VISION-BASED HAND GESTURE RECOGNITION FOR HUMAN-ROBOT INTERACTION

Md. Hasanuzzaman*, M.A. Bhuiyan***, V. Ampornaramveth*, T. Zhang*, Y. Shirai**, H. Ueno*
*Intelligent System Research Division, National Institute of Informatics, 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430, Japan; **Department of Computer Controlled Mechanical Systems, Osaka University, Suita, 565-0871 Japan; ***Jahangirnagor University, Dhaka-1342, Bangladesh.

Keywords: Computer vision, Hand gesture, Human-robot interaction (HRI)

Abstract: This paper presents a vision-based hand gesture recognition system for interaction between human and robot. A real-time two hands gestures recognition system has been developed by combining three larger components analysis based on skin-color segmentation and multiple features based template-matching techniques. Gesture commands are generated and issued whenever the combinations of three skin-like regions at a particular frame match with the predefined gestures. These gesture commands are sent to robots through TCP-IP network for human-robot interaction. A method has also been proposed to detect left hand and right hand relative to face position, as well as, to detect the face and locate its position. The effectiveness of our method has been demonstrated over the interaction with a robot named ROBOVIE.

1 INTRODUCTION

As robots increase in capabilities and are able to perform more humanoid tasks in an autonomous manner, we need to think about the interaction that human will have with robots. There are several ways to communicate with human being and intelligent machine (e.g. robot, vehicle, etc,), with text commands, speech commands, gesture commands, and so on. Text commands based approach is robust but it is not natural compared to human-human communications. Although verbal commands based human-robot interaction system is employed based on few key words (such as move left, move right, stop, etc.) there are so many difficulties to generalize human speech. In this paper we present a method of developing a gestures based nonverbal interaction system between robots and human being. So our first attention is focused on vision based hand gesture recognition and then to interact with robot using gesture commands.

Two approaches are commonly used to interpret gestures for human machine interaction. One is gloved based approach (Vladimir, 1997) that requires wearing of cumbersome contact devices and generally carrying a load of cables that connect the device to a computer. Another approach is vision based technique that does not require wearing any of contact devices with human body part, but uses a set of video cameras and computer vision techniques to interpret gestures. Gesture recognition based on vision technology has been emerging with the rapid development of computer hardware of vision system in recent years and in future it will dominate in both Human-Computer and Human-Robot interactions. For gesture interpretation system gestures modelling is the first step that mainly depend on the intended application of those gestures. Gesture modelling can follow appearance based or model based approach. Model based approach is very hard to implement in real time because they usually use very complicated algorithms to extract accurate joint angles. An appearance-based algorithm is a strong tool for object recognition. Here, a variety of object appearances are stored as a statistical model and used in the recognition task. The gestures are modelled by relating the appearance of any gesture to the appearance of the set of predefined, template gestures. In this work we have used appearance base model.

Takahiro Watanabe et. al. (Watanabe, 1996) used maskable template based on minimum distance between template and partial block of an input image for gesture recognition. Hitoshi Hongo et. al. (Hongo, 2000) has developed a system that can track multiple faces and hands by using multiple cameras to focus on face and gesture recognition. Akira Utsumi et. al. (Utsumi, 2002) detected hand using hand shape model and tracked using extracted color

Hasanuzzaman M., A. Bhuiyan M., Ampornaramveth V., Zhang T., Shirai Y. and Ueno H. (2004).
VISION-BASED HAND GESTURE RECOGNITION FOR HUMAN-ROBOT INTERACTION.
In *Proceedings of the First International Conference on Informatics in Control, Automation and Robotics*, pages 440-445
DOI: 10.5220/0001137904400445

and motion. They also used multiple cameras with individual processor.

