Extraction of Hand Features for Recognition of Sign Language Words

Nobuhiko Tanibata
tanibata@cv.mech.eng.osaka-u.ac.jp

Nobutaka Shimada
shimada@cv.mech.eng.osaka-u.ac.jp

Yoshiaki Shirai
shirai@cv.mech.eng.osaka-u.ac.jp

Computer-Controlled Mechanical Systems,Graduate School of Engineering,Osaka University

Abstract

This paper proposes a method to obtain hand features from sequences of images, where a person is performing the Japanese Sign Language (JSL) in a complex background and to recognize the JSL word.

At the first frame, we find a person's region, and then search for a face, hands in order to determine a range of skin color and search for elbows to determine the position of a wrist.

At each frame, we track the face, the hands by using the decided skin color and track the elbows by matching the template of a elbow shape. When face and hands overlap, they are extracted by matching texture templates of the previous face and hands. Hand features such as the hand direction, the number of fingers, etc. are extracted from the hand regions and the wrist.

In order to recognize JSL words, we use a sequence of the hand features as an input to HMM. We first select words which reach the final state of HMM, and then determine one with the highest probability. We made an experiment with real images of a professional JSL interpreter and recognized 65 JSL words successfully.

1 Introduction

Because Japanese Sign Language (JSL) is the most popular gesture language in Japan and common people cannot understand JSL, a system to interpret JSL is useful. Such a system requires two functions; translating Japanese into JSL and vice versa. Although the former has already developed, but the latter has not for practical use. The difficulties in the latter are extraction of hand shapes and motion and recognition of a sequence of them.

For recognition of a JSL word, we should extract hand shapes as well as motions to obtain hand features because there is a JSL word with the same motions as the other. The CyberGlove is often used[1] to extract them. However it is trouble for users to put on and take off, or wear it. Thus we focus on extracting hand features from a sequence of images without any devices on hands.

On the other hand, vision-based extraction of the hand features has studied in measuring hand shape as well as a field of sign language recognition. The former work[2] obtains hand shapes by matching hand silhouette with hand 3D CG in a simple background. But it is difficult in a complex background. The latter works[3, 4] extract skin regions from a range of skin color. However that cannot be completely done because skin color depends on each user or each situation.

Vision-based extraction of the hand features often includes errors due to image noises or background whose color is similar to skin color. For recognition of a sequence of hand features robustly, Hidden Markov Model (HMM) is one of effective methods[8]. This is often used for recognitions of facial expressions[5], gestures[6] and human behavior[7]. However it is difficult to recognize the JSL words only with probabilities obtained by HMM. Because some JSL words are similar to a part of other words and the probabilities of them are also high.

In our method, skin regions, clothes region and elbows are roughly extracted at the first frame by matching the template of a initial pose to a person's region. The template includes face and hand regions, clothes region and the positions of elbows. The rough skin regions include the parts of clothes region. We remove them from the rough skin regions using color information of clothes region. The skin color depends on the user and it is easily affected by lighting, backgrounds, and reflected light from the user's clothes. In order to cope with this problem, we determine the range of the

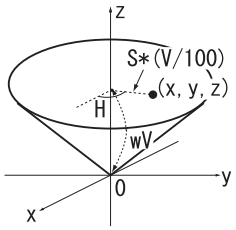


Figure 1: *The relation between xyz and HSV.*

skin color in HSV color space from the extracted skin regions. At each frame, our system tracks the face, the hands and the elbows by using the range of the skin color and the templates of shape. When the face and the hands overlap, we divide them by matching the texture template of the previous face and hand. We obtain features of left and right hands from the face, each hand and each elbow respectively.

We assume that each JSL word sample is segmented from the sentence. Thus we can obtain two segmented sequences of features about left and right hands. We build a HMM for each hand using each sequence respectively.

Sometimes the whole of a word is similar to the beginning part of another word. It is difficult to distinguish only with the highest probability of HMMs. Thus we select a word with highest probability among these which reach the final state of HMM at the end of the sequence.

In the following sections, initialization of tracking of the hand, the hands and the elbows is first explained. Then it is followed by the details of the tracking, feature extraction and recognition of the JSL words. Finally experimental results are shown.

2 Initialization of tracking

We assume that a background image is given by a fixed camera to extract a person's region and that each JSL word starts from a common initial pose. We initialize the positions of a face, hands and elbows by matching the initial pose template to the person's region. Then we determine the range of the skin color using the color of the extracted face and hands.

