Recognizing multiple moving objects by foveated vision

Yasuhiko Kiuchi[†], Yasuo Kuniyoshi[‡],
Taketoshi Mishima[†], Hiroshi Mizoguchi[†], Takaomi Shigehara[†]
[†] Saitama–University., Department of Information and Computer Sciences,
255, Simo-ookubo, Urawa, Saitama 338–8570,JAPAN
[‡] Electrotechnical Laboratory, Intelligent Systems Division,
1–1–4, Umezono, Tsukuba, Ibaraki 305–8568, JAPAN
Tel: +81–298–61–5180 Fax: +81–298–61–5971
Email: {kiuchi,kuniyosh}@etl.go.jp, {mishima,hm,shigehara}@ics.saitama-u.ac.jp

Abstract

Foveated vision has the big advantage of exhibiting a wide field of view, along with a high resolution fovea. However, in the case of using optical flow, foveated vision has one demerit. The demerit is a concentrate of optical flow. For foveated vision, an object moves almost only around the center of the field. In this paper, we suggest how to segment motion of some objects, and how to discriminate a hand and another object. In the future, the method we suggested may be useful for recognizing human actions by foveated vision.

1 Goal of the research

We use foveated vision in biometic vision research because foveated images exhibit a wide field of view, along with a high resolution fovea.

S.Rougeaux and Y.Kuniyoshi have investigated robust tracking vision, which doesn't need any prior knowledge of the targets shape or texture.[1] They used a high performance active binocular camera head named 'ESCHeR', which can move very quickly and has foveated wide-angle lenses. Unfortunately this tracking vision pursues only one target. If two moving objects are in sight, this tracking vision selects one target only, and the other object will be ignored. For ESCHeR's lens characteristic (see the section '**ESCHeR**'), optical flow is generated more, near the center of the field. We would like the tracking vision to memorize this 'other object', and sometimes saccade to see it.

In achieving the first step of a framework for computer vision which recognizes human actions, we have designed and implemented a vision system which pursues a hand, with the ability to memorize multiple targets' position as an additional function to the vision system of ESCHeR (see figure 1). This function is summarized as follows:

The system uses optical flow of images from a camera, and some template hand images as *a priori* knowledge of target.

1. Optical flow is computed using a method proposed by Lucas and Kanade(1981). IIR recursive filter is used for the progress of accuracy.
2. In order to select a region which represents principal flow, this flow field is inputed into a competitive learning network.
3. The gazing direction is computed based on the result of process 2, and memorized. Therefore, the camera can gaze at each point.
4. ESCHeR saccades to see each object vividly, and the object images are cropped in order to be compared with the template images using PCA, etc.

In order to think about the problem simply, we don't consider the following cases. For example, in the case that two objects move together in the same location. These two objects are regarded as one object by our system. Or in the case that one object moves around the center of the field, and the other object moves around the periphery. In this case, the periphery object's flow would be suppressed by the center object's strong flow.

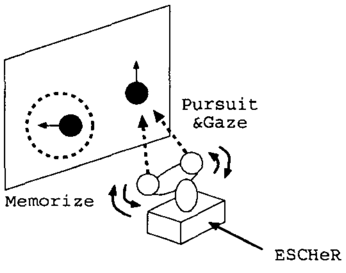


Figure 1: image figure: there are two moving objects in the sight of ESCHeR. One object may be pursued, and the other may be memorized.

2 ESCHeR

The design of ESCHeR[2](**E**tl **S**tereo **C**ompact **He**ad for **R**obot vision) has been mostly inspired by the properties of biological visual systems. It has four DOFs: four DC motors for left and right vergence, and a common tilt supported by a common pan (figure 2). It can perform motions with peak velocity and acceleration comparable to human capabilities(400 deg/s for vergence velocity).

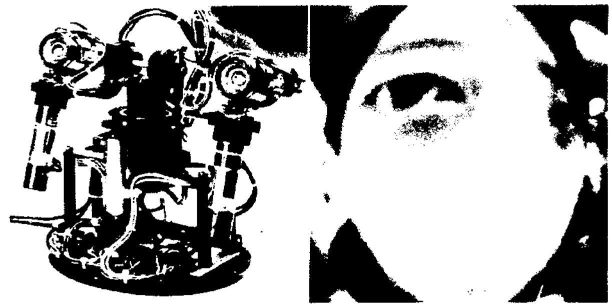


Figure 2: ESCHeR: a high performance stereo camera head(left) with foveated wide-angle lenses(right).

