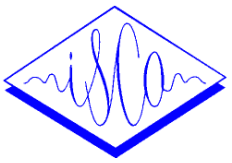

6th International Conference on Spoken Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

A PARALLEL MULTI-STREAM MODEL FOR SIGN LANGUAGE RECOGNITION

Jiyong Ma and Wen Gao

Institute of Computing Technology
Chinese Academy of Sciences
Beijing 100080, China

ABSTRACT

In this paper, the sub-units in each stream are used and embedded in the multi-stream model. In this framework, sign language recognition system was implemented and evaluated. Experiments were carried out for 5177 Chinese signs. The real time isolated recognition rate is 95.1%. For continuous sign recognition, the word correct rate is 91.8%. This has shown that parallel multi-stream model is powerful for sign language recognition considering the problem of scalability.

1. INTRODUCTION

Sign language, a kind of structured gestures, is one of the natural means of exchanging information for the hearing impaired. It has many potential applications including a mode for human-computer interaction, supporting the communication between deaf and hearing society, controlling the motion of a human avatar in a virtual environment (VE), automatic simple task learning of robot tasks through a DataGlove interface by human demonstration, etc.

Hand gestures are physical positions or movements of a person's fingers, hands, arms or body used to convey information. The earliest attempt to automatically recognize started with finger-spelling recognition conducted by Grimes[1]. The recognition algorithms for posture recognition include: inductive algorithms such as ID3, NewID, C4.5, CN2, and HCV, RIEVL [3], neural network approach such as the hybrid approach of radial basis functions (RBF), inductive decision trees [4] and a fuzzy min-max neural network [5]. Fels and Hinton's[6] and Fel's[7] developed a system using a VPL DataGlove Mark II with a Polhemus tracker as input devices. In this system, the neural network was employed for classifying hand gestures. Takahashi and Kishino [8] investigated understanding the Japanese Kana manual alphabets corresponding to 46 signs using a VPL DataGlove. The system could correctly recognize 30 of the 46 signs, while the remaining 16 could not be reliably identified. Murakami and Taguchi [9] made use of recurrent neural nets for sign Language recognition. They trained the system on 42 handshapes in the Japanese finger alphabet using a VPL Data Glove. The recognition rate is 98 per cent. W.Kadous [10] demonstrated a system based on Power Gloves to recognize a set of 95 isolated Auslan signs with 80% accuracy using fast match methods. Tung and Kak[11]described automatic learning of robot tasks through a DataGlove interface. Kang and Ikeuchi[12] designed a system for simple task learning by human demonstration. Kisti Grobel and Marcell Assan [13] 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. Charaphayan and Marble [14] investigated a way using image processing to understand American Sign Language (ASL). This system can recognize correctly 27 of the 31 ASL symbols.

For the case of continuous sign recognition, Starner[15] reported that a color camera was used and the users wore a yellow glove on their right hand and orange one on their left. In this case, a correct rate was achieved 91.3 per cent. By imposing a strict grammar on this, it was shown that accuracy rates for 40 signs in excess of 99 per cent were possible with real-time performance. R.H.Liang and M.Ouhyoung[16] used HMM for continuous recognition of Tainwan Sign language with a vocabulary between 71 and 250 signs based data gloves as input devices. However, the system required that gestures performed by the signer be slow to detect the word boundary. This requirement is hardly ensured for practical applications. C.Vogler and D.Metaxas[17] used HMMs for continuous ASL recognition with a vocabulary of 53 signs and a completely unconstrained sentence structure. C.Vogler and D.Metaxas[18] described an approach to continuous, whole-sentence ASL recognition that uses phonemes instead of whole signs as the basic units. They experimented with 22 words and achieved similar recognition rates with phoneme and word based approaches.

From the review of previous research works above we know that most researches on sign language recognition were made on small test vocabulary. For large vocabulary recognition, Ho-Sub Yoon recently [19] reported that a sign vocabulary consisting of 1,300 alphabetical gestures were recognized using HMMs.

