Sign Language Recognition using HMM which Accepts Various Movements

Tadashi Matsuo*, Yoshiaki Shirai*
*Ritsumeikan University, 1–1–1 Noji–higashi, Kusatu, Shiga, Japan

***Abstract*— We propose a method to automatically construct a transitional structure(topology) of a Hidden Markov Model for recognizing a word in sign language from a sequence of images. 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. A motion is divided into consistent segments, which correspond to states of the model. The topology is constructed from an initial topology by modifying it according to a learning sequence of the segments. With experiments, we show the effectiveness of the proposed method.**

I. Introduction

For recognition of sign language words, a framework with Hidden Markov Model (HMM)[1] has been used, where the model usually corresponds to a word[2]. The model consists of states and transitional structure(topology). The state has parameters representing motions. The topology determines the transitional probabilitiy between states. Automatic generation of models for sign language has not been studied.

For generating models, we have to consider the following problems.

(1) How to determine states for representing motions?
(2) How to determine the topology among the states?

For problem (1), the number of states were generally fixed for all words[2]. Accordingly, the models can not reflect the difference of the complexity of motions for words. In [3], the number of states was estimated for each word. However, the thresholds of hand speed for the estimation were manually adjusted for each speaker.

In this paper, a state corresponds to a consistent partial motion(segment) such as raising hands, spreading hands etc. The segments are extracted from each motion for a word. The segmentation does not require thresholds manually adjusted for each speaker because it is based on directions of a motion, which are more independent on each speaker than the speed.

For problem (2), the topology of models was generally fixed for all words to a model such as Fig. 1 [2]. In [3], the number of states were adaptively determined for each word, but the topology was limited to linear such as Fig. 2. Such models reflect an only type of motions because they have only one transition path. A word in sign language generally has multiple types of motions. Although they can be represented by a set of models, partial motions common to the models are trained as different motions.

In this paper, a topology reflecting multiple types of motions is constructed for each word. It has branches and junctions to represent a flexible structure. The automatic construction is based on similarity between a state and a segment.

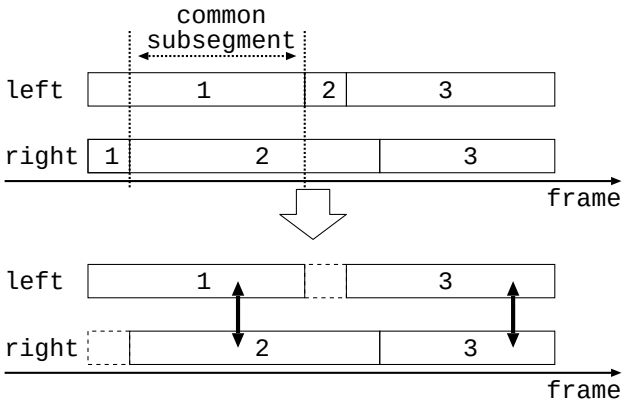


Fig. 3. Pairing segments of both hands

II. Segmentation of a training sample

The word in sign language is represented by one hand or both hands.

In case of one hand, the segmentation of a motion is based on the direction and speed of the motion. We divide a motion of a hand into segments. The segments are classified into the following fundamental types;

(1) Stationary.
(2) Moving: the hand moves sufficiently fast. The moving segment is further divided into two classes;
  (2a) Straight: the hand moves in an almost fixed direction.
  (2b) Turn: the hand changes the direction of motion in a short time.

Each segment has the property of the corresponding trajectory, which is used in the construction of the topology.