2.1 Segmentation of face, hand, clothes and elbow

We extract a person's region using a background subtraction in HSV color space often applied to color representation. In this color space, (1)H (hue) is not stable if S (saturation) is small, and (2)H and S is not stable if V (value) is small. (3)Shadow regions are also

(a) Original image. (b) Person's region.

Figure 2: *Obtaining a person's region.*

obtained because V of shadow regions is smaller than that of a background. We should make the background subtraction not affected by these characters. We first convert HSV space into a x-y-z coordinate cone in Fig. 1 by Eq. (1).

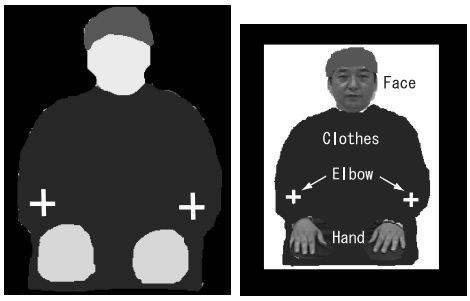
$$\begin{aligned} x &= S*(V/100)*cos(H) \\ y &= S*(V/100)*sin(H) \\ z &= w*V \\ & 0 \le H \le 2\pi, 0 \le S \le 100, 0 \le V \le 100 \\ & 0 \le w \le 1 \end{aligned} \quad (1)$$

where w is a normalization weight for V's change. We obtain a person's region using the background subtraction in the cone. When S is small, the background subtraction is hard to affected by change of H because such a color is mapped around the axis of the cone. When V is small, it is also hard to affected by change of H and S because such a color is mapped around the bottom of the cone. Smaller w is, harder to be affected it is by change of V because the height of the cone is small. In this research, we use $w = 0.5$. The result of extracting the person's region is shown in Fig. 2. We ignore under knees because there are not skin regions.

We use the template of the initial pose including the face, the hands and the elbows as shown in Fig. 3(a). The face and hand regions are roughly extracted by matching the template to the person's region as shown in Fig. 3(b). A clothes region is defined as the matched region except the face and hand regions. The positions of elbows are also obtained. The rough face and hand regions include the parts of the clothes region. Thus we remove the parts with the same color as the clothes region from the face and hand region.

2.2 Skin color range

We obtain the skin color and the clothes color from the face and hand regions and clothes region respectively.



(a) Template of initial pose (b) Matched region

Figure 3: *Matching result of the template of a initial pose to a person's region*

Assuming that each color distribution is a normal distribution in H-S color space, an ellipse including 90% of samples from each distribution is regarded as the color range of each region. A and B in Fig. 4 is the ranges of clothes and skin color respectively.

However, it is difficult to extract tips of fingers exactly as shown in Fig. 5(a), because the skin color is mixed with the clothes color around the outline of hands when hands and clothes overlap.

Therefore, we additionally define the range of a middle color between the skin and clothes color. As shown in Fig. 4, it is defined as the minimum circle E, whose center C is on the line D connecting the center of the clothes color to that of the skin color, which is tangential to those two regions. We redefine the range of skin color as the union of B and E. The tips of the fingers are extracted by using the new range of the skin color as shown in Fig. 5(b).

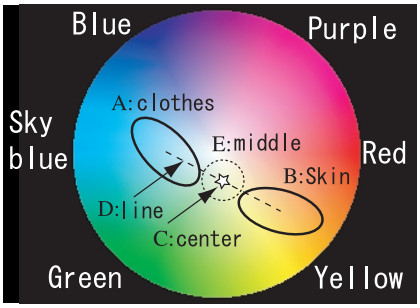
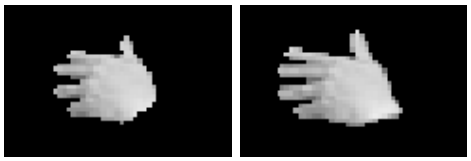


Figure 4: *The ranges of colors in H-S space.*



(a) Only with the range of skin color (b) With the range of the skin and middle color

Figure 5: *Exact extraction of hand outline with a range of middle color.*

3 Tracking of face, hand and elbows

3.1 Basic process for face and hands

At the first frame, the regions of the face and hands are already extracted. In the next frame, the search region of the face is determined as the person's region around the previous face region. For both hands, we predict the position of each hand from the previous hand position, velocity and acceleration. The search regions of both hands are determined as the person's region around the predicted position. If we extract multiple skin regions in each search region, we select the nearest to the center of each search region. Finally, we register the textures of the face and the hands to be used in cases of overlap (see two following sections).

