A Study of Visual Attention System

Based on Recognition Feedback for Image Sequence

Makoto Ito
Nagoya University
Nagoya City, Japan
itou@okuma.nuee.nagoya-u.ac.jp

Yoshikazu Yano
Nagoya University
Nagoya City, Japan
yano@soec.nagoya-u.ac.jp

Shigeru Okuma
Nagoya University
Nagoya City, Japan
okuma@okuma.nuee.nagoya-u.ac.jp

Abstract

The research focuses on approach for robot's vision system. We propose the selective visual attention system which realizes high-speed processing flexible to environmental change. In order to obtain visual attention point, proposed system extracts three kinds of visual features: still image feature, blinking feature and motion feature. Proposed system introduced intention maps to attend interested area. Intention maps are updated dynamically to represent the existence of attention target. As the result of experiments, the proposed system can confirm visual transition to the interest object.

1 Introduction

In recent years, many entertainment robots such as pet robots and humanoid robots have been developing. In order to work at unfamiliar places, they have to understand the surroundings, to recognize the objects and to obtain the information rapidly. Traditional scan-based vision systems have to process huge amount of data for these tasks and are not suitable for those robots with their poorer processor. Contrary, visual attention system can pick up the points to draw attention. When they are in unfamiliar places, they can find remarkable points to recognize. Meaningful gaze points will reduce recognition process which need huge amount of calculation power.

Many studies [1][2] have been made on visual attention system. The vision model proposed by Itti et al [1] builds on a biologically-plausible architecture. This system generates three kinds of “feature maps” from an input image. Feature maps show the amount of visual features: intensity, colors, and orientations. Combined feature maps determine a gaze point.

The existing models like the above select gaze point with visual feature. They are efficient to pick up a gaze point at the first sight, since some of targets which should be attend to have the strong visual feature. However, not all of targets have strong features. There are important targets with weak features. There are unnecessary targets to gaze with strong features either, such as light sources.

Therefore, we proposed visual attention system which will adapt to the environmental condition. Proposed system is designed for movie and has intention maps for each visual feature. Proposed system can select a gaze point using visual features (bottom-up approach) and features of purpose object (top-down approach).

2 Proposed System

The following Fig 1 represents the proposed system. Proposed system determines a gaze region as follows. This system extracts three visual features from input images and creates feature maps. Created feature maps are combined with each weight determined by corresponding intention maps. The combined map determines the visual attention region. Each intention map represents the intention of attention to important target which is constructed according to the result of recognition to the selected region.

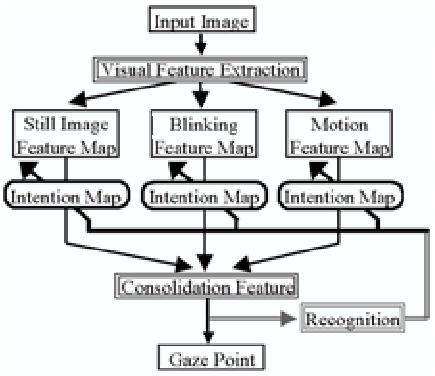

Fig 1. Proposed System

2.1 Visual feature map

Most of traditional systems such as [1] are applied to a still image, and don’t use motion information. In order to watch several targets such as stationary one and moving one, proposed system uses not only still image feature map but also motion feature map and blinking feature map.

- still image feature map

Propose system creates still image feature map just like [1], such as intensity feature, color feature, and orientation feature. The details of those features are mentioned in [1].

- motion feature map

Moving objects are existed at the robot surroundings. Moving objects are more dangerous for the robot. Robots should gaze at the moving object. Therefore, we introduced motion feature. The objects moving nearly or rapidly are more dangerous than objects moving far or slowly. The motion feature of those objects should be large value. Therefore, motion feature map is generated with the length of each optical flow vector.

- blinking feature map

Humans used to gaze at blinking object, so blink alarm is good way to call someone's attention. Blinking object doesn't generate optical flow, because it doesn't move. The blinking feature of the object blinking intermittently should have larger value than that which blinks once or twice. Therefore, blinking feature is generated with several of frame differences.

2.2 Intention map

Human interacts two processes in detecting the object. One is the process to select Region-of-Interest (ROI) in the retina, primary visual cortex and so on. The other is the process to recognize the selected region in parietal association area.

The existing studies showed that visual transition to the same image depends on task by physiology experiment [3]. The experimental results indicate that only visual features don't select ROI, that is humans select ROI changing visual feature gains depending on their experience. In other words, in order to select ROI, not only the bottom-up approach which selects ROI from input images deterministically but also the top-down approach based on feedback from recognition should be implemented.

