MVA2007 IAPR Conference on Machine Vision Applications, May 16-18, 2007, Tokyo, JAPAN

Gesture Recognition using Temporal Templates with disparity information

Kazunori Onoguchi and Masaaki Sato
Hirosaki University
Faculty of Science and Technology
3, Bunkyo-cho, Hirosaki, 036-8561, Japan
onoguchi@eit.hirosaki-u.ac.jp

Abstract

This paper presents a gesture recognition method extending Temporal Templates so that they can contain not only vertical and horizontal motion but also depth information obtained from a binocular stereopsis. The proposed method can discriminate gestures with depth motion. At first, a disparity image generated from stereo images is divided into several disparity levels and in each disparity level, a grayscale feature image (Temporal Template) is created by assigning the intensity according to the frame number to the area where motion has been detected. Next, a gesture model is generated from learning feature images acquired in each disparity level by SVM. A gesture is recognized by checking feature images generated from input stereo images against SVM model. Experimental results have shown the effectiveness of the proposed method for recognizing gestures with depth motion.

1. Introduction

A gesture plays an important role in a communication between human beings. In controlling a robot or a computer, it is important to realize a human computer interaction using a gesture because it can give a human's intention more than a verbal communication under certain circumstances. Many contactless methods recognizing gestures by an image processing have been proposed because sensors attached on a user body, such as a glove type sensor or a magnetic sensor, restrict a user. They are divided into two types.

One is the method using a 3D body model[1][2]. This method can be applied in many ways because a position and a pose of each body part can be obtained. However, it is difficult to fit a 3D body model to an image stably if input images contain a lot of noise or a part of human body is occluded in images.

The other is the method using an appearance-based model. This method does not need to estimate 3D parameters and its computational cost is low. Therefore, a lot of methods have been proposed, such as based on DP matching[3], based on Hidden Markov Model[4] and so on. However, it is difficult to discriminate gestures with different motion toward a depth direction because these methods recognize a gesture in image sequences obtained from a single viewpoint. Though a method using image sequences obtained from multiple viewpoints have also been proposed[5], they need large-scale observation systems arranging cameras around a user. In order to construct an interaction system available easily, it is desirable to recognize a gesture by using only cameras located on a computer or a robot.

This paper presents an appearance-based method which can discriminate gestures with depth motion by using depth information obtained from a conventional binocular stereopsis. The proposed method extends Temporal Templates[5] so that they can contain not only vertical and horizontal motion but also depth information. At first, a disparity image generated from stereo images is divided into several disparity levels and in each disparity level, a grayscale feature image (Temporal Template) is created by assigning the intensity according to the frame number to the area where motion has been detected. Next, a gesture model is generated from learning feature images acquired in each disparity level by SVM. A gesture is recognized by checking feature images generated from input stereo images against SVM model.

In Sect.2, an outline of Temporal Templates is described. In Sect.3, Temporal Templates with disparity information is described in detail. In Sect.4, a gesture recognition method is described. In Sect.5, experimental results including gestures with different depth motion are discussed. Conclusions are presented in Sect.6.

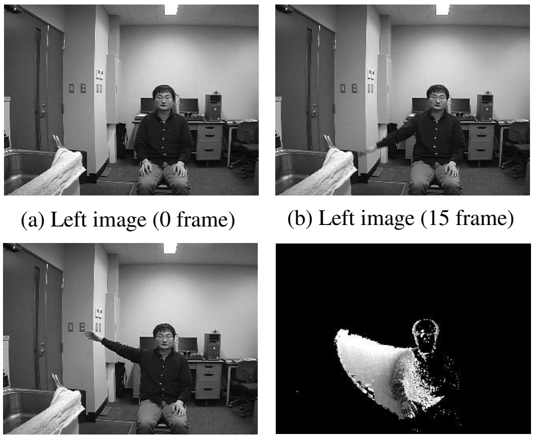
2 Temporal Templates

Temporal Templates represent a history of motion by a gray-scale image[5]. They are created by assigning the intensity according to the frame number to the area where motion is detected by a frame differential method or a background subtraction method. Figure 1(d) shows the example of Temporal Template created from a gesture raising a right arm horizontally(Fig. 1(a)(b)(c)). In this image, motion areas detected near the time starting a gesture are indicated as darker areas and ones detected near the time finishing a gesture are indicated as brighter ones.

