A Human-machine Cooperative System for Generating Sign Language Animation Using Thermal Image Processing, Fuzzy Algorithm, and Simulated Annealing

T. Asada and Y. Yoshitomi

Div. of Environmental Sciences, Graduate School of Life and Environmental Sciences, Kyoto Prefectural University
1-5 Nakaragi-cho, Shimogamo, Sakyo-ku, Kyoto 606-8522, Japan
E-mail: t_asada@mei.kpu.ac.jp, yoshitomi@kpu.ac.jp

Abstract: We propose a method for sign language animation by skin region detection applied to a thermal infrared image. In a system incorporating the proposed method, a 3D CG model corresponding to a person's characteristic posture while using sign language is generated automatically by pattern recognition of the thermal image, and then a person's hand in the CG model is set. The hand part is made manually beforehand. If necessary, the model can be replaced manually by a more appropriate model corresponding to training key frames and/or the same generated model can be refined manually. In our experiments, three hearing-impaired people, experienced in using sign language, recognized the Japanese sign language gestures of 70 words expressed as animation with 94.3% accuracy. We further improved the system by correcting the position and the direction of the hand of the automatically generated model through the use of a fuzzy algorithm and simulated annealing.

Keywords: Japanese Sign Language, Thermal Image Processing, Computer Graphics, Model Fitting, Fuzzy Algorithm, Simulated Annealing.

I. INTRODUCTION

Sign languages enable hearing-impaired people to communicate with each other. However, it is inconvenient for communication with non-hearing-impaired people because there are not enough sign language interpreters. Therefore, there is a strong need for an automatic sign language translation system. In addition to a sign language recognition function, an animation function is necessary for such a system. Several systems for sign language animation have been studied [1-4]. For animations expressing personality and/or emotion appropriately, conventional systems need many extra manual operations. To lower the workload of sign language systems, one of the most promising approaches is to improve the animation automatically acquired from dynamic images of real motion, with only the refinement performed manually. We previously proposed a method for expressing a sign language gestures as computer graphics (CG) animation that uses a human thermal infrared image taken without placing special restrictions on the person providing the image [5].

In this study, we evaluated a system implementing our method [5] for generating sign language animation from a dynamic image of a hearing-impaired person by measuring the recognition accuracy of other hearing-impaired people. Then, we further improved the system by correcting the position and the direction of the hand of the automatically generated model through the use of a fuzzy algorithm and simulated annealing.

II. PROPOSED METHOD

The flowchart for generating sign language animation from an input dynamic image is shown in Fig. 1. In the next sections, important elements in the total procedure are described.

1. Thermal image generation

The thermal image is produced by a thermal video system (NEC Avio Infrared Technologies Co., Ltd., Neo Thermo TVS-700), then transformed into a digital signal by an A/D converter (Thomson Canopus ADVC-300), and input into a computer with an IEEE1394 interface board (IO Data Device 1394-PCI3/DV6). Then, the image for each signed word is recorded as an AVI file on a PC. The images that are input into the computer have a spatial resolution of 720 × 480 pixels and 256 possible gray levels.

2. Selection of key frames from training images

After erasing the noise on each image frame in the AVI file, the sum of the differences of the gray levels of all pixels between the present frame and the previous frame is calculated. On the assumption that the characteristic postures in sign language gestures correspond to frames showing slower movement, a frame having a small gray level difference from the previous frame is considered to be a candidate of

