Automatic Generation of HMM Topology for Sign Language Recognition

Tadashi Matsuo, Yoshiaki Shirai, Nobutaka Shimada
Ritsumeikan University
matsuo@i.ci.ritsumei.ac.jp, shirai@ci.ritsumei.ac.jp, shimada@ci.ritsumei.ac.jp

Abstract

Sign language is used for communicating to people with hearing difficulties. Recogntion of a sign language image sequence is challenging because of the variety of hand shapes and hand motions.

We propose a method to automatically construct a transitional structure(topology) of a Hidden Markov Model(HMM) for recognizing sign language words. Unlike conventional HMM, the constructed topology has branches and junctions in order to represent a flexible structure. The proposed method consists of segmentation of a motion, and construction of the topology from segments. The topology is constructed from an initial topology by modifying it. With experiments, we show the effectiveness of the proposed method.

1 Introduction

Sign language recognition from a image sequence requires feature extraction and feature interpretation. Generally features consist of the position, velocity, and shape of hands. For interpretation of features, a framework of Hidden Markov Model (HMM) has been used, where each model corresponds to a word[3, 5, 6]. The model consists of states and transitional structure(topology) among the states. A state corresponds to consistent partial features such as raising hands, spreading hands etc., and it has parameters representing the features. The topology determines the transition possibilitiy between states.

Each state in HMM corresponds to a segment of features in a image sequence. An important problem is that features for the same word may be different depending on situations or signers.

In [5], the number of states were generally fixed for all words.

In [4], the number of states was estimated for each word. However, the thresholds of hand speed for the estimation were manually adjusted for each speaker.

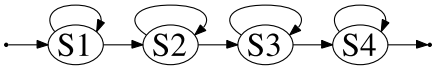


Figure 1. A linear topology

Topology was limited to linear such as Figure 1.

In [2, 1], the variation is resolved by introducing new recognition layers in addition to HMM. However, the topology of HMM is fixed for every word.

One method to overcome the problem is to generate multiple models. However, this may require many samples for learing HMMs. Our method is to learn a HMM for a word so that it may have branches and junctions to represent a flexible structure. In order to generate the HMM automatically, the sequence of images is segmented into states, and by comparing the initial model and segments, states or transitions are added to the model if required.

2 Segmentation of a training sample

We extract features from images by a method similar to [4]. To segment a sequence of frames, we use the direction and velocity of motions because the other features such as the position or shape of hands are different for each speaker. First, each frame is classified as stationary or moving by the hand velocity in the frame. Then, a series of stationary frames is grouped as a stationary segment. A series of moving frames with almost straight motion is grouped as a straight segment. A series of moving frames, where the direction of the hand motion changes in a short time, is grouped as a vibrating segment.

3 Representation of features for HMM

In sign language, important actions are generally performed near the face. Therefore, we use the coordinate system centered at the face and the logarithmically transformed coordinate to represent the position

978-1-4244-2175-6/08/$25.00 ©2008 IEEE

of hands. The transformation is defined as

$$L\left(\begin{bmatrix} x \\ y \end{bmatrix}, r_0\right) = \frac{\log\left(1 + \frac{\sqrt{x^2+y^2}}{r_0}\right)}{\sqrt{x^2+y^2}} \begin{bmatrix} x \\ y \end{bmatrix}, \quad (1)$$

where r_0 is a constant(here, the initial distance from the face). It is assumed that speakers initially move hands from their waist. We represent the hand positon $\boldsymbol{x}$ by $\boldsymbol{y}$ defined as:

$$\boldsymbol{y} = L\left(\boldsymbol{x}, r_0\right). \quad (2)$$

The velocity itself is not effective to recognize the motions because the velocity of motions highly depends on speakers. However, if the hand is stationary, the fact is very important. Therefore, we use the logarithmically transformed velocity vector $\tilde{\boldsymbol{v}}$ of $\boldsymbol{v}$ to roughly distinguish between moving and stationary. The direction of $\tilde{\boldsymbol{v}}$ is the same as $\boldsymbol{v}$ and the length of $\tilde{\boldsymbol{v}}$ is proportional to $\log \|\overline{\boldsymbol{x}}\|$, where $\overline{\boldsymbol{x}}$ is the average of $\boldsymbol{x}$.

