Proceedings of the 17th World Congress
The International Federation of Automatic Control
Seoul, Korea, July 6-11, 2008

A Tele-operated Gesture Recognition Mobile Robot using a Stereo Vision

H.C. Shin*, Y.G. Kim , J.I. Cho*** and Y.J.Cho******

**Electronics and Telecommunications Research Institute, 305-700, DaeJeon, Korea (Tel: +82-42-860-6140; e-mail: creatrix@etri.re.kr).*
*** Korea University of Science and Technology, 305-700, DaeJeon, Korea (e-mail:placeo@etri.re.kr)*
**** Electronics and Telecommunications Research Institute, 305-700, DaeJeon, Korea, (e-mail: jicho@etri.re.kr)*
***** Electronics and Telecommunications Research Institute, 305-700, DaeJeon, Korea, (e-mail: youngjo@etri.re.kr)*

Abstract: In this paper, a tele-operated gesture recognition mobile robot using a stereo vision is represented. We propose a real-time face and hand classification and 3D position extraction using a stereo vision embedded system. To obtain the disparity image, we used the stereo vision process on FPGA and developed the embedded system to obtain the image sequence for basic image processing. Then we applied the simple and reliable algorithm we developed to detect and classify the head and hand in real-time on Tele-operation server. The arm posture and hand trajectory were used for gesture recognition. We also show the experimental result to support validity of the system and the algorithm.

1. Introduction

Recently a service robot market shows rapid growth for vacuum cleaning, home security and content service of education and entertainment. For useful home service robot, a robot must have high reliability, low price, low power consumption, simple structure and high performance. An embedded robot system can be proposed for these requirements. But because the embedded system is generally planed for specified function and small system, an embedded robot system may have insufficient computing power. To solve this problem, URC (Ubiquitous Robotic Companion) was proposed [1]. In this system the high computing power jobs such as face detection, face recognition, voice recognition, SLAM (Simultaneous Localization and Mapping), TTS (Text to Speech) are processed on a remote server and the results are informed to mobile robot. As a service robot, the face detection and gesture recognition are essential function.

2. System Configuration

The developed hardware system is mainly divided into four parts. The FPGA board produces disparity image from two CMOS cameras. The embedded system obtains image from FPGA board and executes basic image processing. These devices are equipped on robot platform and the tele-operation server controls whole robot system.

Fig.1 Developed robot system

2.1 Stereo vision FPGA

The stereo vision has wide range of usage and many applications have been developed. Due to recursive simple computation, it is desirable to be implemented by ASIC. In ETRI, stereo ASIC is being developed which is based on Trellis DP algorithm [2]. Now, it is verified the algorithm implemented on FPGA. The used number of gate is 2,000,000 gates. Developed FPGA outputs disparity image that has 320x240 pixels and frame rate of 30fps shown as Fig. 2.

Fig.2 Stereo vision result

2.2 Embedded main board for image acquisition, and basic processing

The embedded system controls CMOS camera and receives the image sequence from two cameras, then sends to FPGA. After stereo vision processing in FPGA, the embedded system receives computed data from FPGA. We adopted i.MX21(350MHz) embedded processor from Motorola's 6th generation Dragon Ball and it was optimized. With i.MX21 main processor, we developed main control board, MIM (Multi-modal Interface Module) which operates at Linux 2.4.20 shown as Fig. 3.