Because Temporal Template represents the feature of the gesture by one grey-scale image, it is easy to create gesture models and check input image sequences against gesture models. Therefore, it is well suited for constructing a human computer interaction system. However, in the case of gestures which change toward a depth direction, previous motion areas could disappear by overwriting and it is difficult to identify gestures which have similar motions at the time finishing gestures. Figures 6(a) and 6(b) show the examples of gestures whose Temporal Templates are similar. Figure 6(a) shows the gesture named "Comeon" which raises a right arm from the front and Fig. 6(b) shows the gesture named "Stop" which pushes up a hand in front. Temporal Templates of these gestures are very similar because these gestures have the similar vertical and horizontal motion. Bobick[5] avoids this problem by creating Temporal Templates from images taken at several cameras whose view directions are different. However, it is difficult for this method to construct a HCI system by only cameras installed on a robot or a computer because this method needs cameras located away each other.

3 Temporal Templates with disparity information

This paper presents a gesture recognition method which can discriminate the difference of motion toward the depth direction by adding Temporal Templates to the disparity information calculated by a stereopsis. The conventional methods using Temporal Templates create a gesture model from images taken at several cameras located around a user, but the proposed method uses an ordinary binocular stereo whose camera distance is several ten centimeters. Therefore, it can recognize a gesture by using two cameras installed on a robot or a computer. The proposed method assumes the application that a human instructs a robot or a PC interactively. Thus, it is assumed that a user stays in front of stereo cameras, no moving object without a user exists in stereo images and a gesture is composed of continuous motions.


(a) Left image (0 frame) (b) Left image (15 frame)

(c) Left image (30 frame) (d) Temporal Template

Figure 1. Example of Temporal Template

Figure 2 shows a block diagram of the proposed method. At first, moving areas in a left image are detected by the frame differential method and the binarization. Next, a mask image representing moving edge areas is created by AND operation between a frame differential image and a left edge image detected by a sobel operator. Figure 3 shows the example of the gesture named "cheer" which raises both hands above a head. Figure 3(g) shows the frame differential image between Fig. 3(b) and Fig. 3(c), Fig. 3(h) shows the edge image of Fig. 3(c) and Fig. 3(i) shows the mask image.

A disparity image is created by finding corresponding points to only pixels in a mask image. Figure 3(j) shows the disparity image generated from Fig. 3(c). Until a gesture finishes, disparity images are created in each stereo image. It is judged that a gesture finishes when moving areas in a frame differential image becomes smaller than a predetermined threshold.

In order to detect a gesture area which a gesture has occupied in left images, an integrated image is created by accumulating frame differential images in order until a gesture finishes. After thickening and shrinking an integrated image, the largest areas in the image is detected as a gesture area(Fig. 3(k)). A rectangle area containing a circumscribed rectangle of a gesture area is extracted from each disparity image. An aspect ratio of the rectangle is adjusted to the same ratio as an input image. In order to normalize the location and the size of a user in stereo images, a disparity image inside the extracted rectangle is enlarged to the size of an input image. Then, an average disparity D_m is

