Expanding Training Set for Chinese Sign Language Recognition

Chunli Wang[1,2] Xilin Chen[2] Wen Gao[2]
[1]School of Electronic and Information Engineering, DUT, Dalian, 116023, China
[2]Institute of Computing Technology, Chinese Academy of Sciences, Beijing, 100080, China
clwang@jdl.ac.cn; xlchen@jdl.ac.cn; wgao@jdl.ac.cn

Abstract

In Sign Language recognition, one of the problems is to collect enough training data. Almost all of the statistical methods used in Sign Language Recognition suffer from this problem. Inspired by the crossover of genetic algorithms, this paper presents a method to expand Chinese Sign Language (CSL) database through re-sampling from existing sign samples. Two original samples of the same sign are regarded as parents. They can reproduce their children by crossover. To verify the validity of the proposed method, some experiments are carried out on a vocabulary of 2435 gestures in Chinese Sign Language. Each gesture has 4 samples. Three samples are used to be the original generation. These three original samples and their offspring are used to construct the training set, and the remaining sample is used for test. The experimental results show that the new samples generated by the proposed method are effective.

1. Introduction

Hand gesture recognition, which contributes to a natural man-machine interface, is still a challenging problem. Closely related to gesture recognition is sign language recognition. Sign language is one of the most natural means of exchanging information for the deaf people. It is a kind of visual language via hand and arm movements accompanying facial expression and lip motion. Growing public approval and funds for international projects, such as SignPS[1], VisiCast[2] and WISDOM[3], emphasize the importance of sign language.

The reports about gesture recognition began to appear at the end of 80's. T.Starner [4] achieved a correct rate of 91.3% for 40 signs based on the image. By imposing a strict grammar on this system, the accuracy rates in excess of 99% were possible with real-time performance. Fels and Hinton [5][6] developed a system using a VPL DataGlove Mark II with a Polhemus tracker as input devices. Neural network was employed for classifying hand gestures. Y. Nam and K.Y. Wohn [7] used three–dimensional data as input to Hidden Markov Models (HMMs) for continuous recognition of a small set of gestures. R.H.Liang and M. Ouhyoung [8] used HMM for continuous recognition of Tainwan Sign language with a vocabulary between 71 and 250 signs by using Dataglove as input devices. HMMs were also adopted by Kisti Grobel and Marcell Assan to recognize isolated signs collected from video recordings of signers wearing colored gloves, and 91.3% accuracy out of a 262-sign vocabulary was reported [9]. C.Vogler and D.Metaxas [10] described an approach to continuous, whole-sentence ASL recognition, in which phonemes instead of whole signs were used as the basic units. They experimented with 22 words and achieved similar recognition rates with phoneme-based and word-based approaches. Wen Gao[11] proposed a Chinese Sign language recognition system with a vocabulary of 1064 signs. The recognition accuracy is about 93.2%. C. Wang [12] realized a Chinese Sign Language (CSL) recognition system with a vocabulary of 5100 signs. R. Martin McGuire [13] realized a mobile one-way American Sign Language translator. The recognition results of 94% accuracy on a 141 sign vocabulary are gotten.

For signer-independent recognition, Vamplew [14] reported the SLARTI sign language recognition system with an accuracy of around 94% on the signers used in training, and about 85% for other signers. It used a modular architecture consisting of multiple feature-recognition neural networks and a nearest-neighbor classifier to recognize 52 Australian sign language hand gestures. All of the feature-extraction networks were trained on examples gathered from 4 signers, and tested on both fresh examples from the same signers and examples from 3 other signers. Akyol and Canzler [15] proposed an information terminal that can recognize 16 signs of German Sign Language from

video sequences. 7 persons were taken for training the HMMs and the other three for testing. The recognition rate is 94%.

In the above reports, the numbers of signers are still small. It is hard to build signer-independent models. Data collection for both training and testing is a laborious but necessary step. Almost all of the statistical methods used in Sign Language Recognition suffer from this problem. However, sign language data cannot be collected as easily as speech data. We must invite the special persons to perform the signs. The lack of the data makes the research, especially the large vocabulary signer-independent recognition, very difficult.