The most outstanding characteristic of ESCHeR however lies in its foveated wide-angle lenses[3],which exhibits a wide field of view($\approx$120 *deg*), along with a high resolution fovea ($\approx$20*pix/deg*) for facilitating both detection and close observation. The role of ESCHeR in this research is to exhibit a very wide field of view, with a high resolution target image, and smooth target tracking.

3 Competitive learning network

In this research, we use a competitive learning network[4]. Figure 3 is a model of a competitive learning network, implemented with an inhibition connection. Each unit inhibits all other units. The learning process is done according to the following formula:

$$A_j(t) = G\left(A_j(t-1) + \frac{c}{M}A_j(t-1) - \frac{d}{M}\sum_{i \neq j} A_j(t-1)\right) \tag{1}$$

where t is a time(for example, $t = 0, 1, 2, ...$), $A_j(t)$ is an activity level of each unit on the time t. c and d are the small constants ($\ll$ 1.0), M is the summation of the activity level of all the units in the competitive layer, G is a transmission function which has 0 and 1 as the limited value(gain is 1). In this model, time goes stepwise. The function G has a form indicated in figure 4.

The role of competitive learning network in this research, is suppressing noisy flow, and making principal flows concrete around the center part of the object.

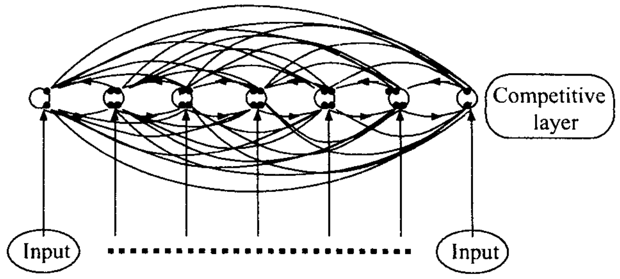


Figure 3: Competitive learning network

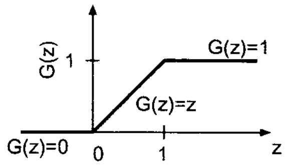


Figure 4: The ramp function G(z)

4 Image discrimination by PCA

In order to find a hand, we use some template images(of course, shot by ESCHeR) and PCA[5].

- Make matrix X from intensity of template images.

$$X = [\mathbf{x}_1\mathbf{x}_2\mathbf{x}_3 \dots \mathbf{x}_{n_X}] \tag{2}$$

where $\mathbf{x}_i = (x_1x_2\dots x_m)^T$ is a vector of a hand image.

- Reduce dimension of template images by PCA. And memorize translation matrix A.

$$Y = XA \tag{3}$$

- Process the images(W) which can be discriminated by matrix A, and discriminate by the distance from average of template images.

$$V = WA \tag{4}$$

$$d = \sqrt{(\beta_1 - \alpha_1)^2 + (\beta_2 - \alpha_2)^2 + \cdots + (\beta_{n_Y} - \alpha_{n_Y})^2} \tag{5}$$

where α_i is an average of each column of matrix Y, and β_i is an average of each column of matrix V.

The size of template images are 10[*pixel*] x 10[*pixel*], If we prepared 80 template images, we can make 100 x 80 matrix. The images which detected as moving objects will be reduced to 10[*pixel*] x 10[*pixel*] size before discrimination. And we use 25 percent of the obtained principal components to discriminate.

In order to show the role of PCA, we describe some figures. Figure 5 is template images of a hand.

Hand forms are, peace sign, closed, "*Inaka Choki*" [1], and opened. We prepared 20 template images on each form. The size of these images are 32[*pixel*] x 32[*pixel*].

Figure 6 is the images of the principal components obtained by PCA. These images are sorted from first to last principal component, from upper-left to bottom-right. The higher principal components have almost no information. To discriminate, we use the first two rows in this figure. From these figures, we can see that PCA has succeeded extracting form characteristics of hand images. For example, the 1st principal component is an opened hand, and the 2nd principal component is a closed hand.