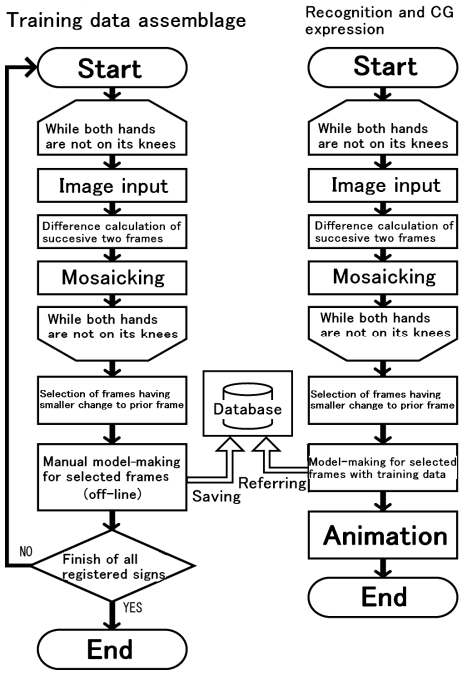

Fig. 1. Flowchart for generating sign language animation

a suitable frame (hereafter called a key frame) for making a 3D CG model. That is, the sum of the differences of the gray levels for the previous frame is used for picking out several key frames. As the selection criterion, we first select the top β % frames of the inverse value of the sum. However, similar successive frames can be selected. To select the key frames, we remove all frames except the first and last frames (when three or more successive frames are selected) with the same value of β for the inverse value of the sum of the gray level differences (Fig. 2). We use 25, 50, and 75 as the values of β. All frames selected with each of these three values for a sign are designated as key frames. The feature vector used for pattern recognition is made from the mosaic image after smoothing.

3. CG model generation for thermal images

The CG model has a hierarchical structure of 34 joints (Figs. 3 and 4) and is described with rotation angles for the corresponding joints. The 3D CG model corresponding to each key frame for a sign is made manually (Fig. 5). We store the feature vectors of the key frames and the corresponding 3D CG models as training data.

4. Animation

To animate a sign, key frames are selected according to the process mentioned in Section II-2, followed by

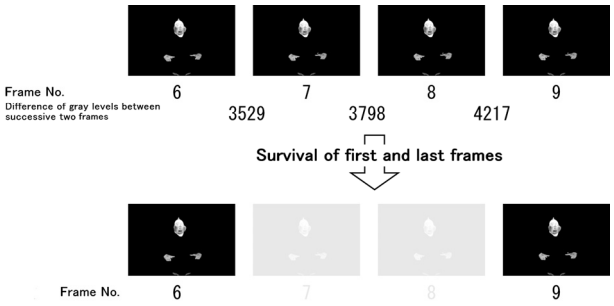

Fig. 2. Selection of key frames from candidates [5]

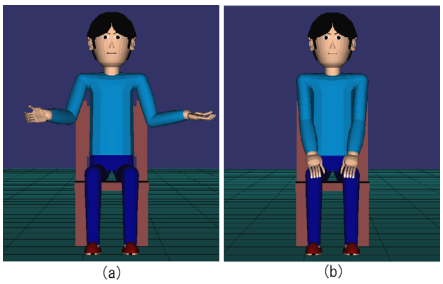

Fig. 3. Human model (a), standard posture (b) [5]

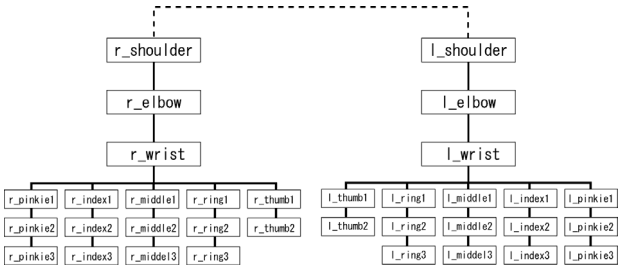

Fig. 4. Structure of the human model [5]