In case of both hands, a meaningful unit of segmentation is a pair of two simultaneous motions of both hands rather than a motion of each hand. Accordingly, we suppose that a state corresponds to a pair of simultaneous motions of both hands. A motion of each hand is segmented in the same way as that of one hand. Then simultaneous segments of both hands are paired. However, both hands do not synchronize generally. The segments are paired if they have a common subsegment with many frames or long trajectory (Fig. 3). If a segment

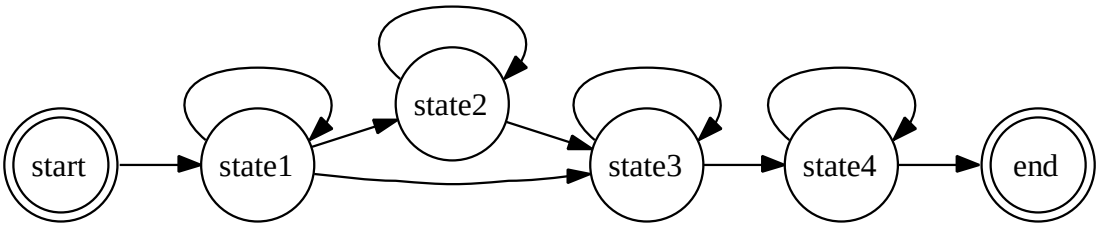


Fig. 1. A topology with a skip

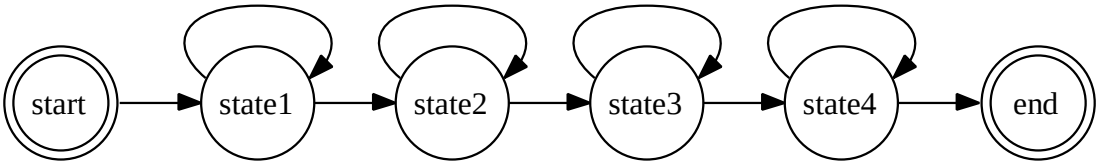
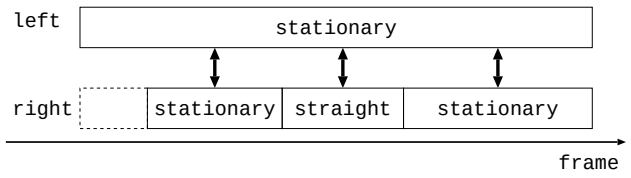
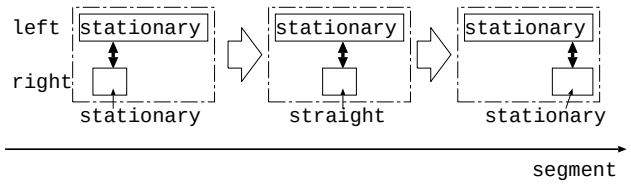


Fig. 2. A linear topology



(a) The segment corresponding to multiple segments



(b) The sequence of the paired segments

Fig. 4. The pairing of the segment corresponding to multiple segments

corresponds to multiple segments, they consistitute individual pairs as shown in Fig. 4.

III. Construction of topology

In this section, we describe how to generate topology adapted for a variety of motions for a word. 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.

A. Initial topology

We select the shortest series of segments. Each segment of the series corresponds to a state of the initial topology. The state is classified by the type of the corresponding segment. The states are arranged in the same order as the series of segments. Every state has two transitions; the one is to the state itself and the other is to the next state as shown in Fig. 2.

B. Integration of a series of segments into topology

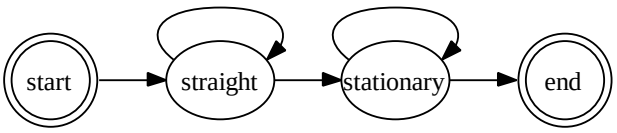
The integration of a new sample into the current topology is divided into the two parts:

(1) Determine the correspondence between the segments in the sample and the states in the topology.
(2) Add new states into the topology so that each segment has a corresponding state.

The matching is based on the similarity between a segment and a state. The similarity for a stationary state is determined by the segment type and that for a moving state is determined by the direction of the segment. We suppose that the states in the correspondence constitute a path from the initial state to the final state of the topology. We also suppose that the order of corresponding segments is preserved as the original series. With the supposition, we take the "best" correspondence that maximizes the total sum of the similarity. In this paper, we take the following similarity.