From the temporal and spatial analysis, for each time instant, hand shape, hand position and hand orientation are three measurable factors forming a hand spatial unit in whole sign space. The basic spatial units in hand gestures should include the following six data streams: right handshape, right hand position , right hand orientation, left handshape, left hand

position and left hand orientation. The six data streams are observable and synchronous. It is advantageous to constructively combine the six streams. There are two ways to fusion the multiple data streams. One is fusion at feature level that means features of the six data streams are combined in a single spatial vector. It usually required to assume that the different features are independent (e.g., by assuming diagonal covariance matrices). The other way is the fusion at classifier level. The multistream model is a typical example of the latter case. The original multi-stream model was proposed in speech recognition community by Bourlard [21]. The model processes multi streams independently. Thus, they can also be trained independently, and do not require consideration of the different combinations at training time. It would be possible to permit some stream asynchrony. When streams are not frame synchronous, the complexity that the decoding algorithm required may be considerably greater than that for a standard recognizer. Results to date in speech recognition have indicated that allowing asynchrony among streams does not give any signification performance improvement [22]. Therefore, synchronous data stream model is used in this paper.

One assumption in the multi-stream model proposed by Bourlard is that the sub-units in whole stream space are well defined. This assumption is valid for speech recognition. Since there are lexical dictionaries available for each word in natural spoken language. But this assumption might be invalid for sign language recognition. Because there is no any lexical dictionary available for each phrase in sign language and the number of possible pattern combinations of all streams is too large, it seems impossible to obtain a few sub-units in the whole sign space. Therefore, the conventional multi-stream model should be modified so that it can be applied into the sign recognition. In this paper, the sub-units in each stream are used and embedded in the multi-stream model. In this framework, sign language recognition system was implemented and evaluated.

2. CONVENTIONAL MULTI-STREAM MODEL

In a HMM –based recognizer the mathematical formalism of multistream model [22] is as the following: assume that the' i input stream is X(i) and the model for a pattern *P* is composed of J sub-unit models P(j). To process each stream independently up to the defined sub-unit level, each sub-unit model P(j) is composed of parallel models P(j,i) (possibly with different topologies) that are forced to recombine their respective segmented scores at some temporal anchor points. The resulting statistical model is illustrated in Fig.1.

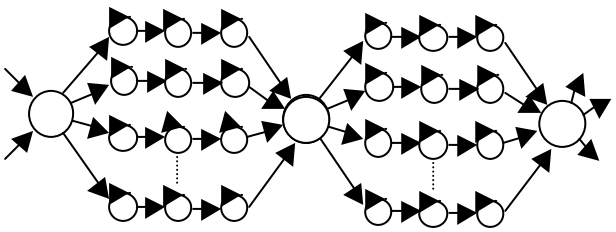

Figure 1.Conventional Multi-stream Diagram

The sub-unit in speech recognition can be a phoneme or a syllable. And the pattern can be a word that can be described with sub-units. The sub-units are well defined in dictionary. However sub-units in a sign are not well defined in sign language books. It is difficult to find such a kind of sub-units in the whole sign space. Because the number of possible pattern combinations of all streams is too large, it seems impossible to obtain a few sub-units in the whole sign space. Therefore, the conventional multistream approach should be modified for sign language recognition.

3. PARALLEL MULTI-STREAM MODEL

For sign language recognition based on datagloves and position trackers as input devices, the following six data streams are measurable :right handshape, right hand position, right hand orientation, left handshape, left hand position and left hand orientation. And they forms a vector denoted by $(X_1(t),X_2(t),X_3(t),X_4(t),X_5(t),X_6(t))$.

For a HMM –based recognizer the mathematical formalism of parallel multistream model is described as the following:

Assume that the' i input stream is X(i) and the model for a pattern *P* is composed of I parallel stream models P(i) that are forced to recombine their scores at the final anchor point. Each stream model P(i) is composed of sequential models P(i,j) (possibly with different topologies). The resulting statistical model is illustrated in Fig.2.

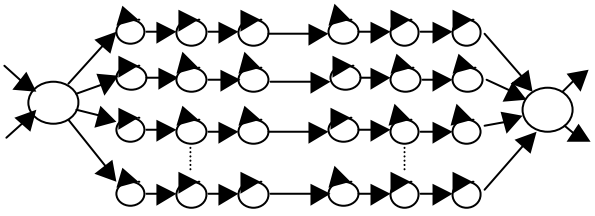

Figure2 Parallel Multi-stream Diagram

The topology structure of the parallel multistream model is a little bit different from that of the traditional multistream model.

Finding sub-units in each stream is easier than finding sub-units in all streams. The sub-units in each stream can be found by automatic clustering approaches. We could not expect to code signs by manual, because the task is tedious and time consuming.

The basic idea of the parallel multistream model can be summarized as follows

Extract appropriate feature vectors for each stream

Normalize feature vectors of each stream

Training independent recognizers for each stream.

