A Continuous Chinese Sign Language Recognition System

Jiyong Ma [1], Wen Gao [1,2], Jiangqin Wu [2], Chunli Wang [3]

[1] Institute of Computing Technology, Chinese Academy of Science, Beijing 100080,China

[2] Department of Computer Science, Harbin Institute of Technology, Harbin, China

[3] Department of Computer Science, Dalian University of Technology, Dalian, China

Email: mjy@cti.com.cn, wgao@ict.ac.cn

Abstract

In this paper, we describe a system for recognizing both the isolated and continuous Chinese Sign Language (CSL) using two Cybergloves and two 3SAPCE-position trackers as gesture input devices. To get robust gesture features, each joint-angle collected by Cybergloves is normalized. The relative position and orientation of the left hand to those of the right hand are proposed as the signer position independent features. To speed up the recognition process, a fast match and a frame predicting techniques are proposed. To tackle epenthesis movement problem, context-dependent models are obtained by the Dynamic Programming (DP) technique. HMMs are utilized to model basic word units. Then we describe training techniques of the bigram language model and the search algorithm used in our baseline system. The baseline system converts sentence level gestures into synthesis speech and gestures of 3D virtual human synchronously. Experiments show that these techniques are efficient both in recognition speed and recognition performance.

1. Introduction

Sign language as a kind of structured gesture is one of the most natural means of exchanging information for most deaf people. This has spawned interesting in developing machines that can accept sign language as the means for human-computer interaction and communication between deaf people and hearing society. In fact, a new field of sign language engineering is emerging that attempts to make use of the computer power to enhance deaf people to hearing society communication or human-computer interfaces. The aim of recognizing sign language is to provide an efficient and accurate mechanism to transcribe human sign language into text or speech.

To date, there are two ways to collect gesture data in sign language recognition, one is the vision-based approach. This technique utilizes cameras to collect the images of hand gestures. Hand gesture features are extracted from the images. In order to robustly extract the hand gesture features, a special glove with areas painted on it to indicate the positions of the fingers or skin color information to segment hand [2] is often used for input, for example, bright points on the edge of the fingers, etc. The approach has the advantage that the signer is not necessary to wear any complex input devices. However, the approach to extracting hand gesture features often suffers from instability due to poor illuminant conditions. Furthermore, hand features extracted from images have poorer discriminant powers for large vocabulary sign language recognition task. The device-based measurement techniques measure hand gestures using devices such as Datagloves, position-trackers and so on. The advantage of device-based approaches manifests that the time and space information of hand gestures is directly measured. The feature discriminant power is higher than that of the vision based feature. So it is suitable for large vocabulary sign recognition task. However, the disadvantage is its high cost.

Attempts at machine sign language recognition began to appear in the literature in 90's. Charaphayan and Marble [1] investigated a way using image processing to understand ASL. This system can recognize correctly 27 of the 31 ASL symbols. Starner [2] reported that the word correct rates for hands wearing colored gloves and hands without gloves were 99% and 92%, respectively, using a color camera as input device. Fels and Hinton's[3] and Fel's[4] Fels developed a system using a VPL DataGlove Mark II with a Polhemus tracker attached for position and orientation tracking as input devices. In this system, the neural network was employed for classifying hand gestures. Wexelblbat[5] developed a gesture recognition system. In the system three Ascension Flock-of-Bird position trackers together with a CyberGlove on each hand. Takahashi and Kishino[6] 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 [7] 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. James Kramer and his supervisor, Larry Leifer [8] worked on a method for communication between deaf individuals. W.Kadous[9] demonstrated a system based on Power Gloves to recognize a set of 95 isolated Auslan signs with 80% accuracy, with an emphasis on fast match methods. R.H.Liang and M.Ouhyoung used HMM for continuous recognition of Tainwan sign language with a vocabulary between 71 and 250 signs based Dataglove as input device. However, the system required that gestures performed by the signer be slowly to detect the word boundary. This requirement is hardly ensured for practical applications. 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. C.Vogler and D.Metaxas[14] used HMMs for continuous ASL recognition with a vocabulary of 53 signs and a completely unconstrained sentence structure. C.Vogler and D.Metaxas[15-16] described an approach to continuous, whole-sentence ASL recognition that used phonemes instead of whole signs as the basic units. They experimented with 22 word vocabularies and achieved similar recognition rates with phoneme-and word-based approaches.