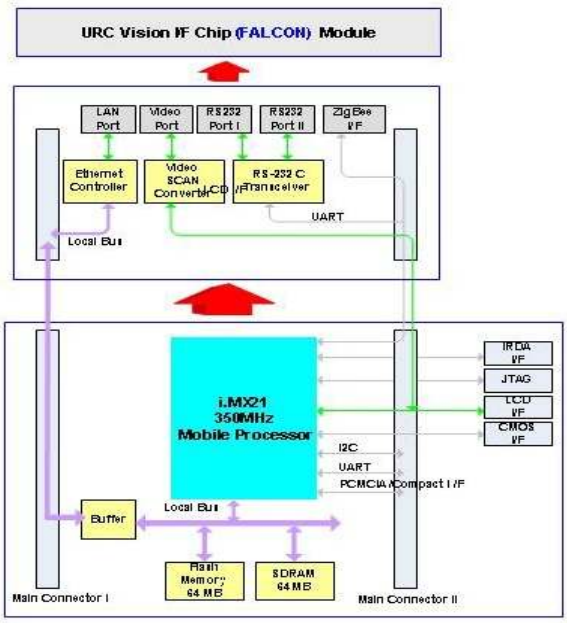

Fig.3 Embedded system structure

2.3 Robot platform

The robot platform has 2 wheels, camera pan-tilt actuator, 4 DOF two arms. The embedded main board controls the robot platform

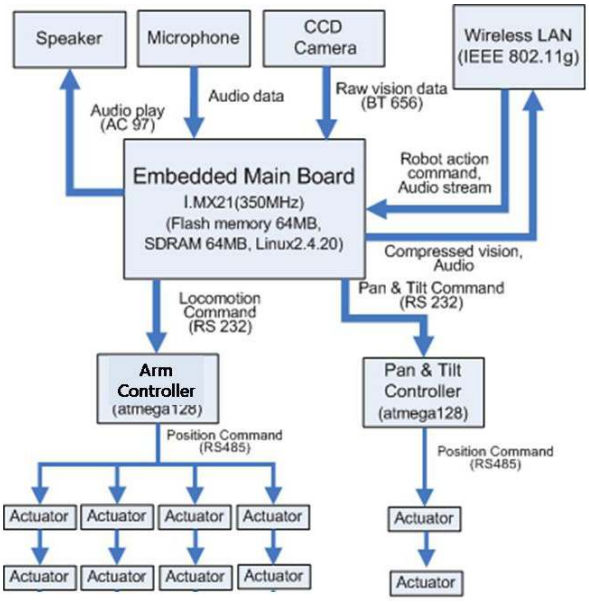

Fig.4 Mobile robot platform

2.4 Tele-operation Server

The transferred data from robot through wireless LAN (IEEE 802.11g) flows to data handler and classified into vision, sensor, and audio data. The data handler hands vision stream over to H.263/MPEG-4 decoder and still image sampler. The server main controller controls robot locomotion, pan-tilt actuation, robot arm control, audio command, robot battery management, face detection result, navigation controller and user command. The server main controller commands locomotion and pan-tilt control to track and approach human with face detection result. Deciding robot approached human enough with face size and proxy sensor data, the server commands robot to present various audio and visual contents and arm gesture. If there is no human, the navigation controller offers locomotion data for free navigation in a given space. The audio stream generator generates audio stream from TTS, music file and microphone of server. The server main controller transfers the generated audio stream to robot and robot plays the downloaded audio stream.

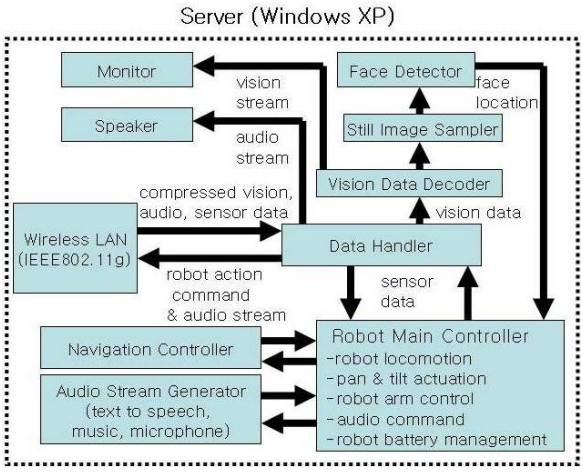

Fig. 5 Tele-operation server structure

3. Face-hand Detection and Classification

The skin color detection is widely used for face detection, recognition, gesture recognition and etc. But the skin color has defects that can be easily affected by illumination, background, camera characteristic and ethnicity. Illumination and camera characteristic is affected by the performance of camera, but background color problem can be solved by using stereo vision[3]. In this research, we eliminated background by using stereo vision then detected face and hand region by using skin color. In case abnormal detection, we performed adaptive color compensation.

