \exp(-d_v(o_t))}$, where $d_{v_t}(o_t)$ denotes the Euclidean distance between the vector o_t and the centroid v_t.

Through this method, the context information of sign frames is fully utilized to supervise the vector quantization, so we can alleviate the effect of the noise data and get the consistent quantization results.

3.2.2 Pattern extraction and pruning

Pattern extraction is performed as follows. After the previous step – fuzzy vector quantization – processing, the quantized sequences are so regular that the classification patterns can be directly extracted from the quantization results according to the duration of hand shape. If the duration is greater than four frames, then this hand shape is regarded as one of the pattern states, otherwise regarded as the noise and discarded.

However, the extracted pattern number is very large so that the generalization of patterns becomes delicate and the corresponding classification deviation is getting large. To solve this problem, the pruning is performed on those patterns. From the data of FVQ, if all the patterns are kept, the total number of the classification template for all the vocabulary is about 1860. If the first three hand-shape states are kept, the number is about 1340. If the first two hand-shape states are kept, the number is about 450. If the first hand-shape state is kept, the number is about 75. For making the extracted patterns have better generalization, the number of classification pattern cannot get too big. Compromising the generalization and the classification time, the first two hand shape states are kept, which is in accord with the fact that most of the signs consist of two hand shape states. For example, three quantized word sequences: 1 1 1 1 1 1 2 2 2 2 2 3 3 3 3 3 3 4 4 4 4, 1 1 1 1 1 2 2 2 2 2 2 3 3 3 3 3 3 3 and 1 1 1 1 1 1 1 2 2 2 2 2 4 4 4 4 4 4 are respectively extracted as the patterns 1 2 3 4, 1 2 3 and 1 2 4 (see in Figure 2, where 1, 2, 3 and 4 denote basic hand shape state). They will be pruned as one pattern 1 2, which will be used as the classification template. For the words with the long frame data, the pruning can reduce the classification time, because only previous parts are used to distinguish rather than the whole sequence. After the pruning, the patterns are regarded as the classification templates of FSM.

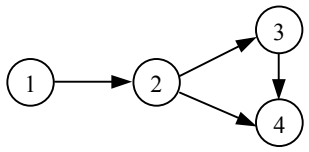

Figure 2. The examples of hand shape pattern

Classification: Input data are processed by fuzzy vector quantization, pattern extraction and pruning, then classified into the corresponding pattern branch through FSM.

Similar methods are experimented with position, orientation and motion features. However, since the information of these features is not very stable, the extracted patterns for the same sign are not very consistent. Though the recognition time is reduced, the

performance cannot be improved. Thus, one- or two- handed, left hand shape, and right hand shape are chosen as three attributes of the hierarchical decision tree.

3.3 SOFM/HMM Classifier

Aiming at the two difficulties of signer-independent SLR—the model convergence difficulty caused by mass data and noticeable distinctions among different people data, and the lack of effective features extracted from different signers' data, SOFM/HMM classifier is presented in the paper [17].

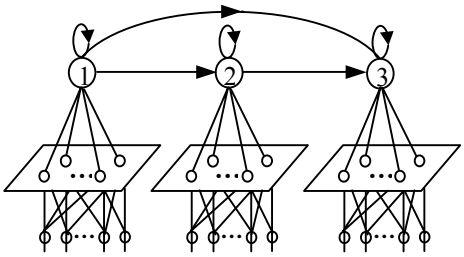

Figure 3. The architecture of SOFM/HMM

The proposed method uses the self-organizing feature maps(SOFM) as an implicit different signers' feature extractor for continuous HMM and its parameters are trained simultaneously with a global optimization criterion. SOFM transforms input sign representations into significant and low-dimensional representations that can be well modeled by the emission probabilities of HMM. Figure 3 shows its architecture.

4. HIERARCHICAL DECISION TREES FOR SIGN LANGUAGE RECOGNITION

A general decision tree [18], [19], [20] consists of root nodes, non-leaf nodes and leaf nodes, where each leaf node denotes a class. The input data include the values of different attributes and these values are initially put in the root node. By asking questions about the attributes, the decision tree splits the values into different child nodes. At last, which class the input data belongs to is decided at the leaf node. Decision trees are, by their nature, readily interpretable and well-suited to classification problems. They are also remarkable for their ability to combine diverse information sources.