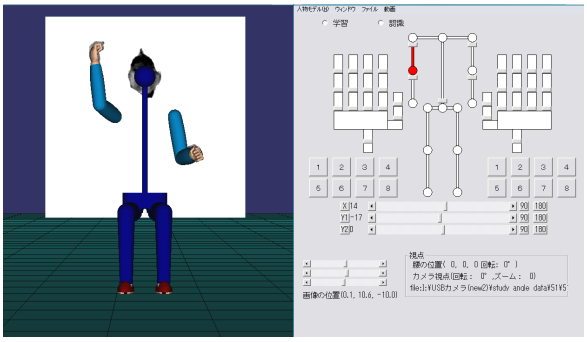

Fig. 5. Manual model fitting [5]

recognition using the "nearest neighbor" criterion for the feature vectors in the training data. When the user judges that the CG model corresponding to the key frame acquired from the input dynamic thermal image is not appropriate, the model can be replaced manually by a more appropriate model corresponding to the one made for the training key frames (Fig. 6) and/or the same model can be refined manually. We assume that we know the meaning of each sign for making the animation. The hands in all CG models are replaced by those made manually as training data. Then, the animation is generated with the CG models

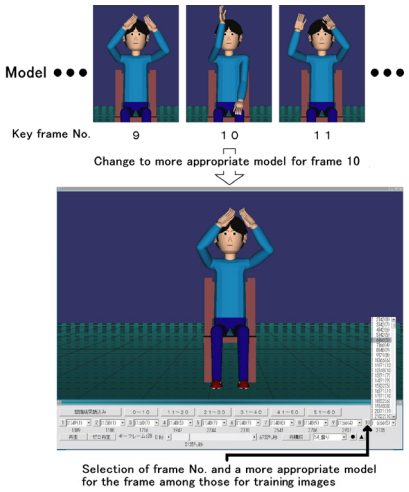

Fig. 6. Modification of model for key frame [5]

corresponding to the key frames by smoothing each rotation angle around the axis for each joint according to the appropriate time sequence.

III. EXPERIMENTS AND DISCUSSION

1. Conditions

In the experiments, a personal computer, Dell Dimension 8300 (CPU: Pentium IV 3.2 GHz, main memory: 2.0 GB, OS: Windows XP), was used, and Microsoft Visual C++ 6.0 was used for programming. The temperature range for detection was 302.3 to 306.7 K.

First, 70 signs were selected for the training data and for investigating the performance of our system according to the following necessary sign conditions: a noun that includes a large motion observable by the listener or the camera and does not include a meaning movement of the head, a movement in the direction of the listener or the camera, the crossing of hands, or the need for an initial pose. With the thermal video system for obtaining the thermal images and a CCD camera for obtaining the visible images, 70 signs by subject A, who was hearing-impaired and knew Japanese sign language, were recorded twice as an AVI file for each word. The first sign for the word was used to make the training data, and the second sign for the word was used for recognition and animation generation. In our experiments, the sign language was Japanese.

2. Results

For the evaluation of our system, the training image expressing the same meaning as that used for recognition was not used. Therefore, 69 signs which did not involve the sign having the same meaning as that for recognition were used to make the training images and test each sign. Three subjects (B, C, and D), who were hearing-impaired and knew Japanese sign language, evaluated the sign language animation produced by our system. The three subjects wrote down the meaning of each sign and rated it according to one of the following three categories—Level 1: instantly understandable, Level 2: understandable through checking the sign language gesture expressed by the visible ray image for training, Level 3: not understandable (no answer). The animations were produced under two conditions—Condition 1: with neither the modification nor the manual refinement of the models, Condition 2: with both the modification (Fig. 6) and the manual refinement of the models. Table 1 shows the correct answer rates at each level. As shown in Table 2, some misunderstandings occurred because subjects B and C did not often check the sign language gestures expressed by the visible ray image for training. The modification of the models, an example of which is shown in Fig. 6, and the manual refinement of the models improved the animation, especially at Level 2, as shown in Table 1. The average correct answer ratio of subjects B, C, and D, except for the misunderstandings caused by their lack of referring to the sign language gesture expressed by the visible ray image for training, was 51.3% and 94.3% under Conditions 1 and 2, respectively. The time required was approximately 20 minutes to make an animation for the sign language gesture using the training image, whereas it took approximately 25 seconds and 15 minutes to make an animation for the sign language gesture using the test image under Conditions 1 and 2, respectively.