This paper focuses on this problem. In face detection and recognition field, researchers employ some methods to generate new samples to swell the face database [16]. This idea can be used in Sign Language Recognition. In this paper, re-sampling is presented to enlarge the sign language database. Inspired by genetic algorithms, the idea of crossover is used to generate more samples from existing ones. Each sign is composed of limited types of components, such as hand shape, position and orientation, which are independent of each other. Two original samples of the same sign cross at one and only one component to generate two children. To verify the validity of the proposed method, some experiments are carried out on 2435 gestures in Chinese Sign Language. These gestures can be regarded as etyma. The words in Chinese Sign Language are composed by one or more etyma. For example, the gestures "Education" and "Room" form the word "Classroom". The experimental results show that the generated data are effective.

The rest of this paper is organized as follows. In Sect. 2, we introduce how to represent a sign in computer. The re-sampling method based on genetic algorithms is proposed in Sect. 3. In Sect. 4, the experimental results are reported. Finally in Sect. 5, we give the conclusions.

2. Representing a Sign

Two CyberGlove and a Pohelmus 3-D tracker with three receivers positioned on the wrist of CyberGlove and the back are used as input device in this system. The input equipments are shown in Figure 1.

Each sample of a sign is a sequence of frames. The number of frames is from 10 to 20. One frame of raw gesture data, which in our system are obtained from 36 sensors on two datagloves, and three receivers mounted on the datagloves and the body, are formed as 48-dimensional vector. An algorithm based on geometrical analysis for the purpose of extracting invariant feature to signer position is employed. Each element value is normalized to ensure its range 0-1.

Figure 1. **The Dataglove and the 3-D tracker used in our system. Three receivers are fixed on two hands and the back respectively.**

Each hand is represented by 24 dimensions data. 18 dimensions data represent the hand shape, 3 dimensions data represent the position of hand, and 3 dimensions data represent the orientation of hand. We can split up one sign into a number of channels. Each channel can be considered as a gene of the sign. Inspired by the genetic algorithms, crossover can be used to generate new samples of the sign. Intuitively, we split a sign into three channels: Position & Orientation, Left Hand Shape and Right Hand Shape.

3. Generating New Samples

In order to get more training data, we can generate new samples from the existing ones. The idea of genetic algorithms, namely crossover, can be employed.

3.1. Basic Idea

Genetic algorithms take their analogy from nature. Two initial samples of the same sign are regarded as parents. They can reproduce their children by crossover.

Each sign is composed of limited types of components, such as hand shape, position and orientation. We can split up one sample into three channels, namely Position & Orientation (P&O), Left Hand Shape (LH), and Right Hand Shape (RH). These three parts of the same sign in different samples may be variable. One sample gives one demonstration of each part.

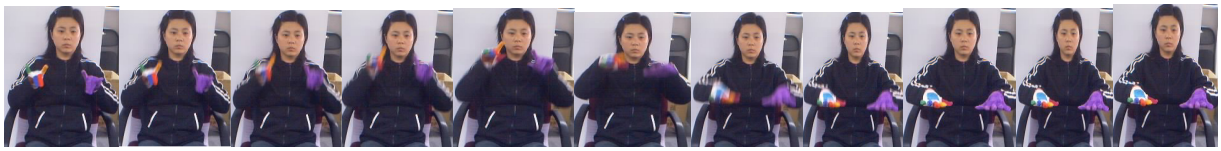
$S_1 = \{P\&O_1, LH_1, RH_1\}$ and $S_2 = \{P\&O_2, LH_2, RH_2\}$ denote two samples of the same sign. The hand shape, orientation and position of the first sample are different from that of the second one. All of them are correct for this sign. So a gesture with the position & orientation of the first sample and the hand shape of the second one, namely $S = \{P\&O_1, LH_2, RH_2\}$, is a possible

sample of the sign. But it is different from S_1 and S_2. Therefore S may not match the model well and may not be recognized correctly. If we re-combine these parts from different samples, the HMMs trained by them can have better generalization performance.

3.2. Warping Two Samples

Two samples of the same sign are picked up from the initial training set randomly. Each sample is a sequence of frames. In order to cross, the length of Parent2 should be warped to that of Parent1.