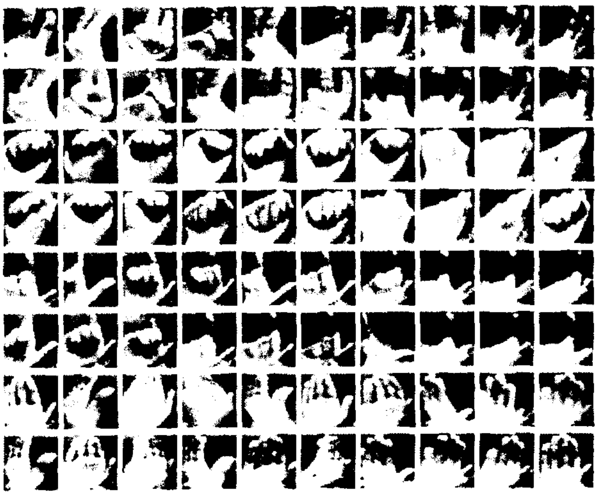


Figure 5: Template images of hand

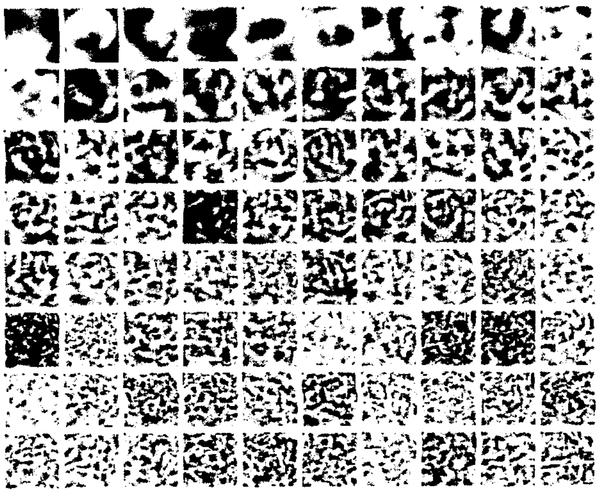


Figure 6: The images of principal components

5 Experiment and results

We describe a result of simulation , to show how to implement a system in recognizing multiple targets.

The simulation is summarized as follows:

1. Prepare a pair of source images captured by ESCHeR that are contained in a continuous frame.
2. Calculate optical flow and each flow power with Lucas and Kanade method.
3. Input each flow power into the competitive learning network.
4. Calculate the averages of the flow vector in each block of flows. And record left, top, right and bottom points of each block.
5. Calculate gazing direction for each detected object.
6. Make ESCHeR saccade to see each object vividly.
7. Crop object images, and compare with some template images by PCA, etc.

At this time, we have omitted process 6, and implemented processes 1 to 5, and also 7. Accordingly, at process 7, we used detected images from process 5 instead of the obtained images for process 6. There are some ways to crop images, we cropped images rectangular. To make ESCHeR pursuit a hand, we implemented a masking process as the next step. In this process, the source image is masked, except for the part which is most similar to a hand, based on the result of PCA.

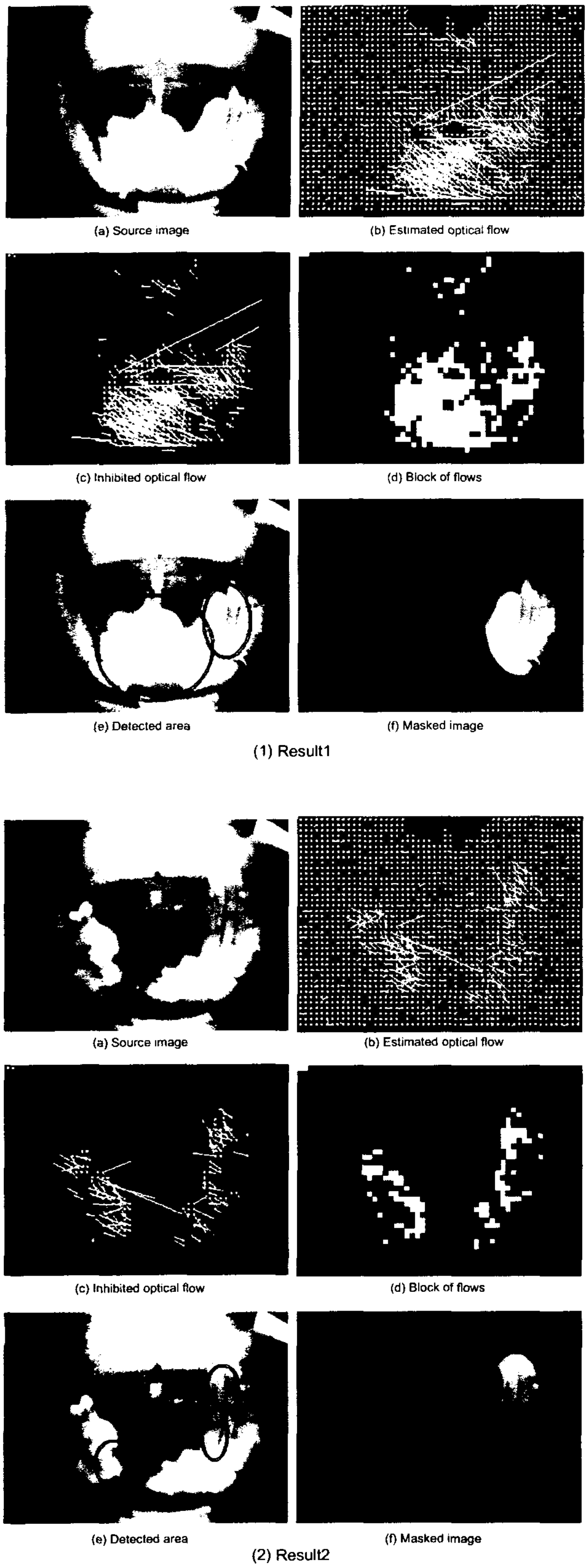
Figure 7 is a result of the object detecting simulation. Figure (a) is the source image. The source images are scenes of a hand and a rubik-cube going up and down. (b) indicates the estimated optical flow. From this figure, we can see that estimated flow field contains many noisy flow. (c) indicates the inhibited flow by competitive learning. From this figure, we can see that many minor flows are inhibited by competitive learning. (d) indicates the block of flows. This figure indicates flow generating points only. Points' intensity are the same, although the flow powers are different. The value c and d of a competitive learning are 0.7 and 0.015. The learning is done for 50 times on each frame. (e) indicates the detected areas as a moving object. Circles in this figure indicate detected area. (f) is the masked image except for the part which is most similar to hand. Judge of the similarity is based on the result of PCA.

