Subunit Modeling for Japanese Sign Language Recognition Based on Phonetically Depend Multi-stream Hidden Markov Models

doi: 10.1007/978-3-642-39188-0_59(http://dx.doi.org/10.1007/978-3-642-39188-0_59)

Subunit modeling for Japanese sign language recognition based on phonetically depend multi-stream hidden Markov models

Shinji Sako and Tadashi Kitamura

Nagoya Institute of Technology, Gokiso-cho, Showa-ku, Nagoya, 466-8555, JAPAN
{s.sako, kitamura}@nitech.ac.jp

Abstract. We work on automatic Japanese sign Language (JSL) recognition using Hidden Markov Model (HMM). An important issue for modeling sign is that how to determine the constituent element of sign (i.e., subunit) like "phoneme" in spoken language. We focused on special feature of sign language that JSL is composed of three types of phonological elements which is hand local information, position, and movement. In this paper, we propose an efficiently method of generating subunit using multi-stream HMM which is correspond to phonological elements. An isolated word recognition experiment has confirmed the effectiveness of our proposed method.

Keywords: Hidden Markov models, Sign language recognition, Subunit, Phonetic systems of sign language

1 Introduction

Sign language is the visual language of deaf people. It is also natural language, different in form from spoken language. To resolve a communication problem between hearing people and deaf, projects for automatic sign language recognition (ASLR) system is now under way[1, 6]. Sign is represented by various combinations of posture or movement of the hands, eyes, mouth, and so on. These representations of sign are happen both sequentially and simultaneously. There remains a need for an efficient method that can modeling such simultaneous events.

In this paper, an important issue to modeling sign is that how to determine the constituent element of sign (i.e., subunit) as similar as phoneme in spoken language. We addressed the method to generate optimal subunit automatically. In order to achieve it, we adopt the technique for statistical modeling based on hidden Markov models (HMMs). In addition, sign language has own system and it is therefore desirable to take account its language structure as well as possible. We also focused on some linguistic feature of sign to classify subunit more effectively by using phonetically depend HMMs. An isolated sign recognition experiment using RWC video database has confirmed the effectiveness of our proposed method.

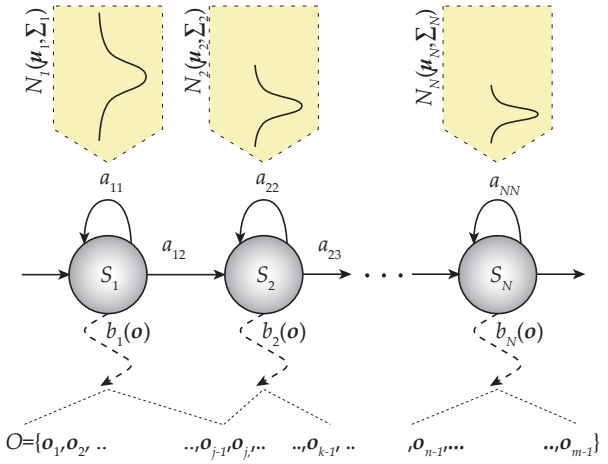

Fig. 1. An example of left-to-right hidden Markov model

2 Methodology

2.1 Hidden Markov Models

Goodness has been proven with the ability of HMM modeling time-varying signals such as speech. Due to similarity between speech and sign language as time-series pattern, HMM is often employed in automatic sign language recognition [9, 10, 12]. HMM is a model composed of N states $\mathrm{S} = (\mathrm{S}_1, \mathrm{S}_2, \ldots, \mathrm{S}_N)$ connected with each other based on the state-transition probability $a_{ij} = P(q_t = i | q_{t-1} = j)$ from state S_j to S_i. Each state outputs observation vector $\boldsymbol{o}$ based on probability distribution $b_i(\boldsymbol{o})$. The parameter of HMM λ is defined as $\lambda = (A, B, \pi)$ using $A = \{a_{ij}\}_{i,j=1}^{N}$, $B = \{b_i(\cdot)\}_{i=1}^{N}$, and initial state-transitional probabilities $\pi = \{\pi_i\}_{i=1}^{N}$. As the output probability distribution, a single Gaussian distribution was employed. Thus the mean vector $\boldsymbol{\mu}_i$ and covariance matrix $\mathbf{U}_i$ of Gaussian distribution are state parameters of HMM.

