Static Gesture Quantization and DCT Based Sign Language Generation

Chenxi Zhang[1], Feng Jiang[1], Hongxun Yao[1], Guilin Yao[1], Wen Gao[1,2]

[1]School of Computer Science and Technology, 15 00 01
Harbin Institute of Technology, P.R.C
{cxzhang, fjiang, yhx, glyao}@vilab.hit.edu.cn
[2]Institute of Computing Technology, CAS, 10 00 85
Beijing, P.R.C
wgao@ict.ac.cn

Abstract. To collect data for sign language recognition is not a trivial task. The lack of training data has become a bottleneck in the research of singer independence and large vocabulary recognition. A novel sign language generation algorithm is introduced in this paper. The difference between signers is analyzed briefly and a criterion is introduced to distinguish the same gesture words of different signers. Basing on that criterion we propose a sign word generation method combining the static gesture quantization and Discrete Cosine Transform (DCT), which can generate the new signers' sign words according to the existed signers' sign words. The experimental result shows that not only the data generated are distinct with the training data, they are also demonstrated effective.

1 Introduction

The purpose of Sign Language Recognition (SLR) is to provide an effective and accurate mechanism to translate sign words to texts or common language, to make it more convenient to communicate between the deaf and the normal by computers. Many researchers have documented methods for recognizing sign language from instrumented gloves at high accuracy while these systems suffer from notable limitations: signer-dependent and small vocabulary [1-5].

The main methods of Chinese Sign Language Recognition are based on HMM. Later ANN/HMM [6], DGMM/HMM [7] SOFM/HMM [8], DTW/HMM [9] recognition systems were presented. These systems implemented the signer-independent SLR with a large vocabulary. Although the systems are signer independent, there are only 7 signers' gesture words during the training the testing process which restricts the signer-independent SLR's effectiveness. In SLR, one of the problems is to collect enough data. Data collection for both training and testing is a laborious but necessary step. All of the statistical methods used in SLR suffer from this problem. However, sign language data cannot be gotten as easily as speech data. We must invite the special persons to pantomime. If Datagloves are used to collect the data, the difficulty will increase enormously. Data gloves are extremely expensive, so there are maybe

only one or two pairs in a research institution. Signers have to pantomime one by one. Besides, the sensors on the data gloves are brittle. The lack of data makes the research, especially the large vocabulary signer-independent recognition, very formidable. Due to the very large Chinese sign vocabulary, one more signer to sample the training data, much more time and money it would cost. Therefore, generating new signers' sign words from the existed signers' sign words is an imperative job.

In order to achieve effective generation performance, the critical problem is Signal Analysis. This paper first analyzes the sign word signals of different signers, and presents a sign word generation approach that comes from static gesture quantization and DCT method. We also propose a criterion to measure the common and distinct features of the same sign word from different signers. The experiment demonstrates that the generated sign word data not only differ from the existed training data, they are also demonstrated correct after recognizing by our recognition system.

The remainder of this paper is organized as following. In section 2, sign word data and the features that different signers perform are analyzed; Section 3 describes the sign word generation approach basing on static gesture quantization and DCT in detail. Experimental result is presented in the last section.

2 Sign language data analysis

Data gloves are adopted in this paper as the input equipment. American Virtual Technologies Company's CyberGlove with 18 sensors and three Polhemus FASTRAK 3-D position trackers are utilized as input devices. The 51-dimension vectors got from two CyberGloves and position trackers in every moment, function as the final input data.

Fig. 1. The three part of the word of "aunt"

Communicative gestures can be decomposed into three motion phase: preparation, stroke, and retraction. Psycholinguistic studies show that stroke may be distinguished from the other gesture phases, since stroke contains the most information. Generally speaking, stroke is composed of three parts: the beginning gesture, the terminative gesture, and the transition movement from the beginning to the terminative gesture in Chinese sign language. The variation process of the word 'aunt' can be obviously saw from figure 1, the first sub-figure is the sign word's beginning static gesture, the third sub-figure is the sign word's terminative static gesture, the middle stage is the partition period from the beginning gesture to the terminative gesture