3.2 Segmentation of overlapping face and hand

When the candidate of the hand is the same as the face and the face area increases by more than a half of the previous hand area, we estimate that the hand and the face overlap. We first determine the position and rotation of the face template (X_F, Y_F, Θ_F) using Eq. (2).

$$
\begin{aligned}
(X, Y, \Theta) &= \arg \min_{\substack{x_0, y_0 \in A \\ -15^\circ < \theta < 15^\circ}} Sum(x_0, y_0, \theta) \\
Sum(x_0, y_0, \theta) &= \sum_{x, y \in T_\theta} D_{comp}(T_\theta(x, y), O(x_0, y_0)) \\
D_{comp}(T(x, y), O(x_0, y_0)) &= \\
&\quad \{R_T(x, y) - R_O(x + x_0, y + y_0)\}^2 \\
&\quad + \{G_T(x, y) - G_O(x + x_0, y + y_0)\}^2 \\
&\quad + \{B_T(x, y) - B_O(x + x_0, y + y_0)\}^2 \quad (2)
\end{aligned}
$$

where, T_θ is the θ rotated template, A is the overlapping region, O is an original image, R,G and B are compo-

nents(red, green and blue) of image, $D_{comp}(T(\cdot), O(\cdot))$ is the difference of pixel values.

We determine search regions to match the template of the hand. As shown in (1) of Fig. 6, we remove the face region from the overlapping region if $D_F(x, y)$ in Eq. (3) is smaller than a threshold.

$$D_F(x, y) = D_{comp}(F_{\Theta_F}(x, y), O(X_F, Y_F)) \quad (3)$$

where, F_{Θ_F} is the Θ_F rotated face template, O is the overlapping region. The rest regions are regarded as the search regions. Then, the hand region is obtained using Eq. (2) in the search regions.

Because the shape of hands may change during the overlapping, however, the hand region is not perfectly extracted by matching the template of hand. Therefore, regions which is not the face region around the hand region are regarded as parts of the hand region as shown in (2) of Fig. 6. The hand region can be obtained unless hand shape changes largely. The segmentation result of overlapping face and hand is shown in Fig 8(a).

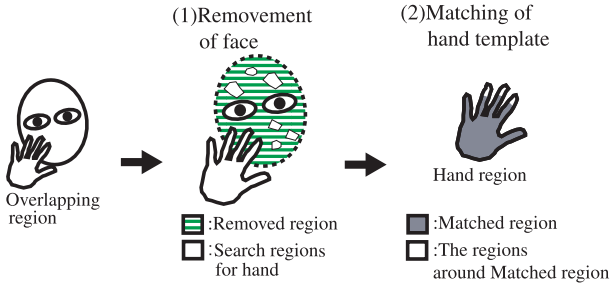


Figure 6: *Segmentation of face and hand*

3.3 Segmentation of overlapping hands

When a candidate of a hand is the same as the other hand and the hand area increases by more than a half of the previous hand area, we estimate that both hands overlap. First, the position and rotation (X_H, Y_H, Θ_H) of each hand template is obtained by Eq. (2) respectively. Then, we determine that the hand template with smaller D_H obtained by Eq. (4) is front.

$$D_H = \sum_{x,y \in A} D_{comp}(H_{\Theta_H}(x, y), O(X_H, Y_H)) \quad (4)$$

where, H_{Θ_H} is the Θ_H rotated template of the hand, O is the original image, A is the overlapped region as shown in Fig. 7.

The front hand template is renewed by the texture of the current image. Not overlapping part of the back hand template is also renewed. The segmentation result of overlapping hands is shown in Fig 8(b).

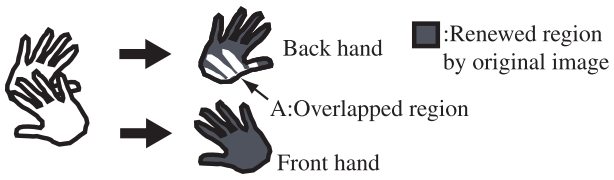
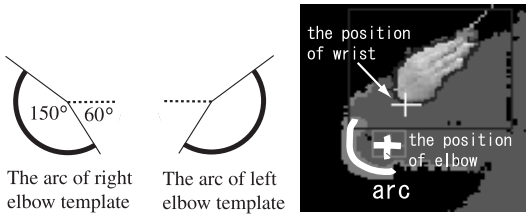


Figure 7: *Segmentation of hands*

3.4 The position of elbow

We assume that the direction from the elbow to the hand is toward around the user's face or body. Thus we determine the template of an elbow as an arc which has a thickness of some pixels as shown in Fig. 9. In order to find the position of elbow, we first put the arc onto the outline of the person's region obtained in section 2.1 around the previous position of the elbow. Next we count the number of outline points overlaped with the arc. The center of the arc with the largest number is defined as the position of the elbow.