Similarly, the relative position of the right hand from the left hand should be distinguished in detail when they are close. Therefore, we represent the relative position by $\boldsymbol{y}_{\mathrm{rel}}$ defined as:

$$\boldsymbol{y}_{\mathrm{rel}} = L\left(\boldsymbol{x}_{\mathrm{right}} - \boldsymbol{x}_{\mathrm{left}}, r_1\right), \quad (3)$$

where r_1 is a constant(here, the initial distance between both hands), $\boldsymbol{x}_{\mathrm{left}}$ is the position of the left hand, and $\boldsymbol{x}_{\mathrm{right}}$ is that of the right hand.

4 Construction of topology

Here, we start from an initial topology generated from a sample of a word and then integrate the other samples of the same word one by one.

4.1 Initial topology

We select the shortest sequence of segments as the initial topology, where the states correspond to the segments. The number of states is equal to the number of segments. The topology is linear as shown in Figure 1.

4.2 Integration of a series of segments into topology

The integration of a new sample into the current topology is divided into the two stages:

1. Determine the correspondence between the segments in the sample and the states in the topology.
2. If necessary, add new states or transitions into the topology so that each segment has a corresponding state.

Table 1. The similarity between a state and a segment

	segment					
	1	2	3	4	5	6
S_1	+0.4	-1.7	-6.8	-1.4	-1.4	-0.6
S_2	-2.9	-13.9	-117.7	-3.8	-0.9	0.1
S_3	-7.8	-5.1	0.5	-20.1	-14.5	-11.0
S_4	-1.4	-2.2	-2.5	0.4	0.7	-1.0

In the stage 1, the matching is based on the similarity between a segment and a state. The "best" correspondence is determined as the one which maximizes the total sum S of the similarity C;

$$S = \sum_k C\left(S_{i(k)}, \mathrm{seg}_{j(k)}\right), \quad (4)$$

where S_i is the i-th state, seg_j is the j-th segment, and $S_{i(k)}$ and $\mathrm{seg}_{j(k)}$ are the matched pair. The best correspondence is found by DP matching with skips. We take the following similarity $C(S_i, \mathrm{seg}_j)$.

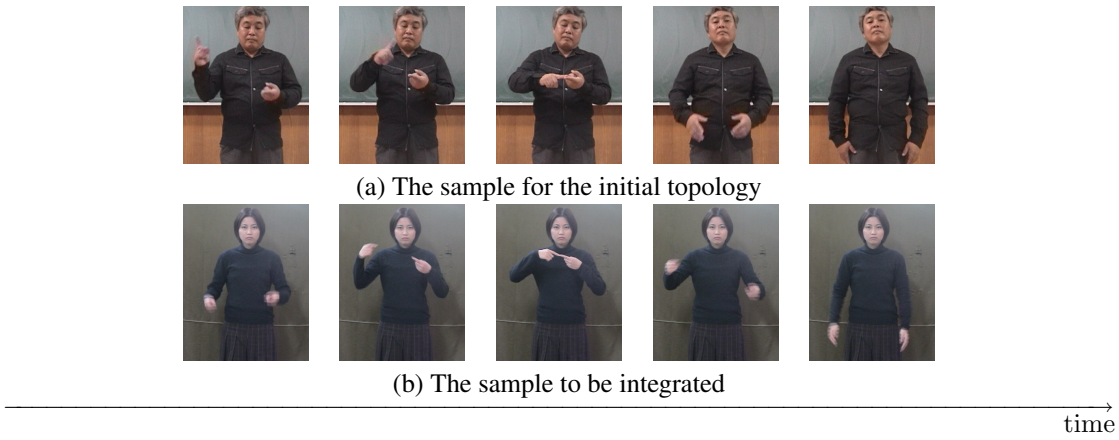
$$\begin{aligned} C(S_i, \mathrm{seg}_j) &= 1 - \frac{1}{T_j} \sum_{t \text{ in } \mathrm{seg}_j} \left\{\frac{d_i\left(\boldsymbol{f}_t\right)}{\sigma}\right\}^2, \\ d_i\left(\boldsymbol{f}_t\right) &= \sqrt{\left(\boldsymbol{f}_t - \boldsymbol{\mu}_i\right)^{\mathrm{T}} \Sigma_i^{-1} \left(\boldsymbol{f}_t - \boldsymbol{\mu}_i\right)}, \end{aligned} \quad (5)$$

