A human-machine cooperative system for generating sign language animation using thermal image

T. Asada[1], Y. Yoshitomi[2], and R.Hayashi[3]

1: Dept. of Environmental Information Graduate School of Human Environment Science Kyoto Prefectural University, Shimogamo Sakyo-ku Kyoto 606-8522 JAPAN, E-mail: t_asada@aoi.kpu.ac.jp
2: Dept. of Environmental Information Faculty of Human Environment Kyoto Prefectural University Shimogamo Sakyo-ku Kyoto 606-8522 JAPAN, E-mail: yoshitomi@kpu.ac.jp
3: Kyoto Shinkin Bank, 7 tachiuri-higashi-cho Shimogyo-ku Kyoto 600-8005 JAPAN

Abstract: We have started a new approach aimed at sign-language animation by skin region detection on an infrared-ray image. For making several kinds of animations expressing personality and/or emotion appropriately, the conventional systems need much more manual operations. We think that to manually refine the animation made automatically with dynamic image of real motion is one of the most promising ways for realizing less workload. In the method, a 3D CG model corresponding to a characteristic posture in sign language is made automatically by pattern recognition for thermal image, whereas the hand in CG model is made manually. If necessary, the model can be replaced manually by more appropriate model corresponding to one of those made for training key frames and/or the model can be refined manually. In our experiments, the person who can use sign language recognized Japanese sign language of 71 words expressed as the animation with 87.6% accuracy, and they also recognized the sign language animation representing each of 3 emotions (neutral, happy and angry) with 88.9% accuracy.

Keywords: Sign language, Thermal image, Computer graphics, Model fitting

I. INTRODUCTION

Sign languages are usual for hearing impaired people to communicate together. However, it is inconvenient for them to communicate with other people through a sign-language interpreter, because the interpreter is very few. Therefore, there is a strong need for an automatic sign-language translation system. As well as a function of sign-language recognition, an animation function is also expected to be involved in the system. Several systems for sign language animation have been studied [1,2]. For making several kinds of animations expressing personality and/or emotion appropriately, the conventional systems need much more manual operations. We think that to manually refine the animation made automatically through a dynamic image of real motion is one of the most promising approaches to realizing less workload. We have proposed a method for expressing a human motion as CG animation with human thermal image taken under no special restrictions on human [3]. In this article, our approach to sign-language animation system with thermal image is introduced and then we discuss the performance of the system.

II. ANIMATION GENERATION METHOD

An example of an input image is shown in Fig.1. The flowchart for generating sign-language animation is shown in Fig.2. In this section the important parts in the total procedure are described.

Fig.1 Input image

1. Thermal-image generation

Under the condition that emissivity is set as 1, thermal images extracting regions of human skin are produced by a thermal video system (Nikon LARD-3ASH) with infrared rays, and the image for each word is recorded as an AVI file. The detected temperature range is 302.3 to 306.7 K. The input-images in a computer have a spatial resolution of 720×480 pixels and a gray level of 256.

2. Selection of key flames from training images

To find the beginning of sign language on an AVI file easily, a subject wearing a jacket with long sleeves sits on a seat and puts its left and right hands on the corresponding knees before and after performing sign language. After segmentation of input image, the area having the value of 1 on a specific region of the subject's knee-parts is measured. Then, in frame B,

