Construction of General HMMs from a Few Hand Motions for Sign Language Word Recognition

Tadashi Matsuo[1] Yoshiaki Shirai[2] Nobutaka Shimada[3]
Ritsumeikan University
1-1-1 Noji-higashi, Kusatsu, Shiga, JAPAN
matsuo@i.ci.ritsumei.ac.jp[1] shirai@ci.ritsumei.ac.jp[2] shimada@ci.ritsumei.ac.jp[3]

Abstract

We propose a method to construct a Hidden Markov Model (HMM) for sign language recognition with a topology which is suitable for a variety of hand motions. First, candidate HMMs are generated from sub-motions extracted from training samples. If we have many and various samples of motions, an optimal HMM can be selected from candidates by the maximum likelihood (ML) method. However, it is difficult to collect many real samples and the ML method with a small number of samples may select a HMM too much specialized for the training samples. The proposed method selects the best HMM for each word by evaluating the performance using real samples and virtual samples generated by an HMM made from real training samples. Virtual samples are generated from a HMM estimated from given real samples. On evaluation of HMMs, they bring a HMM that accepts not only given real samples but also their variations sufficiently similar to the real samples. With experiments, we show the effectiveness of the proposed method.

1 Introduction

Sign language recognition consists of extraction of features such as hand shapes, positions and velocity, and recognition of extracted feature sequences. For the latter, Hidden Markov Model (HMM) has widely been used [1, 2]. The HMM consists of states corresponding to sub-motions of hands, and transitions between states.

Starner et al. proposed a method using HMMs with a common fixed topology [1, 3]. However, it is difficult to appropriately learn state parameters if a sign language word can be expressed by partially different motions because multiple motions correspond to a single state as shown in Figure 1. Starner [1, 3] and Gaolin [4, 5] proposed methods where a word corresponds to multiple HMMs in order to accept various motions. It is, however, difficult to recognize various motions by a fixed topology for all words. Because each word may have a variety of hand motions, the topology of the HMM should reflect the variety.

Kawahigashi et al. proposed a method generating linear topology HMMs where number of states is automatically determined for each word [6]. However, if a word can be expressed by partially different motions, a sub-motion may not correspond to a state in an HMM generated from another type of motions due to the restriction of topology as the 3rd interval of Motion B in Figure 1. It may be difficult to train such models for a word with various motions.

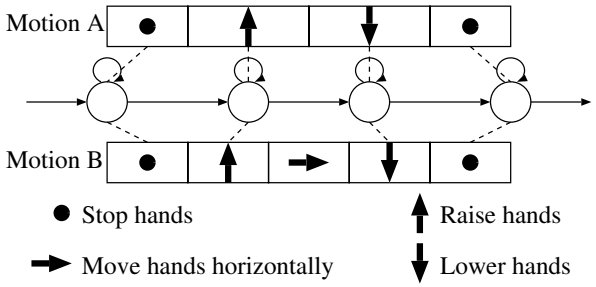


Figure 1. Correspondence between states and sub-motions.

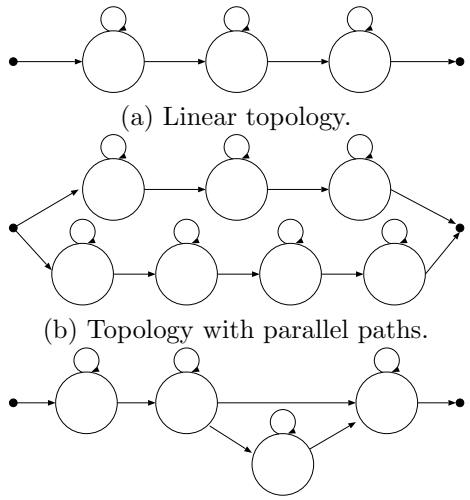
Matsuo et al. proposed a method generating a HMM with junctions and branches [7]. But it generates only one topology for each word and does not consider the possibility that simpler topology may have the same or better performance.

Generally, the HMM should have a topology which brings high likelihood to training samples. However, if the HMM over-fits to the training samples, it may bring a low likelihood to test samples. If we have sufficiently many and various samples, the maximum likelihood method(ML method) can give us a suitable model to describe the motion variety. However, collecting many sign language samples is not easy.