From the result, we can understand that this system detected a hand and a rubik-cube as moving objects. From the figure of masked images, we also understand this system succeeded in detecting the location of a hand.

6 Conclusion and future work

In the process of the whole system, we implemented the part of processing flow, detecting objects, and discrimination of detected objects. By this system, foveated

[1] state that thumb and index finger stands



(a) Source image (b) Estimated optical flow

(c) Inhibited optical flow (d) Block of flows

(e) Detected area (f) Masked image

(1) Result1

(a) Source image (b) Estimated optical flow

(c) Inhibited optical flow (d) Block of flows

(e) Detected area (f) Masked image

(2) Result2

Figure 7: Result of experiment: (a) is a source image, and (b) indicates the estimated optical flow. (c) indicates the inhibited optical flow by competitive learning, and (d) indicates the flow generating points. (e) indicates the detected areas as a moving object, and (f) is the masked image except for the part which is most similar to hand.

vision is able to find the moving object, not only at the center of field, but also at other position. Our suggestion succeeded to segment motion of some objects, and discriminate a hand and another object.

There are some future works.

- Putting to the real machine and the real time processing
  We have experiment without real time processing at this time, thus having a real time experiment with real machine is the one of the future works. Especially, we have to realize the process 6 (see the section '**Experiment and results**'). We will also think about processing speed, accuracy, and so on.

- Tuning of competitive learning
  At this time, the parameters of competitive learning are determined experimentally. In future, it might be improved to determine the parameters systematically.

- Application of ICA
  ICA (Independent Component Analysis) is known as one of multivariate analyses. This method is well known as a method of measurement of brain such as MEG or EEG. By this method, we expect that the accuracy of discrimination will be improved. We think about using ICA in addition to PCA.

In the future, this research will be developed to recognize human actions.

References

[1] Sebastein Rougeaux : "Real-Time Active Vision for Versatile Interaction" ,Philosophical Doctor thesis(1999).

[2] Y.Kuniyoshi, N.Kita, S.Rougeaux, and T.Suehiro : "Active stereo vision system with foveated wide angle lenses", *2nd Asian Conference on Computer Vision, Singapore*, I:359–363, 1995.

[3] Y.Kuniyoshi, N.Kita, K.Sugimoto, S.Nakamura, and T.Suehiro : "A foveated wide angle lens for active vision", *IEEE International Conference on Robotics and Automation, Nagoya, Japan*,2:2982–2985, 1995.

[4] J.Dayhoff, H.Katsurai(translate): "Neural network architecture handbook", Morikita publication Co.,Ltd., 1992

[5] S.Arima, S.Ishimura : "The story of multivariate analysis", Tokyo Tosho Co.,Ltd., 1987