Model parameters λ, which is trained by the given training sequence $\boldsymbol{O} = (\boldsymbol{o}_1, \boldsymbol{o}_2, \ldots, \boldsymbol{o}_T)$ based on the equation $\lambda_{\max} = \underset{\lambda}{\operatorname{argmax}}\, P(\boldsymbol{O} \mid \lambda)$ is normally calculated by Baum-Welch algorithm. The maximum likelihood state sequence concerning the $\boldsymbol{O}$ is represented as $\boldsymbol{Q} = (\boldsymbol{q}_1, \boldsymbol{q}_2, \ldots, \boldsymbol{q}_T)$. It is given by maximizing $P(\boldsymbol{O}, \boldsymbol{Q} \mid \lambda)$, and is effectively calculated by Viterbi algorithm. Vector sequence $\boldsymbol{O}$ is segmented to be assigned to each state, and $P(\boldsymbol{O}, \boldsymbol{Q} \mid \lambda)$ is the evaluation value about $\boldsymbol{O}$ calculated from λ.

2.2 Generation of subunits composing sign HMMs

Each sign follows a variety of motion, but similar short motions are partly included. In this paper, these common elements referred to above are called subunits, and are regarded as fundamental component for all signs like the phoneme dealt with in a spoken language.

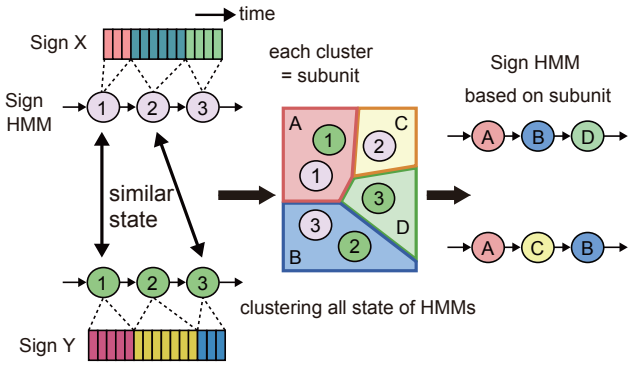

Fig. 2. How to obtain the subunits

Figure 2 shows the process of generating the subunit by using HMMs. Roughly, our method consists of two step in order to obtain subunits. In the first step, several isolated sign HMMs are trained separately as initial model. A sign HMM makes approximation to the sequence of sign motion as a chain of small number of states. Because a particular segment in the sign corresponds to each state, the similar states over different signs can be regarded as a common element of the sign. In the second step, subunits are obtained by applying clustering technique on all states of isolated sign HMMs. At this stage, inter-cluster distance is defined as the maximum within the distance between the two states belonging to two different clusters. And inter-state distance is defined as Eq. 1. i and j are state number, V is the number of dimensions, μ and σ are mean and variance of state parameter.

$$d(i, j) = \sqrt{\frac{1}{V}\sum_{k=1}^{V}\frac{(\mu_{ik} - \mu_{jk})^2}{\sigma_{ik}\sigma_{jk}}} \tag{1}$$

We adopted tree based clustering algorithm in order to tie the states. Finally, the output distributions across different sign HMMs are shared with each other when they has visual pattern similarity.

2.3 Phonetic depend subunit

In order to handle simultaneous events, we adopt some linguistic feature of Japanese Sign Language (JSL). Dr. Kanda said that manual behavior in JSL can be represented by three types of phonological elements, which are hand position, movement and shape[7, 8]. These three elements are happened simultaneously. For this reason, it was difficult to separate subunit correctly if sign feature parameter was represented as a single stream. In order to solve this problem, a multi-stream HMM are applied to build subunit. The feature vector of each frame is splited into three phonetic stream, and a multi-stream isolated sign HMM are trained for each word. Phonetic depend subunits are obtained by clustering each state of multi-stream HMMs as same manner as described in section