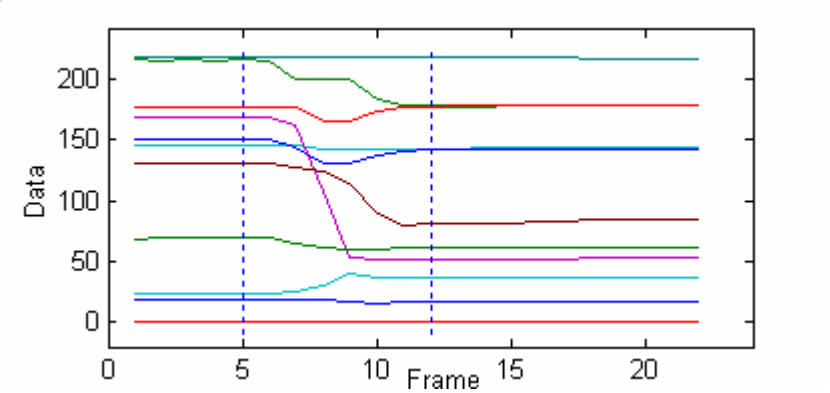

Fig. 2. Partial original data of the sign word 'aunt' obtained from the sensors. The lines represent the values changing with the time of the sensors. The two vertical lines divide all the curves into three parts. In the first part, the curves change gently, which means the signers are in the beginning static period; all the curves in the third part that change little indicate that the signers are in the terminative static gesture period; the curves between them are in the middle variation period.

Different signers have their own rhythms which include time length, range, and data change when they are performing sign words. This paper considers these three characteristics as the criterion to measure the differences of the sign words of different signers.

2.1 Time length analysis

Time length is the length of time that the signers perform. It reflects the speed of individual signer. The time length we define here is just the middle part in figure 2. We omit the beginning and the terminative time because the errors that data collection procedure brings are very big. In the ideal situation, the beginning and the terminative static gesture are both one frame. Thus in order to find the middle period and compute the time length in the middle stage of each sign word, we must dissect the word first. To compute time length is to find the critical points, namely the X-coordinates of the two vertical lines. The steps are composed of forward and backward searching. Forward Searching is to find backward from the first frame in every dimension. If the subtract of current frame and the average of the following two frames is smaller than 1, stops searching and record the current X-coordinate. Or continue searching. The Backward Searching is similar to the Forward, which searches every dimension from the back to the beginning. In this paper K and K' represent the forward and backward searching result.

$$\text{AverageTime} = \frac{\sum_{i=1}^{N} K'^{(i)} - K^{(i)} + 1}{N}, \quad N = 4942$$

AverageTime is the average time length of the sign word an individual signer performs in average. Here the time is measured by the frame number of the data collec-

tion. $K'^{(i)} - K^{(i)'} + 1$ denotes the varying frame number in the middle stage of i th sign word. Compute the AverageTime as below.

Table 1. The comparison of the average time of one sign word 5 signers perform in average. The 4942 sign words performed by 5 signers independently are used as training sample.

Names of the signers	pfz	lwr	ljh	mwh	ygy
AverageTime（frame)	18.59	19.18	19.68	20.25	21.61

We can see from table 1 that the average performing speed of different signers are different. Thus time length can be used to measure the differences among different signers.

2.2 Range analysis

The same sign word's curves' changing trend is the same, whereas every sensor's value range is different. For instance, the value of a certain sensor of the data glove can reflect the bending extent of the thumb. Apparently, every signer has different bending extent. Therefore we could find all the value ranges of the 51-dimension data and their average value to measure the range feature of every signer. This difference measuring function $Distance(i, j)$ of each signer is presented later. The algorithm is as following:

Definition 1: The form of a word is $O = \{A_1, A_2, \cdots, A_i, \cdots, A_N\}$, where N represents the number of the dimension (in this paper $N = 51$), A_i is the 51-dimension vector: $A_i =< a_1, a_2, a_3, \cdots, a_{51} >^T$, $A_i(j) = a_j$ denotes that the value of time j for the ith sensor is a_j, $O_k(A_i)$ denotes the kth word of the i th sensor.

