RECOGNITION OF HUMAN HAND GESTURE FROM A MONOCULAR IMAGE SEQUENCE FOR HUMAN-MACHINE COMMUNICATION SYSTEM

Satoru Odo[†], Kiyoshi Hoshino^{†,‡}

[†]Faculty of Engineering, University of the Ryukyus

[‡]PRESTO and SORST, Japan Science and Technology Corporation

odo@bme.ie.u-ryukyu.ac.jp, hoshino@ie.u-ryukyu.ac.jp

Abstract The goal of this study is to construct a human-machine communication system. In the present study, we propose a technique to estimate the motion and position of human hand from a monocular camera, so as to extract motion parameters and recognize human gestures. Sequential images which are recorded with a monocular camera are transformed into log-polar mapped images to reduce the data size of the images without accuracy decreasing near the target. We carried out experiments to demonstrate the effectiveness of the method proposed. The experiments were carried out to actual sequential images. As the experimental results, the system performed the tracking and gesture recognition with high accuracy and with processing time of about 40 milliseconds per a frame, with an available personal computer without extra peripheral equipments.

Keywords: monocular camera, sequential images, log-polar mapping, gesture recognition

1. Introduction

It is not easy for PC beginners to master how to use the computer. In the human-human communications, for example, nonverbal messages such as facial expression, gesture, voice tone are very important cues. One of the typical examples are “gesture” which performed mainly with arms and hands. It is sometimes quite effective more than verbal language in the human communication, and mainly used as the means of communication among auditory-handicapped persons or foreigners. If the computers can recognize the human gestures, these kinds of the system may be a good interface between human and computer.

The original version of this chapter was revised: The copyright line was incorrect. This has been corrected. The Erratum to this chapter is available at DOI: [10.1007/978-0-387-35660-0_65](https://doi.org/10.1007/978-0-387-35660-0_65)

R. Nakatsu et al. (eds.), *Entertainment Computing*

© IFIP International Federation for Information Processing 2003

For measurement of human motions, there are three types of motion capture systems; magnetic, electro-mechanical and optical, respectively. The magnetic system has a centrally located transmitter and sensors that output the positions and directions. But the system is affected by any metals in the measurement space. The electro-mechanical system, however, does not have the problems that the magnetic system has. But the system requires test subjects to put on some human-shaped sensors, restricting the range of the movement. The optical system, to the contrary, has no restriction to the performers.

For gesture recognition from video images, it is necessary first to extract features from motion images and then to fit the input data to the models prepared beforehand in a computer. As the feature extraction methods, two techniques are popular; one is 3D model-based methods, which represent objects with three-dimensional models and extract parameters, and the other is appearance-based methods, which extract features from changes in appearance. The latter type is easier for implementation because a smaller amount of computation is required, but it is difficult for real-time recognition.

But in the human visual system, there is enough rapidity for recognition. Human sight has high resolution in the central part of visual axis which is called fovea centralis, and also has low resolution in the circumference part. These are big differences compared to usual artificial image-input devices which have uniform resolution. These relations can be approximated by the log-polar mapping proposed by Schwartz [1], through which the differentiated functions of central vision and peripheral vision are realized. The central part keeps high resolution, and the circumference part catches rough conditions.

In this paper, we propose a technique to estimate the human hand motion and position from a monocular camera, so as to extract motion parameters and recognize human gestures.

2. Visual characteristics

In biological vision systems, particularly referring to primates, broadly respond to the light sensitive elements (cones) on the retina are distributed in a non homogeneous. There are two areas; one is fovea centralis which is located at the center of visual axis and has high resolution, and the other is the peripheral where the resolution decreases according to the distance from the fovea. This relationship between distance and visual resolution is known as the log-polar mapping, which realizes two different functions in central and peripheral visions respectively. The object is focused and recognized with high resolution in the central vision,

but the area is small. However, the objects in a large area are roughly recognized in peripheral vision with low resolution.

3. Tracking system

3.1. Outline of the object tracking process

The system consists of a monocular camera and a personal computer(CPU:Pentium III(1 GHz)). It can track a human hand that moves under a real environment in real-time. The system tracks a human hand as follows: Firstly, sequential images are recorded with a monocular camera, and then transformed into log-polar mapped images by log-polar mapping. Secondly, the motion edge image is extracted by integrating an obtained a frame subtraction image and an edge image, and then that image limited with color information. Thirdly, the optical flows are extracted which is calculated by the generalized gradient method, and then typical flows is calculated by average flows in three frames before current time. Finally, the next gaze point is estimated based on the current position and the typical flows.

