A Novel Approach to Automatically Extracting Basic Units from Chinese Sign Language

Gaolin Fang[1], Xiujuan Gao[1], Wen Gao[1, 2], Yiqiang Chen[2]
[1]*Department of Computer Science, Harbin Institute of Technology, Harbin, 150001, China*
[2]*Institute of Computing Technology, Chinese Academy of Sciences, Beijing, 100080, China*
{glfang, xjgao, wgao, yqchen}@jdl.ac.cn

Abstract

In sign language recognition, using subwords instead of whole signs as basic units will scale well with increasing vocabulary size. However, there are no subwords defined in the signs' lexical forms. How to automatically extract subwords is a challenging issue. In this paper, a novel approach is proposed to automatically extract these subwords from Chinese sign language(CSL). Signs can be broken down into several segments using hidden Markov models in which each state represents one segment. Temporal clustering algorithm is presented to extract subwords from these segments. The 238 subwords are automatically extracted from 5113 signs, and they can be used as the basic units for large vocabulary CSL recognition with good performance.

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 is to provide an efficient and accurate mechanism to transcribe sign language into text or speech. Sign language recognition (SLR), as one of the important research areas of human-computer interaction (HCI), has spawned more and more interest in HCI society. Attempts at machine sign language recognition began to appear in the 90's. Previous work mainly focused on using signs as basic units and hidden Markov models(HMM) as the recognition method. Grobel [1] used HMM to recognize isolated signs with 91.3% accuracy out of a 262-sign vocabulary. Liang[2] employed HMM for the continuous Taiwan SLR with the accuracy of 80.4% for 250 signs based on a Dataglove. Starner[3] used a view-based approach for continuous American SLR with 40 signs. They used single camera to extract two-dimensional features as the input of HMM. Vogler[4] coupled the computer vision methods and HMM to recognize continuous American sign language with a vocabulary of 53 signs.

The sign-based system cannot scale well with increasing vocabulary size. In large vocabulary SLR, the huge search space arisen from a variety of recognized classes leads to the training and recognition difficulties. To overcome this problem, the subword-based systems are proposed.

Stokoe[5] proposed that each sign can be broken down into three parameters: location, handshape, and movement. He used this observation to devise a transcription system. This system assumes that three parameters are performed simultaneously. However, Liddel and Johnson [6] do not agree with this approach because this system cannot describe the sequential contrast in sign language. They proposed segmental models where the sequential feature is emphasized instead of the simultaneous occurrence. In segmental models, each sign is partitioned into the sequences of movements and holds. Movements are characterized as those segments during which some aspects of the signer's configuration changes. Holds are characterized as those segments during which all aspects of the signer's configuration remain stationary. Based on this model, Vogler[7] [8] broke down signs into subwords for SLR. They used subwords as basic units, experimented with 22 words and achieved similar recognition rates with word-based approach.

However, it is difficult to transcribe sign language into the composition of subwords. No unified lexicon of transcription exists for sign language, hence the transcription has to be undertaken manually, which is almost infeasible for large vocabulary. Bauer[9] used k-means clustering algorithm to self-organize the subunits for sign language. Using these subunits as basic units, they developed HMM-based continuous SLR system. The accuracy of 80.8% was achieved in the corpus of 12 different signs and 10 subunits.

Bauer's clustering is built on every frame of signs that cannot directly include temporal information. However, subwords are the sequence signal in which the sequential feature is very important for sign performance. In this paper, a novel approach is proposed to automatically extract subwords from CSL. Signs can be broken down several segments using HMM in which each state represents one segment. Temporal clustering algorithm is presented to extract subwords from these segments. The

0-7695-2128-2/04 $20.00 (C) 2004 IEEE

proposed algorithm efficiently utilizes both spatial information and temporal information of subwords.

2. Sign segmentation

Unlike speech, there are no subwords defined in the signs' lexical forms. Manually segmenting a sign sequence into subwords is difficult for large vocabulary. However, in HMM, each sign can be broken down several states. In the duration of one state, the variability is very small. Intuitively, each state is associated with one basic unit. Different states in all signs exist the same or similar basic units. Thus, we can regard each state as the subword of sign language.