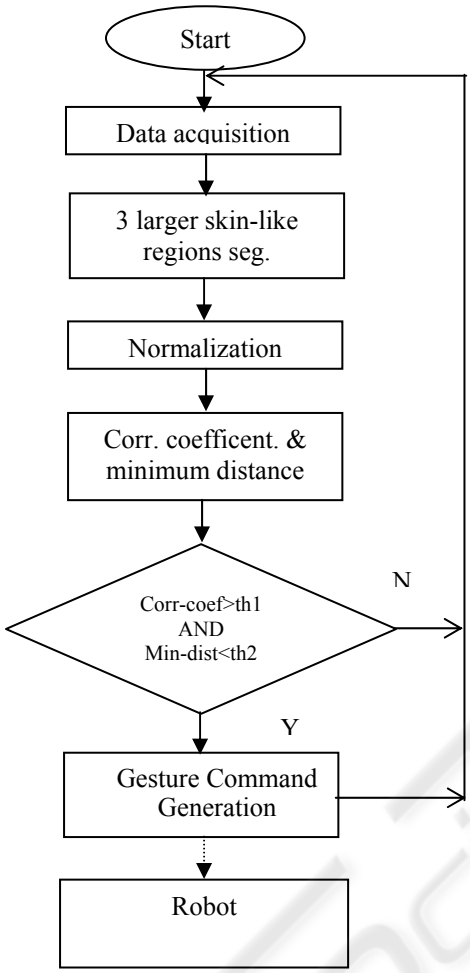


Figure 1: Flowchart of gesture based human robot interaction

In this paper we present a simple and faster method for recognizing gestures with skin-color segmentation and multiple features-based template matching techniques as shown in Figure 1. In this method three larger skin like regions are segmented from the input images assuming that a face and two hands may be present in the image frame at the same time. Skin-like region is segmented using color segmentation technique from YIQ color space. Segmented blocks are normalized and compared with template images for finding best match. For template matching we have used combination of two features: correlation coefficient and minimum (Manhattan distance) distance qualifier. If the combinations of three skin-like regions at a particular frame match with our predefined gesture then corresponding gesture command is generated. In this experiment we have recognized three gestures, TwoHand, LeftHand and RightHand as shown in Figure 2. Gesture commands are being sent to robots through TCP-IP network and their actions are being accomplished according to users defined action for that gesture. A method has also been developed to detect left hand and right hand relative to face position, as well as, to detect the face and locate its centre position. As an application of our method, we have implemented real time human-robot interaction systems using a robot named ROBOVIE with the commands, such as: “Raise Two Arms”, “Raise Left Arm” and “Raise Right Arm”, according to three gestures TwoHand, LeftHand and RightHand respectively.

The remainder of this paper is organized as follows. In section 2, we have briefly described skin-like regions segmentation, filtering, normalization, multiple feature based template-matching techniques for face and hand pose detection and gestures recognition system. Section 3 presents our experimental results and discussions. Section 4 we concludes this paper.

2 HAND GESTURE RECOGNITION

Figure 1, shows the hand gesture recognition system flowchart. This system uses video camera for data acquisition. The system first segmented into three larger skin-like regions by using skin-color information from the input images. The selected skin-like regions are normalized and resized as template size in order to match with templates. The gestures are recognized according to matching results of three segmented blocks at a particular image frame.

2.1 Skin Color Segmentation, Filtering and Normalization

This section introduces a color segmentation based approach for determining skin parts (mainly face and hands) from color images. YIQ (Y is luminance of the color, I, Q chrominance of the color) color model is used for skin color segmentation, since color footprint is more distinguishable and less sensitive to illumination changes in the YIQ space than the standard RGB color space. YIQ color representation system is typically used in video coding and provides an effective use of chrominance information for modelling the human skin color.

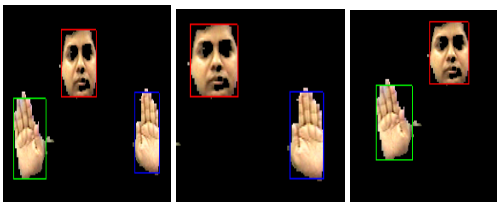


Figure 2: a) TwoHand, b) LeftHand, c) RightHand gestures

To detect hands or face regions, the RGB image taken by video camera is converted to YIQ color representation system and threshold it by the skin color range (Bhuiyan, 2003). Values of Y and I play an important role to distinguish skin like regions from non-skin like regions. Values of Y and I vary on person's body colors as well as lighting conditions. If body color is black then Y and I values decrease and if body color is white then Y and I values increase. We have computed our threshold values using several persons from Bangladesh, Japan, China and Thiland.

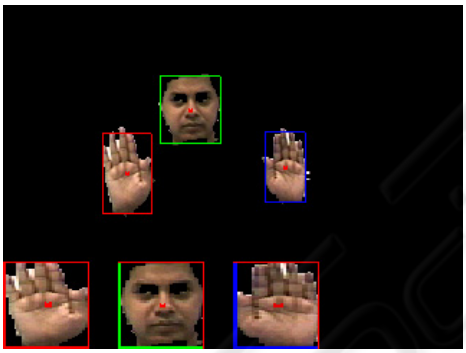


Figure 3: Output of skin color segmentation and normalization.