Table 1. Number of correct answers at each level

		Level 1	Level 2	Total
Subject B	Condition 1	27/37 (73.0%)	3/7 (42.9%)	30/70 (42.9%)
	Condition 2	52/57 (91.2%)	6/6 (100%)	58/70 (82.9%)
Subject C	Condition 1	24/33 (72.7%)	10/26 (38.5%)	34/70 (48.6%)
	Condition 2	51/59 (86.4%)	8/10 (80.0%)	59/70 (84.3%)
Subject D	Condition 1	22/26 (84.6%)	15/18 (83.3%)	37/70 (52.9%)
	Condition 2	54/55 (98.2%)	10/13 (76.9%)	64/70 (91.4%)

Table 2. Number of misunderstandings

	Number of incorrect answers	Reference shortage			Poor CG		
		Level 1	Level 2	Level 3	Level 1	Level 2	Level 3
Subject B	12	2	0	6	3	0	1
Subject C	11	6	1	0	2	1	1
Subject D	6	0	2	1	1	1	1

3. Estimation of the position errors in CG space

The accuracy rate under Condition 2 was much higher than that under Condition 1, whereas the workload to generate an animation under Condition 2 was larger than that under Condition 1. To find some strategies for improving the accuracy rate under Condition 1, the position errors in several sign languages gestures using only the movement of one hand (the right hand) were evaluated in CG space. We investigated the causes of the poor animation under Condition 1 when Condition 2 provided a correctly understandable animation but Condition 1 did not. To find the cause of the position errors, we attached markers to the hand. We compared each position of the markers attached on a subject's hand in physical space with the corresponding position on the model in CG space. For simplicity, we selected animations in which only the right hand moved while performing the sign language gesture.

A. Inspection Method

We used the motion capture system named Radish (Library) for measuring the position of the markers attached on the subject. We attached 15 markers each on subject A and the model in CG space (Fig. 7). Then the model was manually refined to fit the human posture. The markers on the model were set as accurately as possible on the positions in CG space so that they had the same positions as those of subject A in physical space.

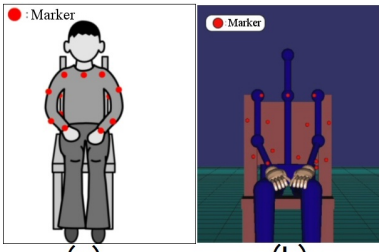

Fig. 7. (a) Subject with attached markers, and (b) model with attached markers

B. Results

Fig. 8 shows the position errors of the markers attached to the model in CG space. When an animation under Condition 1 was correctly understood, the position errors were, on the whole, smaller than those in the case of misunderstanding the animation. However, even when the position error was small, the animation was sometimes misunderstood. This phenomenon was primarily caused by the unwanted movement of the other hand (left hand). However, another phenomenon