Chinese Sign Language (CSL) is the primary mode of communication for most deaf people in China. CSL consists of about 5500 elementary vocabularies including postures, hand gestures. In this paper we describe a system for continuos Chinese Sign Language (CSL) recognition.

How do we extract singer position invariant features? This is very important to practical applications because it is not necessary to restrict a singer to a certain position when the signer is gesturing. How do we tackle the phenomenon of movement epenthesis during continuously gesturing? This is a very important problem for accurate recognition of continuous sign language. How do we prune efficiently during the tree search of Viterbi decoding? This is a very important problem for speeding up recognition procedure and reducing memory resources. How do we use language model in the decoding process? This is an important problem to prune unlikely hypothesis as soon as possible and to enhance the recognition accuracy. The paper will address the four problems.

The organization of this paper is as follows, in Section 2 we describe a gesture feature extraction approach. Statistical approaches in sign language recognition are discussed in Section 3. Section 4 describes the system outline. Section 5 briefly describes Chinese sign language synthesis. Section 6 describes the synchronously driving of speech and gestures. The performance evaluations of the system are presented in Section 7. Section 8 contains the summary and discussion.

2. Feature Extraction

In the following, we address modeling of the relative 3-D motion of receivers with respect to a transmitter. 3-D motion of receivers can be viewed as rigid motion. It is well known that 3-D displacement of a rigid object in the Cartesian coordinates can be modeled by an affine transformation as the following,

$$X' = R(X - S) \qquad (1)$$

Where R is a 3×3 rotation matrix as the following

$$R = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\alpha & -\sin\alpha \\ 0 & \sin\alpha & \cos\alpha \end{pmatrix} \begin{pmatrix} \cos\beta & 0 & \sin\beta \\ 0 & 1 & 0 \\ -\sin\beta & 0 & \cos\beta \end{pmatrix} \begin{pmatrix} \cos\gamma & -\sin\gamma & 0 \\ \sin\gamma & \cos\gamma & 0 \\ 0 & 0 & 1 \end{pmatrix} \qquad (2)$$

$X = (x_1, x_2, x_3)^t$ and $X' = (x'_1, x'_2, x'_3)^t$ denote the coordinates of an object point with respect to the Cartesian coordinate systems of the transmitter and receiver ,respectively, S is the position vector of the receiver with respect to Cartesian coordinate systems of the transmitter. The receiver output the Eulerian angles, namely, the α, β, γ, which are angles of rotation about X_1, X_2 and X_3 axes, respectively. However, they can not be directly used as features because of their variations when the position of the transmitter is not fixed at a certain position during training and testing. This situation often happens when the system is moved. Therefore, it is necessary to define a reference point so that the features are invariant whenever the positions of the transmitter and a signer are changed. To meet this requirement, we propose the following approach when two receivers are available.

For the case of two receivers available, firstly, the reference Cartesian coordinate system of the receiver at left hand is chosen. Secondly, the position and Cartesian coordinate system of the receiver at right hand with respect to reference Cartesian coordinate system of the receiver at left hand are calculated as invariant features to the positions of the transmitter and the signer. The algorithm is described as follows. Suppose that S_r , S_l are the position vectors of the receivers at both hands which are measured by the position tracking system. R_l is the rotation matrix of the receiver at the left hand respect to Cartesian coordinate systems of the transmitter. R_r^t is the transpose matrix of R_r that is the rotation matrix of the receiver at the right hand respect to Cartesian coordinate systems of the transmitter. They can be calculated according to the Eulerian angles measured by the orientation tracking system. Firstly, the product $R_l R_r^t$ is invariant to the positions of the transmitter and the signer. The reason is that each element of the matrix is a dot product between a unit directional vector of axis of the receiver at the left hand and a unit directional vector of axis of the receiver at the right hand. The relative angle is invariant to the position of the transmitter. Secondly, the relative position vector $R_l(S_r - S_l)$ is also invariant. This approach can be generalized for the case of the number of receivers over two.

The raw gesture data, which in our case are values of 18-joint angles collected from the Cyberglove for each hand and 12 positions and orientations collected from two receivers. For two hands, they are formed as a 48 dimensional vector. However, The range of each angle value is within 0-255. The dynamic range of each component is different. Each component value is normalized to ensure its dynamic range is 0-1.