The hierarchical decision tree for sign language recognition is constructed as follows: 1) In the training set, all the training samples are classified using GMM, the candidate words associated with one-handed sign and two-handed sign are generated. Those words cannot be robustly classified will appear both in the candidate words of one-handed signs and in the candidate words of two-handed signs. 2) For the candidate words of one-handed signs, their training samples are inputted into the right hand shape classifier. After all the training samples are classified using FSM, the candidate words associated with each pattern are generated. Those candidates with common elements will be used as the candidate words of SOFM/HMM classifier. 3) For the candidate words of two-handed signs, the processing is the same as Step 2, first into left hand shape classifier, and then into right hand shape classifier. However, the classification results of left shape are used as the candidate words of right hand shape classifier, and the classification results of right shape are regarded as the candidate words of SOFM/HMM classifier.

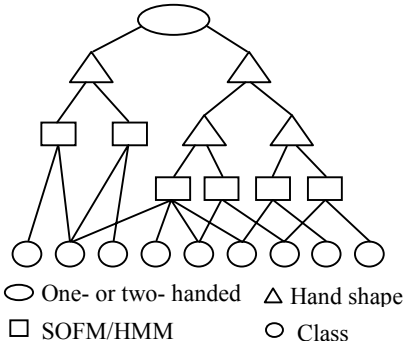

Figure 4. The hierarchical decision tree for sign language recognition

After the hierarchical decision tree for sign language recognition is constructed, the architecture of the hierarchical decision tree and the candidates associated with every node are produced. Figure 4 illustrates the diagram of the hierarchical decision tree for sign language recognition, where each non-leaf node denotes a classifier associated with the corresponding word candidates, and each branch at a node represents one class of this classifier. There are common elements among adjacent branches under one node, and their intersection is learned by a large amount of training samples. The input data are first fed into one- or two-handed classifier, then into left hand shape classifier and right hand shape classifier (no left hand shape classifier for the one-handed sign branch), and at last into the SOFM/HMM classifier only with few candidates in which the final recognition results are gotten. To illustrate the idea more concretely, consider the sign "ai-hu" with a vector sequence of 48-dimensional features. The sign is firstly fed into GMM with its 6-dimensional feature sequence of left hand position and orientation information for judging one- or two- handed sign, and then into FSM with its 18-dimensional feature sequence of left hand shape information for classifying into one of the branches, and along this branch into FSM with its 18-dimensional feature sequence of right hand shape information for classifying, along the assigned branch into SOFM/HMM with its 48-dimensional features for decision-making, and the final classes are gotten among the candidates of this SOFM/HMM node.

5. EXPERIMENTS

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

In order to extract the invariant features to signer's position, the tracker at signer's back is chosen as the reference Cartesian coordinate system, and the position and orientation at each hand with respect to the reference system are calculated and can be

taken as invariant features. By this transformation, the data are composed of a relative three-dimensional position vector and a three-dimensional orientation vector for each hand, which don't change with the signer position and orientation. In the case of two hands, a 48-dimensional vector is formed, including the hand shape, position and orientation vector. The data from different signers are calibrated by some fixed postures performed by each signer. In our experiments the 14 postures that can represent the min-max value ranges of the corresponding sensor are defined. As each component in the vector has different dynamic range, its value is normalized to [0,1].

All experiments were carried on the large vocabulary with 5113 signs. Experimental data consist of 61356 samples over 5113 signs from 6 signers with each performing signs twice. The vocabulary is taken from the Chinese sign language dictionary. One group data from 6 signers represented by A-F are referred to as the registered test set (Reg) and the other 11 group data are used as the training samples. Using the approach of cross validation test, 10 group data samples from 5 signers are used as the training samples and the other signer data represented by A-F are referred to as the unregistered test set (Unreg).

The first experiment is to test the recognition performances on large vocabulary signer-independent SLR respectively with HMM, SOFM/HMM and decision tree. SOFM/HMM is special case of hierarchical decision tree, that is, only SOFM/HMM classifier is used to recognize sign language.