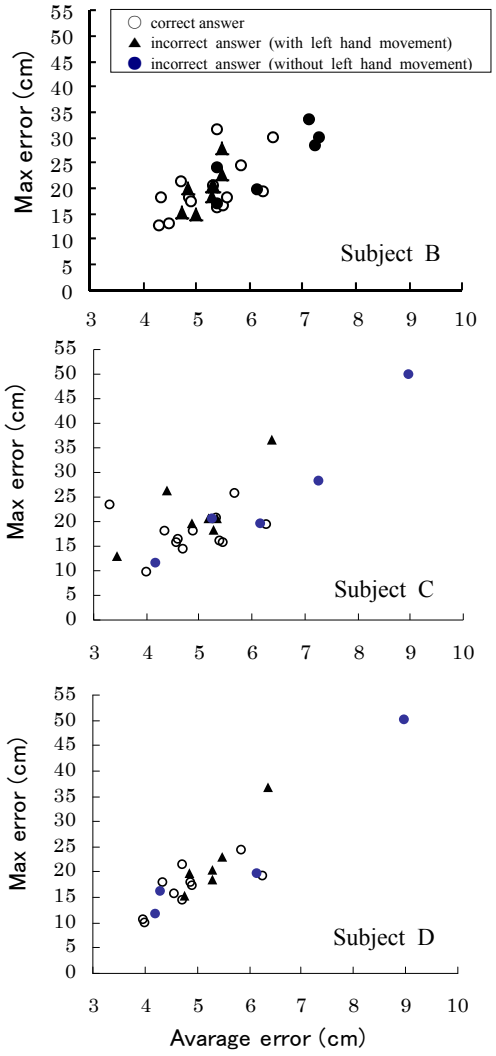

Fig. 8. Position errors of markers on subjects B, C, and D

causing the poor animation was found, as discussed in Section III-4.

4. Discussion

The sign language animation for "Meiji Era" generated under Condition 1 was compared with that generated under Condition 2. Though the difference between the animations generated under Conditions 1 and 2 was relatively small with regard to the hand orbit, the rotation of the elbow under Condition 1 was remarkably different from that under Condition 2 at time instances important for recognizing the animation (Fig. 9). In the case of the sign language animation for "younger brother", the same phenomenon was observed (Fig. 10). Therefore, we thought that a technique to correct the direction of the arm including the elbow was necessary. We used the angular information and the vector expressing the hand direction for the key frame

selected in the training images to produce a fuzzy system to improve the hand orientation. When fitting a model, as described in Section II-3, the hand direction and the center of gravity coordinates of the three markers (A, B, and C) in the 3D CG model corresponding to the key frame for the training image were recorded. A hand direction is expressed as the vector from marker A to marker B (hereinafter called vector B) and the vector from marker A to marker C (hereinafter called vector C) (Fig. 11).

Using cluster analysis, 487 key frames for the training images of sign language gestures in which only one hand moved were investigated. For this analysis, Ward's method was applied to 18 elements composed of the angles of the shoulder, elbow and wrist, vector B, vector C, the center of gravity coordinates of markers A, B, and C. The number of groups divided by the cluster analysis was used as the number of rules of the fuzzy algorithm. This value was experimentally decided as 24.

The fuzzy rules are described in (1). The fuzzy algorithm is described in (2) to (5).

$$\left.\begin{aligned}&\text{IF } x_1 \text{ is } A_{i1} \text{ and } x_2 \text{ is } A_{i2}\\&\text{and } x_3 \text{ is } A_{i3} \text{ and } x_4 \text{ is } A_{i4}\\&\text{and } x_5 \text{ is } A_{i5} \text{ and } x_6 \text{ is } A_{i6}\\&\text{and } x_7 \text{ is } A_{i7} \text{ and } x_8 \text{ is } A_{i8}\\&\text{and } x_9 \text{ is } A_{i9}\\&\text{THEN } y_1 \text{ is } B_{i1} \text{ and } y_2 \text{ is } B_{i2}\\&\quad\text{and } y_3 \text{ is } B_{i3} \text{ and } y_4 \text{ is } B_{i4}\\&\quad\text{and } y_5 \text{ is } B_{i5} \text{ and } y_6 \text{ is } B_{i6}\\&\quad\text{and } y_7 \text{ is } B_{i7} \text{ and } y_8 \text{ is } B_{i8}\\&\quad\text{and } y_9 \text{ is } B_{i9}\end{aligned}\right\} \quad (1)$$

$$(i = 1,2,\ldots,24)$$

$$\mu_{B_{ij}^*}(y_j) = w_i \ \mu_{B_{ij}}(y_j) \quad (2)$$
$$(i = 1,2,\ldots,24,\ j=1,2,\ldots,9)$$

$$w_i = \min_j \mu_{A_{ij}}(x_j^*) \quad (3)$$
$$(i = 1,2,\ldots,24,\ j=1,2,\ldots,9)$$