3.2. Log-polar mapping

Cartesian coordinate is transformed to polar coordinate. Then, the polar coordinate is expressed in logarithm.

We explain the detail of processing as follows: Let $I(x, y)$ be the original image and $L(p, q)$ be the log-polar mapped image to which $I(x, y)$ is transformed by log-polar mapping. Treating each pixel coordinate (x, y) and (p, q) as a complex number $z = x + iy$ and $w = p + iq$ respectively, log-polar mapping is

$$w = \log(z + \alpha) \quad (1)$$

where α is the shift parameter to avoid the singularity at the origin.

The object is focused and recognized with high resolution in the central vision, but the area is small. However, the objects in a large area are roughly recognized in peripheral vision with low resolution. An example of transformation is shown in figure 1. (a) is log-polar mapping from the cartesian plane to the log-polar plane. (b) is original image;the circle indicate the gaze point. (c) is an example of log-polar mapped image.

In this paper, we used the transform processing on computer to evade the use of extra hardwares. Calculation of log-polar mapping was needed in every image due to usage of usual CCD camera. We relationally calculated original image and log-polar mapping image, and made lookup table to shorten computation time at stage of preprocessing[2]. Calcu-

Figure 1. Constructing the log-polar mapped image.

lation of log-polar mapping was carried out at stage of preprocessing for once and lookup table was referred from next time.

Lookup table is shown as follows:

$$L(p, q) = LUT(x, y) \quad (2)$$

where LUT means the lookup table for an original image $I(x, y)$.

3.3. Estimate of position

We extracte the motion edge by integrating an obtained frame subtraction image and an edge image, and then that image is limited with color information.

Let $L(p, q, t)$ denote the pixel value at a point (p, q) in the log-polar mapped image. A frame subtraction image $L_1(p, q, t)$ is calculated by frame subtraction[3] as follows:

$$L_1(p, q, t) = \begin{cases} 1, & |L(p, q, t) - L(p, q, t-1)| \geq th_t \\ 0, & otherwise \end{cases} \quad (3)$$

An edge image $L_2(p, q, t)$ is calculated by Sobel operator as follows:

$$L_2(p, q, t) = \begin{cases} 1, & \sqrt{L_{HS}(p, q, t)^2 + L_{VS}(p, q, t)^2} \geq th_s \\ 0, & otherwise \end{cases} \quad (4)$$

Parameters in RGB color coordinate system are easily influenced by brightness baseline of the image.

Therefore, in this paper, $L^*u^*v^*$ color coordinate system (CIE1976 color coordinate system) is adopted in our system in spite of RGB color system, and then we use the u^*v^* plane in $L^*u^*v^*$ color coordinate system.

A skin color image $L_3(p, q, t)$ is calculated as follows:

$$L_3(p, q, t) = \begin{cases} 1 & , \quad (L_c - M)^T C^{-1} (L_c - M) \geq th_c \\ 0 & , \quad otherwise \end{cases} \quad (5)$$

where M is the mean vector of skin color and C is the variance-covariance matrix of skin color.

A motion edge image $L_d(p, q, t)$ is calculated as follows:

$$L_d(p, q, t) = \begin{cases} 1 & , \quad \sum_{i=1}^3 L_i(p, q, t) = 3 \\ 0 & , \quad otherwise \end{cases} \quad (6)$$

The human hand displacement $P(t)$ is determined by the center of the motion edge image.

3.4. Estimate of motion

Optical flow is calculated by the generalized gradient method[4]. we obtain the next equation:

$$f_x(x, y, t)u + f_y(x, y, t)v + f_t(x, y, t) = 0 \quad (7)$$

where $u = \frac{dx}{dt}$, $v = \frac{dy}{dt}$, (u, v) is flow vector. This is the constraint equation of the gradient method.

By applying two different spatial filters g , h to the log-polar mapped image, the following two constraint equations are derived.

$$g * f_x(x, y, t)u + g * f_y(x, y, t)v = g * f_t(x, y, t) \quad (8)$$

$$h * f_x(x, y, t)u + h * f_y(x, y, t)v = h * f_t(x, y, t) \quad (9)$$

where $*$ represents convolution. The flow vector (u, v) is obtained by solving these simultaneous equations. But, we use the reliability of the flow by the angle between two lines corresponding to equation [5]. If the reliability is small, the flow vector is not calculated.