Definition 2: Matrix $D = \{min, max, mean\}$, min, max and $mean$ are all 51-dimension vectors. $min(i)$, $max(i)$, $mean(i)$ respectively denote the minimum, the maximum and the average of the i th sensor.

1. Compute $min(i)$ and $max(i)$. $frame(k)$ is the frame number of the k th gesture word. N is the overall number of the training words 4942. $i = 1, 2, \cdots 51$

$$min(i) = \min\{O_k(A_i(j)) \mid j = 1, 2, \cdots frame(k); k = 1, 2, \cdots, N\}$$

$$max(i) = \max\{O_k(A_i(j)) \mid j = 1, 2, \cdots frame(k); k = 1, 2, \cdots, N\}$$

2. Compute $mean(i)$.

$$M = \sum_{k=0}^{N} frame(k), \qquad mean(i) = \frac{1}{M} \sum_{k=1}^{N} \sum_{j=1}^{frame(k)} O_k(A_i(j)).$$

where $i = 1, 2, \cdots 51$.

Construct a feature Matrix for each signer. $Matrix_{3\times 51} = \{min, max, mean\}$. Now define the formula $Distance(1,2)$ that measures the two signer's $Matrix1 = \{min1, max1, mean1\}$ and $Matrix2 = \{min2, max2, mean2\}$ (1, 2 refers to two person):

$$Distance(1,2) = \alpha_1 \left\| min1 - min2 \right\| + \alpha_2 \left\| max1 - max2 \right\| + \alpha_3 \left\| mea1 - mean2 \right\|$$

where $\alpha_1, \alpha_2, \alpha_3$ are weights.

Table 2. *Distance* comparison of 5 signers . Consider the 4942 sign words that five different signers perform as the training sample.

Name / Name	pfz	lwr	ljh	mwh	ygy
pfz	63.04	255.27	332.60	346.59	365.13
lwr	—	91.19	372.29	362.57	428.13
ljh	—	—	54.01	339.31	341.15
mwh	—	—	—	64.54	360.69
ygy	—	—	—	—	57.04

We can see from Table 2 that the values on the diagonal are the result of the same signer who performs twice, '—' means the value is the same with the value that is symmetrical to the diagonal and are thus omitted. It is clearly shown the *Distance* of the same signer is smaller than 100, and the *Distance* of different signers is at least 255. It is arrived that the *Distance* of the same signer is much smaller than the *Distance* of different signers. By this token, *Distance* can be used to measure the range feature of different signers.

2.3 Data variation analysis

Data variation this paper analyzes is the varying period in the middle stage shown in Figure1 and Figure 2. The curves are almost the same as for the value of the same sign word of a certain dimension. Just as shown in Figure 3.

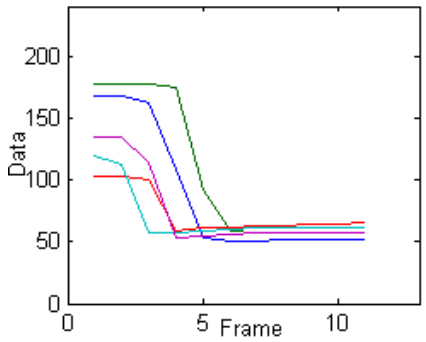

Fig. 3. Value variation curves of 'aunt' that is performed by 5 signers of the same dimension

We can clearly see that the curves' changing trends are the same: changing descendingly. Their initial values and terminative values are distinctive. The initial and terminative values are determinative and the middle data variation is aid to them. We are going to use this variation to aid the change of the initial and terminative values when generating sign words in the following Section.

3. Generating sign language

According to the above analysis, now we are proposing a sign language generation method basing on the static gesture quantization and DCT methods. The principle is to generate the beginning static gesture and the terminative static gesture for every sign word respectively, later modulate the data variation curve in the middle process without changing their curves to make them satisfy both the beginning and the terminative static gestures.