where d_i is the Mahalanobis distance of the feature $\boldsymbol{f}_t$, T_j is the number of frames in the segment seg_j, σ is a constant (here, 3.0), $\boldsymbol{f}_t$ is the feature vector of hands in the frame t, $\boldsymbol{\mu}_{state}$ is the mean of the feature vectors that are already aligned to the state, and Σ_i is their covariance matrix. The feature vector $\boldsymbol{f}_t$ consists of $\|\tilde{\boldsymbol{v}}_{\mathrm{hand}}\|$ and each element of $\tilde{\boldsymbol{v}}$.

In the stage 2, states and/or transitions are added if a segment has no corresponding state. Such a segment consists of the motion which is not yet included in the initial topology. To construct the topology including such motions, a state is added for each segment without a corresponding state. In the path, we put the states in the same order of corresponding segments. In addition, each inserted state has a transition to the state itself.

4.3 Example of integration

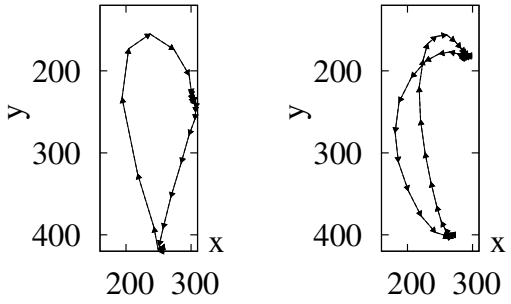
We show three examples of ingegration. As the example 1, we consider integration of two samples shown in Figure 2. The initial topology such as Figure 1 is constructed from the sample in Figure 2(a). The segments extracted from the sample in Figure 2(b) is integrated into the initial toplogy. The right hand motions of the



(a) The sample for the initial topology

(b) The sample to be integrated

Figure 2. The motion of the word "match"



(a) The sample for initial topology

(b) The sample to be integrated

Figure 3. The trajectory of right hand motions

samples are shown in Figure 3. The similarity between a state and segment is displayed in Table 1. The boxed cells in Table 1 compose the "best correspondence" that has the largest sum of similarity. From Table 1, we can find the best correspondence as Figure 4. In this example, the segment 2 and 6 have no corresponding state because the segment has negative similarity for all states. From the correspondence in Figure 4, we have the integrated topology shown in Figure 5. The result topology reflects that the two motions share intermediate stationary state and the word has variation in beginning and finishing motions.

As the example 2, we show that the proposed method allow positional variations because the similarity is based on the direction and velocity of motion. Two sample motions for the word "warm" are shown in Figure 6(a) and (b), where both hands are moved up to the front and then moved up and down as rotated. Although the positions and trajectories of the motions are different, the estimated topology shown in Figure 6(c) reflects correctly the segments of the samples.

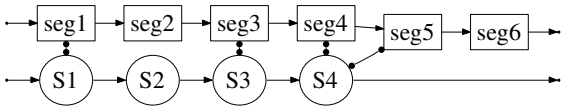


Figure 4. The matching result

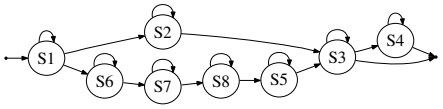
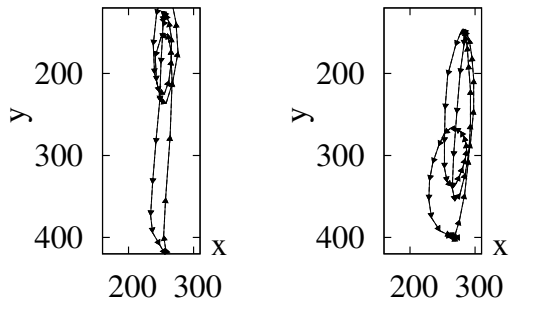