$$B_j^0 = \bigcup_{i=1}^{24} B_{ij}^* \quad (j=1,2,\ldots,9) \quad (4)$$

$$y_j^* = \frac{\int \mu_{B_j^0}(y_j)\, y_j\, dy_j}{\int \mu_{B_j^0}(y_j)\, dy_j} \quad (j=1,2,\ldots,9) \quad (5)$$

where $x_1, x_2, \cdots, x_9$ denote the parameters of the antecedent; $y_1, y_2, \cdots, y_9$ denote the parameters of the consequent; $A_{i1}, A_{i2}, \cdots, A_{i9}, B_{i1}, B_{i2}, \cdots B_{i9}$ denote the corresponding fuzzy labels; w_i denotes the fitness value of the i th rule to the input of $x^*_1, x^*_2, \cdots, x^*_9$;

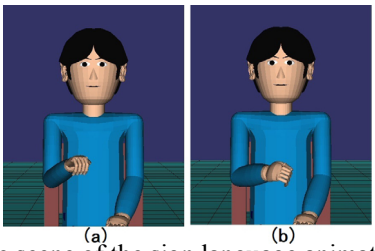

Fig. 9. One scene of the sign language animation for the Meiji Era under (a) Condition 1, (b) Condition 2

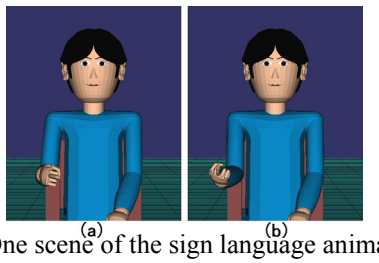

Fig.10. One scene of the sign language animation for "younger brother" under (a) Condition 1, (b) Condition 2

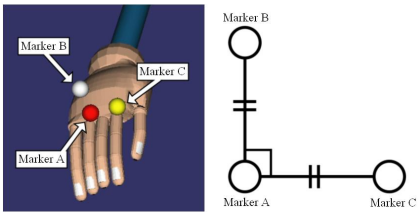

Fig. 11. Hand model with attached markers

and y^*_i denotes the output for y_i. In this study, the coordinates of the center of gravity for markers A, B, and C attached on the model were assigned for (x_1, x_2, x_3); the elements of vectors B and C were assigned for (x_4, x_5, x_6) and (x_7, x_8, x_9), respectively; and the rotation angles on the shoulder, the elbow, and the wrist in the model were assigned for $y_1, y_2, \cdots, y_9$, respectively. Each norm of vectors B and C had the value of 1. Fig. 12 shows a schematic diagram of the membership functions used for both the antecedent and the consequent, where C_i denotes a fuzzy label of the antecedent or consequent, for example, A_{11}, whereas Z corresponds to a variable of the antecedent or consequent, for example, x_1. Each membership function had a pentagonal shape and a value of 1 at the corresponding measured mean value and a value of 0.5 at the corresponding measured minimum and maximum values. We used a fuzzy

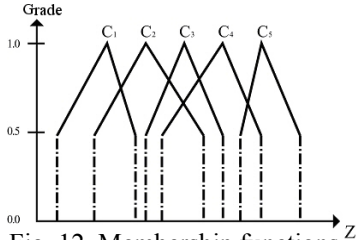

Fig. 12. Membership functions

algorithm having 24 rules, 9 antecedent valuables, and 9 consequent variables. The rotation angles of the shoulder, elbow, and wrist are inferred using this fuzzy algorithm.