Furthermore, performing a gesture is a time-dependent process. Several utterances of the same word are likely to have different durations, and utterances of the same word with the same duration will differ in the middle, due to different parts of the words being performed at different rates. Figure 2 demonstrates two samples. Because the data gathered from datagloves are not visual, we use the video to illuminate the change of frames. Here an image is corresponding to a frame.



(a) Sample 1 of the sign "Australia"

(b) Sample 2 of the sign "Australia"

Figure 2. **Two dynamic samples of the same sign. They are performed at different speed.**

From the figure, we can see that the lengths of these two sequences are different and it is obviously incorrect if crossing over at frame 5. Therefore, we need align two dynamic samples to map each frame in one sample to the similar frame in the other sample. Dynamic Programming (DP) can resolve this problem. When applied to template-based dynamic pattern recognition, it is often regarded as Dynamic Time Warping (DTW).

DP is guaranteed to find the shortest-distance-path through a matrix, with minimal amount of computation. The DP algorithm operates in a time-synchronous manner: each column of the time-time matrix is considered in succession (equivalent to processing the input frame-by-frame) so that, for a template of length N, the maximum number of paths being considered at any time is N [17]. Here, we use it to get the mapping table of frames between two samples.

d(i,j) denotes the distance of Frame *i* in Sample 1 and Frame *j* in Sample 2. N_1 is the number of Sample 1 and N_2 is the number of Sample 2. Here we use the Euclidean distance of two frames. *D(i,j)* is the global distance up to *(i,j)* . *D(i,j)* can be computed by this expression:

$$D(i,j) = \min\{D(i-1,j-1), D(i-1,j), D(i,j-1)\} + d(i,j) \quad (1)$$

Given that *D(1,1) = d(1,1)* (this is the initial condition), which is the basis for an efficient recursive algorithm for computing *D(i,j)*. The final global distance $D(N_1, N_2)$ gives us the overall matching score of the two series. During the procedure, backtrace array (or backpointer array) must be kept with entries in the array pointing to the preceding point in that path. And then we can trace back along the best-matching path and get the mapping table.

The mapping table of the two samples given in Figure 2 is shown in Figure 3.

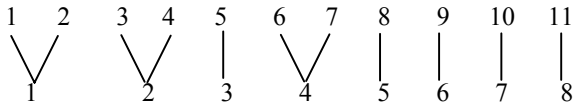


Figure 3. **The mapping table. The first line denotes the frames in Sample 1 and the second line denotes those in Sample 2.**

Each frame in Sample 2 is mapped to one or more frames in Sample 1. If there are several frames in sample 1 mapped to one frame in Sample 2, the average of these frames mapped to the same frame are used to cross with the frame in Sample 2. For example, f_i^j denotes Frame *i* in Sample *j*. According to the mapping table in Figure 3, f_1^1 and f_2^1 are mapped to f_1^2, so the average of f_1^1 and f_2^1 crosses with f_1^2. The lengths of children samples are equal to Sample 2. If

two samples are exchanged, that is, Sample 1 has 8 frames and Sample 2 has 11 frames, the lengths of the generated children samples are equal to 11. And then, Frame 1 of the shorter sample crosses with Frame 1 and Frame 2 of the longer sample respectively.

3.3. Crossover

Each frame is divided into three parts: position & orientation, left hand shape and right hand shape. Every pair of parent crosses at one and only one part, that is, all frames of the parent cross over at the same part during one crossover operator. The parent can cross three times and generate 6 new samples. The process of crossover is shown in Figure 4.