Top-down approach enables the visual attention system to select regions using recognition results and its purpose to watch the target.

2.2.1 Intention map

We introduced intention maps in order to select an important region. Intention maps should be modified by recognition results. They are updated dynamically in order to represent the existence of interests. The value on the intention map represents the gain of corresponding visual feature. The intention map consists of two maps, Purpose Conformity Map (PCM) and Reference History Map (RHM). Therefore, intention map represents following distribution.

- How important the region is (record in PCM)
- How often the region referred (record in RHM)

In order to reduce referring the same region, intention map is calculated with eq(1).

(Intention Map)=(PCM) − (RHM) (1)

2.2.2 Purpose conformity map

PCM will help this attention system not to choose useless regions which unfortunately hold attractive visual features. This map also helps to make attention to significant regions with poor visual feature.

In order to explain the role of PCM, take pedestrian detection, for example. If the visual feature of the region in the sky has the largest amount, the visual attention system selects the sky region and the region is recognized. The system reduces the gain of the region, since the sky region is not the target for detection, that is, doesn't include a pedestrian. After that, even though the sky region have large amount of visual feature, visual attention system will choose not the useless sky region, but another useful part. On the other hand, once the system chooses the pedestrian region, the region will get the higher gain. Even though the pedestrian regions have smaller amount of visual feature, the higher gain of the region where pedestrian often appears will often help to attract the pedestrian. Since recognition results are marked on the map, well-learned map will understand the concept of pedestrian existence, such as 'a pedestrian exists on the road'. Thus, accumulated information expresses spatial concepts of purpose objects, and proposed system takes advantage of it.

Recognition results are accumulated into PCM. When selected region fits for the purpose, the value of corresponding region on PCM will increase (Fig 2 – point A). Since the gain of visual feature for corresponding region become higher, the system chooses the region as a gaze point more easily. On the other hand, when selected region is not an important part, the value will decrease (Fig 2 - point B). Since the gain of visual feature becomes lower, the corresponding region will be hardly chosen using neglected visual feature.

Recognition results are obtained not from a certain area around the gaze point, but from the object region indicated by gaze point. Recognition results should affect the object region in PCM. In order to estimate the object region, the object region is defined as follows.

Def1) feature value of pixels belonging to the object region is similar with each other.

Def2) the object region is estimated from each feature map respectively.

Snakes is the famous method to extract the object, however, this method takes to much calculation time. Proposed System need to reduce calculation amount and doesn't need high accuracy boundary for the gaze point determination. In this paper, rapid region tracking method is proposed.

Proposed method determined the same region by two basis: "distance from the recognized region (D(r))", "feature's similarity to the recognized region ($F_s(x,y)$)". At a certain point (x,y), attribution probability to the object region is smaller, when the region is farther. To represent this relationship, gaussian function is applied as a probability function D(r) where r is distance to the gaze point. The feature similarity to gaze point is evaluated at a certain point. When the point has the same property of the gaze point, similarity function $F_s(x,y)$ represent 1. Therefore, adequacy value A(x,y) which denotes a certain point (x,y) belongs to the object region according to the distance and the similarity of features, are shown in eq(2)

$A(x,y)=D(r)* F_s(x,y)$ (2)

The update value of PCM U(x,y) is denoted using R and A(x,y) in eq(3), where R is distributed recognition results according to contribute for the gaze point selection.

$U(x,y)=R* A(x,y)$ (3)

Using A(x,y) makes it possible to update largely the value of PCM in the same region as the recognition object. So PCMs express attention to objects.

Proposed system creates PCMs which is suitable for the surroundings dynamically without prior information of them. PCMs enable proposed system to select the region according to the task using surroundings understandings.

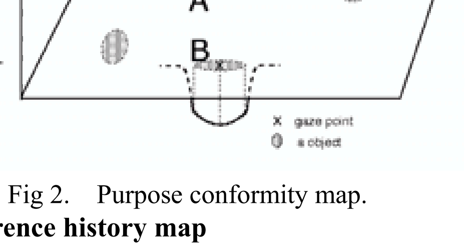

Fig 2. Purpose conformity map.

2.2.3 Reference history map

RHM shows the reference history to each attractive region. The system intends to select the same region again and again where a purpose object exists, because of PCM. However, the selected region need not to process so often, because useful information about the region was already obtained at first recognition. On the other hand, there are the regions which aren't selected in spite of existing a purpose object because the region doesn't have strong visual feature. In order to refer to those regions, we introduced RHM. RHM is updated by putting gaussian function on the selected region (Fig3 - point A). In order to gaze at the region again after some period, these reduction effects by gaussian function should be reduced with time (Fig 3 - point B). Therefore, RHM will help the attention system to choose the best attractive region, not every time but each some periods. This map also makes a chance to focus second (or more) attractive region. RHM makes it possible to gaze at various points, and enables proposed system to do visual transition like saccade.