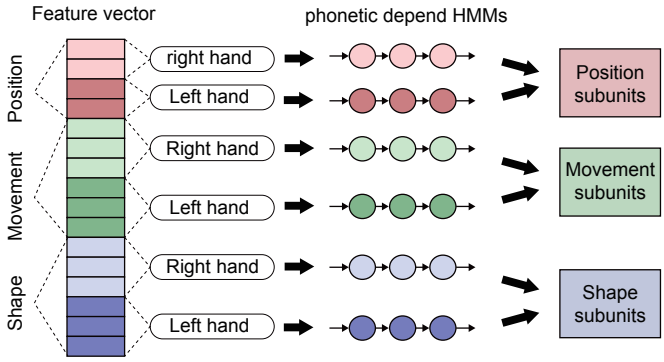

Fig. 3. Phonetic depend subunit modeling

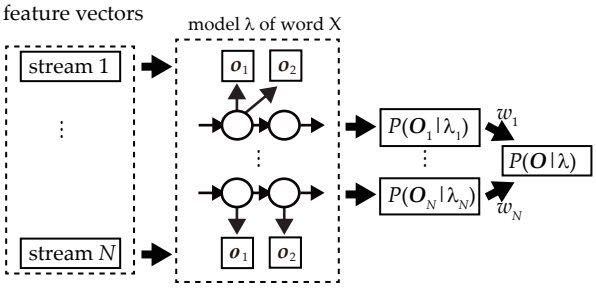

Fig. 4. Recognition method by integrating likelihoods calculated from each stream

2.2. Because clustering are applied for each phonetic element, partially similar state will be merged and generate subunit effectively.

Recognition process using multi-stream HMM is performed as shown in Figure 4. The integrated likelihood is calculated as in Equation 2. It is noted that $P(O_s|\lambda_s)$ means the likelihood in the stream s against feature vector O_s. And w_s represents the weight parameters between the streams.

$$P(\boldsymbol{O} \mid \lambda) = \prod_{s=1}^{S} P(\boldsymbol{O}_s \mid \lambda_s)^{w_s}, \sum_{s}^{S} w_s = 1 \tag{2}$$

3 Experiment

As an evaluation experiment, isolated sign recognition is conducted. We used three kind of modeling method:

– **Conventional method**: Typically left-to-right HMM for each sign which are trained by several utterances. We also used single-stream HMM and multi-stream HMM.

Table 1. Details of RWC multi-modal database (1998 version)

Number of signer	4 (2 males and 2 females)
Number of sign	308
Number of data	8 times per sign (4 × 2 takes)
image size	320 × 240 pixel
frame rate	29.7 frame / sec (NTSC)
Number of frame per sign	56 ~ 144 frame

Table 2. Details of experiment conditions

	Conventional	Proposed 1	Proposed 2
Number of sign	100（3 set）		
Train Test	train: 600 sample（3 person × 2 takes × 100 sign) test: 200 sample（1 person × 2 takes × 100 sign) leave-one-out cross validation (4 set）		
Feature parameters	location: hand position(4 dim.) movement: Δ, Δ^2 (4 + 4 dim.) hand shape: PCA coefficient of hand region (first 16 components) $+\Delta + \Delta^2$ (16 × 3 dim.)		
Stream weight	$position : movement : shape = 0.25 : 0.60 : 0.15$ $left hand : right hand = 1 : 1$		
Number of state	5 ~ 35	25	25 per stream
Number of subunit	–	100 ~ 1,200	100 ~ 1,200 per stream

- **Proposed method 1**: Each sign is represented as subunits concatenation
- **Proposed method 2**: Each sign is represented as phonetic depend subunits concatenation

3.1 Database and experimental settings

The database used in the experiment is the RWC video database of the isolated signs of JSL[13]. Table 1 shows the details of the database. The database contains the data of 4 signers, each signer performs the same set of 308 signs, and recording is individually done twice with respect to each signer. In this experiment, data for 3 signers (6 utterances for every sign) are used for training HMMs, whereas data for a remaining single signer are used for recognition by taking up all the combinations available within this constraint. The number of the subunits generated by the clustering is changed from 100 to 1,200 with 100 increments in between. Other experiment details are shown in Table 2.

In order to obtain hand position and movement, the hand tracking algorithm finds the centroid of a moving hand region. Hand shape feature is also obtained by the well-known Eigenface method as applied to detected hand region images[16]. We used 60

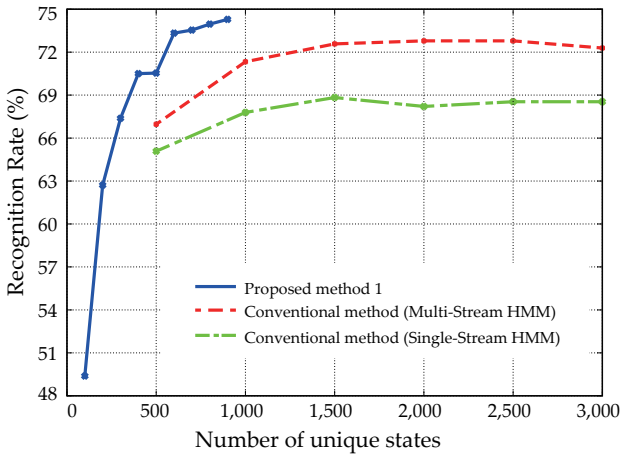

Fig. 5. An experiment result of conventional method and proposed method 1

dimensional feature vector sequence as sign representation. Δ is calculated by Eq. 3. Where $c(t)$ is a coordinate of hands at time t, θ is fixed to be 1 in this experiment. The same formula is applied to Δ obtain Δ^2. These parameters called dynamic feature are generally used in speech recognition.