The sign segments can be gotten through the following two steps: First, one HMM model can be built for each sign. We use four samples of each sign to train the parameters for HMM, in which each sign has 1 mixture component, the state number (From 3 to 5) depends dynamically on the lengths of signs. Second, one samples of each sign is used for segmentation. The sample is recognized through the corresponding sign's HMM, and the state sequence for each frame can be obtained. According to the state sequence, the duration of each state represents one segment. All these segments are regarded as the training samples for temporal clustering.

3. Temporal clustering

In order to obtain the subwords for CSL, we need to cluster all the segments gotten from sign segmentation. Since the segments are the time sequence of the vector, the clustering algorithm is required to handle not only the spatial vector but also the temporal sequence information. Furthermore, there is no criterion to describe how many clusters are very rational, so we must dynamically cluster the vector sequence according to the data distribution.

The k-means clustering algorithm cann't handle the temporal data because its distance measure only builds between the two spatial vectors. Wilpon[10] proposed modified k-means algorithm(MKM) for producing the robust matching templates for speaker-independent speech recognition. However, MKM cannot dynamically cluster the data. In this paper, temporal clustering algorithm based on MKM is proposed to cluster the temporal sequence of the vectors. DTW is employed as the distance computation criterion because it can measure the distance between two temporal sequences by aligning different time signals and normalizing them to a warping function. In the algorithm, the corresponding skills are proposed to solve the issues of cluster splitting and combination. The proposed algorithm can automatically split and combine the centroids according to the data distribution to obtain the more reasonable cluster number and centers. The following subsection will discuss DTW-based distance computation and temporal clustering algorithm in detail.

3.1. DTW-based distances computation

Dynamic time warping(DTW) is to search the best warping function using the dynamic programming technique so as to minimize the distance between the two temporal signs. Let two sign segments $X=(X_1,X_2,...,X_{T_X})$, $Y=(Y_1,Y_2,...,Y_{T_Y})$, where X_i and Y_i are the 48-dimensional vectors, and T_x, T_y are frame numbers. Define the warping function $\phi=\{\phi(1),\phi(2),...,\phi(N)\}$, where N is the "normal" duration of the two segments on the normal time scale, and $\phi(n)=(\phi_X(n),\phi_Y(n))$, $\phi_X(n)\in\{1,\cdots,T_X\}$, $\phi_Y(n)\in\{1,\cdots,T_Y\}$. The n-th matching pair $\phi(n)$ consists of the $\phi_X(n)$ vector in X and the $\phi_Y(n)$ vector in Y.

The measure $d(\phi_X(n),\phi_Y(n))$ is defined as the Euclidian distance. The aim of DTW is to search the minimal accumulating distance and the associated warping path, that is:

$$D(X,Y)=\min_{\phi}\sum_{n=1}^{N}d(\phi_X(n),\phi_Y(n)) \qquad (1)$$

The warping functions used in our experiment satisfy endpoint constraints, monotony constraints and one-step local continuity constraints.

The minimum partial accumulated distortion along a path from (1,1) to (i_X,i_Y) is defined as:

$$D(i_X,i_Y)=\min_{\phi,T'}\sum_{n=1}^{T'}d(\phi_X(n),\phi_Y(n)), \qquad (2)$$

where $\phi_X(T')=i_X$ and $\phi_Y(T')=i_Y$.

The auxiliary parameter $\psi(i_X,i_Y)$ is defined to record a point before the point (i_X,i_Y) in the local optimal path. The recursive relations according to the constraints are given as follows:

$$D(i_X,i_Y)=\min_{(i_X',i_Y')}[D(i_X',i_Y')+d(i_X,i_Y)] \qquad (3)$$

$$\psi(i_X,i_Y)=\arg\min_{(i_X',i_Y')}[D(i_X',i_Y')+d(i_X,i_Y)] \qquad (4)$$

where $(i_X',i_Y')\in\{(i_X-1,i_Y),(i_X-1,i_Y-1),(i_X,i_Y-1)\}$.

Through the dynamic programming search, the minimal distance $D(X,Y)$ between the two segments and the associated warping function pair ϕ are finally obtained.

3.2. Temporal clustering algorithm

Let $\Pi=\{O_1,O_2,...,O_V\}$ be a data set for V sign segments to be clustered. Temporal clustering algorithm is to dynamically cluster the c centers $\{\Gamma_j;j=1,2,...,c\}$,

and get $\Pi = \bigcup_{j=1}^{c} \Gamma_j$.