Figure 5. The generated topology

As the example 3, we show a more complex topology in Figure 7, which is generated for the word "winter clothing". In the motion for the word, both hands are vibrated near the face and then moved up. Although a state should be generated from a single vibration, the three states, S_2, S_3, and S_4 are generated in the topology. This is caused by the variation of motions among speakers. Since there are samples where hands move faster or slower in comparison with states in the initial model, multiple states are added for similar motions. Although the estimated topology may include unnecessary transitions, it accepts motions with vibration.

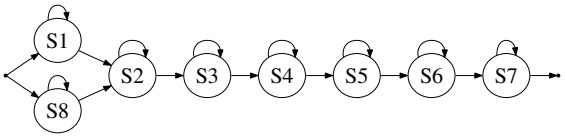
5 Experiment of recognition

We take the following features defined in Section 3 for recognizing sign language; $\boldsymbol{y}$, $\|\tilde{\boldsymbol{v}}\|$ and each element of $\tilde{\boldsymbol{v}}$ for each hand and $\boldsymbol{y}_{\mathrm{rel}}$ for relation of both hands.



(a) The right hand trajectory of sample 1

(b) The right hand trajectory of sample 2



(c) The estimated topology

Figure 6. The motions and the estimated topology for the word "warm"

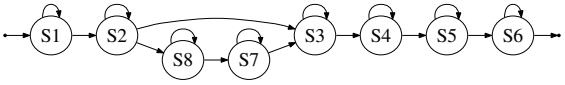


Figure 7. The estimation result for the word "winter clothing"

Considering the above features for each hand, we take 12-dimensional feature vectors for both hands.

In this experiment, we ask 2 speakers to perform 3 times for each word. We take 43 words with either hand and both hands. For each word, one of the samples is recognized and the others are used for training. The words have various motions. The recognition results by the previous method[4] and the proposed method are shown in Table 2. Although the models are automatically estimated without threshold adjusted for each speaker or word, the ratio of success is over 80% in most cases. The result of the proposed method is comparable to that of the previous method with thresholds adjusted for each speaker.

Table 2. The number of success in recognition for words with both hands

	Speaker 1	Speaker 2
previous method[4]	104(80.6%)	107(82.9%)
proposed method	117(90.7%)	109(84.5%)

43 words, 3 motion for each word

6 Conclusion

We proposed the method to automatically generate models for recognizing a sign language word. The proposed method consists of segmentation and integration. The former divides a motion into meaningful segments and the latter constructs a topology from multiple series of segments. By the proposed methods, the models can be automatically adapted for various motions for a word.

In addition, it is possible to tune the training of the model according to the property of the states because the states are classified by the proposed methods.

References

[1] G. Fang, W. Gao, and D. Zhao. Large vocabulary sign language recognition based on fuzzy decision trees. *Systems, Man and Cybernetics, Part A, IEEE Transactions on*, 34(3):305–314, May 2004.

[2] W. Gao, J. Ma, J. Wu, and C. Wang. Sign language recognition based on HMM/ANN/DP. *International Journal of Pattern Recognition and Artificial Intelligence*, 14(5):587–602, 2000.

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

[4] K. Kawahigashi, Y. Shirai, N. Shimada, and J. Miura. Segmentation of sign language for making HMM. *IEICE technical report*, 105(67):55–60, 20050513. (in Japanese).

[5] T. Starner, J. Weaver, and A. Pentland. Real-time american sign language recognition using desk and wearable computer based video. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(12):1371–1375, 1998.

[6] L.-G. Zhang, Y. Chen, G. Fang, X. Chen, and W. Gao. A vision-based sign language recognition system using tied-mixture density hmm. In *ICMI '04: Proceedings of the 6th international conference on Multimodal interfaces*, pages 198–204, New York, NY, USA, 2004. ACM.