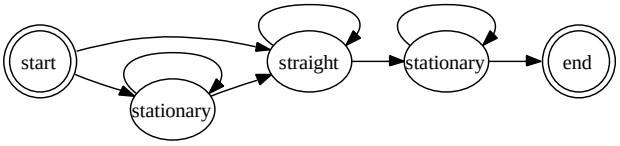
$$C(\text{state}, \text{segment}) = \begin{cases} w_p, & \text{(a stationary state and a stationary segment)} \\ -w_n, & \text{(a stationary state and a moving segment or a moving state and a stationary segment)} \\ \cos(\theta_{\text{segment}} - \theta_{\text{state}}), & \text{(a moving state and a moving segment)} \end{cases}, \quad (1)$$

where w_p and w_n denote the weights of consistency and inconsistency of stationary state, θ_{segment} denotes the direction



(a) The initial topology with one path

(b) The series with a preceding stationary segment



(c) The result of integration

Fig. 5. The integration of a series of segments into the initial topology

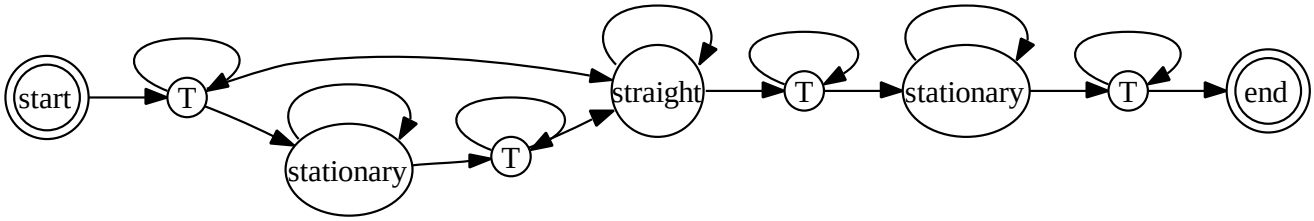


Fig. 6. Insertion of temporary states

of total motion in the segment, and θ_{state} denotes the average direction of the frames in the segments assigned to the state. In this paper, we take $w_p = w_n = 1$.

For each series of segments without corresponding state, the same number of states are inserted as an additional path. The path starts from the state corresponding to the segment preceding the series and ends at the state corresponding to the following segment. 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.

We show an example of integration in Fig. 5. The initial topology does not have stationary states before a moving state as shown in Fig. 5(a). We integrate the series of segments that includes a preceding stationary segment as shown in Fig. 5(b). As shown in Fig. 5(c), the integration insert a state and transitions as an additional path of the initial topology. The additional path allows the topology to accept a series which includes a preceding stationary segment, though the initial topology does not accept such a series.

C. Temporary states for frames between segments

For words with both hands, the frame between two paired segments may have features different from both the previous and next segment because the motions of a hand may not be synchronized with that of the other hand as shown in Fig. 3. The feature in such frames may reflect the previous segment for a hand and the next segment for the other hand. Or the feature may have the transitional property different from both the adjacent segments. The model should accept such frames because they are temporary but inevitable.

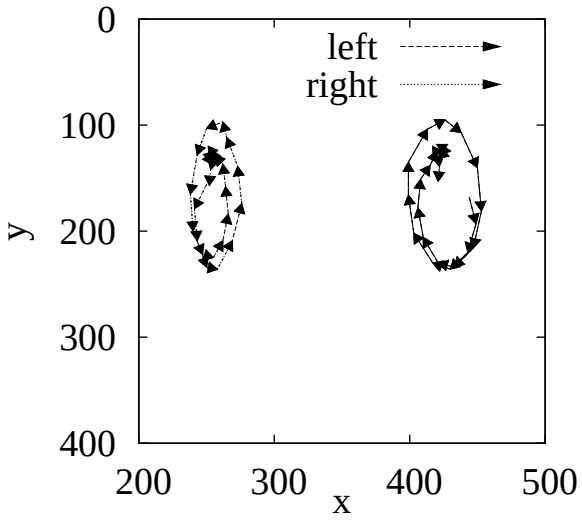
To accept them, we add temporary states as the next states of every state as shown in Fig. 6, where the label "T" means a temporary state. To accept the unstable features, the temporary state has large variances of features. The state also has a recursive transition with low probability in order to reject long motions.