We have a prior knowledge that two motions can be considered as the same word if their difference is less than an admissible difference estimated from the resolution of the screen. Niyogi et al. proved that utilizing "virtual examples" on training a model is equivalent to incorporating a prior knowledge in some context [8]. We introduce the prior knowledge of motions to evaluate models. We propose a method that generates candidates of HMMs from training samples of a word and selects an appropriate HMM by evaluating them with not only real samples but also virtual samples. Evaluation of HMMs with them brings a HMM that accepts not only given real samples but also their variations similar to the real samples. This is effective especially in the case where only a few real samples are available. The best topology determined by the proposed method reflects both the real training samples and samples with a little different motions. The proposed method improves construction of HMMs in the case where only a few real samples can be used. We show the effectiveness of the proposed method by experiments with real sign language word motions.

2 Generation of candidate HMMs

In order to obtain the best HMM of each word, we generate candidate HMMs. First we define two HMM



(a) Linear topology.

(b) Topology with parallel paths.

(c) Topology with branches and junctions.

Figure 2. Examples of topology of a composite model.

models:

- Basic model: the HMM trained for a single type of motions, which is generated directly from training samples.
- Composite model: the HMM trained for multiple types of motions, which is generated from basic models or other composite models.

Then we consider the following 3 types of composite models:

- HMM with a linear topology as shown in Figure 2(a)
- HMM with parallel independent paths for each motion type as shown in Figure 2(b)
- HMM with junctions and branches as shown in Figure 2(c).

2.1 Generation of basic models

A sequence of frames of a real sample is divided into intervals corresponding to simple motions such as "raising a hand" , and the intervals are classified into stationary, linearly moving, or transition interval by the method proposed in [6]. If samples of a word have the same sequences of the classes, they are considered to be the same type of motions and classified into a same group. For each group, a basic model is generated which has a linear topology with the states correspond to the intervals. The state parameters of the HMM are estimated from features in a corresponding interval.

2.2 Generation of composite models

A composite model is at first generated from basic models. Then a further structured composite model is generated by combining the composite model and another basic model. Given two HMMs A and B, we consider 4 composite models with the following topologies.

(a) The topology of HMM A.

(b) The topology of HMM B.

(c) The topology including HMM A and B in parallel in order to accept both samples.

(d) The topology generated by integrating common states in the topology of (c) by the method in [7] in order to avoid too much specialization.

First, two basic models are combined into 4 types of composite models. Then, each composite model is combined with a new basic model to generate 4 new composite models. By repeating this process, we generate a set of composite models that depend on all basic models.

The above process still requires many combinations; if we have M basic models, the total number of candidate HMMs is $(M!/2)4^{M-1}$. Training all of the HMMs requires high computation cost. Therefore, we further restrict the order of the combination. Concretely, we combine the two basic models with the most samples into 4 composite models, and then combine each of the composite models with the basic model with the next most samples. The combination of a basic model and a composite model is repeated in the descending order of the number of corresponding samples. This is because basic models corresponding to the more samples are the more important. By this restriction, the number of candidate HMMs becomes 4^{M-1}.

3 HMM selection with virtual samples

3.1 Trade-off between specialization and generalization

The HMM of a word should give high likelihood only for the word while it should give high likelihood for a variety of motions corresponding to the word. There is a trade-off between specialization and generalization.

If only a single type of motions is valid for a word, the word can be learned by an HMM with linear topology like Figure 2(a). If multiple types of motions are valid for a word and they are learned by an HMM with linear topology, it may give high likelihood for the other words because a state trained with different motions may have a large variance for a feature. Such a word can be learned by an HMM including parallel paths like Figure 2(b). However, the trained HMM may over-fit to the training samples, and give low likelihood for other similar samples. Since appropriate topology differs for each word, the trade-off between generalization and specialization should be solved for each word.

The simple criterion of likelihood does not take account of generalization. Therefore, it tends to select an HMM specialized for a small number of samples.

To solve the trade-off, Minimum Description Length (MDL) [9] and Akaike's Information Criterion(AIC)[10] have been proposed. However, the above methods should be applied with many samples. because the degree of freedom of an HMM is as high as

$$O((\text{number of states}) \times (\text{dimension of feature vector})^2). \quad (1)$$

For example, if a very simple motion consists of 3 sub-motions and the feature vector consists of only two-dimensional position and velocity for each hand, the number of parameters exceeds 150.

Difficulty in collecting many training samples is an important problem on evaluating candidate HMMs. So, we propose a method to avoid too much specialization by introducing virtual samples probabilistically generated from real samples.

3.2 Generation of virtual samples

Virtual samples of a word should be valid for the word, and on the other hand they should be different from real samples. We first build a HMM for virtual samples and generate virtual samples from the HMM.

A word has multiple types of valid motions. Each type can be learned by an HMM with linear topology. We make the HMM for virtual samples by parallely connecting the basic models generated in Section 2.1. But, if the samples of a certain motion type are not so many, the learned variances may be smaller than difference clearly admissible as similar motions on the screen. So, we artificially increase the variances of feature vectors in order to generate various virtual samples.