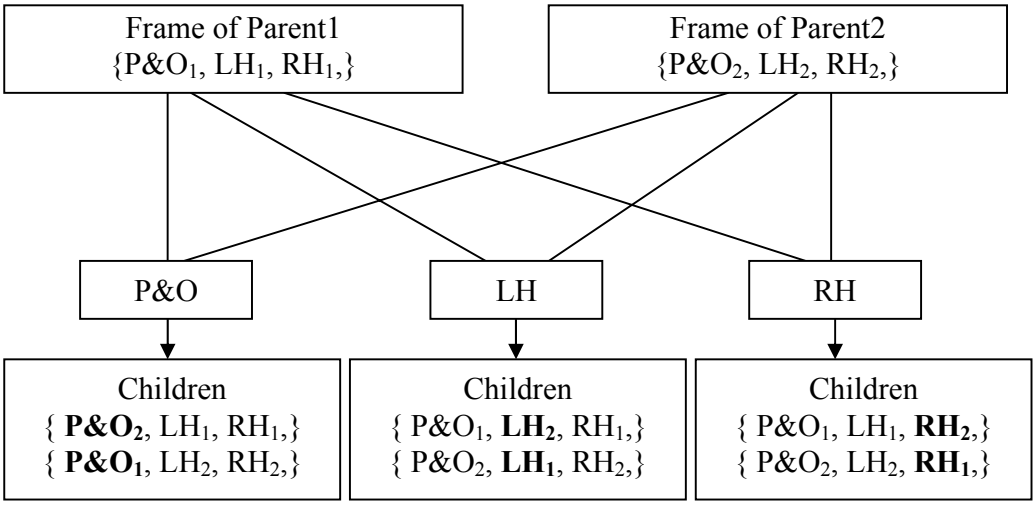


Figure 4. **Crossover operator during the re-sampling. Each frame is broken down into three independent parts. During a crossover, one and only one part is exchanged. All frames of the parent cross over at the same part.**

To verify the validity of the proposed method, we compare it with parallel HMMs (PaHMMs).

3.4. PaHMMs

As mentioned in Sect. 3.1, if the data of the channel 1 of a test sample are similar to a training sample of the same sign, and the data of channel 2 are similar to another training one, the test sample may be different from all the training samples and is probably recognized as another sign. To verify the validity of the above method, we assume the other methods that can resolve this situation. Christian Vogler [18] proposed a framework of PaHMMs, which builds a HMM for each channel respectively. For example, PaHMM-CN3 shown in Figure 5 models 3 channels with 3 independent HMMs. Three channels are: Position & Orientation, Left Hand Shape, and Right Hand Shape.

The data of each channel are used to train the HMM for the corresponding channel. And then it is not necessary to generate more samples by crossover. If the results of PaHMMs are better than those of HMMs trained by generated samples, the method proposed here is useless. Contrarily, the method is effective.

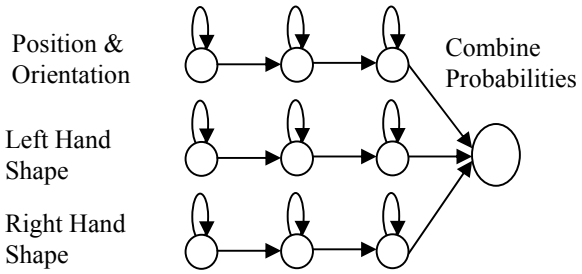


Figure 5. **Structure of PaHMMs with 3 independent channels.**

In Sect. 4, the comparisons between the results of PaHMMs and HMMs trained by generated samples are reported.

4. Experiments

To verify the generalization capability of the proposed method, some experiments are performed. We carry out experiments based on a vocabulary of 2435 gestures. These signs are performed by the same signer. We invite a deaf signer to collect data for us. Each sign has 4 samples. The traditional leave-one-out cross-validation is employed. Three samples are used to construct the original training data, and the

remaining one is used for testing. So there are four groups of training sets and test samples.

HMMs are trained based on different training sets. In our system, HMM is left-to-right model allowing possible skips. The number of states is set to be 3, and the number of mixture components is set to be 2. We fix the values of variances of covariance matrix. The variances of the feature data in the dimensions representing Position and Orientation are set to be 0.2 and those in the dimensions representing Hand Shape are set to be 0.1. The above values are obtained by experiments.

With 3 original samples, 18 new samples can be generated by crossover. The recognition results are shown in Figure 6. The horizontal axis indicates different group of training set. From Figure 6, it can be seen that the results based on the generated training set are better than those based on the original training data. Generating new samples improves the accuracy.