IV. Experiment of estimation of topology

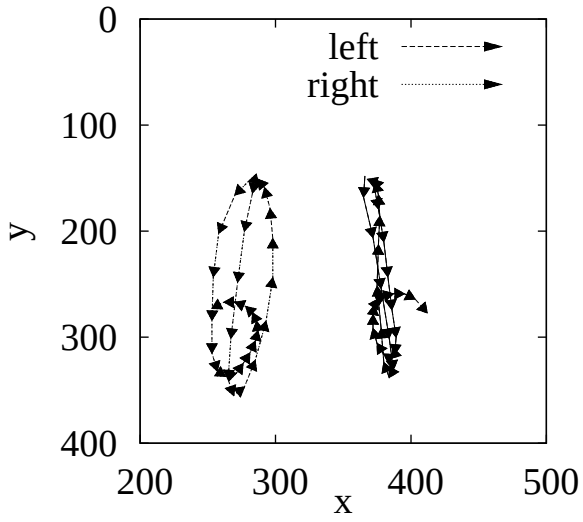
We estimated topology from samples extracted from real images for 23 words. For each words, the estimated topology reflects a variety of motions well. In this section, we show typical results that indicate adaptivity of the proposed method. In the figures for the following example, motions of hands are indicated in the corresponding states. The temporary states are omitted for convenience.

The proposed method does not depend on the position of motions because it is based on the segments. As an example, in Fig. 7(a) and (b), we show two samples for the word "warm", where both hands are moved up to the front and then moved up and down as rotated. Although the positions and trajectories of the samples are different, the estimated topology shown in Fig. 7(c) reflects correctly the segments of the samples.

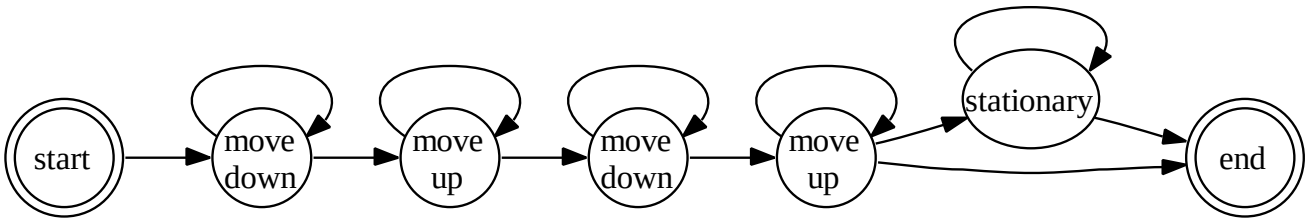
We mention the case where the preparing and terminating motion are not correctly omitted. As an example, we take



(a) The trajectory of sample 1



(b) The trajectory of sample 2



(c) The estimated topology

Fig. 7. The motions and the estimated topology for the word "warm"

the word "short", where one hand is moved down from the high position. For the word, the extra motions may be omitted successfully as shown in Fig. 8(a) or may not be omitted as shown in Fig. 8(b). The extra motions remain if the preparing or terminating motion consists of multiple segments. The estimated topology accepts flexibly the extra motions because it has two essential states and three avoidable states as shown in Fig. 8(c).

As an example of a more complex word, we show the topology estimated for the word "winter clothing" (Fig. 9), where both hands are vibrated near the face and then moved up. In the vibration, the segmentation of motions is unstable because some frames may be considered as stationary frames. Although we should define a state type specific for such vibration, the estimated topology accepts motions with vibration.