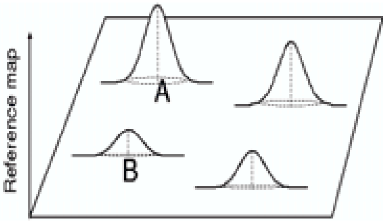

Fig 3. Reference history map.

2.3 Consolidation features

The proposed system creates intention maps for every visual feature. The updating value given from recognition system is distributed to each visual feature's intention map according to the amount of each visual feature. Each map is updated on the basis of distributed value. For example, when the system gazed at the object with strong still image feature and with weak motion and blinking feature, PCM of still image feature is updated with a large amount. Since each intention map is updated using assigned value independently with each other, some intention map is not changed when its visual feature does not affect the gaze point decision. Therefore, when a new object appears at the region where recognition system once recognized a useless region, the system can gaze at the region by the rest of visual features which didn't affect in previous gaze point decision. For instance, when the background region with strong still image feature is selected, the still image feature gains in the region get lower. After that, when moving object appeared in the region, the system can gaze at the moving object using gains of motion and blinking feature. This means the system realized attention which doesn't express one kind of spatial concepts of purpose objects.

Proposed system processes each visual feature in parallel. So we can easily add another feature to proposed system, such as distance to the object.

3 Experimental Results

3.1 Purpose and method

Detecting person with small calculation amount is in high demand for the entertainment robots. So we show the results of the person detection using proposed system. This system needs external recognition system for detecting a person. It may safely be assumed that at robot surroundings the moving object is a human. Recognition system treats the region as a human by one basis:"the amount of movement". The amount of movement is generated with the ratio of the number of the changing pixels to that of the whole pixels in the certain region obtained by frame difference. The value of the amount of movement is normalized so that the range is [-1, 1]. Recognition system gives the normalized value as evaluation value of Human Likeness.

Input Image was taken by stationary camera (SONY EVI- D100). Input scene is shown in Fig 4. Proposed system selected 3 visual attention points at every flame, which interval was about one second.

3.2 Express of attention

PCMs created by feedback from recognition system are shown in Fig 5,6,7, which is corresponding to still image feature, motion feature, and blinking feature. White regions in PCM are where the system detected a pedestrian and feature gain became high. Black regions are where the system didn't detect a pedestrian and feature gain became low. Gray regions are the regions which have been never selected as a gaze point, feature gain is not modified.

In this experiment, most parts in PCM for still image feature are rejected shown in Fig 5. There are a lot of remarkable feature points in the background. However, this system does not need to gaze those unsuitable points. Therefore, the gain in most parts which support the background area is decreased. Contrary PCM for motion and blinking feature are not modified in the background area shown in Fig 6 and 7, because these features don't recommend the gaze point. On the other hand, those features represents human's motion remarkably, so that the gain of area is increased.

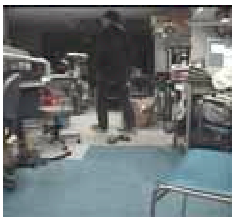

Fig 4. Input scene image.

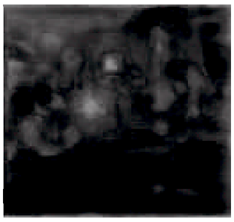

Fig 5. Created PCM for still image feature.

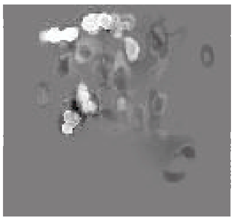

Fig 6. Created PCM for blinking feature.

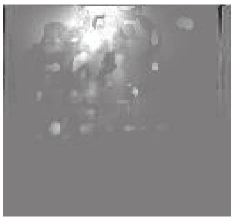

Fig 7. Created PCM for motion feature.

Fig 8 and 9 show the visual transition on scenes at same experimental as previous. The white line shows the visual transition.