$$\Delta c(t) = \frac{\sum_{\theta=1}^{D_w} \theta(c(t+\theta) - c(t-\theta))}{2\sum_{\theta=1}^{D_w} \theta^2} \tag{3}$$

3.2 Result

Figure 5 shows the result of the word accuracy for conventional and proposed method 1. The horizontal axis represents the model complexity. It means the granularity of classification for state of HMMs. If this value is small, small number of subunit and more generalized subunit are obtained. The experimental result shows that subunit model can get a better recognition performance than conventional method as phonetic independ model. In addition, higher recognition performance is also achieved by small number of subunits.

Figure 6 also shows the result of the word accuracy for proposed method 1 and 2. It is noted that the result of method 1 in both figure 5 and 6 is identicle. As the result, our proposed method 2 is superiorto the method 1.

We also conducted a survey with varying number of subunits for each stream. The number of the subunits generated by the clustering is changed from 100 to 900 for each stream (the number of possible combination is 729). Figure 7 shows the result of recognition acuracy in case of hand shape subunit number at 200, 400, 600, and 800. This result shows the importance of hand shape information. Therefore, It is necessary to optimize the number of sub-units in each stream.

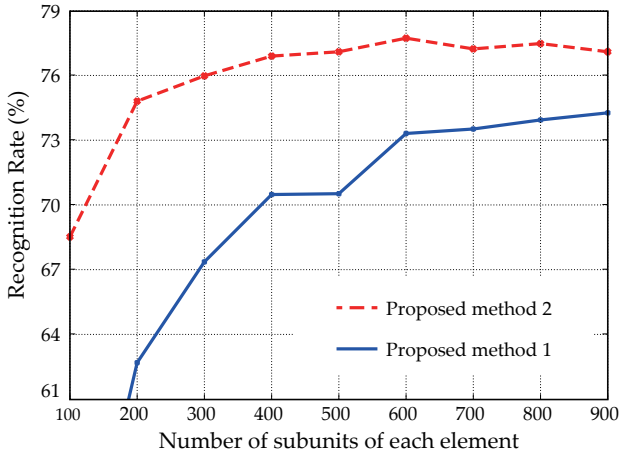

Fig. 6. Sign recognition accuracy of proposed method 1 and 2

4 Conclusion

In this paper, we presented a subunit modeling in order to take into account simultaneous events of JSL. The key point of our method is automatic subunit generation technique using phonetic depend multi-stream HMM. Experimental results on RWC JSL video database showed effectiveness of our proposed method and improvements in isolated sign recognition accuracy. In addition, higher recognition performance is also achieved by small number of model parameters.

Our feature work is stream weight optimization in order to adjust a balance of each sign. And, not only hand movement but hand shape, facial expression etc. are important for sign language essentially, thus more appropriate feature of sign should be used.

5 Acknowledgment

This research is financially supported in part by Grants-in-Aids for Scientific Research (22500506).

References

1. Y. Okazawa, M. Nishida, Y. Horiuchi, and K. Ichikawa, “Sign Language Recognition Using Gesture Component Involved Transition Part, ” Technical Report of IEICE (WIT), vol.103, no.747, pp.13–18, 2004. (*in Japanese*)
2. H. Sawada, S. Hashimoto, and T. Matsushima, “A Study of Gesture Recognition Based on Motion and Hand Figure Primitives and Its Application to Sign Language Recognition, ” IPSJ Journal, vol.39, no.5, pp.1325–1333, 1998. (*in Japanese*)
3. K. Kanayama, Y. Shirai, and N. Shimada, “Recognition of Sign Language using HMM, ” Technical Report of IEICE (WIT), vol.104, no.93, pp.21–28, 2004. (*in Japanese*)