The temporal clustering algorithm is described as follows:

1. Initialization:

Calculate all distances $d(O_i, O_j)$ using DTW. Set the initial parameters: c - the number of clusters, C - the expected number of clusters, θ_N - the minimum number of samples in each cluster, θ_C - the threshold of the intercluster distance that determines whether to combine or not, t - the number of iteration, and $t_{\max}$ - the maximum iterations.

2. Initialize the cluster centers:

The method described in [10] is employed to set the initial cluster centers. It splits the clusters from one to the expected number C step by step.

3. Classification:

According to the minimum DTW distance rule, each sample is classified to the corresponding center.

For each cluster, if its sample number is less than θ_N, then this cluster is discarded, and set $c = c - 1$, and re-classify the samples in this cluster.

4. Recalculate the cluster center:

The recalculation is described by the following two steps:

First, find the pseudo-average center O'. A particular segment in the cluster has the largest population of segments (subset of the cluster) whose distance to the particular sample falls within a threshold. If several patterns have the same largest count of segments with distances below the threshold, then the segment that has the smallest average distance to all segments in the subcluster is chosen as the pseudo-average center.

Second, all samples in Γ_j are warped to the pseudo-average center O'. We then group the samples according to their individual warping paths with respect to O'. The vectors that are aligned to the same index i are then averaged to produce an average vector for the new cluster. The resultant sequence with vectors indexed from 1 to $T_{O'}$ (duration for O') is the average cluster center $m(\Gamma_j)$.

5. If $t \bmod 2 = 0$ **or** $c \geq 2C$ **, then goto step 7, else goto step 6.**

6. Cluster splitting:

Calculate intracluster distance λ_j for each cluster j:

$$\lambda_j = \frac{1}{\|\Gamma_j\|} \sum_{O \in \Gamma_j} d(m(\Gamma_j), O), \quad j = 1,2,..,c \qquad (5)$$

Find the cluster $\Gamma_{j\max}$ with the maximum intracluster distance, if $\|\Gamma_{j\max}\| \geq 2\theta_N$ or $c \leq C/2$, then split $\Gamma_{j\max}$ as follows. Find two segments O_{p1} and O_{p2} satisfying $d(O_{p1}, O_{p2}) \geq d(O_{p3}, O_{p4})$ for any other pair O_{p3}, O_{p4} in $\Gamma_{j\max}$. The two segments O_{p1} and O_{p2} are used as the new cluster centers to replace original cluster, and set $c = c + 1$, then goto step 8

7. Cluster combination:

For all the cluster centers, calculate the intercluster distances $d(m(\Gamma_i), m(\Gamma_j))$ between all the pairs. Find the pair with the minimum interclass distance $d(m(\Gamma_p), m(\Gamma_q))$, if $d(m(\Gamma_p), m(\Gamma_q)) < \theta_C$, then combine Γ_p and Γ_q. Using DTW the optimal path between the sequences Γ_p and Γ_q is gotten. Let T be the warping path length for ϕ, and the new cluster $\overline{m}$ is calculated:

$$\overline{m}_k = \frac{1}{2}(m(\Gamma_p)_{\phi_x(k)} + m(\Gamma_q)_{\phi_y(k)}), \quad k = 1,2,...,T \qquad (6)$$

Replace these two clusters with new cluster $\overline{m}$, and set $c = c - 1$.

8. $t = t + 1$ **, if** $t < t_{\max}$ **, then return to step 3, otherwise, save the clusters data and exit.**

4. 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. 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. To extract the invariant features to signer position, the trackers 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 as invariant features. By this transformation, the data consist of a relative three-dimensional position vector and a three-dimensional orientation vector for each hand. In the case of two hands, a 48-dimensional vector is formed, including hand shape, position and orientation vector. As each component in the vector has different dynamic range, its value is normalized to [0,1].

Two experiments are performed: one is to judge the clustering validation of the temporal clustering algorithm, the other is to automatically extract subwords from CSL.

The first experiment validates that the proposed algorithm can cluster the similar segments into one class. Database consists of 1268 samples from 317 signs which are random selected among 5113 signs, each having four samples. Because the corresponding classes are known beforehand, the clustering validation can be judged. The expected cluster center is set to 317. After the processing of temporal clustering algorithm, the 309 cluster centers

can be gotten. The 301 centers are the same as the sign data, and each has four samples. The rest 8 centers are the sample combination of two signs.