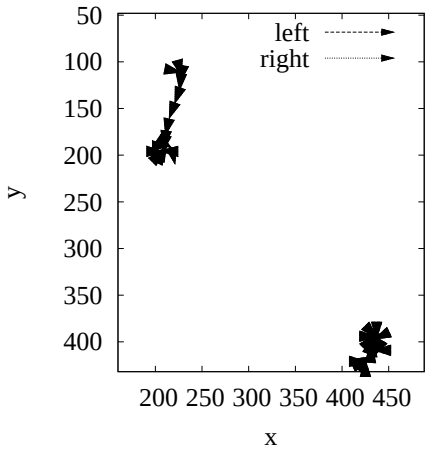
V. Experiment of recognition

In this section, we describe features for recognizing a word in sign language. Then, we show the experimental result of the recognition.

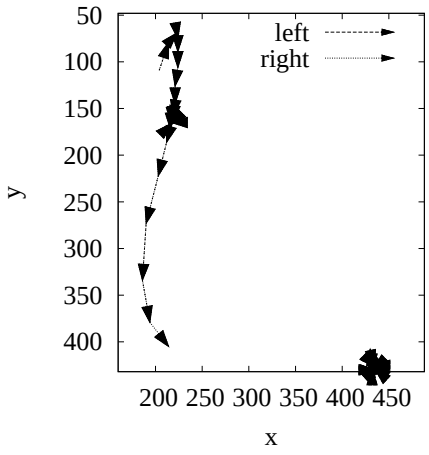
In sign language, the positions of the hand should be distinguished in detail when the hand is near to the face. However, the positions should be distinguished roughly when the hand is far from the face. To reflect such property, we take the following features including logarithmic positions for recognizing sign language;

- the position $\vec{x}_{\mathrm{loghand}} = \mathrm{logvec}\,(\vec{x}_{\mathrm{hand}} - \vec{x}_{\mathrm{face}})$, where $\mathrm{logvec}\,(\vec{x}) = \left\{\log\left(1 + \frac{\|\vec{x}\|}{R}\right)\right\}\vec{x}$,
- the direction of the motion $\vec{v}_{\mathrm{direction}} = \frac{1}{\|\vec{v}_{\mathrm{hand}}\|}\vec{v}_{\mathrm{hand}}$, and
- the speed $\log\|\vec{v}_{\mathrm{hand}}\|$,

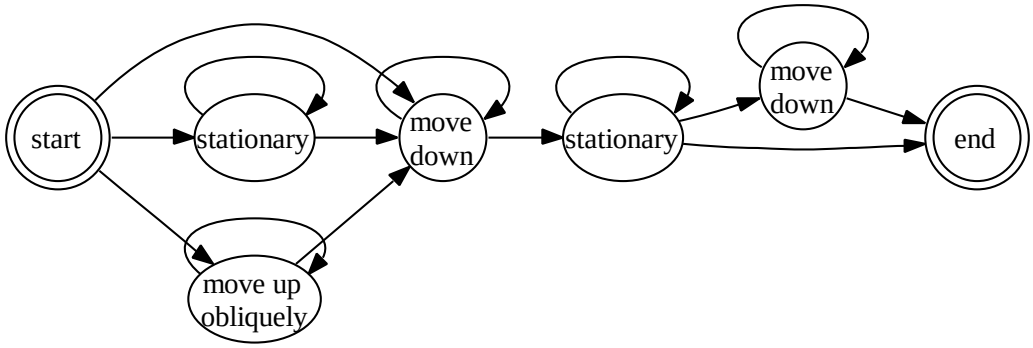
where $\vec{x}_{\mathrm{face}}$ and $\vec{x}_{\mathrm{hand}}$ are the center of gravity of the face and that of the hand, respectively. Additionally, $\vec{v}_{\mathrm{hand}}$ and $\|\vec{v}_{\mathrm{hand}}\|$ mean the velocity vector and the speed of the gravity center of the hand, respecitively. $\mathrm{logvec}(\vec{x})$ is nearly logarithmic if $\|\vec{x}\|$ is sufficiently larger than R and it is nearly linear if $\|\vec{x}\|$ is not so large. Therefore, the vector $\vec{x}_{\mathrm{loghand}}$ is roughly distinguished when the hand is far from the face. The vector $\vec{x}_{\mathrm{loghand}}$ is distinguished in detail when the hand is near to the face. Considering the above features for each hand, we take 5-