Proposed system often gazed at pedestrians. However, when the display screen changed or the lamp lighted, the system has ability to draw attention not only to pedestrians, but also to display screen and to lamp, because of strong blinking feature (Fig 8). Those object aren't the target, so that proposed system decreases Purpose Conformity of those objects. Actually, we can confirm that PCM for blinking feature was small in the region of display screen and lamp (Fig 6). As a result, as often as the system gazed, the interest of the object is weak, proposed system makes it hard to make attention to display screen and so on (Fig 9). On the other hand, Purpose Conformity of target objects is enlarged. So proposed system makes it easy to make attention to pedestrians. It means that proposed system realize attention gazing at target objects.

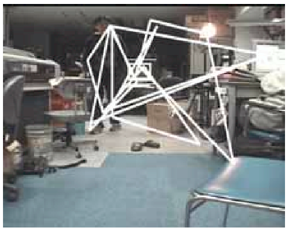

Fig 8. Results of visual transition before gain reduction.

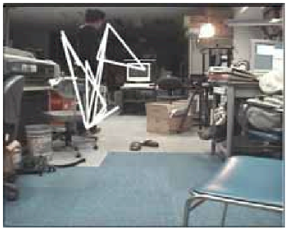

Fig 9. Results of visual transition after gain reduction.

The other experimental results obtained at outside and the staircase are shown in Figure 10 and 11. There are white region in PCM for still image feature, that is the gain of still image feature is increased. In this case, at the area where pedestrian are passing, PCMs for motion and for still image features get increase. Especially, the characteristic of the man with red clothes affected to a PCM for still image feature. PCM is created according to environment condition using recognition results. So proposed visual attention system can adapt to various environment condition.

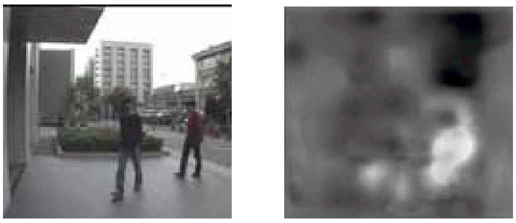

Fig 10. PCM for still image feature created in outside: scene image(left) and created PCM(right)

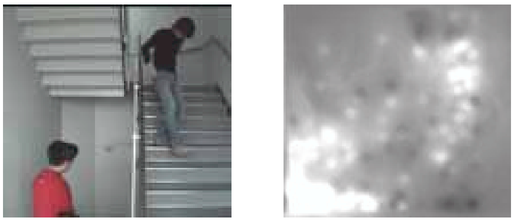

Fig 11. PCM for still image feature created in staircase: scene image(left) and created PCM(right)

3.3 Effect of visual attention with feedback

Table 1 shows the number of frames that a pedestrian appeared and that the system detected a pedestrian. The existing bottom-up system using only visual feature selected pedestrian region as gaze region in about 55% probability. The system using only visual feature gazed at the object with strong feature, so the system continued to select the unwanted object such as light-source. But proposed bottom-up and top-down system realizes attention because of using recognition feedback. As a result, proposed system detected pedestrians in about 95%, even though recognition system verifies only 3 points per frame which proposed system picked up from whole area. Proposed visual attention system learns the surrounding of vision system according to the given recognition results and finds some appropriate candidates for next recognition. This system reduces the amount of recognition process, therefore it is suitable for robot vision.

Table 1. The number of flames pedestrian is detected.

	the number of flames
a pedestrian appeared	65
a pedestrian is detected (Selected gaze points using only visual feature)	36
a pedestrian is detected (Selected using visual feature and intention map)	62

4 Conclusion

This paper proposed the visual attention system which does not need information about the environment and its tasks in advance, and picks up some appropriate targets to recognize. Intention maps help to gaze at various regions according to the task. Intention maps updated by recognition results represent attention. Proposed system can select a gaze point by top-down and bottom-up approach using intention maps. Experimental results show the effectiveness of this system.

This system calculates the visual features of the whole image by software. Therefore, it needs much times. These feature maps can be generated by hardware vision chips more rapidly. We will propose the system, which can select a gaze point at high speed, and recognition system according to purpose with software in order to maintain the flexibility of the design. We can extract feature at high speed by applying the technology of an existing vision chip. The frequency of recognition, which has slow processing speed, can decreased, proposed system will be applied to vision system that can't process too much large calculation amount.

References

[1] Itti L, Koch C, Niebur E, "A Model of Saliency-Based Visual Attention for Rapid Scene Analysis", *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol.20, no.11, pp1254-1259, 1998.

[2] S. Frintrop, A. Nuchter, H. Surmann, " Visual Attention for Object Recognition in Spatial 3D Data" 2nd International Workshop on Attention and Performance in Computational Vision (WAPCV 2004) , pp. 75-82, May 2004

[3] A.L.Yarbus, "Eye movements and vision" Plenum Press, New York, NY(1967)