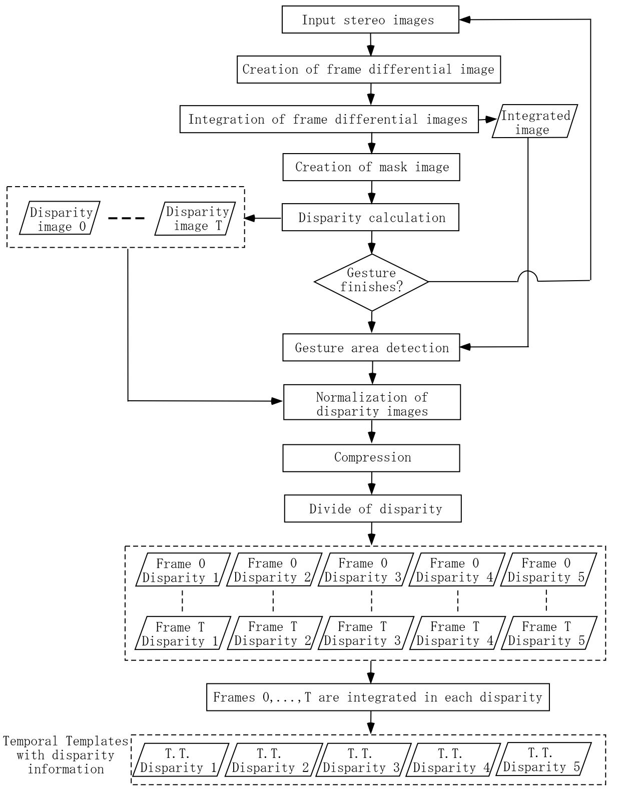


Figure 2. Flow of the proposed method

calculated in each enlarged disparity image and the disparity value in a range from $D_m - dd$ to $D_m + dd$ is normalized to the value in a range from 1 to 255. dd is a displacement value of a disparity determined experimentally. If a distance from stereo cameras to a user scarcely changes, dd can be fixed. Figure 3(l) shows a normalized disparity image.

Next, normalized disparity images are compressed to images whose size are 20 × 15. Normalized disparity images are divided into 16 × 16 grid regions and in each grid region, an average intensity of pixels whose disparities are more than 1 is calculated. This is used as an intensity of each pixel in a compressed image. This compression decreases both the influence of pixels whose disparities are calculated wrongly and the computation cost drastically. Then, each compressed image is divided into five images in accordance with the disparity level. A disparity is normalized to the value in a range from 1 to 255. Therefore, the binary image in the first disparity level is generated from binarizing a compressed image in a range from 1 to 50, the binary image in the second disparity level is generated from binarizing a compressed image in a range from 51 to 100, the binary image in the third disparity level is generated from binarizing a compressed image in a range from 101 to 150, the binary image in the fourth disparity level is generated from binarizing a compressed image in a range from 151 to 200 and the binary image in the fifth disparity level is generated from binarizing a compressed image in a range from 201 to 255. Figure 4 shows a compressed image and its binary images in each disparity level. In each disparity level, Temporal Templates are created by assigning the intensity according to the frame number to each binary image. As shown in Fig. 5, five images obtained from the above process are Temporal Templates with disparty information.

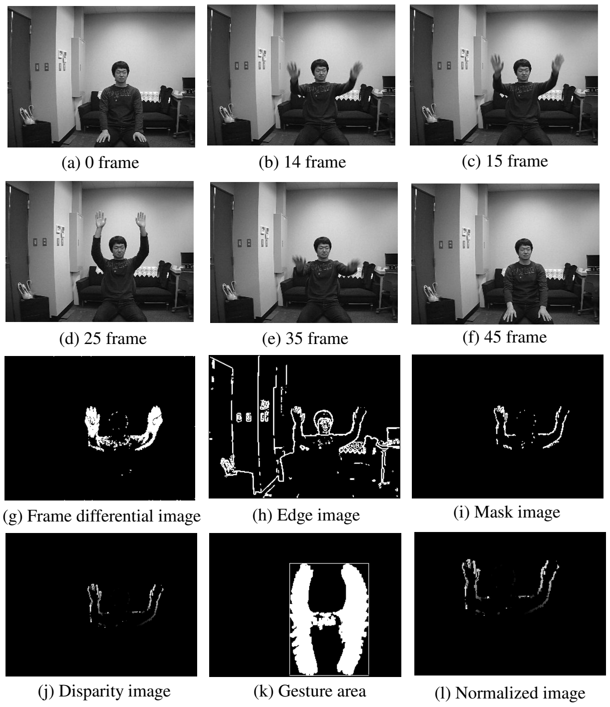

(a) 0 frame (b) 14 frame (c) 15 frame
(d) 25 frame (e) 35 frame (f) 45 frame
(g) Frame differential image (h) Edge image (i) Mask image
(j) Disparity image (k) Gesture area (l) Normalized image

Figure 3. Normalized disparity Image

4 Gesture discrimination

The proposed method uses the Support Vector Machine(SVM) to discriminate gestures. Temporal Templates in all disparity levels are treated as 1500 (20 × 15 × 5) dimensional feature vector and the SVM model are learned for each gesture. Temporal Templates with disparity information created from input stereo images are discriminated by the SVM and the gesture whose discrimination rate is maximum is selected as the recognition result.