(a) Templates of elbow (b) The result of matching

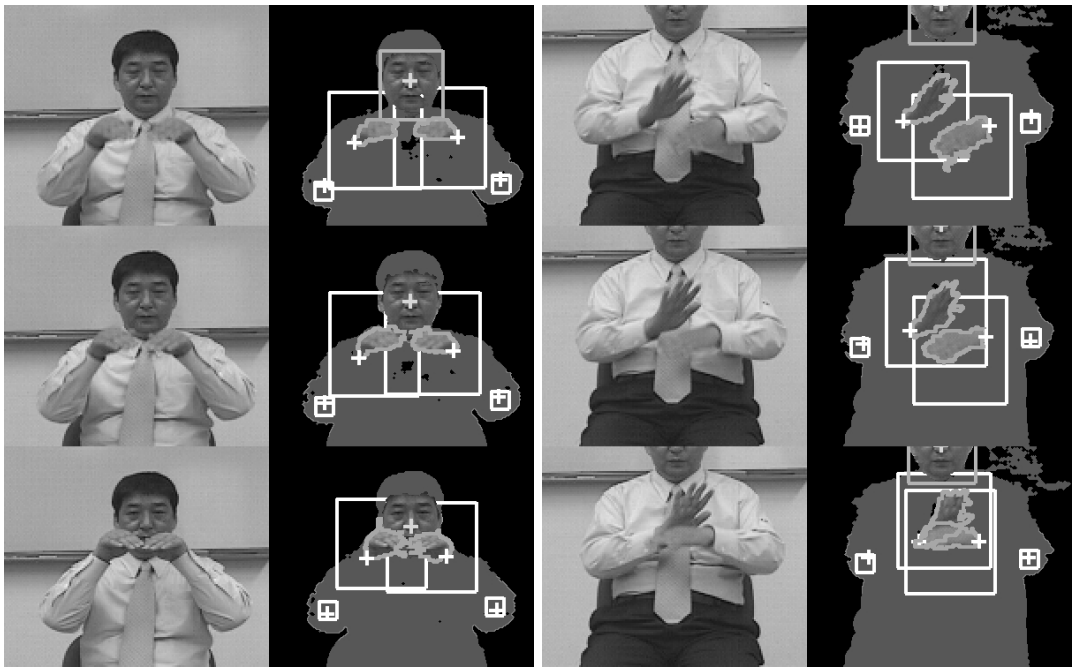
Figure 9: *Finding the elbow*

4 Feature extraction

We define 6 features for recognition of JSL words.

1. r : The flatness of hand region
2. (x_{hand}, y_{hand}) : The gravity center position of the hand region relative to that of face region
3. A : The area of the hand region
4. θ_{motion} : The direction of hand motion in the image coordinate
5. θ_{hand} : The direction of hand region in the image coordinate
6. N_p : The number of protrusions

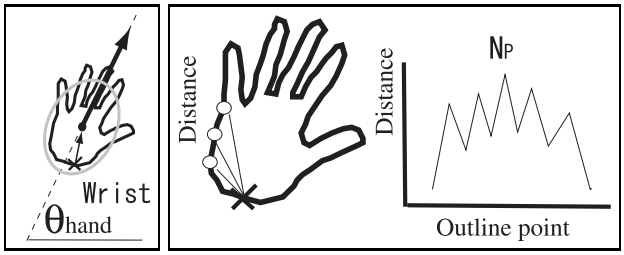
$(x_{hand}, y_{hand}), A, \theta_{motion}$ are easily obtained from the face and hand regions. The hand ellipse is defined as the ellipse of inertia of the hand region as shown in



(a) Overlapping of hands and face (b) Overlapping of a hand and the other hand

Figure 8: The result of extracting skin region:each search region of hand region is a quadrilateral, the center of hands and face and the positions of elbows are cross, and hand outline is drawed.