The characteristics of the algorithm are as follows：

The parallel HMMs associated with each of the input streams do not necessarily have the same topology. Different recognition strategies might ultimately be applied in each stream.

The synchronous control state illustrated by the symbol S in Fig.2 is not a regular HMM state since it will be responsible for recombining the probabilities and controlling synchrony of the decoding process.

The fusion is taken at the final state of each sign model. The transition at sign boundary is determined by the fusion score at that boundary, including the language model, this enables the linguistics model can be used as early as possible and the search path is more reasonable.

The separated stream has little state space, this enable it is possible to use few states to describe large number of signs.

One characteristic of the search is that all streams are synchronous at sign boundary. It is ensured that the search is synchronous at sign level. It permits that the number of states in HMMs of different streams is different.

As Viterbi decoding is taken in each stream, a lot of memory resources are needed. Since the decoding is taken synchronously for six streams, the computation load is higher. But because the fusion is taken at sign boundary, the computation load of the observation probabilities is little.

4. EXPERIMENT

The baud rate for both CyberGlove and 3-D tracker is set 38400. The number of states in HMM of each sign is 3 or 5. The raw gesture data, which in our case are values of 18-joint angles collected from the Cyberglove for each hand, the range of each angle value is within 0-255. For two hands, they are formed as a 48 dimensional vector appended with hand and position and orientation features. The dynamic range of each component is different. Each component value is normalized to ensure its dynamic range is 0-1. [24].

The HMM structure for each sign is left to right without skip. The hardware environment is Pentium III 450Hz.

4.1 Isolated sign recognition evaluation

For the case of isolated sign recognition 5177 signs in Chinese sign language were used as evaluation vocabularies. Each sign was performed 5 times by a sign language teacher. 4 times were used for training and one for test. Using the approach of cross validation test, the test times for each word is 5. For different numbers of different patterns in each stream, the off-line recognition rates are listed in the Table 1. Where Lp, Lo, Ls, Rp, Ro, Rs, are the number of different patterns in data stream of left/right hand position, left/right hand orientation, left/right hand shape, respectively. The number of states in HMM is 3 or 5. When the number of states in HMM is 5,the results are listed in the Table 2. This shows that 3 states are better than the 5 states considering the computation load and recognition accuracy. Therefore 3 states for each HMM are used in the continuous recognition.

Table 1.The Recognition rates (3 states)

4.2 Continuous sign recognition evaluation

For the case of continuous recognition, the database of gestures consists of 5177 signs and 500 sentences. In general, each sentence consists of 2 to 15 signs. No intentional pauses were placed between signs within a sentence. Sign transition model discussed in section 5.2 was used to model the movement epenthesis.

To test the recognition performance at sentence level, one test was carried out described as following. When 500 sentences were not used for any portion of the training. The 5177 words are used as basic units. Within 500 sentences, 264 sentences can be correctly recognized, the left 236 sentences have deletion (D), insertion (I), and substitution (S) errors, D=186,I=302, S=332,N=5162, N denotes the total number of signs in the test set, the word correct rate is 84.1%.

This shows that the movement epenthesis has affected on recognition performance at sentence level. To take into account this effect, the sign transition HMMs were trained by the sentences. For the left 264 sentences, the training procedure for sign transition models was used for each of sentence. In the collected sentence samples, four in five are used for training and one for test. The word correct rate is 91.26%, where D=98, I=182, S=171, N=5162. The accuracy measure is calculated by subtracting the number of deletion, insertion, substitution and errors from the total number of signs and divided by the total number of signs. The result shows that sign transition models are necessary for sentence level recognition. The recognition speed is about 2 times of real-time. When the number of sign transition models is set to 3 for all sign transition model, the word correct rate is 91.4%, where D=97, I=180, S=167, N=5162. The recognition rates are summarized in the Table 3.

5. CONCLUSION

We have presented a framework for large vocabulary sign recognition by parallel multistream model. Experiments have shown that this model is powerful for sign language recognition considering the problem of scalability.

6. CONCLUSION

[1] G. Grimes. Digital Data Entry Glove interface device. Patent 4,414,537, AT & T Bell Labs, November 1983.

[2] Jong-Sung Kim,Jung-Bae Kim,Kyung-Joon Song, Byungeui Min and Zeungnam Bien, On-line motion control of avatar using hand gesture recognition. Journal of the Institute of Electronics Engineers of Korea C, vol.36-C, No.6, June 1999, pp.52-62.