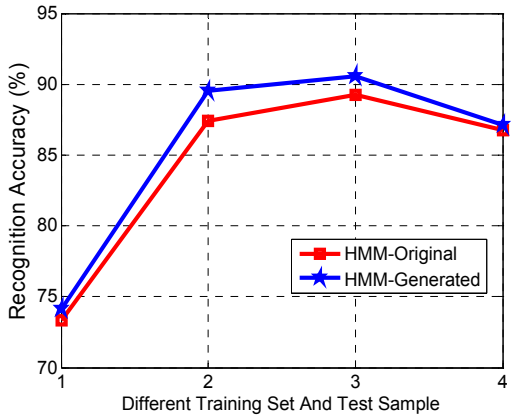


Figure 6. **Comparison of the recognition results based on the original training data and the generated training data. Leave-one-out cross-validation is employed.**

The accuracy of the first group is lower than those of the others. The possible reasons are as follows. The test data in this group were collected for the first time. The signer was not very accustomed to perform gestures with unwieldy dataglove and tracker on body. Some signs are not up to the standard. Besides, when we collected data for the first time, some details of the input devices, such as the effective range of the tracker, are neglected. The data are somewhat affected.

Figure 7 shows the average recognition rates. The horizontal axis indicates the number of candidates, represented as rank. When rank = 1, the average accuracy based on the original training data is 84.19%, and Generated Training Data Strategy achieves the best accuracy of 85.35%. According to the relative accuracy improvement computing formula (2), the relative accuracy improvement of 7.3% is achieved. This result is encouraging. The experimental results show that the data generated by the proposed method are effective.

$$\Delta\lambda = \frac{\lambda_1 - \lambda_2}{1 - \lambda_2} \tag{2}$$

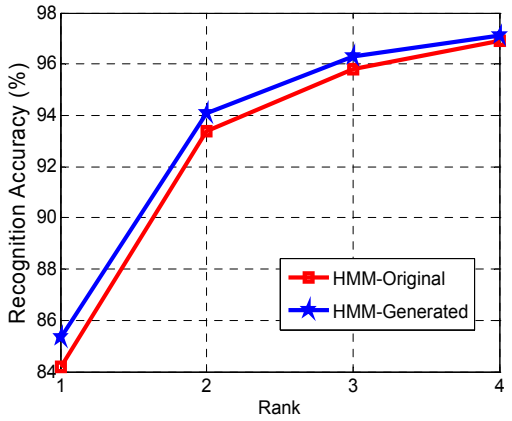


Figure 7. **The average recognition rates based on the original training set and generated training set at rank up to 4.**

The possible reasons for the above results are as follows. The generated new samples may be different from the original training samples but similar to the unknown test sample. So by this method, the system can get better generalization performance with the limited training data.

To verify the validity of the re-sampling method, we carry out some experiments on PaHMMs. The recognition rates on cross-validation tests are given in Figure 8.

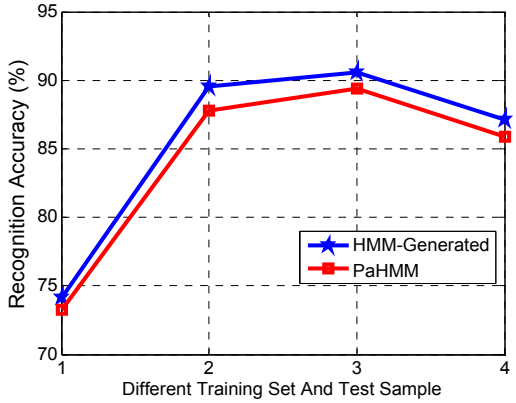


Figure 8. **The comparisons between PaHMMs and HMMs trained by generated training sets based on cross-validation tests.**

Table 1 summarizes the average accuracies of HMMs trained by original training set, HMMs trained by generated training sets and PaHMMs. It can be seen that generating new samples may achieve better accuracy than PaHMMs.

Table 1. **The average recognition rates of HMMs trained by original training set, HMMs trained by generated training sets and PaHMMs**

Model	HMM-Orignial	HMM-Generated	PaHMMs
Accuracy rate	84.19%	**85.35%**	84.06%

5. Conclusions and Discussions

In this paper, a re-sampling method based on the crossover of genetic algorithms is proposed to swell the sign language database and improve the recognition accuracy of gestures. Before crossover, we use DTW to align two parent samples. By the re-sampling method, a number of new samples can be generated from the existing ones. These new samples and original samples construct training set to build HMMs. Experiments conducted on a sign language database containing 2435 gestures (9740 gesture samples) show that the recognition accuracy is improved by applying the proposed method.