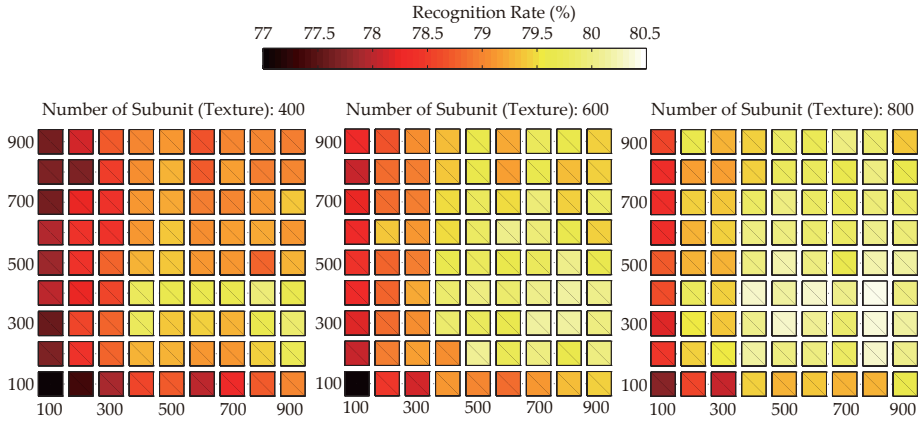

Fig. 7. Research of the number of subunits for each phonological category

4. K. Imagawa, S. Lu, H. Matsuo, and S. Igi, "Real-Time Tracking of Human Hands from a Sign-Language Image Sequence in Consideration of Disappearance by Skin Regions, " IEICE Journal (D), vol.J81-D-2, no.8, pp.1787–1795, 1998. (*in Japanese*)
5. U. vonAgris, M. Knorr, and K.-F. Kraiss, "The significance of facial features for automatic sign language recognition," Proc. 8th IEEE Int. Conf. on Automatic Face and Gesture Recognition, pp.1–6, 2008.
6. M. Maebatake, M. Nishida, Y. Horiuchi, and S. Kuroiwa, "Sign Language Recognition Based on Position and Movement Using Hidden Markov Model, " Technical Report of IEICE (PRMU), vol.108, no.94, pp.7–12, 2008. (*in Japanese*)
7. K. Kanda, Study on the characteristics of the sign language — Architecture of an electronic sign language dictionary, Fukumura Shuppan Inc., 2010. (*in Japanese*)
8. K. Kanda and H. Naka, "Phonological Notational System for Japanese Sign Language, " Journal of JASL, vol.12, pp.31–39, 1991. (*in Japanese*)
9. Y. Toyokura, Y. Nankaku, T. Goto, and T. Kitamura, "A-4-5 Approach to Japanese Sign Language Word Recognition using Basic Motion HMM, " Annual conference of IEICE, p.72, 2006. (*in Japanese*)
10. B. Bauer and K.F. Kraiss, "Video-Based Sign Recognition Using Self-Organizing Subunits," Proc. 16th Int. Conf. on Pattern Recognition, vol.2, pp.434–437, 2002.
11. C. Vogler and D. Metaxas, "Handshapes and Movements: Multiple-Channel American Sign Language Recognition," Lecture Notes in Computer Science, vol.2915, pp.247–258, 2004.
12. M. Nishida, M. Maebatake, I. Suzuki, Y. Horiuchi, and S. Kuroiwa, "Sign Language Recognition Based on Position and Movement Using Multi-Stream HMM, " Journal of IEEJ, vol.129, no.10, pp.1902–1907, 2009. (*in Japanese*)
13. H. Yabe, R. Oka, S. Hayamizu, T. Yoshimura, S. Sakurai, S. Nobe, T. Mukai, and H. Yamashita, "RWC Database –Gesture Database–, " Technical Report of IEICE (PRMU), vol.100, no.181, pp.45–50, 2000. (*in Japanese*)
14. K. Ariga, S. Sako, and T. Kitamura, "Sign Language Recognition Considering Signer and Motion Diversity Using HMM, " Technical Report of IEICE (WIT), vol.110, no.53, pp.55–60, 2010. (*in Japanese*)

15. Y. Hamada, N. Shimada, and Y. Shirai, “Shape Estimation of Quickly Moving Hand under Complex Backgrounds for Gesture Recognition, ” IEICE Journal (D), vol.J90-D, no.3, pp.617–627, 2007. (*in Japanese*)
16. M. Turk and A. Pentland, “Eigenfaces for recognition,” Journal of Cognitive Neuroscience, vol.3, no.1, pp.71–86, 1991.
17. H. Yonehara, Y. Nagashima, and M. Terauchi, “A Measurement of Fixation Point Distribution of Native Signer, ” Technical Report of IEICE (TL), vol.102, no.254, pp.91–95, 2002. (*in Japanese*)
18. “Hidden Markov Model Toolkit (HTK) version 3.4.1”. http://htk.eng.cam.ac.uk/