Fig. 10 and r is defined as the ratio of the major axis to the minor axis. θ_{hand} is define as the direction of the major axis as shown in Fig. 11(a). N_p is defined as the number of local maxima of the distance between the wrist and outline points of the hand region as shown in Fig. 11(b). The wrist position is defined as the hand region point nearest to the elbow position as shown in Fig. 9(b).



(a) θ_{hand} (b) N_p

Figure 11: Feature Extraction

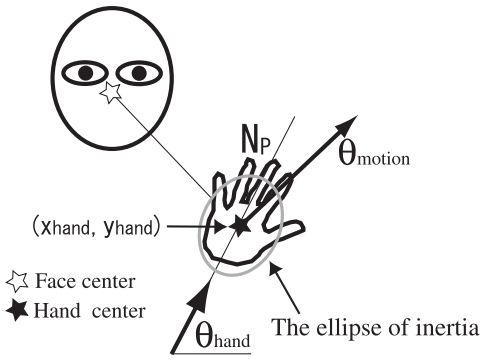


Figure 10: *Extraction of features.*

5 Recognition of JSL words

5.1 Modeling

A JSL word is composed by hand motions and shapes. Thus each component can be matched off against a state of HMM, for instance, a hand motion to some direction is a component, posing some hand shape is other one as shown in Fis. 12.

In order to determine HMM of a JSL, we first detect the borders of states from a sequence of sample

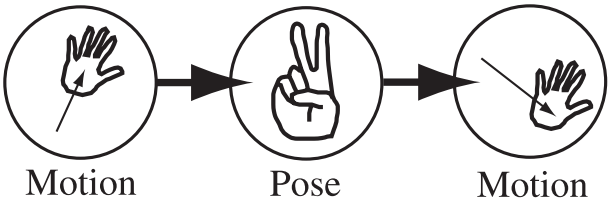


Figure 12: *The correspondence of JSL components and HMM states.*

for modeling. Considered from the character of JSL words, the hand motion is important when its velocity is large, while the hand shape is important when its velocity is small. Thus the borders are detected when Eq. (5) is satisfied. This equation is to find borders between motions and poses. In addition, the borders are also detected when the direction of the hand motion changes largely by Eq. (6).

$$\begin{cases} V(t) \leq \Theta_V & \text{if } \bar{V} > \Theta_V \\ V(t) \geq \Theta_V & \text{if } \bar{V} < \Theta_V \end{cases}$$

$$V(t) = \sqrt{\begin{array}{l}\{x_{hand}(t) - x_{hand}(t-1)\}^2 \\ +\{y_{hand}(t) - y_{hand}(t-1)\}^2\end{array}} \quad (5)$$

where

Θ_V:threshold for the change of the hand velocity.

$\bar{V}$:the average velocity from $t-1$ to $t-n$

$$I(t) > \Theta_I$$
$$I(t) = |\theta_{motion}(t) - \theta_{motion}(t-1)| \quad (6)$$

where

Θ_I:threshold for the change of the direction of the hand motion.

After we determine the initial parameters of HMM using the borders, we build HMMs of left and right hands for each JSL word using Baum-Welch algorithm[8]. To confirm whether HMM is correctly modeled, we check manually whether the transitions of the samples occur around the borders using Viterbi algorithm[8] and whether automatically determined parameters of HMMs agree with our decision, by illustrating the parameters of each states such as the average of hand features, the covariance ellipse of (x_{hand}, y_{hand}) and the range of θ_{motion}.

5.2 Recognition

We assume that the JSL word samples are segmented into each word. Thus we do not have to find the start and the end of the words.

First, we obtain two sequences of features for left and right hands from one JSL word sample. Next we obtain the probability of each sequence using each HMMs of one word. We obtain the probability by multiplying both probabilities as shown in Fig. 13. In the same way, we obtain the probabilities of other JSL word HMMs. However there are JSL words whose beginning parts are similar to another JSL word. It is difficult to recognize only with the highest probability because such a word's probabilities may be also high. Thus, we find the candidates of the word which reach the final state using Viterbi algorithm[8]. Then we select one with the highest probability.

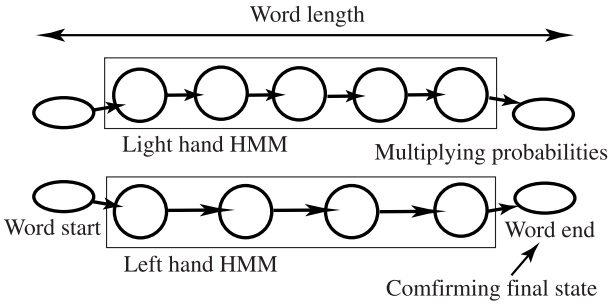


Figure 13: *The method to multiply the probabilities*

6 Experiments

6.1 JSL word samples

Our goal is the recognition of the JSL words in many kinds of situations: shop, city office etc. Now we focus on the situation of buying clothes.