5 Experiments

An experiment has been conducted to recognize twenty kinds of gestures whose contents are shown in

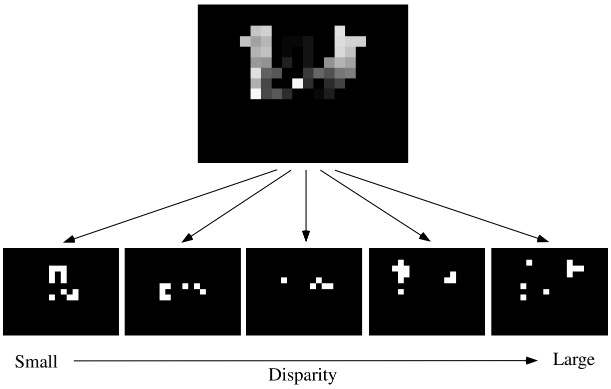


Figure 4. Divide of compressed image

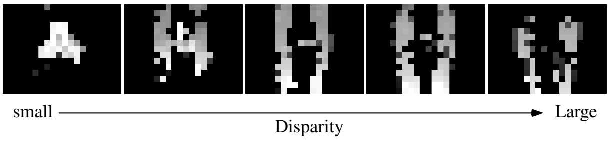


Figure 5. Temporal Template with disparity information

Table 1. Optical axes of two cameras are almost parallel each other. The distance between two cameras is 28 centimeters, the focal length of each camera is 8 millimeters and the image size is 320 × 240. It is difficult for the method using conventional Temporal Templates to discriminate these gestures because Temporal Templates shown in Fig. 6 (a)∼(c) and Fig. 6 (d)∼(f) are similar each other.

In order to evaluate recognizing performance for practical use, a subject's gestures were descriminated by SVM model created from images excluding a subject. In Table 2, experimental results conducted by four subjects are shown. Each subject performs each gesture five times. Each row in Table 2 shows the number of false recognition and the false recognition rate (error rate) in each gesture. False recognition rates of "Comeon" (Fig. 6(a)), "Stop" (Fig. 6(b)) and "Push" (Fig. 6(c)) which show similar Temporal Templates each other are less than 10%. And, false recognition rates of "Stop-w" (Fig. 6(d)), "Push-w" (Fig. 6(e)) and "No" (Fig. 6(f)) which also show similar Temporal Templates each other are less than 5%. These results show that the proposed method is effective to recognize gestures with depth motion. The average recognition rate in this experiment was 94.24%.

In Table 2, false recognition rates of "Out" and "Bad" are much higher than those of other gestures. Though "Out" is a gesture that pushes out a fist in front after raising it up to a shoulder, its three-dimensional trajectory is similar to that of "Push" or that of "Stop". Therefore, in order to recognize "Out" correctly, it is necessary to make the depth resolution higher. In the case of "Bad", the result of the subject 4 is extremely worse than those of other subjects because a gesture "Bad" of the subject 4 is quite different from those of other subjects. This false recognition can be avoided by using the SVM model created from sample gestures of more persons.

6 Conclusion

A method has been proposed which can easily discriminate gestures with depth motion by adding depth information to Temporal Templates. The future work is to improve the recognition rate by collecting gesture samples of more persons.

References

[1] M.Yamamoto and K.Koshikawa, "Human motion analysis based on a robot arm model," Proc. of CVPR'91, pp.664-665, 1991.

[2] N.Krahnstoever, "Automatic Acquisition and Initialization of Kinematic Models," Proc. of CVPR2001, 2001.

[3] S.Seki, K.Takahashi and R.Oka, "Gesture Recognition from Motion Images by Spotting Algorithm," Proc. of ACCV'93, pp.759-762, 1993.

[4] J.Yamato, J.Ohya and K.Ishii,"Recognizing Human Action in Time Sequential Images using Hidden Markov Models," Proc. of CVPR'92, pp.379-387, 1992.