3.1 The generation of static gesture

We have found every sensor's value range of each signer: the minimum $min(i)$ and the maximum $max(i)$ from Section 2.1. Now quantize $min(i)$ and $max(i)$ in every dimension, and get the value in the quantization table for every sign word's beginning and terminative static gesture. We choose five signers' data.

1. Compute the quantization range of every dimension, RANGE is the quantization step length.

$$List_i(\mathrm{j}) = \mathrm{j} \times (max(\mathrm{i}) - min(\mathrm{i}))/\mathrm{RANGE} + min(\mathrm{i})$$

where $j = 1, 2, \cdots \mathrm{RANGE}$, $i = 1, 2, \cdots, 51$

2. Get the quantization value of the beginning and the terminative static gesture of every sign word: $O_k(frontG_i)$ and $O_k(backG_i)$ denote the beginning and the terminative static gesture quantization value of the i th dimension, the k th sign word. $O_k(frontG_i)$, $O_k(backG_i)$ have the following relationship:

$$List_i(\mathrm{j}) < List_i(O_k(frontG_i)) \le List_i(\mathrm{j}+1)$$

$$List_i(\mathrm{j}) < List_i(O_k(backG_i)) \le List_i(\mathrm{j}+1)$$

where $k = 1, 2, \cdots, 4942$.

3. The beginning and the terminative static gesture for each sign word are $O_k^{\,j}(frontG_i)$, $O_k^{\,j}(backG_i)$ ($k = 1, 2, \cdots, 5$. denotes 5 persons) Find the repeated values from the five $O_k^{\,j}(frontG_i)$ and give it to $O_k(generateFG_i)$ **.** If the five numbers differ from each other, let $O_k(generateFG_i)$ is the median of the five numbers.

Generate $O_k(generateBG_i)$ in the same way. $O_k(generateFG_i)$, $O_k(generateBG_i)$ denote the beginning and the terminative quantization value of the k th sign word respectively.

4. Generate the original values with reference to $O_k(generateFG_i)$ and $O_k(generateBG_i)$. $O_k(gestureF(i))$, $O_k(gestureB(i))$ denote the generation of the i th dimension value of the k th sign word's beginning and terminative static gesture, where $max'(i)$ and $min'(i)$ are the given maximum and minimum value.

$$O_k(gestureB(i)) = (max'(i) - min'(i)) \times (O_k(generateBG_i) + rand()) / \text{RANGE} + min'(i)$$

$$O_k(gestureF(i)) = (max'(i) - min'(i)) \times (O_k(generateFG_i) + rand()) / \text{RANGE} + min'(i)$$

where $i = 1, 2, \cdots, 4942$, $rand() \in [0,1]$. The algorithm stops here.

We choose the repeated value when computing $O_k(generateFG_i)$ in the third step, because when making statistics for the 4942 sign words, the probability of the number repeating of these 5 numbers is 64% (RANGE is chosen 50). The repeatability fully exhibits the common features when different signers are performing the same sign word. In the fourth step, $max'(i)$ and $min'(i)$ can be decided by ourselves, and this work is the very crucial point when generating data. $max'(i)$ and $min'(i)$ can reflect the range value of the newly generated virtual signer's data.

3.2 Generation of data variation curve

We generated the static gesture in the above section, now we are going to generate the varying data in the middle stage. From Figure 3 in Section 2, it is seen that the trend of the same sensor's value variation curves are the same when different signers are performing the same sign word, so the main purpose of this section is to find this sameness. The steps of this algorithm are as follows:

1. Dissect each sign word using the dissecting algorithm discussed in Section 2.1, find the stage of the middle variation.

2. Process Discrete Cosine Transformation for every dimension data. Find the DCT variation range of five signers' training data. $O_k(DCT\text{coefficient}_i(j))$ denotes the j th cosine coefficient of the i th dimension data, the k th sign word in the middle variation stage. Find its maximum value $O_k(DCTmax_i(j))$ and the minimum value $O_k(DCTmin_i(j))$ for every j. To simplify the algorithm, it is supposed the middle varying frames of each sign word's five data are the same. $j = 1, 2, \cdots, frame$

3. Generate new cosine coefficients. $O_k(DCTgenerate_i(j))$ denotes the generated j th cosine coefficient of the i th dimension data, the k th sign word in the middle variation stage.

$$O_k(DCTgenerate_i(j)) = (O_k(DCTmax_i(j)) - O_k(DCTmin_i(j)) \times rand() + O_k(DCTmin_i(j)$$

Do inverse discrete cosine transform (IDCT) to $O_k(DCTgenerate_i(j))$ to get the final result.