Figure 5 shows the test results of HMM, SOFM/HMM and decision tree, where solid lines and dashed lines respectively denote the results of Reg and UnReg. HMM have 3 states and 5 mixture components and SOFM/HMM has 3 states and 5 initial SOFM neurons. The average recognition rates of 87.3% for HMM, 90.5% for SOFM/HMM and 91.6% for decision tree are observed for Reg. For UnReg, the average recognition rates of 80.0%, 82.9% and 83.7% are obtained, respectively. SOFM/HMM is more suited for signer-independent SLR because SOFM implicitly extracts the different signers' features. Therefore, SOFM/HMM has better performance than conventional HMM.

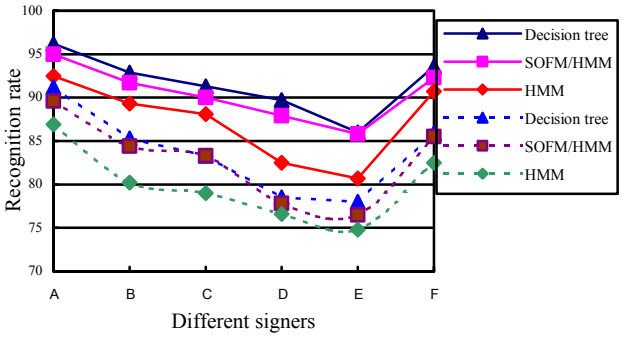

Figure 5. The recognition results of different methods

On basis of SOFM/HMM, hierarchical tree increases the recognition accuracy by 1.1% on the registered test set and by 0.8% on the unregistered test set in the experiments. This may be due to the following reasons. First, in an integrated hierarchical decision tree framework, different features can be further researched individually, and their discriminations can be fully utilized through different feature classifiers. Second, fuzzy classification alleviates the loss of crisp classification of decision tree through allowing the partitions with common elements.

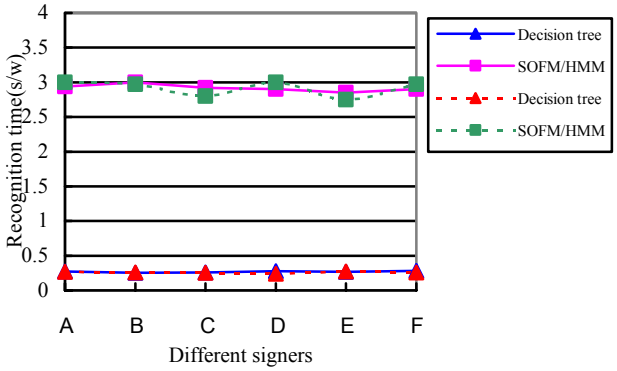

Figure 6. The recognition time of different methods

The second experiment is to test the recognition time on large vocabulary signer-independent SLR with SOFM/HMM and decision tree. The approach of cross validation test is employed both in the registered test set and in the unregistered test set. All experiments are performed on the PIV1600 (512M Memory) PC.

Figure 6 shows the recognition time of SOFM/HMM and decision tree on the vocabulary of 5113 signs, where solid lines and dashed lines respectively denote the results of Reg and UnReg. We define s/w as second per word. For the registered test set, the average recognition time of 2.922 (s/w) for SOFM/HMM and 0.268 (s/w) for decision tree are observed. For the unregistered test set, the average recognition time of 2.910 (s/w), 0.258 (s/w) are obtained, respectively. The average recognition time of SOFM/HMM and decision tree are respectively 2.916 second per word and 0.263 second per word in the registered and unregistered test sets. Experiments illustrate that hierarchical tree dramatically reduces the recognition time by 11 times over single SOFM/HMM. In the hierarchical decision tree, the feature classifiers of one- or two-handed and hand shape with little computational cost are first employed to eliminate the impossible candidates, and then the complex classifier of SOFM/HMM is performed on the previous candidates. Thus, this coarse-to-fine hierarchical decision leads to the dramatic reduction of computational complexity and recognition time.