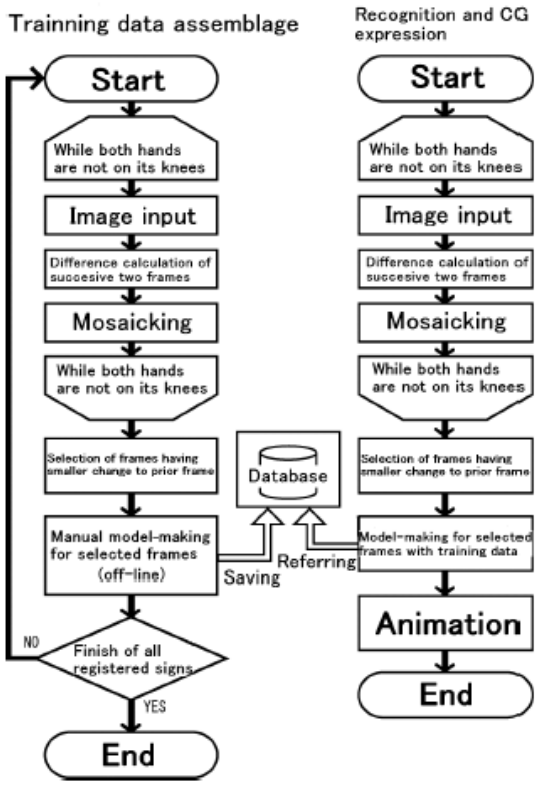


Fig.2 Flowchart for generating sign-language animation

where this area first has a value less than the threshold α, which has been experimentally decided beforehand, the processing of inputting the static images is started. Frame B is judged as the first frame of sign language. After processing several frames, at frame F, where the area on the specified region of the subject's knees again has a value larger than the threshold, the processing of inputting the static images is terminated. Frame F is judged to be the final frame of the sign language.

After erasing some kinds of noise on each image-frame in an AVI file, the sum of gray level difference between the present and previous frames is calculated for all pixels. On the assumption that the characteristic postures in sign language corresponds to the frames showing slower movement, the frame having the small difference for the previous frame is acquired as an suitable frame (hereinafter referred as a key frame) for making a 3D CG model. The sum of gray level difference for the previous frame is used for picking up several key frames. We first select the top β% frames on the inverse value of the sum. However, as a result, some successive frames which are very similar can be selected. For surviving as key frames the representative frames among the selected frames, we remove the frames except the first and last frames when more than three successive frames for one value of β are selected in using the above criterion on the difference for the previous frame (Fig.3). As the values of β, we use 25, 50, and 75. All frames survived with each of three values for a sign are exploited as the key frames. The feature vector used for pattern recognition is made from a mosaic image after smoothing (Fig.4).

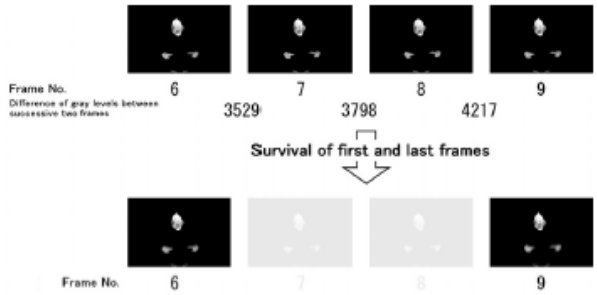


Fig.3 Selection of key frames from candidates

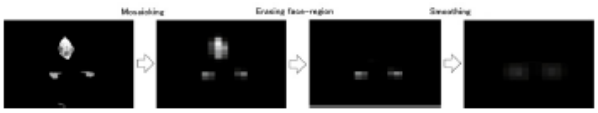


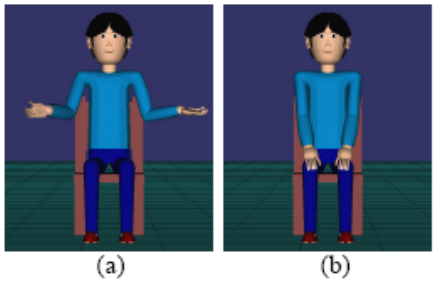
Fig.4 Process for getting mosaic image after smoothing

3. CG models generation for thermal images

The CG model has the hierarchical structure of 34 joints(Figs. 5 and 6). The CG model is described with rotation angles for their corresponding joints. The 3D CG model corresponding to each key frame in a sign is made manually (Fig.7). We store as training data the feature vectors of key frames and those corresponding 3D CG models in a computer.