Because the changing trend of the same dimension's sensor is the same, randomly select the coefficients in $[O_k(DCTmax_i(j)), O_k(DCTmin_i(j))]$ will not change the curve's changing trend. We have generated all the three parts of a sign word till now, but these three parts may not be continuous as is shown in Figure 4.

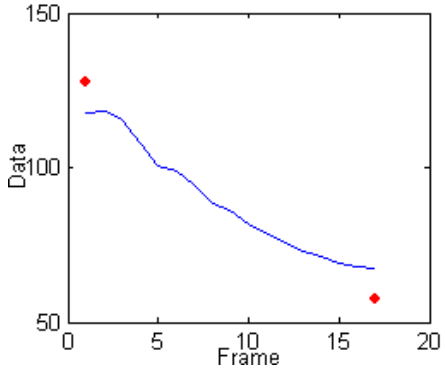

Fig. 4. Discontinuity of IDCT transformed curves and the static gesture.

IDCT transformed curves and the static gesture values are discontinuous, how to solve this problem? Considering the characteristics of DCT coefficient, let's make some modifications to DCT coefficient in order to meet the needs of continuity of the curves and the discrete points without changing the shape of the figure. The algorithm is described below:

1. Shift the curve to make the curve's two ends located in the middle or outside of the two discrete points, also make the distances from the curve's two ends to the two discrete points the same(as shown in Figure 4). This step just indicates changing the direct current sub-value of the DCT coefficient $O_k(DCTgenerate_i(0))$.

2. If the ends of the curve locate in the middle part of the two discrete points, change the value of $O_k(DCTgenerate_i(1))$. δ is the speed factor.

$$O_k(DCTgenerate_i(1)) = O_k(DCTgenerate_i(1)) + \delta .$$

3. If the ends of the curve locate outside of the two discrete points, change the value of $O_k(DCTgenerate_i(1))$.

$$O_k(DCTgenerate_i(1)) = O_k(DCTgenerate_i(1)) - \delta$$

4. When the distances from the ends of the curve to the two discrete points are the same, finish this algorithm; Or else, repeat Step 2 or Step 3.

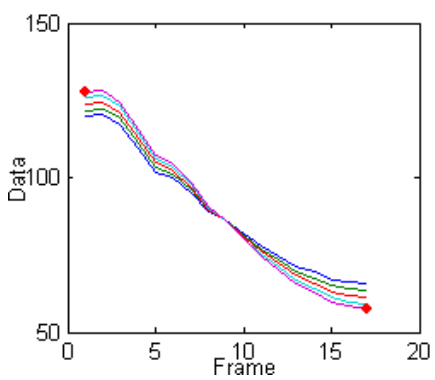

Fig. 5. The curve after modulating is continuous with the static gesture points.

From Figure 5 we can clearly understand the process shifting the curve closer and closer to the discrete points, which did not change the shape. According to compressible characteristic of DCT, the higher the DCT's frequency, the closer to 0 the coefficient is, thus we are able to change the frame number by increasing or reducing the high frequency coefficient without changing the shape.

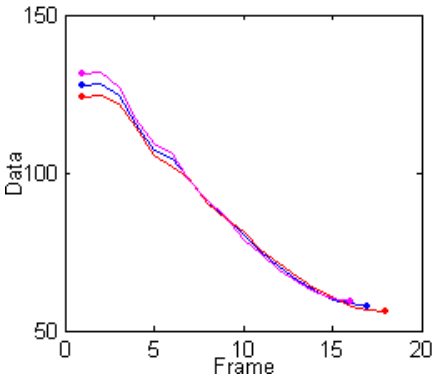

Fig. 6. Change the number of the high frequency coefficients in order to change the length of the curve. This is the picture after doing DCT transformation for one dimension's middle transitional stage, increasing a 0 to the high frequency coefficient, reducing the last high frequency and doing inverse transformation. We can see that even though the frame number changed, the shape did not change.