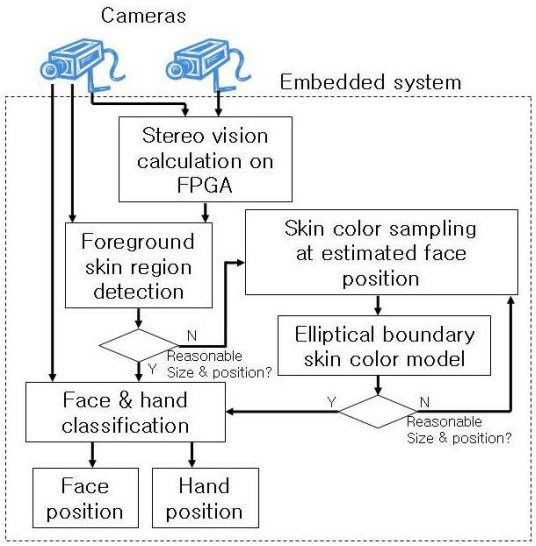

Fig.6 Face and hand detection procedure

3.1 Background elimination, skin color extraction and segmentation

Shown in Fig. 6, initially the foreground skin color region detection is executed by using actual color image and disparity image. First, the skin color region is extracted by using the chrominance components of actual image. YUV color space is mainly used because chrominance components of this space don't be affected by illumination much [5]. Using disparity map, background subtraction is performed. Then, foreground skin region is extracted by comparing with formerly extracted skin region. To reduce noise, Gaussian filtering is performed to foreground skin region. 5x5 Gaussian mask is used to our implementation. Next, remaining regions are segmented. To segment extracted region, we used seed method. Seeds are sprayed to the Gaussian filtered image with regular interval, and then volumes are enlarged in case of extracted region, if not, volume growing is stopped. After that, each volume is labeled. Finally, the depth, size and position are calculated to the foreground skin region divided after labeling process in the pixel domain. This calculation is performed by using the trigonometry. By calculating the average disparity value of extracted region, the disparity value of object is obtained. And this disparity value is converted to the distance value. By using the triangulation, the height and width of an object is calculated as shown in Fig. 5. Finally these values are transmitted to next stage. Foreground skin region detection procedure is depicted in Fig. 7.

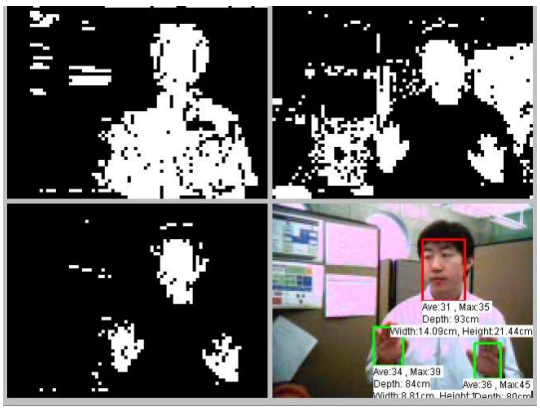

Fig.7 Segmented and classified skin region

3.2 Skin color compensation

In the Foreground skin detection, dominant problem is the case of a man wearing the cloth which is similar to the skin color. In this case, since it becomes the segmentation and labeling about the whole body of a man, the physical size of extracted region is appeared largely like an a and b compared with normal case c and d as shown in Fig. 8. To extract the skin color region only, the candidate area is set by referencing the height of all labeled objects. We assumed that the head height is 22cm. because the head is usually positioned in the top, we can only get the head candidate region using trigonometry. After sampling skin color from the candidate region, elliptical boundary model is applied

using the chrominance components. Elliptical boundary model is defined as [7]

$$\Phi(c) = [c - \Psi]^T \Lambda^{-1} [c - \Psi]$$
$$\Psi = \frac{1}{n}\sum_{i=1}^{n} c_i, \Lambda = \frac{1}{N}\sum_{i=1}^{n} f_i (c_i - \mu)(c_i - \mu)^T$$