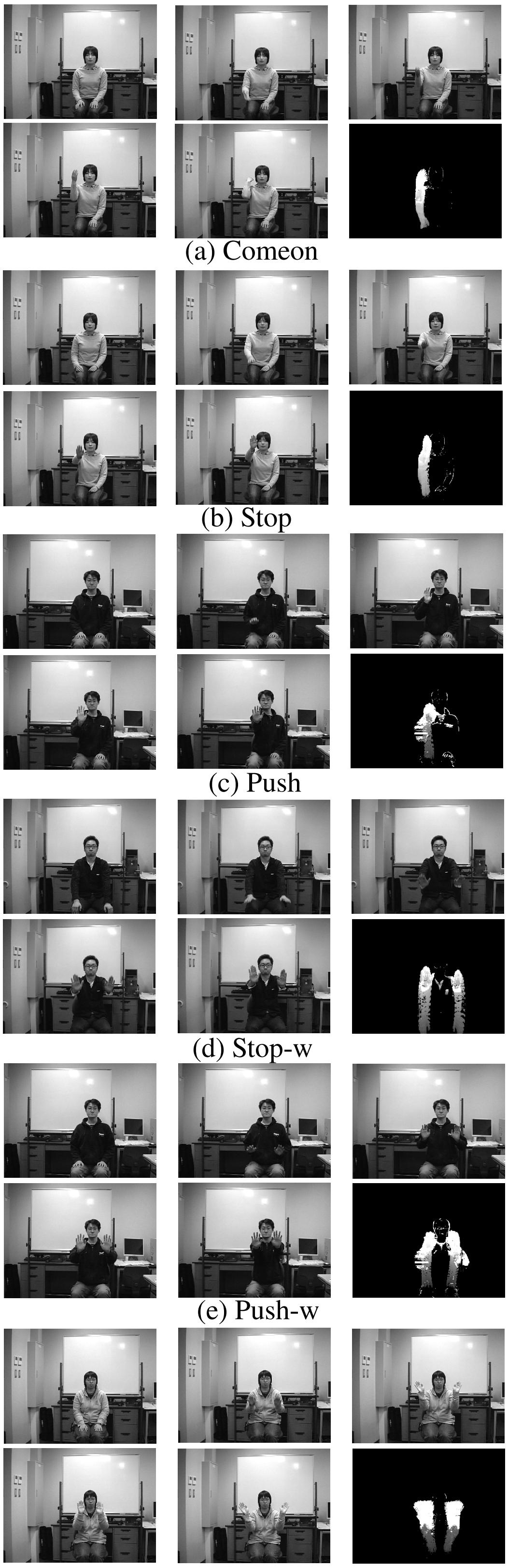
[5] A. F. Bobick and J. W. Davis, " The Recognition of Human Movement Using Temporal Templates, " IEEE Trans. on PAMI, vol.23, no.3, pp.257-267,2001.

Table 1. Gestures in experiment

Bad	Cross both arms before a chest
Bow	Bow
Bye	Raise a right arm to the side of a head and wave it
Cheer	Raise both hands above a head
Comeon	Raise a right arm from the front
Left	Raise a left arm horizontally
Right	Raise a right arm horizontally
Less	Open arms and join them before a body
Stop	Push up a hand in front
Stop-w	Push up both hands in front
More	Raise both arms before a body and open them
Moreleft	Push out a left hand and swing it to the left
Moreright	Push out a right hand and swing it to the right
No	Raise both hands and wave them small
Ok	Raise both hands from the side and join them above a head
Out	Raise a right fist up to a shoulder and push out it in front
Push	Push out a hand in front after raising it
Push-w	Push out both hands in front after raising them
Salute	Salute
Safe	Open arms from below to the sides

Table 2. Recognition result

	Subject 1	Subject 2	Subject 3	Subject 4	Error rate
Bad	1	0	1	3	25%
Bow	0	0	0	0	0%
Bye	0	0	2	0	0%
Cheer	0	0	0	0	0%
Comeon	0	0	0	0	0%
Left	0	1	0	0	5%
Right	1	0	0	0	5%
Less	0	0	0	0	0%
Stop	0	0	0	2	10%
Stop-w	0	0	0	0	0%
More	0	0	0	0	0%
Moreleft	0	0	0	0	0%
Moreright	0	1	0	0	5%
No	0	1	0	0	5%
Ok	0	0	0	0	0%
Out	2	0	4	4	50%
Push	0	0	2	0	10%
Push-w	0	0	0	0	0%
Salute	0	0	0	0	0%
Safe	0	0	0	0	0%



(a) Comeon

(b) Stop

(c) Push

(d) Stop-w

(e) Push-w

(f) No

Figure 6. Examples of gestures in experiments