The typical flows $EF(t)$ is calculated by average flows in three frames before current time as follows:

$$EF(t) = \frac{1}{3} \sum_{i=0}^2 F(t-i) \quad (10)$$

3.5. The gaze point search

We extract the object and estimate the next gaze point. Let $EF(t)$ denote the typical flows in the image. A next gaze point $gp(t+1)$ is defined by the next formula.

$$gp(t+1) = P(t) + EF(t) \quad (11)$$

4. Estimation of hand gestures

A gesture trajectory can be represented as a list of vectors which consist of the central points of the motion edge images. In the continuous time domain, the sequences of centroid points are detected by the tracking system, and thus, the gesture trajectory is produced by connecting centroid points. In the matching process, weighted sum of the difference of above features are utilized between the observed gesture and model gestures that are registered in a database a priori. The hand gesture is determined by selecting the model of minimum trajectory error.

5. Experimental results

We show some experimental results of the object tracking to demonstrate the effectiveness of the method proposed. The experiments were carried out on both pseudo object and actual image sequences. We describe the following making procedure of pseudo image sequence. First, we took the stationary image which has 320×240 pixel size and 256 levels by a monocular camera. Next, the two pseudo image sequence were generated on the basis of the stationary image. One is the image sequence including the object which had movement with triangular wave from the left direction to the right direction on the image, and the other is the image sequence including the object which revolved counterclockwise around the central point of stationary image. Both objects were a hand model and moved with uniform velocity.

Figure 2 and 3 show the experimental results of our method. In these figures, the thick lines indicate the actual measurements, and the broken lines indicate theoretical points. These figures suggest that the tracking task is successfully accomplished, because the actual measurements overlap completely with theoretical points.

The experiments were carried out on actual sequential images. We took the sequential image by a monocular camera. Each frame in the sequence has 320×240 pixel size and 256 levels. The experiment was conducted under the condition of normal illumination. Four gestures

Figure 2. Triangular-wise movement of the target.

Figure 3. Counterclockwise movement of the target.

(a) Raise hand.

(b) Bye-bye.

(c) To left.

(d) To right.

Figure 4. Standard pattern of four gestures.

Figure 5. Extracting results and traces of action "bye bye".

were used in experiment: (1) Raise hand, (2) Bye-bye, (3) To left, (4) To right. Figure 4 shows snapshots of each gesture.

Hand gesture were made on a planar space in front of a camera. The four gestures are drawn by one stroke gesture and then the start point and the last point are the same place. Note that performer was instructed to stop his gesture just a moment before starting a meaningful gesture and after stopping it. Using this rule, the standstill area responds to spotting area, and the gesture between two spotting area can be considered as a gesture. Figure 5 indicates the extracting results and traces of action "bye bye".

6. Conclusion

This study proposed a new human-hand tracking method in sequential images with a monocular camera using log-polar mapping. In the

system, input image was transformed into log-polar mapped image. The current position of the human hand was obtained as the center of gravity of motion edge image through difference image method and selection by color information. The next gaze point was estimated by the current gaze point and the average flow. Experiments were carried out to demonstrate the effectiveness of the proposed method. The results showed that the tracking task was successfully accomplished with high accuracy. The system can track the target with processing time of about 40 ms per a frame with an available personal computer without extra peripheral equipments.

Acknowledgments

This research was supported in part by the Ministry of Education, Culture, Sports, Science and Technology(MEXT), Grant-in-Aid for Scientific Research on Priority Areas (A), Research for Comprehensive Promotion of Study of Brain.

References

- [1] E.L. Schwartz, "Computational anatomy and functional architecture of striate cortex: a spatial mapping approach to perceptual coding," *Vision Research*, vol.20, no.8, pp.645-668, 1980.
- [2] Naoki Oshiro, Noriaki Maru, Atsushi Nishikawa, Fumio Miyazaki, "Binocular Tracking using Log Polar Mapping," *IEEE/RSJ International Conference on Intelligent Robots and Systems*, vol.2, pp.791-798, 1996.
- [3] Kameda Yoshinari, Minoh Michihiko, Ikeda Katsuo "Three Dimensional Motion Estimation of a Human Body Using a Difference Image Sequence," *Proceedings of Second Asian Conference on Computer Vision '95*, pp.181-185, 1995.
- [4] M.V.Srinivasan, "Generalized gradient schemes for the measurement of two-dimensional image motion," *Biological Cybernetics*, vol.63, pp.421-431, 1990.
- [5] Ryuzo Okada, Yoshiaki Shirai and Jun Miura "Object Tracking based on Optical Flow and Depth," *Proc. of IEEE/SICE/RSJ Int. Conf. on MFI*, pp.565-571, 1996.

5. MUSIC INFORMATICS