We have included an off-line program to calculate threshold values of Y and I if person's color and lighting condition is not fitted with current threshold values. In that case we need to select skin part of the new person from the images and run our threshold calculation program, then it will give the threshold values of that person and we update our threshold values accordingly in system. Locations of the probable hands and face are determined from the image with three larger connected regions of skin-colored pixels. In this experiment, 8-pixels neighbourhood connectivity is employed. In order to remove the false regions from the isolated blocks, smaller connected regions are assigned by the values of black-color. Noise and holes are filtered by morphological dilation and erosion operations. Normalization is done to convert the segmented

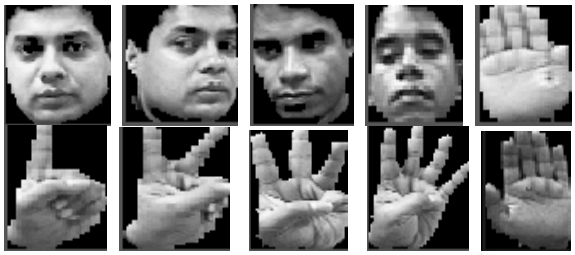


Figure 4: Sample template images

images to gray images and the image is resized as template images size. Sample output of the skin-color segmentation and normalization part are shown in Figure 3, where lower part squares show the resized images. Red points on the image shows the centre points or location of each segmented region. If hand or face is detected then corresponding centre point is used as the hand and face location.

2.2 Face and Hand Detection Using Template Matching

As template-matching approach is a more natural approach of pattern recognition, we use it to recognize gesture from an unknown input image. First we have prepared noise free version of faces and hands palm of different persons in different lighting conditions as template images as shown in Figure 4. To support small rotation we have included some slightly rotated images within our template images. For template matching we have considered two features: one is maximum correlation coefficient and another is minimum distance classifier (Manhattan distance) between two same size images. Correlation coefficient is calculated using following equation,

$$\alpha_t = M_t / P_t \ (0<\alpha_t \leq 1) \qquad (1)$$

where M_t is total number of matched pixels (white pixels with white pixels and black pixels with black pixels) with t-th template, P_t is number of total pixels in t-th template and t is a positive number. For exact matching α_t is 1, but for practical environment we have chosen threshold for α_t through experiment for optimal matching.

Minimum distance can be calculated by using following equation,

$$\delta_t = \{\sum_{1}^{x \times y} | I - G_t |\} \quad (2)$$

where, I(x, y) is the input image and $G_1(x,y)$, $G_2(x,y)$,----------,$G_t(x,y)$ are template images. There are more than one way to define δ_t corresponding to different ways of measuring distance. Two of the most common are: Euclidean metric and Manhattan metric. In our experiment we have used Manhattan metric. For exact matching δ_t is 0 (zero) but for practical purpose we have used a threshold value through experiment for finding optimal matching.

We have combined output of these two matching methods to make our system more accurate to recognize gestures. In our method we have grouped template images as face class (C_1), left hand class (C_2) and right hand class (C_3). Face class includes different person's frontal faces with some of them slightly rotated. Left hand class includes left hand palms and right hand class include right hand palms (frontal) of different persons. If max $\{\alpha_t\}$>th_1 is true then corresponding class (C_α) is identified, similarly if min $\{\delta_t\}$<th_2 is true then corresponding class (C_δ) is identified, where th_1 and th_2 are thresholds for correlation coefficient and minimum distance qualifier respectively. If both methods identified the same class then corresponding class is detected, otherwise ignored. Using similar way detected poses for three segments from an image and calculate face and hands location from the centre coordinates of

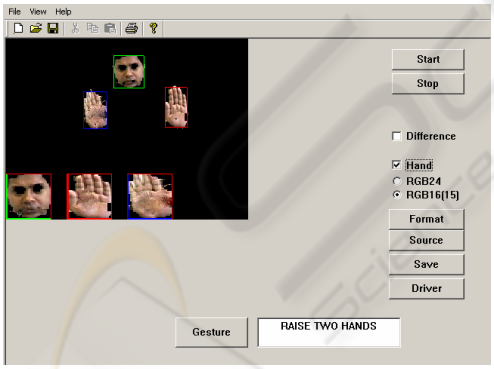


Figure 5: Sample output for two hands gesture

2.3 Gesture Recognition

In this experiment we have recognized static gesture for human robot mimic operations that means robot imitates human actions. Gestures are recognized using rule-based system from the combinations of pose detection output of three segments for a particular image frame. If two hands and one face are present in the input image then the system recognizes it as "TwoHand" gesture. If one face and one hand are present in the input image frame then the system recognizes it as either left hand or right hand depending on its position with respective to face position using following equation,

$$\varepsilon = f_x - h_x \quad (3)$$

where, f_x is the x-coordinate of the centre position of face segment and h_x is the x-coordinate of the centre position of hand segment (when one hand is present). If the distance ε is negative then it is detected as right hand gesture (RightHand) and if it is positive then it is detected as left hand gesture (LeftHand). According to the gesture recognized, corresponding gesture command is generated and transferred to interact with robot through TCP-IP network. Our approach has been implemented on a communicative humanoid robot named ROBOVIE. The commands employed for the interaction are, "Raise Two Arms", "Raise Left Arm" and "Raise Right Arm" corresponding to gestures "TwoHand", "LeftHand" and "RightHand".