4. Experimental result

In this paper, the sign words performed by 5 signers are considered as training data. Sign words are the 4942 words chosen from《Chinese Sign Language Dictionary》. We generate three different signers' 4942 sign words by using the method described

in Section 3, and measure the newly generated sign words by using the criterion given in Section 2.

Table 3. Comparison of the AverageTime and Recognition accuracy of the newly generated 3 signers' data .

New data	A	B	C
AverageTime(frame)	19.54	20.42	21.29
Recognition accuracy in %	80.19	80.43	80.01

We can know from Table 3 the AverageTime of the newly generated 3 signers' data are different, because we changed the frame number in the middle transitional stage by changing DCT high frequency coefficient. We tested the newly generated sign word data by using the DGMM Recognizer in reference 7. The experiment shows the newly generated three signers' data all have the recognition accuracy above 80%, thus the method to generate sign word data presented in this paper is correct.

Table 4. Comparison of *Distance* of the newly generated 3 signers' data and the training data

Training data \ New data	A	B	C
pfz	377.86	363.80	325.53
lwr	407.23	396.51	276.81
ljh	339.45	426.75	378.17
mwh	298.14	342.69	297.98
ygy	317.61	387.55	416.44

Table 4 indicates all the *Distance* of the newly generated data and the training data are far above 255 which fully demonstrates the newly generated data can be distinguished from the training data. We can also see that the maximum number in Table 4 is 426.75 that is between signer ljh and the generated signer B. That means the feature difference between the generated signer B and signer ljh is the biggest.

We have generated another 3 person's data, which deviate the original signers' rhythm feature in the training data on the premise of accuracy, and successfully arrived at the purpose of generating sign language.

5. Conclusion

This paper analyzes the features of the rhythm of the same sign word performed by different signers, presents the formula for measuring the characteristics of time length, range and data variation of different signers. The sign language generation method basing on static gesture quantization and DCT which can generate new signer's sign word data according to the sign word data performed by the existed signers is then

given. This method is demonstrated by the experimental result that the new data are accurate and differ from the training data.

References

1. Charayaphan C, Marble A. Image processing system for interpreting motion in American Sign Language. Journal of Biomedical Engineering, 1992, 14 (15): 419-425
2. S. S. Fels and G. Hinton, “Glove Talk: A neural network interface between a DataGlove and a speech synthesizaer”, IEEE Transactions on Neural Networks, 1993, Vol. 4, pp.2-8.
3. M. W. Kadous, “Machine recognition of Auslan signs using PowerGlove: Towards large-lexicon recognition of sign language”, proceeding of workshop on the Integration of Gesture in Language and Speech, Wilmington, DE, 1996, pp.165-174.
4. 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, Nara, Japan,1998, pp. 558-565.
5. C. Vogler, D. Metaxas, “Toward Scalability in ASL Recognition: Breaking Down Signs into Phonemes”, In Proceedings of Gesture Workshop, Gif-sur-Yvette, France,1999, pp. 400-404.
6. Wu Jiangqin and Gao Wen. Sign Language Recognition Method on ANN/HMM. Computer science and application.No.9, pp 1-5.1999.
7. Wu Jiang-Qin,Gao Wen.A Hierarchical DGMM Recognizer for Chinese Sign Language Recognition. Journal of Software. Vol.11.No.11, pp 552-551.2000
8. Gaolin Fang, Wen Gao, Jiyong Ma, “Signer-Independent Sign Language Recognition Based on SOFM/HMM”, IEEE ICCV Workshop on Recognition, Analysis and Tracking of Faces and Gestures in Real-time Systems (RATFG-RTS 2001), Vancouver, Canada, 2001: 90-95
9. Feng Jiang, Hongxun Yao, Guilin Yao. “Multilayer Architecture in Sign Language Recognition”, Proceedings of the 5th International Conference on Multimodal Interfaces,2004:102-104.