After these transforms, the feature vector used for recognition is formed as nine relative direction cosines of the left hand relative to right hand with appended the three relative position components of the left hand relative to the right hand and 36 –joint normalized angles of two hands.

3. Recognition Approaches

The most popular framework for the sign recognition problem is a statistical formulation in which we choose the most probable word sequences from all word sequences that could have possibly been generated. For a sequence of words $W = w_1, w_2, \cdots w_n$, suppose that F is the feature evidence extracted from the data collected by Datagloves that is provided to the system to identify this sequence. The recognition system must choose a word string $\hat{W}$ that maximizes the probability that the word string W was gestured given that the feature evidence of hand gesture F was observed. This problem can be significantly simplified by applying Bayesian approach to finding $\hat{W}$:

$$\hat{W} = \arg\max_{W} P(F|W)P(W) \tag{3}$$

The probability, $P(F|W)$, is typically provided by the spatial models of hand gestures. The likelihood $P(W)$ that denotes the a priori chances of the word sequence W being gestured is determined using a language model.

3.1. Spatial models of hand gestures

Hidden Markov Models (HMMs)[17] have been used successfully in continuous speech recognition, handwriting recognition, etc. A key assumption in stochastic gesture processing is that the gesture signal is stationary over a short time interval. In the case of continuous recognition, one difficult problem is the coarticulation problem, coarticulation means that both the sign in front of it and the sign behind it can affect a sign. If two signs are performed in succession, an extra movement from the end position of the first sign to the start position of the second sign appears. To take into account of the effect of epenthesis, the basic sign HMMs are concatenated to form a large lexical HMM for each training sign sentence. Then, the Dynamic Programming (DP) algorithm [19] is used to segment the training sentence into basic units that are then sorted into individual basic unit files for further re-estimating by Welch-Baum algorithm. Suppose that we have a sequence of training data for a sentence $x(t)$, $t = 1,2,\cdots,T$, T is the number of total frames of training data. The T frames are divided into N segments with boundaries $t_0 = 0$, $t_N = T$. Each segment corresponds to a word in the sentence. The average segmentation probability of each segment is defined as

$$P(t_{n-1}+1,t_n)=\frac{1}{t_n-t_{n-1}}\log P\left(x(t_{n-1}+1),,\cdots,x(t_n)\middle|Word_n\right) \qquad (4)$$

Where $P\left(x(t_{n-1}+1),,\cdots,x(t_n)\middle|Word_n\right)$ denotes the probability of the *word$_n$* HMM. The task is to find the segment boundaries $t_1,\cdots,t_{N-1}$ so that

$$\sum_{n=1}^{N}P(t_{n-1}+1,t_n) \qquad (5)$$

is maximized. Dynamic programming offers an efficient solution [19]. We introduce an auxiliary function $F(n,t)$ which denotes the probability of the best segmentation of the frame interval *[1,t]* into *n* segments. By decomposing the frame interval *[1,t]* into two frame intervals *[1,j]* and *[j+1,t]* and using the optimality in the definition of $F(n,t)$, we can obtain the recurrence equation of dynamic programming:

$$F(n,t)=\max_j\{F(n-1,j)+P(j+1,t)\} \qquad (6)$$

As equation shown above, the best segmentation of the frame *[1,j]* into *n-1* segments is used to determine the partition of the frame interval *[1,t*] into *n* segments. The optimal segment boundaries are calculated along with the maximum log likelihood $F(N,T)$ by recursively applying Eq.6.