The idea of mutation can be added, too. There are some factors that can be mutated, for example, the hand shape, the scope or speed of action, the track of movement, etc. To employ mutation, it is necessary to judge whether a new sample generated by crossover and mutation is effective or not.

6. Acknowledgment

This research is sponsored by Natural Science Foundation of China (No. 60533030).

References

[1] SignPS homepage: http:// www.handicom.nl/english /SignPS/

[2] VisiCast homepage: http://www.visicast.co.uk/

[3] WISDOM homepage: http://www.mobilewisdom.org/

[4] Starner T., Weaver J., Pentland A., "Real-Time American Sign Language Recognition Using Desk and Wearable Computer Based Video", IEEE TPAMI, IEEE CS, Vol 20, Issue 12, Dec 1998, pp. 1371-1375.

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

[6] S.Sidney Fels, *Glove –TalkII: Mapping hand gestures to speech using neural networks-An approach to building adaptive interface,* PhD thesis, Computer Science Department, University of Torono, 1994.

[7] Yanghee Nam, K. Y. Wohn, "Recognition of space-time hand-gestures using hidden Markov model", ACM Symposium on Virtual Reality Software and Technology, ACM, HongKong, July, 1996, pp. 51-58.

[8] R.-H.Liang, M.Ouhyoung. " A real-time continuous gesture recognition system for sign language", In Proceeding of the Third International Conference on Automatic Face and Gesture Recognition, IEEE CS, Nara, Japan, 1998, pp. 558-565.

[9] Kirsti Grobel, Marcell Assan. "Isolated sign language recognition using hidden Markov models", In Proceedings of the International Conference of System, Man and Cybernetics, 1996, pp. 162-167.

[10] ChristianVogler, Dimitris Metaxas, "Toward scalability in ASL Recognition: Breaking Down Signs into Phonemes", In Proceedings of Gesture Workshop, Spinger, Gif-sur-Yvette, France, 1999, pp. 400-404.

[11] Wen Gao, Jiyong Ma, Jiangqin Wu and Chunli Wang, "Large Vocabulary Sign Language Recognition Based on HMM/ANN/DP", International Journal of Pattern Recognition and Artificial Intelligence, World Scientific , Vol. 14, No. 5, 2000, pp. 587-602.

[12] Chunli Wang, Wen Gao, Jiyong Ma, "A Real-time Large Vocabulary Recognition System for Chinese Sign Language", Gesture and Sign Language in Human-Computer Interaction, Springer, London, UK, April 2001, pp. 86-95.

[13] McGuire, R., Hernandez-Rebollar, J., Starner, T., Henderson, V., Brashear, H., "Towards a One Way American Sign Language Translator", IEEE International Conference on Face and Gesture Recognition 2004, IEEE CS, Seoul, Korea, May, 2004, pp. 620-625.

[14] Vamplew, P., Adams, A., "Recognition of Sign Language Gestures Using Neural Networks", Australian Journal of Intelligent Information Processing Systems, Vol. 5, No. 2, Winter 1998, pp. 94-102.

[15] Suat Akyol, Ulrich Canzler, "An information terminal using vision based sign language recognition", ITEA Workshop on Virtual Home Environments, Paderborn, Germany , 2002, pp. 61-68.

[16] Jie Chen, Xilin Chen, Wen Gao, "Expand Training Set for Face Detection by GA Re-sampling", The 6th IEEE International Conference on Automatic Face and Gesture Recognition(FG2004), IEEE CS, Seoul, Korea, May17-19, 2004, pp. 73-79.

[17] Dan Ellis, http://www.ee.columbia.edu/~dpwe/e6820/; Columbia University, Dept. of Electrical Engineering.

[18] C. Vogler, D. Metaxas, "Handshapes and movements: Multiple-channel ASL recognition", Springer Lecture Notes in Artificial Intelligence 2915, 2004. Proceedings of the Gesture Workshop, Genova, Italy, 2003, pp. 247-258.