The actual procedure is as follows:

1. Generate an HMM by connecting basic models with their head and tail. The transition probabilities are assigned a number proportional to the number of corresponding real samples.
2. Train the HMM by the Baum-Welch method with real samples.
3. Artificially increase autovariances. The increment for each dimension is defined as a squared difference of the value of the dimension when hand or face move b[pixel] on the screen.

3.3 Evaluation of candidate HMMs with virtual samples

We select a final HMM from the candidates generated in Section 2 by the ML method using both real and virtual samples. The likelihood function $L(m)$ for a candidate HMM m is defined as follows.

$$L(m) = \sum_{n=1}^{N} \log_2 P\left(O_n \,|m\right) + \sum_{k=1}^{K} \log_2 P\left(\tilde{O}_k \,|m\right),$$

where O_n and $\tilde{O}_k$ mean a real sample and a virtual one, respectively. The number of virtual samples, K is experimentally determined. Note that the virtual samples are used only for evaluating the candidate HMMs as shown in Figure 3.

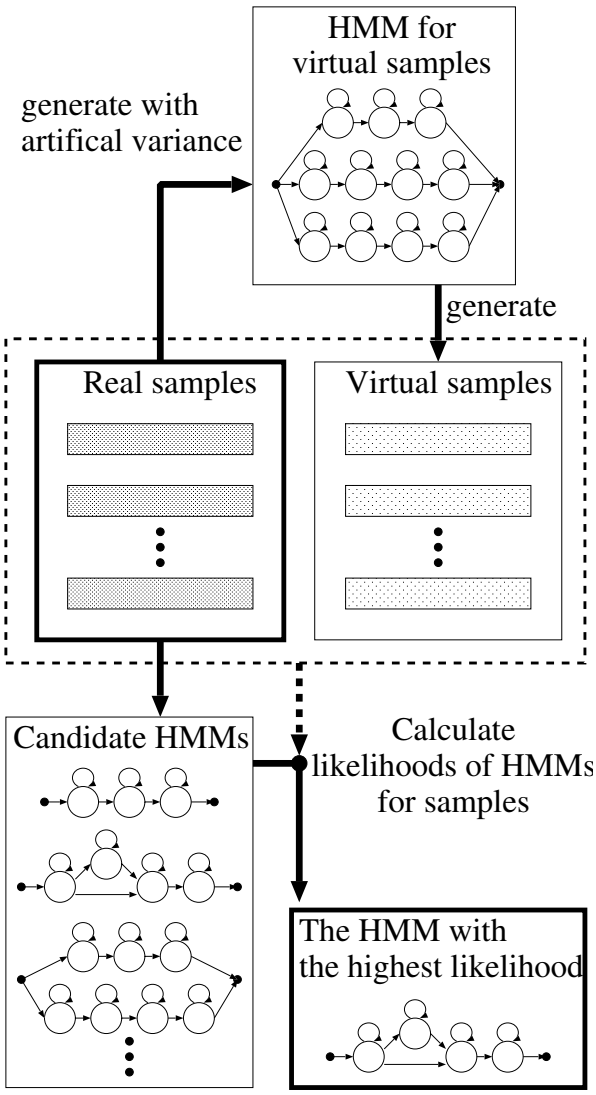


Figure 3. Process for generating an HMM for a word.

4 Experiment

4.1 Recognition of motions by unknown speakers

We perform an experiment recognizing real samples from motions by an unknown speaker. To see the effectiveness in the case where we have only a few real samples, each HMM is generated by a few real samples. We ask 3 untrained speakers to perform motions 3 times for each word and generate a HMM by using the 6 real samples from the two speakers. Then, 3 motions by the other speaker are recognized. The result for the 20 words shown in Table 1 is shown in Table 2.

The result shows that the proposed method generates better HMMs than the fixed topology method and the simple ML method. The proposed method can select more general HMMs.

The result also shows that more virtual samples do not always give better recognition ratios. This means that sufficiently many virtual samples sufficiently represent variation of samples and models selected by them are almost fixed. In our experiments, virtual samples as many as real ones are sufficient.

4.2 Recognition of motions by known speakers

To see the performance of the proposed method, we performed a recognition experiment with leave one out method, where 3 speakers perform each word motion 3 times, and then an HMM by using 8 samples recognizes the rest sample. The result for 20 words in Table

Table 2. Correct recognition ratio for a speaker not used when training HMMs.