3.2. Bigram language models

The language model provides constraints on the sequences of words that are to be recognized. In bigram models, we make the approximation that the probability of a word depends on only the immediately proceeding word. To make $P(w_1|w_0)$ meaningful ,we asumme that the beginning of the sentence with a distinguished token<*bos*>, that is w_0=<*bos*>. For a word strings *W*, the probability over *W* is as

$$P(W)=P(w_1,w_2,\cdots w_n)=\prod_{i=1}^{n}P(w_i|w_{i-1}) \qquad (7)$$

To estimate $P(w_i|w_{i-1})$,the frequency with which the word w_i occurs given that the last word is w_{i-1}, we can simply count how often the bigram occurs in some the training corpus. If the training corpus is not large enough, many actually existing word successions will not be well enough observed, leading to many zero probabilities. So smoothing is critical to make the estimated probability robust for unseen data. In this paper, we use the Katz smoothing [18] to smooth the zero probabilities. The bigrams indirectly encode syntax, semantics and pragmatics by concentrating on the local dependencies between two words. The net result of the techniques is to limit the number of alternatives that must be searched to find the most probable sequence of words. Hence the bigram language model reduces the search space. The corpus to estimate the bigram probabilities consists of about 30 million Chinese words in the Chinese newspapers from the year 1993 to 1995.

3.3. Search algorithms

Viterbi search and its invariant forms belong to a class of breadth-first search techniques. All hypotheses are pursued in parallel and gradually pruned away as the correct hypothesis emerges with the maximum score. In this case, the recognition system can be treated as a recursive transition network composed of the states of HMMs in which any state can be reached from any other sate. In Viterbi beam searches only the hypothesis whose likelihood falls within a beam of the most likely hypothesis are considered for further growth. The best beam size is determined empirically. By expanding the network to include an explicit arc from the end of each word to the start of the next, the bigram language model has been incorporated to improve recognition accuracy. Signs can be classified into one-handed and two handed. For the case of one-handed, the left hand is motionless, the right hand conveys the information of a sign. To reduce the computation load, the right hand shape and position information is used firstly to prune unlikely words.

3.4. Frame predicting

Since some gesture signals change slowly, the observation probabilities do not usually change dramatically from one frame to the next. Therefore, the gesture feature of the preceding frame can be used for predicting the gesture feature of the current frame. If the distance between the two frame features is below a threshold, the observation probabilities of the current frame is assumed to be those of the preceding frame. This technique reduces the computation effort without loss of noticeable accuracy if suitable threshold is chosen.

4. System Outline

The baseline system organization is shown as the following. The gesture-input devices are two DataGloves with 18 sensors measuring hand joint-angles and two

3SPACE-position trackers for each hand. The data collected by the gesture-input devices is fed to the feature extraction module, the feature vectors are input to the fast match module. The fast match finds a list of candidates from 220 words. The Bigram model uses word transition probabilities to assign, *a priori*, a probability to each word based on its context. The system combines the fast match score to obtain a list of no more than 150 candidates, each of which is then subject to a detailed match. The decoder controls the search for the most likely word sequence using the search algorithm described in section 3.When the word sequence is output from the decoder, each word drives the speech synthesis model and 3D-virtual human gesture animation model to produce the speech and gestures synchronously.

5. Chinese Sign Language Synthesis

Our early synthesis system of CSL, speech and the corresponding facial expression driven by the text was reported in [20]. A Chinese sentence is input to the text parser module, which divides the sentence into basic words. The parser algorithm is the Bi-directional maximum matching with backtracking approach. After the words are matched, each word in the sentence is then input to the sign language synthesis module and speech synthesis module. For the time being, the word library consists of above 3163 words.

6. Synchronously Driving

Conversion the text recognized by the sign language system to speech is important to communication between deaf people and hearing society, while text to synchronous gesture and lip movement is useful for long distance communication between deaf and deaf. To synchronously drive speech and gestures, we align the display time of gesture to the playing time of speech so that human perceives it comfortable. The underlying assumption of this approach is that the display speed of gestures is faster than that of playing speech. Fortunately, this requirement is usually satisfied.

7. Experiments

The database of gestures consists of 220 words and 80 sentences. To collect training and test gesture samples, the gesture of each isolated word was performed five times by a sign language teacher, four for training and one for test. In general, each sentence consists of 2 to 15 words. No intentional pauses were placed between signs within a sentence.

For the isolated word, recognition rates are listed in the Table 1.The recognition rate with feature normalization is 99.1 % and the recognition rate without feature normalization is 97.3%. This result shows that feature normalization is necessary to increase system performance.