Where c is the is the chrominance component vector, N is the total number of samples, f_i is the number of samples with chrominance c_i and μ is the mean of the chrominance vectors in the sampled data set. c_i is the sampled chrominance vector. Λ is the covariance matrix. $\Phi(c)$ is called Mahalanobis distance. The pixel chrominance c is

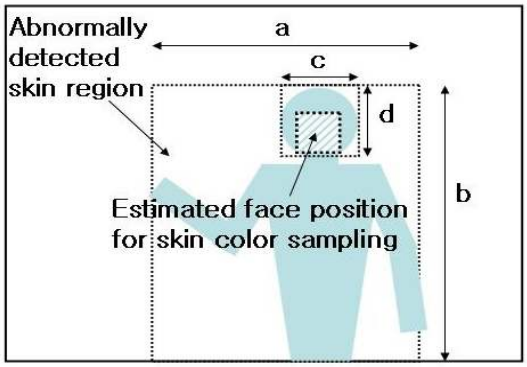

Fig.8 Estimated face position for skin color sampling

classified as a skin pixel,if , where is a threshold value chosen empirically as a trade-off between the true and false positives.

3.3 Face-Hand Classification and Gesture Recognition

After foreground skin region detection, each region is classified as face or hand. Generally, the face height is larger than hand and face has more horizontal patterns such as eyebrow, eye, nose and mouth. For the horizontal pattern detection, we used the horizontal sobel mask for the detected skin region and calculated the amount of horizontal patterns. We defined the face score by multiplying detected skin region height and amount of horizontal pattern as shown in Fig. 8. If face score is larger than the threshold value, the detected skin region considered. Using hand position trajectory, a neural net algorithm classifies the trajectories for gesture recognition. The elbow positions are estimated from stereo vision using inverse kinematics.

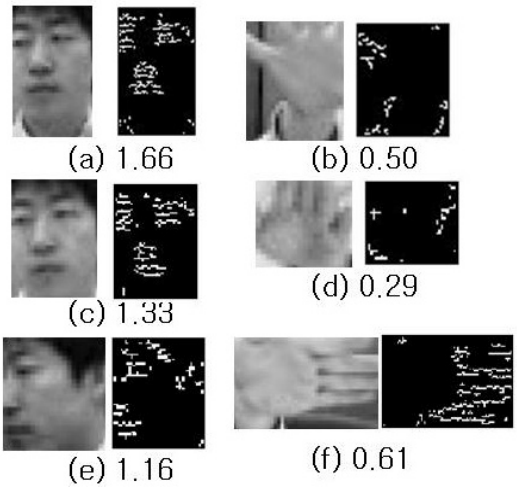

Fig. 9 Face scores for face-hand classification

4. Tele-operation Control

4.1 Tele-operation Structure

The server controls robot locomotion and pan-tilt actuation for face tracking. The server controller commands the robot linear velocity, angular velocity and pan-tilt angle to track and approach to human as shown in Fig.10 (vrobot: robot linear velocity, ω robot: robot angular velocity, φpan, φtilt: robot pan-tilt angle, dIR:proxy sensor data set). Because the face detection process is not periodic, the pan-tilt angle control is event-driven at the time of face detection finish. But time-driven control is more adequate for locomotion control.