3 EXPERIMENTAL RESULTS AND DISCUSSION

This section describes experimental procedures, as well as experimental results of the gestures recognition system and human-robot interaction system. This system uses a standard video camera for image acquisition. Each captured image is digitized into a matrix of 320×240 pixels with 24-bit color. First we have prepared pure templates. All the templates are of 60×60 pixels gray image. The template images are consisted a total of 180 frontal images of faces, left hands and right hands of different people. Figure 4 shows example template images. We have tested our system for real time input images. The sample visual output of our gesture recognition system is shown in Figure 5 for the "TwoHand" gesture.

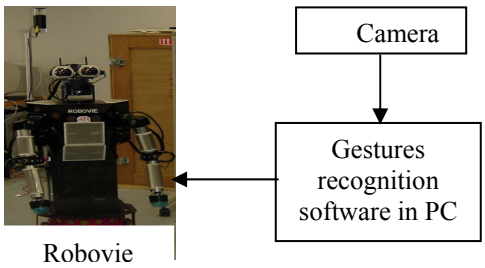
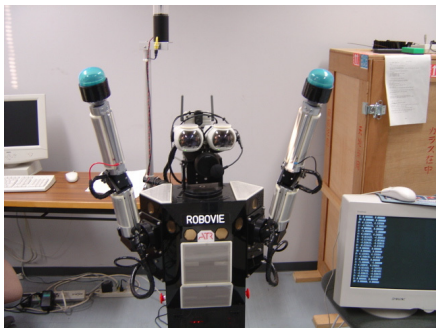
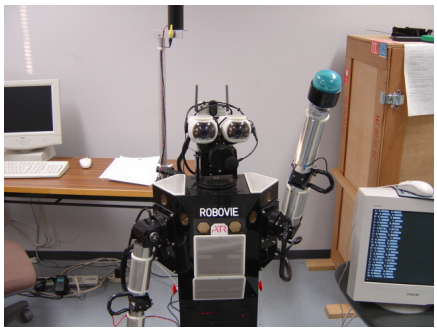


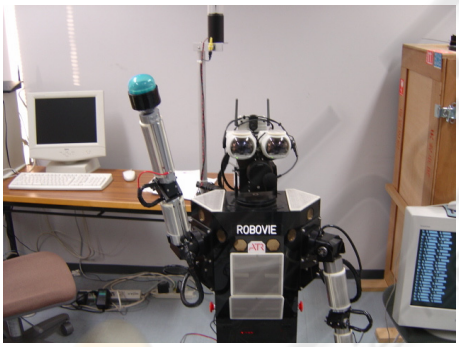
Figure 6: Architecture for human-robot interaction



a) Raise two arms (TwoHand)



b) Raise left arm (LeftHand)



c) Raise right arm (RightHand)

Figure 7: Sample outputs a), b) and c) corresponding to gestures TwoHand, LeftHand and RightHand

It also shows gesture command at the bottom side text box corresponding to matched gesture. In case of no match, it shows "no matching found".

We have also made a comparison among the correlation coefficient based, the Manhattan distance based and their combinations based templates matching approaches for sample still input images (Table-1). In this table E_m, is the total number of wrong detection using Manhattan distance, E_{cc} is the total number of wrong detection using correlation coefficient and E_{com} is the total number of wrong detection using their combination. From the table we conclude that using combinations of two features we can remove errors of two separate methods. In this case we have considered five gestures such as ONE, TWO, THREE, FOUR and FIVE and used five template classes corresponding to those gestures as shown in second row in Figure 4.