Table 1. The recognition rates of isolated word test

Without Feature Normalization	97.3%
With Feature Normalization	99.1%

Table 2.The recognition rates of sentence level test

Without Embedded Training	73.6%
With Embedded Training	98.2%

Table 3. The recognition rates of sentence level test

With Embedded Training (Without Bigram Model)	94.7%
With Bigram Model(Without Embedded Training)	79.4%

To test the recognition performance of sentence level, one test was carried out as described the following. When 80 sentences were not used for any portion of the training. The 220 words were used as basic units. Within 80 sentences, 43 sentences can be correctly recognized, the left 37 sentences have deletion (D), insertion (I), and substitution (S) errors,D=21,I=30, S=53. This shows that the movement epenthesis has greatly affected on the sentence level recognition performance. To take into this effect, the embedded training technique was used for estimating the context dependent HMMs. For the left 37 sentences, the embedded training procedure was used for each of sentence. In the collected sentence samples, four in five were used for training and one for test. The word recognition rate is 98.2%, where D=3, S=2,I=2,N=394, N denotes the total number of signs in the test set. The accuracy measure is calculated by subtracting the number of deletion, insertion substitution and errors from the total number of signs and dividing by the total number of signs. The result shows that context dependent models are necessary for sentence level recognition. To reach such a high recognition rate, bigram model is also necessary, otherwise the recognition performance becomes worse. Table 3 shows this comparison. The recognition rate with embedded training but without bigram language model is 94.7%, where D=9,S=7,I=5.The recognition rate with bigram but without embedded training is 80.7%, D=171,I=21, S=43. This indicates that the effect of context dependent model is higher than that of bigram language model.

8. Summary and Conclusion

A CSL system has been developed using HMMs based recognizers. We have developed a new gesture feature extraction approach and an embedded training approach. The performance of these techniques ware evaluated using the CSL recognition system. Experimental results have shown that these techniques are capable of improving both the recognition performance and speed. The approach for extracting signer position independent features is very powerful for sign language recognition system in practical applications.

References

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

[2] T.Starner.Visual recognition of American Sign Language using hidden Markov models. *Master's thesis, MIT Media Laboratory,* July. 1995.

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

[4] S.Sidney Fels. *Glove –TalkII: Mapping* Hand Gestures to Speech Using Neural Networks-An Approch to Building Adaptive Interfaces. *PhD thesis,Computer Science Department,University of Torono*,1994.

[5] Alan Daniel Wexelblat.A feature-based approach to continuous gesture analysis. *Master thesis,MIT*,1993

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

[7] Kouichi Murakami and Hitomi Taguchi. Gesture recognition using recurrent neural networks *In CHI'91 Conference Proceedings,* pages 237-242. Human Interface Laboratory,Fujitsu Laboratories,ACM,1991

[8] James Kramer and Larry J.Leifer. A "Talking Glove" for nonverbal deaf individual. *Technical Report CDR TR 1990 0312,*Center For Design Research-Standford University, 1990

[9] Mohanmmed Waleed Kadous. Machine recognition of Auslan signed using PowerGlove: Towards large-lexicon recognition of sign language. *In Lynn Messing,editor, Proceedings of WIGLS. The Workshop on the Integration of Gesture in Language and Speech,* pages 165-174,Applied Science and Engineering Laboratories Newwark, Delaware and Wilmington, Delaware, October 1996.

[10] 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*,pages 558-565,Nara,Japan,1998.

[11] 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*,pages1-8,1995.

[12] S.B.Kang and K.Ikeuchi.Robust task programing by human demonstration. *In Proceedings of the Image Understanding Workshop*,1994.

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

[14] Christian Vogler and Dimitris Metaxas. Adapting hidden Markov models for ASL recognition by using three-dimensional computer vision methods. *In Proceedings of the IEEE International Confference on Systems,Man and Cybernetics*,pages156-161,Orlando, FL,1997.

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

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

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

[18] Slava M.Katz,"Eestimation of probabilities from sparse data for the language model component of a speech recognizer. *IEEE Trans. Acoustic, Speech, Signal Processing,* 1987,pages,400-401.

[19] R.Bellman,S.Dreyfus.Applied Dynamic Programming. P*rinceton University, Princeton,* NJ,1962.

[20] Wen Gao,Yibo Song,Baocai Yin, Jie Yan and Ying Liu. Synthesis of Sign Language ,sound and corresponding facial expression driven by text. *In Multimodal Interface.In Proceedings of the First International Conference on Multimodal Interface,* pages 244-248,1996,Beijing, China.