4. Animation

For a sign for making an animation, the key frames are also selected according to the process mentioned in the section 2, followed by the recognition with the nearest neighbor criterion to the feature vectors in 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 may be replaced manually by more appropriate model corresponding to one of those made for training key



(a) (b)

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

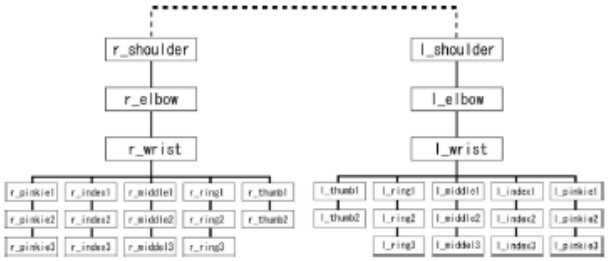


Fig.6 Structure of human model

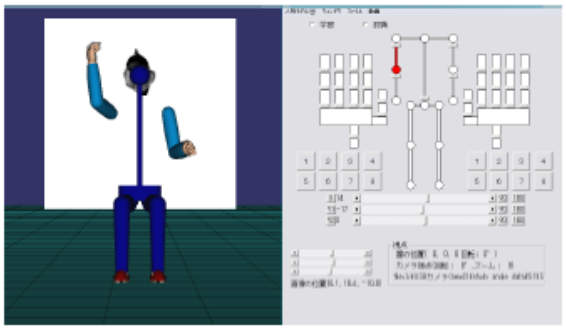


Fig.7 Manual model fitting

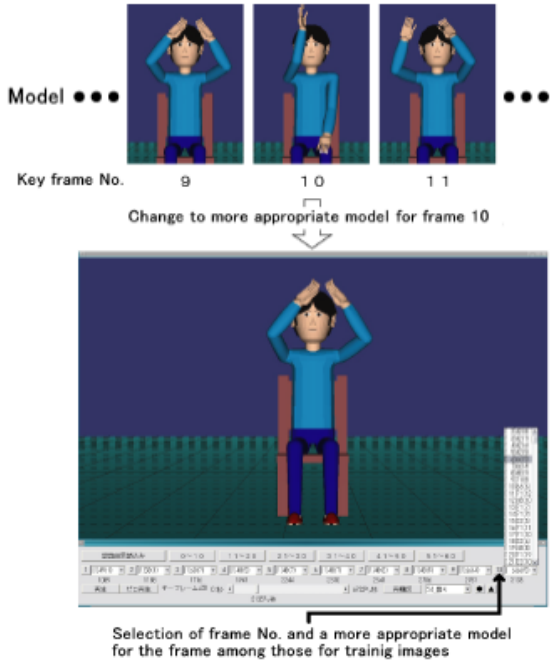


Fig.8 Modification of model for key frame

frames (Fig.8) and/or the model may be refined manually. We assume that we know the meaning of a sign for making an animation. The hand in CG model is replaced to that made manually as training data. Then the animation is generated with the CG models corresponding to key frames through smoothing each rotation angle around axis for each joint in the direction of time.

III. EXPERIMENTS AND DISCUSSIONS

In this experiment, a personal computer, DELL Dimension 8300 (CPU : Pentium Ⅳ 3.2GHz, main memory : 2.0 GB, OS : Windows XP) , was used. For programming, Microsoft Visual C++ 6.0 was used.

At first, 71 kinds of sign were selected for preparing training data and investigating the performance of our system, according to the following necessary conditions; noun which does NOT include a meaning movement of the head, a movement in the direction of the listener or the camera, a large motion observable by the listener or the camera, the crossing of hands, or the need for an initial pose. With a thermal video system for thermal image and a CCD camera for visible image, 71 kinds of sign by subject A, who could use sign language but was not a hearing impaired person, were saved twice as an AVI file for a word. The first sign for a word is used for making training data, while the second sign for the word is used for recognition and animation generation.

1. Word recognition

For the evaluation of our system, the training image expressing the same meaning as that for recognition was not used. Therefore, the 70 words which do not involve the word for recognition were used for the test of each sign. The subject B, who was a hearing impaired person and could use sign language, evaluated sign-language animation produced by our system. The subject B wrote down the meaning of each sign on the following three categories; level 1: instantly understandable, understandable through checking sign-languages expressed by visible-ray image for training, level 3: not understandable (no answer). The animations were produced on the two conditions; Condition 1: without the modification of models (Fig.8), Condition 2: with it. Table 1 shows the correct answer rates at each level. There were some misunderstandings because subject B did not check sign-languages expressed by visible-ray image for training (Table 2). The modification of models as shown in Fig.8 had an effect to improve the animation, especially at the Level 2 (Table 3). I took about 20 minutes to make an animation for a sign language using training image, while it took about 30 second and 10 minute to make an animation for a sign language using test image on the Conditions 1 and 2 respectively.