6. CONCLUSIONS

In this paper, a hierarchical decision tree is first presented for 5113-gesture vocabulary SLR in signer-independent field. As each sign feature has the different importance to gestures, GMM-based classifier and FSM-based classifier are respectively proposed for the features of one- or two- handed and hand shape to eliminate the impossible candidates. SOFM/HMM classifier is employed to get the final results at the last non-leaf nodes that only include few candidates. Experimental results show hierarchical decision tree has an average recognition rate of 91.6% on the registered test set and 83.7% on the unregistered test set over a 5113-sign vocabulary. The average recognition time is 0.263 second per word. Experiments also show the proposed

method drastically reduces recognition time by 11 times and also improves the recognition rate about 0.95% over single SOFM/HMM. Future work will focus on large vocabulary continuous SLR.

7. REFERENCES

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

[2] H. Matsuo, S. Igi, S. Lu, Y. Nagashima, Y. Takata, and T. Teshima, The recognition algorithm with non-contact for Japanese sign language using morphological analysis, Proc. Int'l Gesture Workshop, pp. 273-284, 1997.

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

[4] J.S. Kim, W. Jang, and Z. Bien, A dynamic gesture recognition system for the Korean sign language (KSL), IEEE Trans. Systems, Man, and Cybernetics, vol. 26, no. 2, pp. 354-359, 1996.

[5] K. Grobel and M. Assan, Isolated sign language recognition using hidden Markov models, Proc. Int'l Conf. System, Man and Cybernetics, pp. 162-167, 1997.

[6] B. Bauer and H. Hienz, Relevant features for video-based continuous sign language recognition, Proc. Fourth Int'l Conf. Automatic Face and Gesture Recognition, pp. 440-445, 2000.

[7] R.H. Liang and M. Ouhyoung, A real-time continuous gesture recognition system for sign language, Proc. Third Int'l Conf. Automatic Face and Gesture Recognition, pp. 558-565, 1998.

[8] T. Starner, J. Weaver, and A. Pentland, Real-time American sign language recognition using desk and wearable computer based video, IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 20, no. 12, pp. 1371-1375, 1998.

[9] C. Vogler and D. Metaxas, Adapting hidden Markov models for ASL recognition by using three-dimensional computer vision methods, Proc. IEEE Int'l Conf. Systems, Man and Cybernetics, pp. 156–161, 1997.

[10] C. Vogler and D. Metaxas, Toward scalability in ASL recognition: breaking down signs into phonemes, Proc. Int'l Gesture Workshop, pp. 400-404, 1999.

[11] C. Vogler and D. Metaxas, A framework for recognizing the simultaneous aspects of American sign language, Computer Vision and Image Understanding, vol. 81, no. 3, pp. 358-384, 2001.

[12] W. Gao, J.Y. Ma, J.Q. Wu, and C.L. Wang, Sign language recognition based on HMM/ANN/DP, Int'l J. Pattern Recognition and Artificial Intelligence, vol. 14, no. 5, pp. 587-602, 2000.

[13] W. Gao, J.Y. Ma, X.L. Chen et al., HandTalker: A multimodal dialog system using sign language and 3-D virtual human, Proc. Third Int'l Conf. Multimodal Interface, pp. 564-571, 2000.

[14] 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.

[15] C.S. Lee, G. Park, J.S. Kim, Z. Bien, W. Jang, and S.K. Kim, Real-time recognition system of Korean sign language based on elementary components, Proc. Sixth Int'l Conf. Fuzzy Systems, pp. 1463-1468, 1997.

[16] P. Hong, M. Turk, and T. S. Huang, Gesture modeling and recognition using finite state machines, Proc. Fourth Int'l Conf. Automatic Face and Gesture Recognition, pp. 410-415, 2000.

[17] G.L. Fang, W. Gao, J.Y. Ma, Signer-independent sign language recognition based on SOFM/HMM, Proc. IEEE ICCV Workshop on RATFG-RTS '01,pp. 90-95, 2001.

[18] J. R. Quinlan, Induction of decision trees, Machine Learning, vol. 1, no.1, pp. 81-106, 1996.

[19] J. R. Quinlan, C4.5: Program for Machine Learning. Morgan Kaufmann, San Mateo, CA, 1993.

[20] L. Breiman, J.H. Friedman, R.A. Olshen, and C.J. Stone, Classification and Regression Trees. New York: Chapman & Hall, 1984.