# Automatic Synthesis of Training Data for Sign Language Recognition Using HMM

Kana Kawahigashi[1], Yoshiaki Shirai[2], Jun Miura[1], and Nobutaka Shimada[1]

[1] Department of Mechanical Engineering, Osaka University
[2] Department of Human and Computer Intelligence, Ritsumeikan University

**Abstract.** The paper describes a method of synthesizing sign language samples for training HMM. First face and hands regions are detected, and then features of sign language are extracted. For generating HMM, training data are automatically synthesized from a limited number of actual samples. We focus on the common hand shape in different word. The database hand shapes is generated and the training data of each word is synthesized by replacing the same shape in the database. Experiments using real image sequences are shown.

## 1 Introduction

The paper deal with recognition of sign language from a video sequence. Features of sign language consists of the position, velocity, and shape of hand regions. Sign language words are trained and recognized using Hidden Markov Model (HMM) [1]. For building reliable HMM, many training samples are necessary. However, collecting many samples from different subjects is difficult. Therefore we propose a method of synthesizing training data automatically from a limited number of actual samples.

In speech recognition, speech samples are synthesized from phoneme units [2]. In CG animations of sign language, sign sentences are generated by combining the patterns of words [3]. However, there is no methods for synthesizing training data of sign language. Here, we focus on the hand shape. The database of sign language word and its constituent contour shapes is generated from sign language images. The training data of each word is synthesized by replacing the contour shapes with the same shape in other words.

## 2 Feature Extraction

### 2.1 Extraction of Hand Regions

Face and hand regions are first extracted using a model of skin color, the range of which is determined from the initial image.

When hands and face regions overlap, they are separated using the previous an the succeeding frames. First, the image of the face and hands just before overlapping are saved as templates. Assuming that those images does not change, hands and face regions are extracted by template matching.

Next, assuming that hand shape changes during overlapping, the hand region is extracted. The image of face and hands just after overlapping are saved as templates and the regions are similarly extracted by template matching.

K. Miesenberger et al. (Eds.): ICCHP 2006, LNCS 4061, pp. 623–626, 2006.

Then the timing of the shape change is determined comparing the degree of matching in the forward backward template matching.

### 2.2 Extraction of Hand Features

From the extracted face and hands region, the positions, the velocities and the shapes of hands are obtained.

Because a small position difference of a hand near the face or a small movement near the face is often important, the positions and velocities of hands are:

- The log-distance between the face and each hand region.
- The change of the log-distance between the face and each hand region.
- The direction of each hand from the face region.
- The change of the direction of each hand from the face region.

Because, for two hand gestures, the relative position is important, the relative position of the right hand to the left hand is also included in the features

The shape features of the hand include:

- The number of protrusions of the hand region.
- $\{u(1-r), v(1-r)\}$ , where
  $(u, v)$: The x and the y component of the principal axis of the hand region.
  $r(0 < r < 1)$: The degree of circularity of the hand region.

Because the direction of the principal axis of a circular region is unstable, weight $1 - r$ is multiplied to decrease the value of the direction.

## 3 Generation of Initial HMM

Initial HMM is generated from states corresponding to the motion of hands. The image sequence is first segmented into static and moving periods. Even if a hand moves slowly, the corresponding static period is segmented into two if the moving direction changes significantly. Moreover, if the distance between the face and a hand is small and the direction of the hand from the face changes significantly in a static period, the period is also segmented into two.

The means and the variances of features in each state are calculated to create an initial model.

## 4 Synthesis of Training Data

### 4.1 Making a Hand Shape Database

Fukuda et al. proposed a classification of the hand shape used in sign language [4]. This classification is based on the finger alphabet of the Japanese syllabary. In order to deal with words which are not expressed by those alphabet, we add three hand shapes.

Moreover, since the view of a hand in images changes depending on a hand position and pose, we further extend the classification to include typical variations of the same hand shape. Figure 1 shows the extended classification.

Because the shape is important while a hand is not moving, the static hand shape features are saved in the database.