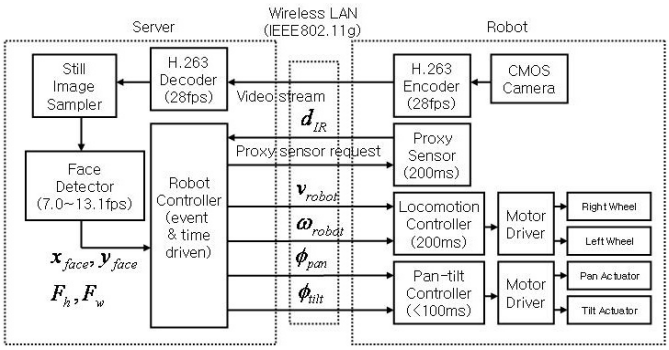

Fig. 10 Tele-operation control structure.

4.2 Robot control for face tracking

With face location and pan-tilt angle information, we can calculate the human face direction with respect to robot camera and robot body as shown in Fig. 8. Because the field of view of CMOS camera is 50 by 40 degree, the face position direction φ_{face_x}, φ_{face_y} and θ_{face} can be determined as shown in Fig.11.

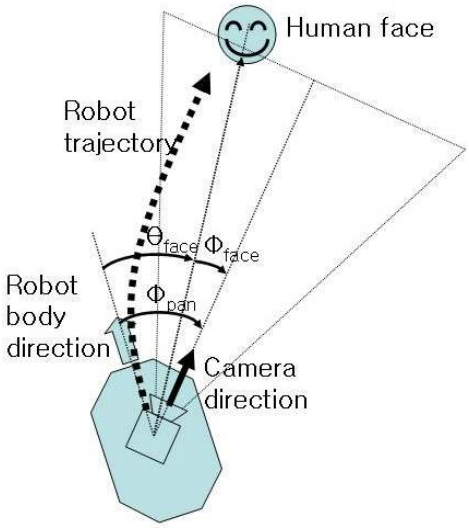

Fig. 11 Face tracking

At the time of face detection finish, the robot controller commands pan-tilt angle change.

5. Experimental results

5.1 Face Detection Performance

The performance of the face detection can be different to people, because every one has a difference of the face size, color, brightness, the hair style, and etc of the face. Moreover, as the distance of a man and camera becomes further, the probability of an error being generated is increased. Fig. 9. shows the face detection success ratio according to the distance of a man and camera from 100cm to 300cm. The face detection performance is measured in the proportion of the whole disposed frames to the frames determined by face.

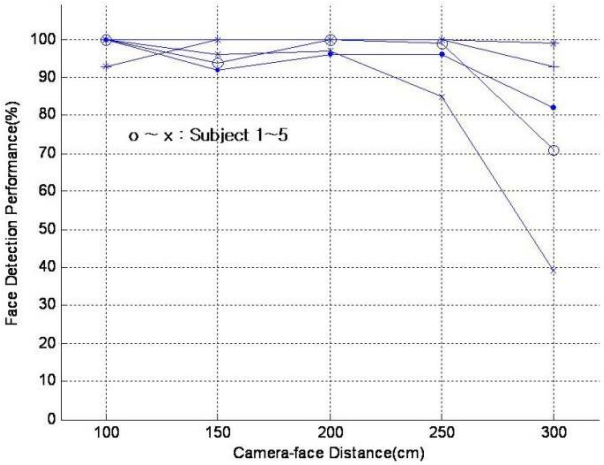

Fig. 12 Face detection performance

In the Table 1, the face-hand discrimination accuracy can be looked at. We determined whether the head and hand are correctly discriminated or not apart 1m and 2m from the camera.

5.2 Position Accuracy of Head and Hand Position

The distance of a camera and the area which is detected by using the average disparity value at the detected area can be calculated. The accuracy of the measured distance according to a distance between the face and a camera is shown at Fig. 10. Usually if the distance between the camera and head is further, error is also increased. In the Fig. 11. the hand position accuracy on perpendicular direction to camera axis can be looked at.