Rank	Linear only	ML	proposed method (number of virtual samples)		
			1	6	12
1st	0.394 ± 0.072	0.444 ± 0.080	0.476 ± 0.075	0.490 ± 0.074	0.487 ± 0.073
2nd or above	0.661 ± 0.102	0.706 ± 0.083	0.719 ± 0.081	0.741 ± 0.088	0.739 ± 0.086
3rd or above	0.800 ± 0.067	0.844 ± 0.055	0.872 ± 0.063	0.864 ± 0.072	0.861 ± 0.068

Each column means (MEAN) ± (STANDARD DEVIATION).

Table 3. Correct recognition ratio of leave one out method.

Rank	Linear only	ML	proposed method (number of virtual samples)		
			1	8	16
1st	0.706 ± 0.107	0.733 ± 0.091	0.743 ± 0.067	0.752 ± 0.063	0.749 ± 0.062
2nd or above	0.872 ± 0.058	0.922 ± 0.048	0.920 ± 0.058	0.923 ± 0.044	0.926 ± 0.042
3rd or above	0.950 ± 0.033	0.972 ± 0.025	0.963 ± 0.034	0.967 ± 0.033	0.966 ± 0.033

Each column means (MEAN) ± (STANDARD DEVIATION).

Table 1. Words used in experiments.

Index	Meaning	Index	Meaning
1	small	11	which
2	big	12	after a long time
3	long	13	talk to
4	short	14	cooking
5	"Thanks"	15	highest
6	happy	16	lowest
7	an elder sister	17	care
8	an elder brother	18	examination
9	brothers	19	stop
10	come together	20	rest

1 is shown in Table 3. "Linear only" means the experiment where linear HMMs generated by [6] are used. "ML" means the experiment where candidate HMMs are generated as Section 2 and an HMM is selected by the ML method for each word. The ratios for "Linear only" and "ML" in Table 3 are averages for 9 ways of selecting the sample to recognize. In the proposed method, virtual samples are probabilistically generated from HMMs. The recognition ratios for the proposed method in the table are averages for 8 different set of virtual samples and 9 ways of selecting the sample to recognize.

The result shows that the proposed method achieves better recognition ratios. The improvement by the proposed method in Table 2 is higher than that in Table 3. This shows that the proposed method is more effective in the case where the given real samples are few.

The recognition ratios reach the ceiling similar to those in Table 2. In our experiments, virtual samples as many as real ones are sufficient.

5 Conclusion

We proposed a method to generate a HMM with the best topology reflecting variety of motions for a sign language word. It is effective especially in the case where only a few real samples are available. To avoid over-fitting, artificially generated virtual samples are used for selecting the best topology. Experiments show that the proposed method generates better HMMs than those with linear topology or those generated conventional ML criteria.

References

[1] Starner, T., Pentland, A.: Real-time american sign language recognition from video using hidden markov models. Computer Vision, International Symposium on **0** (1995) 265

[2] Grobel, K., Assan, M.: Isolated sign language recognition using hidden markov models. Systems, Man, and Cybernetics, 1997. 'Computational Cybernetics and Simulation'., 1997 IEEE International Conference on **1** (12-15 Oct 1997) 162–167 vol.1

[3] Starner, T., Weaver, J., Pentland, A.: Real-time american sign language recognition using desk and wearable computer based video. IEEE Transactions on Pattern Analysis and Machine Intelligence **20** (1998) 1371–1375

[4] Gao, W., Ma, J., Wu, J., Wang, C.: Sign language recognition based on HMM/ANN/DP. International Journal of Pattern Recognition and Artificial Intelligence **14** (2000) 587–602

[5] Fang, G., Gao, W., Zhao, D.: Large vocabulary sign language recognition based on fuzzy decision trees. Systems, Man and Cybernetics, Part A, IEEE Transactions on **34** (May 2004) 305–314

[6] Kawahigashi, K., Shirai, Y., Shimada, N., Miura, J.: Segmentation of sign language for making HMM. IEICE technical report **105** (2005) 55–60 (in Japanese).

[7] Matsuo, T., Shirai, Y., Shimada, N.: Automatic generation of hmm topology for sign language recognition. In: Pattern Recognition, 2008. ICPR 2008. 19th International Conference on. (2008) 1 –4

[8] Niyogi, P., Girosi, F., Poggio, T.: Incorporating prior information in machine learning by creating virtual examples. Proceedings of the IEEE **86** (1998) 2196 –2209

[9] Rissanen, J.: Estimation of structure by minimum description length. Circuits, Systems, and Signal Processing **1** (1982) 395–406

[10] Akaike, H. In: Information theory and an extension of the maximum likelihood principle. Springer-Verlag (1973) 267–281