In the 8 centers, they can be classified into three categories. One is that the two signs have the same action, such as zhu-ren(director) and zhu-chi(preside), where the word before the bracket is the Chinese sign in PinYin, and the word in the bracket is the English meaning. The second is that two signs have the same postures, but only small differences in position, such as zhong-zu(race or tribe) and zhong-lei(category). The third is that two signs have very similar postures, where one has slight movement and the other hasn't. In this experiment, there is only one case, i.e. J and jiu-shi(ninety) in which the sign J is static, and ninety has a slight movement of first finger. Figure 3 shows the description of J and ninety.

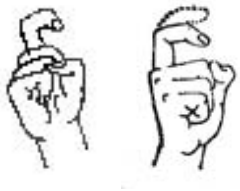

Figure 3. The description of the words "J" and "ninety", left for "J", and right for "ninety"

From this experiment we can know the temporal clustering algorithm can efficiently cluster the segments with high similarity into the same cluster.

Table 1. The result of clustering

C	$t_{\max}$	θ_N	θ_C	c
300	200	2	0.05	238

The second experiment is to extract subwords from CSL. Database consists of 12113 segments from 5113 signs through sign segmentation. The initial parameters are set to the values as listed in Table 1. The 238 clusters are gotten from these segments, which can be regarded as the subwords in CSL. The experiments use these subwords as the basic units for SLR. In 5113 signs, one samples of signs are used for encoding, another samples are used as the test set. The 90.5% recognition rate can be gotten. Using those subwords instead of whole signs as the basic units will scale well with increasing vocabulary size, and it can be extended for larger scale SLR.

5. Conclusions and future work

In this paper, a novel approach to automatically extracting subwords from CSL is proposed. Signs can be broken down into several segments using HMM in which each state represents one segment. Temporal clustering algorithm is presented to extract subwords from these segments. The 238 subwords are automatically extracted from CSL, and they can be regarded as basic units of CSL. The temporal clustering algorithm is not only used in this research but also further extended to the clustering of temporal signals.

There are many issues to be further investigated in the future. These extracted subwords can be used as the basic units for both isolated SLR and continuous SLR. This will alleviate the computational difficulty due to large vocabulary and scale well with increasing vocabulary size. Using those subwords as the basic units for recognition, it can reduce the number of epenthesis subwords between two signs because we can model the epenthesis between two subwords rather than between two signs.

6. Acknowledgment

This work was supported in part by Natural Science Foundation of China (Grant No.60303018) and National High-Technology Development '863' Program of China (Grant No.2001AA114160).

7. References

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

[2] Liang R.H., and Ouhyoung, M. A real-time continuous gesture recognition system for sign language, *Proc. Third Int'l Conf. FG' 98,* pp. 558-565, 1998.

[3] Starner, T., Weaver, J., and Pentland, A. Real-time American sign language recognition using desk and wearable computer based video, *IEEE Trans. PAMI,* 20(12): 1371-1375, 1998.

[4] Vogler, C., and Metaxas, D. ASL Recognition Based on a Coupling between HMMs and 3D Motion Analysis, *Proc. IEEE Int'l Conf. Computer Vision*, 1998, pp. 363-369.

[5] Stokoe, W.C. *Sign Language Structure: An Outline of the Visual Communication System of the American Deaf*, Studies in Linguistics: Occasional Papers 8. Linstok Press, 1960. Revised 1978.

[6] Liddel, S.K., and Johnson, R.E. American Sign Language: The phonological base, *Sign Language Studies*, pp. 64:195-277, 1989.

[7] Vogler, C., and Metaxas, D. Toward scalability in ASL Recognition: Breaking Down Signs into Subwords, *Proc. Int'l Gesture Workshop*, pp. 400-404, 1999.

[8] Vogler C., and Metaxas, D. A framework for recognizing the simultaneous aspects of American sign language, *CVIU,* 81(3): 358-384, 2001.

[9] Bauer, B., and Kraiss, K. Towards an Automatic Sign Language Recognition System Using Subunits, *Proc. Int'l Gesture Workshop*, pp. 64-75, 2001.

[10] Wilpon, J.G., and Rabiner, L.R. A modified K-means clustering algorithm for use in isolated word recognition, *IEEE Trans. ASSP,* 33: 587-594, 1985.