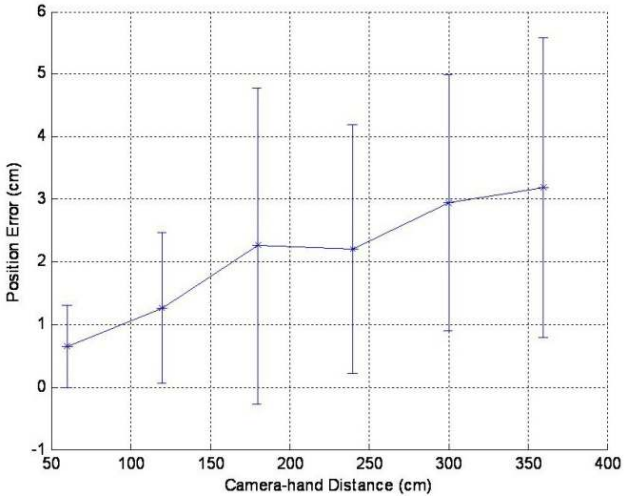

Fig. 13. Detected hand position accuracy

Table 1. Face-hand discrimination accuracy

Distance	1m	2m
Performance	90%	87%

Fig. 14 Detected gesture

6. Conclusion

In this paper, a tele-operated gesture recognition mobile robot using a stereo vision is represented. We propose real-time face and hand classification and 3D position extraction using a stereo vision embedded system. To obtain the disparity image, we used the stereo vision process on FPGA to get the disparity image and also developed the embedded system to obtain the image sequence to perform image processing. Then we applied the simple and reliable algorithm we developed to detect and classify the head and hand in real-

time. This system can recognize abnormal skin region detection such as skin color similar outfits, and extract minute skin color for elliptical skin color model from estimated face position using head shape from stereo vision. To classify face and hand, we defined a head score multiplied horizontal sobel edge density by physical height of detected region. The developed system shows 98% face detection accuracy and 89% face-hand classification accuracy within 2 meters. It also shows ±5cm 3D position accuracy of face and hand within 2 meters. The detected head and hand positions are processed for gesture recognition. The arm posture and hand trajectory were used for gesture recognition. This algorithm was realized on the tele-operation server.

ACKNOWLEDGEMENT

This work was supported in part by MIC & IITA through IT Leading R&D Support Project

REFERENCES

Cho, Y.J. and Oh, S.R.( ) "Fusion of IT and RT: URC (ubiquitous robotic companion) program", *JOURNAL-ROBOTICS SOCIETY OF JAPAN*, **v.23, no.5**, pp.22-25.

Mau-Tsuen Yang, Shih-Chun Wang and Yong-Yuan Lin (2005). A multimodal fusion system for people detection and tracking. *International journal of imaging systems and technology*, **v.15 no.2**, pp.131-142.

Hong Jeong and Yuns Oh (2000). Parallel Trellis Based Stereo Matching Using Constraints. *Lecture notes in computer science*, pp.227-237.

Jojic, N., Brumitt, B., Meyers, B., Harris, S., and Huang, T. (2000). Detection and estimation of pointing gestures in dense disparity maps. *Automatic Face and Gesture Recognition*, pp.468-475

Chai, D. and Ngan, K.N. (1999). Face segmentation using skin-color map in videophone applications. *IEEE Trans. Circuits Syst. Video Technol*. 9 (4)

Kakumanu, P., Makrogiannis, S. and Bourbakis, N. (2007). A survey of skin-color modeling and detection methods. *Pattern Recognition*, **v.40(3)**, pp. 1106-1122.

Quan Huynh-Thu, M. Meguro, M. Kaneko, "Skin-color extraction in images with complex background and varying illumination ," Applications of Computer Vision, 2002 (WACV 2002), pp.280-285

J.Y. Lee, S.I. Yoo, "An elliptical boundary model for skin color detection," Proceedings of the International Conference on ImagingScience, Systems and Technology, 2002.