[3] Salem B. Yates R. Saatchi R. Current trends in multimodal input recognition. IEEE Colloquium Virtual Reality: Personal, Mobile and Practical Applications. IEEE. 1998, pp.311-316. London, UK.

[4] Meide Zhao. Quek FKH. Xindong Wu. RIEVL: recursive induction learning in hand gesture recognition. IEEE Transactions on Pattern Analysis & Machine Intelligence, vol.20, no.11, Nov. 1998, pp.1174-1185.

[5] Gutta S. Imam IF. Wechsler H. Hand gesture recognition using ensembles of radial basis function (RBF) networks and decision trees. International Journal of Pattern Recognition & Artificial Intelligence, vol.11, no.6, Sept. 1997, pp.845-872.

[6] Jong-Sung Kim. Chan-Su Lee. Wong Jang. Zeungnam Bien. Online dynamic hand gesture recognition system for the Korean sign language (KSL). Journal of the Korean Institute of Telematics & Electronics, vol.34C, no.2, Feb. 1997, pp.61-70.

[7] S.S.Fels and G.Hinton, GloveTalk:A neural network interface between a DataDlove and a speech synthesizer, IEEE Transactions on Neural Networks,4(1993):2-8.

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

[9] Tomoichi Takahashi and Fumio Kishino, Gesture coding based in experiments with a hand gesture interface device, SIGCHI Bulletin, 1991,23(2): 67-73.

[10] Kouichi Murakami and Hitomi Taguchi, Gesture recognition using recurrent neural networks, In CHI' 91 Conference Proceedings, 1991, pages 237-242.

[11] Mohanmmed Waleed Kadous, Machine recognition of Auslan signed using PowerGlove: Towards large-lexicon recognition of sign language, In Lynn Messing,editor, Proceddings of WIGLS. The Workshop on the Integration of Gesture in Language and Speech, 1996,pages 165-174.

[12] C.P.Tung and A.C.Kak, Automatic learning of assembly tasks using a dataglove system, In Proceedings of IEEE/RSJ Conference on Intelligent Robots and Systems,1995,pp.1-8.

[13] S.B.Kang and K.Ikeuchi, Robust task programming by human demonstration, In Proceedings of the Image Understanding Workshop,1994,pp.303-308.

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

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

[16] Starner T. Weaver J. Pentland. A Real-time American Sign Language recognition using desk and wearable computer based video. IEEE Transactions on Pattern Anal ysis & Machine Intelligence, vol.20, no.12, Dec. 1998, pp.1371-5

[17] 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,pages,558-565.

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

[19] ChristianVogler and Dimitris Metaxas," ASL recognition based on a coupling between HMMs and 3D motion analysis," I*n Proceedings of the IEEE International Conference on Computer Vision*, Mumbai, India, 1998,pages 363-369.

[20] Ho-Sub Yoon. Jung Soh. Byung-Woo Min. Hyun Seung Yang. Recognition of alphabetical hand gestures using hidden Markov model. IEICE Transactions on Fundamentals of Electronics Communications & Computer Sciences, vol.E82-A, no.7, July 1999, pp.1358-66. Publisher: Inst. Electron. Inf. & Commun. Eng, Japan.

[21] Vogler C. Metaxas D. Parallel hidden Markov models for American Sign Language recognition. Proceedings of the Seventh IEEE International Conference on Computer Vision. IEEE Comput. Soc. Part vol.1, 1999, pp.116-22 vol.1. Los Alamitos, CA, USA

[22] Bourlard,H., Non-stationary multi-channel(multi-stream)processing towards robust and adaptive ASR, Proc. Tampere workshop on robust methods for speech recognition in aderse conditions,1995, pp.1-10.

[23] Mirghafori,N.,A multi-band approach to automatic speech recognition,PhD dissertation, University of California at Berkely, Dec 1998.Reprinted as ICSI Technical report ,ICSI TR-99-04.

[24] L. Rabiner and B. Juang. An introduction to hidden Markov models. *IEEE ASSP Magazine*, Jan. 1996,pages 4-16.

[25] Jiyong Ma,Wen Gao,Jiangqin Wu and Chunli Wang,A Continuous Chinese Sign Language recognition system,IEEE,FG' 2000 accepted paper Mar.,2000.

[26] V.V.Digalakis,P.Monaco and H.Murveit,Genones:Generalized mixture tying in phoneme-based speed recognition,IEEE Transaction on ASSP,vol.4,1996,pp281-289

[27] X.D Huang,A.Acero,H.Hon,and S.Meredith, Spoken language processing, pp553,Prentice Hall, 1999.