Next, we propose a method to improve the rotation angle set to the joint of the elbow by decreasing the difference of the hand direction for the test image to that for the training image, and the difference of the center of gravity coordinates of markers A, B, and C of the models before and after using the fuzzy algorithm. The rotation angle of the elbow joint was acquired by simulated annealing (SA). The amount of the modification in the rotation angle of each rotation axis of the elbow joint is $n_q(q=1,2,3)$. The value of a uniform random number generated with the value of $-T$ to T is $n_q(T)$. We used the experimentally decided (6), shown below, as the cooling schedule.

$$T(s)=\frac{T_0}{\log\left(2.0+\frac{s}{20.0}\right)}, \qquad (6)$$

where T_0 is the initial temperature and s is the number of steps of the iteration operation. As initial values for (6), the angles of the model obtained from the recognition result were used. The total error to be reduced is defined as the sum of the errors in the hand direction to that for the training image, and the error in the center of gravity of the three markers attached on the hand with respect to that obtained from the recognition result. When the total error obtained at an iteration step is smaller than the minimum value obtained before the step, the rotation angles of each rotation axis of the elbow join are updated to those at that step. The iteration operation is terminated at a previously decided step. Condition 3 denotes the condition for the sign language animation generated under Condition 1 followed by improvement using the fuzzy algorithm and SA. The sign language animations generated under Condition 3 are more natural than those under Condition 1 (Figs. 13 and 14).

IV. CONCLUSION

We developed an enhanced system for generating sign language animation using skin region detection on a thermal infrared image. In this system, a 3D CG model corresponding to a characteristic posture used in sign language is generated automatically by pattern recognition of the thermal image, whereas the human hand in the CG model is created manually. In experiments, three people experienced in using sign language recognized with good accuracy the Japanese sign language gestures of 70 words expressed as animations. Then, using the fuzzy algorithm and simulated annealing, we improved the system to correct the position and the direction of the hand of the automatically generated model.

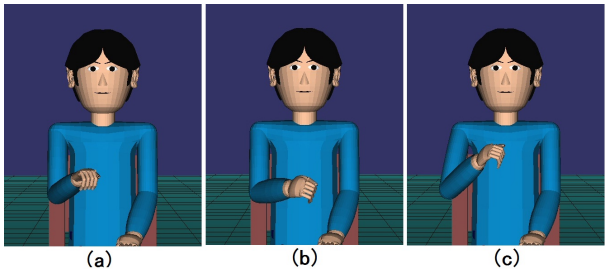

Fig. 13. One scene of the animation of "Meiji Era"; (a) Condition 1, (b) Condition 2, and (c) Condition 3

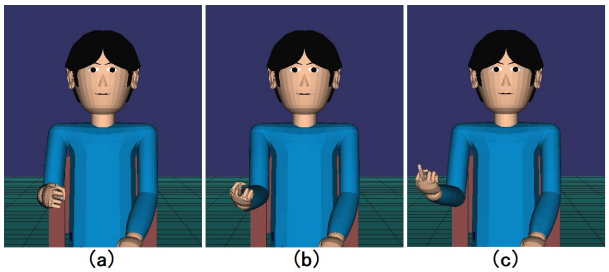

Fig. 14. One scene of the animation of "younger brother"; (a) Condition 1, (b) Condition 2, and (c) Condition 3

Acknowledgment

We would like to thank all the subjects who cooperated with us in the experiments.

REFERENCES

[1] Sagawa H, Sakou H, Oohira E, et al (1994), Sign-Language Recognition Method Using Compressed Continuous DP Matching (in Japanese). IEICE Trans. Inf. & Syst 77(4):753-763

[2] Watanabe K, Iwai Y, Yagi Y, et al (1998), Manual Alphabet Recognition by Using Colored Gloves (in Japanese). IEICE Trans. Inf. & Syst 80(10):2713-2722

[3] Igi S, Watanatabe R, and Lu S (2001), Synthesis and Editing Tool for Japanese Sign Language Animation (in Jpanese). IEICE Trans84 (6):987-995

[4] Kurokawa T (2004), Representation of sign animation for Japanese-into-Japanese sign language translation (in Japanese). Proc. of 32nd Symp. of visualization 24(1):273-276

[5] Asada T, Yoshitomi Y and Hayashi R (2008), A Human-machine Cooperative System for Generating Sign Language Animation Using Thermal Image. Journal of Artificial Life and Robotics 13:36-40