Table 1 Number of correct answers at each level

	Level 1	Level 2	Total
Condition 1	41/47 (87.2%)	5/10 (50.0%)	46/57 (80.7%)
Condition 2	47/54 (87.0%)	6/10 (60.0%)	53/64 (82.8%)

Table 2 Number of misunderstanding

	Level 1	Level 2	Level 3
Condition 1	2	1	1
Condition 2	5	2	4

Table 3 Correct answer rate except misunderstanding by subject B

	Level 1	Level 2	Total
Condition 1	91.1%	53.6%	68.6%
Condition 2	93.9%	75.0%	86.7%

2. Emotion recognition

The Japanese sign-languages of RAIN, MARRIAGE, BYCYCLE, SIGN LANGUAGE, CONSULTATION, FATHER, WIND, PROBLEM, which were expressed by the subject B, were recorded as AVI files of thermal image. The above 8 words were selected by the subject B among 71 words used for word recognition because they could be easily expressed with emotions. The selected emotions were NEUTRAL, HAPPY, and ANGRY. The three subjects (C ,D, E) , who were hearing impaired people and could use sign language, wrote down the meaning of each of 24(=8×3) signs on the three categories mentioned in the subsection of Word recognition. The animations were produced on Condition 3 where the animation was manually improved after the modification of models (Fig.8), in addition to Condition 1. We used six AVI files categorized by 3 emotions and 2 conditions. Each AVI file had 8 signs expressing 8 words. The modification of models as shown in Fig.8 and additional manual operation had an effect to improve the animation (Table 4). The two of three subjects (D, E) perfectly recognized emotions expressed on the 6 animations (Table 5). The subject C was poor at recognizing not only the meaning but also the emotion of signs expressed by animation (Tables 4 and 5). It took about 25 second and 15 minute to make an animation for a sign language using test image on the Conditions 1 and 3 respectively.

VI. CONCLUSION

We have developed a system for sign-language animation using skin region detection on an infrared-ray image. In the method, a 3D CG model corresponding to a characteristic posture in sign language is made automatically by pattern recognition for thermal image, whereas the hand in CG model is made manually. In our experiments, the person who can use sign language recognized Japanese sign language of 71 words expressed as the animation with good accuracy, and they also recognized the sign language animation representing each of 3 emotions (neutral, happy and angry) with good accuracy.

Table 4 Correct answer rate except misunderstanding by subject

		Level 1	Level 2	Total
Subject C	Condition 1	3/3 (100%)	5/7 (71.4%)	8/24 (33.3%)
	Condition 3	18/18 (100%)	3/3 (100%)	21/21 (100%)
Subject D	Condition 1	12/12 (100%)	4/4 (100%)	16/24 (66.7%)
	Condition 3	23/23 (100%)	-	23/23 (100%)
Subject E	Condition 1	16/20 (80.0%)	0/2 (0%)	16/24 (66.7%)
	Condition 3	23/23 (100%)	-	23/23 (100%)

Table 5 Correct answer rate of emotion recognition

		Emotion			Correct ratio (%)
		Neutral	Happy	Angry	
Subject C	Condition 1	○	×	×	33.3%
	Condition 3	×	○	○	66.7%
Subject D	Condition 1	○	○	○	100%
	Condition 3	○	○	○	100%
Subject E	Condition 1	○	○	○	100%
	Condition 3	○	○	○	100%

○=correct recognition, ×=wrong recognition

REFERENCES

[1] IGI S, Watanabe R, and Lu S (2001), Synthesis and editing tool for Japanese sign language animation, Trans. of IEICE, J84-D- I (6):987-995.

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

[3] Asada T, Uragami K, Ikeno Y, Tanijiri T, and Yoshitomi Y (2004), A method for synthesizing computer graphics animation by transforming feature vector of posture on thermal image into 3-dimensional model, Proceedings of 13th IEEE International Workshop on Robot and Human Interactive Communication, pp.325-330.