Table 1: Comparison of two template based methods

Gesture	No.of Inputs	E_m	E_{cc}	E_{com}
1	25	4	0	0
2	25	2	1	0
3	20	3	0	0
4	20	2	1	0
5	20	0	0	0

In this part we have explained a real time human-robot (ROBOVIE) interaction system using recognization gestures. We have implemented this application by off-board configuration that means gesture recognition program was run in client PC, not in ROBOVIE as shown in Figure 6. We have considered our robot as a server and our PC as a client. Communication link has been established through TCP-IP protocol. Initially, we connected the client PC with ROBOVIE server and then gestures recognition program was run in the client PC. As a result the client PC sends gesture commands to the robot (ROBOVIE) through TCP-IP protocol and it acted according to users predefined actions. The results of our interaction program are shown in Figure 7, for "Raise Two Arms", "Raise Left Arm" and "Raise Right Arm" in accordance with gesture TwoHand, LeftHand and Right Hand respectively. After completing each mimic operation the robot goes to its initial position. We have considered for human-robot interaction that gesture command will be effective until robot finishes corresponding action for that gesture.

4 CONCLUSIONS

This paper describes a real-time hand gesture recognition system using skin color segmentation and multiple features based template-matching techniques. For the matching algorithm we have used combinations of minimum distance qualifier (Manhattan distance) and correlation coefficient based matching approaches, that's why it is more robust than any single feature based template-matching techniques. One of the major constrain of this system is that the background should be non-skin color substrate. If we use infrared camera then it is possible to overcome this problem just by a little

modification of our segmentation module, other modules will remain the same.

We have also successfully implemented simple gestures based human-robot interactive system for mimic operation, using a robot named ROBOVIE. We believe that vision system will replace attached physical sensors for human robot interaction in the near future. A particular user may assign distinct commands to specific hand gestures and thus control various intelligent robots using hand gestures.

The significant issues in gesture recognition for our method are the simplification of the algorithm and reduction of processing time in issuing commands for the robot. Our next step is to make the detecting system more robust and to recognize dynamic facial and hand gestures for interaction with different robots such as AIBO, ROBOVIE, SCOUT, MELFA, etc. Our ultimate goal is to establish a symbiotic society for all of the distributed autonomous intelligent components so that, they share their resources and work cooperatively with human beings.

REFERENCES

Vladimir I. Pavlovic, 1997. Visual Interpretation of Hand Gestures for Human-Computer Interaction: A Review. IEEE PAMI, Vol. 19, No. 7, pp. 677-695.

Watanabe, T., 1996. Real-Time Gesture Recognition Using Maskable Template Model. Proc. of the International Conference on Multimedia Computing and Systems (ICMCS'96), pp. 341-348.

Hongo, H., 2000. Focus of Attention for Face and Hand Gesture Recognition Using Multiple Cameras. AFGR00, IEEE, pp. 156-161.

Matthew, T., 1991. Eigenface for Recognition. Journal of Cognitive Neuroscience, Vol. 3, No.1, pp. 71-86.

Utsumi, A., 2002. Hand Detection and Tracking using Pixel Value Distribution Model for Multiple-Camera-Based Gesture Interactions. Proc. of the IEEE workshop on knowledge Media Networking (KMN'02), pp. 31-36.

Bhuiyan, M. A., 2003. Face Detection and Facial Feature Localization for Human-machine Interface. NII Journal. Vol. 5, pp. 25-39.

Huang, Yu, 2002. Two-Hand Gesture Tracking Incorporating Template Warping With Static Segmentation. AFGR'02, IEEE, pp. 260-265.

Bretzner, L., 2002. Hand Gesture Recognition using Multi-Scale Colour Features, Hierarchical Models and Particle Filtering. AFGR'02, IEEE pp. 423-428.

Bhuiyan, M. A., 2004. ON TRACKING OF EYE FOR HUMAN-ROBOT INTERFACE. International Journal of Robotics and Automation, Vol. 19, No. 1, pp. 42-54.

Shimada, N., 1996. 3-D Hand Pose Estimation and Shape Model Refinement from a Monocular Image Sequence. Proc. of VSMM'96 in GIFU, pp.23-428

Grzeszczuk, R., 2000. Stereo Based Gesture Recognition Invariant to 3D pose and lighting. CVPR'00, IEEE, pp. 1826-1833.

Yunato, Cui, 1996. Hand Segmentation Using Learning-Based prediction and verification for hand Sign Recognition. Proc. of the Conference on Computer Vision and pattern Recognition (CVPR'96), IEEE, pp. 88-93.

Yoichi Sato, 2000. Fast Tracking of hands and Fingertips in Infrared Images for Augmented Desk Interface. AFGR'00, IEEE, pp. 462-467.

Charles, J., 2001. A Basic Hand Gesture Control System for PC Applications. Proc. of the 30th Applied Imagery Pattern Recognition Workshop (AIPR'01), IEEE, pp. 74-79

Dong, Guo, 1998. Vision-Based Hand Gesture Recognition for Human-Vehicle Interaction. Proc. of the International conference on Control, Automation and Computer Vision, Vol. 1, pp. 151-155.