We use the JSL samples as shown in Tab. 1. We took the samples performed by a professional JSL interpreter. This samples were taken in a complex background whose color is not similar to the skin color. The clothes of the interpreter is not similar to the skin color and a long sleeves.

This includes three samples of 70 words. The speed of performing the JSL words is for intermediate level users. The performed JSL sentences are segmented into each JSL word manually. There are samples including overlapping of the hands and the face (see Tab. 2).

Table 1: Sign language samples

	words samples
adj	blue,red,long,short,thin,thick,etc
adv	more,really,a little,etc
none	one-piece dress,L,M,S,credit card,hat,etc
verve	search,pay,take select,put in,etc

Table 2: The number of samples including overlaps

	the number of samples
face and hand	22
both hands	7
face and both hands	6
total	35

6.2 Extraction of skin region

We tested whether our system can extract the face and the hands, and segment the overlapping regions into each region. We obtained them exactly in 65 in the 70 samples.

In a complex background, we can obtain the skin regions using our method. We tried to extract skin region with many kinds of situations and users. We could deal with them by determining the range of the skin color at the first frame. However we cannot divide hand and face exactly when the facial expression and direction or the hand shapes change dramatically.

6.3 Recognition of the JSL word

We used 65 JSL words whose face and hands were extracted exactly. We tested whether our system can recognize them robustly with HMM. Each JSL word has three samples. We use two of them to model the JSL word and the rest to recognize. We could recognize 64 in the 65 word samples only with the probability. The words we could not distinguish only with the probability is shown in Fig. 14. The beginning of the word "cut the price" is similar to the whole of the word " little". In such a case, the probability of the latter was highest accidentally. We could distinguish them because the model of the word "cut the price" did not reach the final state.

Figure 14: *Example of misunderstanding*

7 Conclusion

In this paper, we propose a method of extraction of hand features and recognition of JSL words. For tracking the face and hands, we initialize the face and hand regions by matching the initial pose template, and decide the range of skin color at the first frame. We can track the overlapping face and hands by matching the texture template of the previous face and hands. We extract 6 features of the face and hands for recognition by the HMM. We can distinguish a word similar to the beginning part of another word by comfirming whether they reach the final state of the HMM. Our system could recognize 65 JSL words in the experiment with real images in a complex background.

We should test whether our system can recognize JSL words for more JSL words and users. We try to develop a system to recognize JSL sentences applying our method.

References

[1] H.Sagawa,M.Takeuchi, "A Method for Recognizing a Sequence of Sign language Words Represented in Japanese Sign Language Sentence",*Face and Gesture*,pp. 434–439,2000.

[2] N. Shimada, K. Kimura and Y. Shirai, "Real-time 3-D Hand Posture Estimation based on 2-D Appearance Retrieval Using Monocular Camera", *Proc. Int. WS. on RATFG-RTS (satellite WS of ICCV2001)*, pp. 23–30, 2001.

[3] K.Imagawa, "Color-Based Hands Tracking System for Sign Language Recognition", *Face and Gesture (FG1998)*, pp. 462–467,1998.

[4] K.Imagawa,H.Matsuo,R.Taniguch, and D.Arita, "Recognition of Local Features for Camera-based Sign Language Recognition System", *Face and Gesture (FG2000)*,pp. 849–0853,2000.

[5] T.Otsuka,J.Ohya, "Spotting Segments Displaying Facial Expression Sign Image Sequences Using HMM",*Face and Gesture (FG1998)*,pp. 442–447,1998.

[6] Takio Kurita, Satoru Hayamizu, "Gesture Recognition using HLAC Features of PARCOR Images and HMM based Recognizer", *Face and Gesture (FG1998)*,pp. 422–427,1998.

[7] Christopher R. Wren, Brian P. Clarkson, Alex P. Penland, "Understanding Purposeful Human Motion", *Face and Gesture (FG2000)*,pp. 378–383,2000.

[8] Jurgen Kinscher,Holger Trebbe, "The Munster Taging Project - Mathematical Background", *Arbeitabereich Linguistik, University of Munster*,D-58149 Munster,May 19 1995.