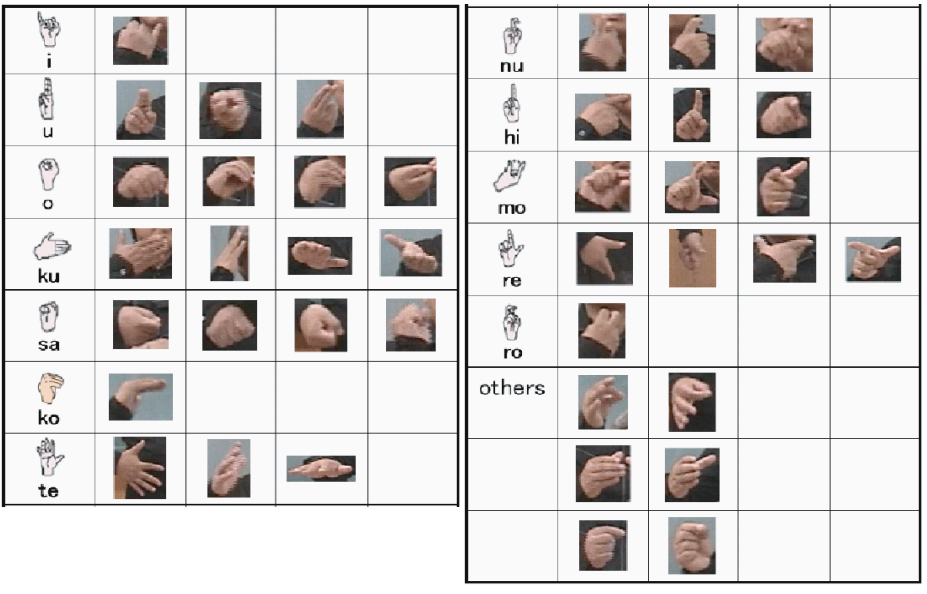


**Fig. 1.** Classification of hand shape

### 4.2 Synthesizing New Training Data

In order to synthesize a training data of a word, the degree of circularity is checked to find static periods. Then, for each hand shape, the database is searched for the similar hand shape. If it is found, the shape features replace the original features, while the other features such as the position and the motion are kept unchanged.

## 5 Experimental Results

Experiments of recognition is performed using the HMM generated from original image sequences and the synthesized data. In the experiments, images of two persons are used. For each word, three sequences of images are obtained and the total of six samples are obtained.

Excluding words whose images are not successfully processed, 14 words expressed by both hands and 21 words expressed by one hand are used. First, only actual data are used for training, where five of the sequences are used for training and the rest is used for the test. By changing the sequence for the test, six experiments are performed.

Next synthesized data sequences are added for training and similar six experiments are performed. Experimental results are shown in Table 1. Recognition rates of the experiments using only actual data and the experiments using synthesized data have little differences.

Before adding synthesized data, the word “like” of subject A is misrecognized as the word “red”, and the word “red” of subject B is misrecognized as the word “dislike”. By adding synthesized data, they are recognized successfully.

**Table 1.** Experimental results

(a) Subject A

| | both hands | | one hand | |
|---|---|---|---|---|
| synthesized data addition | before | after | before | after |
| the number of success | 52/52 | 52/52 | 57/63 | 57/63 |
| recognition rate | 100% | 100% | 90.5% | 90.5% |

(b) Subject B

| | both hands | | one hand | |
|---|---|---|---|---|
| synthesized data addition | before | after | before | after |
| the number of success | 50/52 | 50/52 | 57/63 | 58/63 |
| recognition rate | 96.2% | 96.2% | 90.5% | 92.1% |

(a) black (b) head

**Fig. 2.** Words of similar hand features

On the other hand, the word "head" shown in Figure 2(a) of subject A is misrecognized as the word "black" shown in Figure 2(b). Because the hand shape view of these words are similar, it is easily misrecognized. Even by adding synthesized data, the difference is ambiguous. We need better features to discriminate them.

One of the reason of misrecognition in either case is that the test data is far from training data. It is necessary to collect hand shape data for more words in order to synthesize a variety of training data.

## 6 Conclusion

We proposed a method of automatically synthesizing of training data using a hand shape database for sign language recognition. Hand shapes in sign language words are classified and training data are synthesized by replacing hand shape features.

Although experiments with a small number of sign language words, the effect of using synthesized data is not prominent, the method seems promising if it is applied to larger data set.

The future problem is in addition to extension to a larger data set, application of the method for various persons.

## References

1. J. Kinscher, H. Trebbe, "The Munster Taging Project - Mathematical Background", *Arbeitabereich Linguistik, University of Munster*, D-58149 Munster, 1995.
2. T. Toda, H. Kawai et al, "Unit Selection Algorithm For Japanese Speech Synthesis Based on Both Phoneme Unit and Diphone Unit", *Proc. Int. Conf. Acoustics, Speech and Signal Processing (ICASSP2002)*, pp. 3088-3091, 2002.
3. H. Sagawa, M. Ohki et al, "Pattern Recognition and Synthesis for a Sign Language Translation System", *Jounal of Visual Languages and Computing* Vol. 7, No. 1, pp. 109-127, 1996.
4. Y. Fukuda, H. Kimura et al. "Expression through hand and finger actions used in 'words in continuous signing': A study in reference to the dialogue in Japanese Sign Language by the deaf persons", *Technical Report of IEICE*, Vol. 97, No. 586, pp. 81–85, 1998 (in Japanese).