(a) The trajectory of the hands without preparing motion



(b) The trajectory of the hands with preparing motion



(c) The estimated topology

Fig. 8. The estimation result for the word “short”

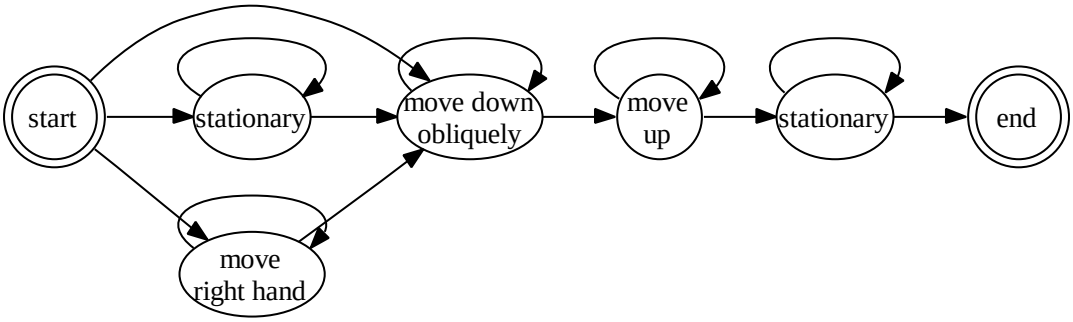


Fig. 9. The estimation result for the word “winter clothing”

dimensional feature vectors for one hand and 10-dimensional feature vectors for both hands.

In this experiment, we ask 2 speakers to perform 3 times for each word. We take 8 words with one hand and 15 words with both hands. For each word, one of the samples is recognized and the others are used for training. The words have various motions as shown in IV.

In [3], each recognized word has an only type of motions

TABLE I

THE RESULT OF RECOGNITION FOR WORDS WITH ONE HAND

	Speaker 1			Speaker 2		
	Motion A	Motion B	Motion C	Motion A	Motion B	Motion C
the number of success	6	6	7	7	5	6
the ratio of success	75%	75%	88%	88%	63%	75%

the total number of words:8

TABLE II

THE RESULT OF RECOGNITION FOR WORDS WITH BOTH HANDS

	Speaker 1			Speaker 2		
	Motion A	Motion B	Motion C	Motion A	Motion B	Motion C
the number of success	14	13	12	13	11	12
the ratio of success	93%	87%	80%	87%	73%	80%

the total number of words:15

because the model does not reflect various types of motions. In this experiment, words with various types of motions are also taken.

The recognition result is shown in Tab. I and II. Although the models are automatically estimated without threshold adjusted for each speaker or word, the ratio of success is over 70% in most cases.

VI. CONCLUSION

In this paper, 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] S. Nakagawa, *Speech recognition by statistical model*. IEICE, 1988, (in Japanese).

[2] 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*, vol. 20, no. 12, pp. 1371–1375, 1998. [Online]. Available: citeseer.ist.psu.edu/starner98realtime.html

[3] K. Kawahigashi, Y. Shirai, N. Shimada, and J. Miura, "Segmentation of sign language for making HMM," *IEICE technical report*, vol. 105, no. 67, pp. 55–60, 20050513, (in Japanese). [Online]. Available: http://ci.nii.ac.jp/naid/10016435220/