A Visual Attention Model Using Depth Information from the Point of Gaze

Takahiro Ogawa, Motoyuki Ozeki, and Natsuki Oka
Kyoto Institute of Technology
Kyoto, Japan
ogawa@ii.is.kit.ac.jp, ozeki@kit.ac.jp

Abstract—We propose a visual attention model in which a novel feature map called the “depth-of-field map” is introduced. The depth-of-field map functions similar to the depth of field effect of human vision by enhancing the saliency of the regions near the point of gaze in the direction of depth and reducing the gaze movements between regions widely separated from each other. In this study, we first counted the relative frequency of three types of human gaze movements (movement to adjacent objects at a similar depth, movement to adjacent objects in the same two-dimensional direction, and movement to other objects) in three situations (matching, memorizing, and no assigned task). The result shows that the percentage of gaze movement to adjacent objects at the same depth is more than chance and the proportions of these three types of gaze movements vary according to the task. We then implement the proposed model using Microsoft Kinect for Windows and input the scenes from the human experiment into the model. The result shows an occurrence distribution similar to that of human gaze movement according to a given task. It indicates that the proposed model could bring robotic gaze movement closer to that of a human.

Keywords—visual attention, depth information, human-robot interaction.

I. INTRODUCTION

Computational models of visual attention, which have been studied in order to shed light on the human brain in the areas of computational neuroscience[1][2][3][4][5][6][7] and cognitive developmental robotics[8][9], have been applied to a wide range of communication robots in recent years[10][11][12][13][14]. Adopting a visual attention model for robots is useful to reduce the computational load on the visual process, to identify important objects in a complex scene, to share attention in communication with humans, etc. However, most of the traditional visual attention models use only a two-dimensional (2D) input image and cannot keep attention from moving between regions widely separated from each other by depth. This causes undesirable behavior of the robot, e.g., repeatedly moving gaze between a human face in front of the robot and a poster on a distant wall behind the human.

In humans, the gaze movement between regions separated by depth is prevented by the depth-of-focus limitation of eyes as well as top-down attentional control. Depth-of-focus limitation causes the regions far from the point of gaze to defocus and the saliencies\footnote{Saliency is a measure of conspicuity to draw the attention of the human/robot.} in these regions to be smaller. To directly provide this function for a robot, its camera would need to have a bright lens with a shallow depth of field. It also should have highly accurate and quick automatic focus so as to maintain focus on the intended target. At this time, such cameras are usually large, heavy, and expensive.

Meanwhile, the availability of commercial robots with a stereo sensor or range sensor has gradually increased[15]. Microsoft Kinect also has become widely used as a motion sensing input device for non-commercial robots. Kinect allows developers to inexpensively acquire scene depth information. We therefore aim to resolve the problem mentioned above by introducing depth information into the visual attention model. Although this approach departs from the biological model, it still allows many communication robots to realize human-like gaze movements at a behavioral level.

Several visual attention models using depth information have been proposed. Ouerhani et al. aimed to recreate the pop-out effect due to contrasts between the depth of objects and their neighborhoods[16]. In their method, stronger saliency was added to a region of depth that is different from its background based on a center-surround mechanism on the depth image. Penaloza et al. proposed a method based on an infant’s imitation behavior, which added stronger saliency to the region in front of detected humans[17]. These are justifiable attention models based on biological or clinical knowledge. However, the problem of gaze movement between regions separated by depth cannot be avoided.

In the area of object tracking, Ozeki et al. use a stereo camera to extract objects at a particular range of depth from a complex scene[18]. Maki et al. proposed a similar method in which the range of depth is dynamically changed based on the previous target depth, while Ozeki’s method statically defines the range of depth by workspace location[19]. These methods discriminate the target object from its background based on a center-surround mechanism on the depth image. Penaloza et al. proposed a method based on an infant’s imitation behavior, which added stronger saliency to the region in front of detected humans[17]. These are justifiable attention models based on biological or clinical knowledge. However, the problem of gaze movement between regions separated by depth cannot be avoided.

In contrast, we propose a novel visual attention model with a depth-of-field map, consisting of a conspicuity measure derived from scene depth relative to the point of gaze. Our method can decrease (control) the gaze movements between regions separated by depth. The depth-of-field map can be integrated into Itti’s model such that the measure of conspicuity...
is controllable and can be used in the same way as other feature maps.

The rest of the paper is organized as follows: Section 2 describes the experimentally determined occurrence distribution of three types of human gaze movements in three situations (matching, memorizing, and no assigned task). Section 3 briefly introduces Itti’s model and Ozeki’s model, which form the basis of our model, and then describes the proposed method. Finally, Section 4 shows an example of output from the proposed model and an experimental result obtained using the proposed model in the same simulations as the experiment of Section 2.

II. OBSERVATION OF HUMAN GAZE MOVEMENTS

To confirm whether the gaze behavior that our study aims to implement is really observed in the case of humans, we first investigated the assumptions listed below:

- Humans tend to move their gaze to objects similar in depth to the previous point of gaze as well as in the 2D direction of the field of view.
- The tendency of this gaze behavior varies according to the given task.

For this purpose, we prepared scenes with ten objects placed on a table. The objects were roughly lined up at 40 cm intervals in three rows, left to right, while the distance between each row was around 100 cm. However, as shown in Fig. 1, the nearest objects of a certain type appear to be objects in different rows but are in the same row. In order to eliminate the potential for bias with respect to the gaze path between a certain set of objects, the places of the objects were changed and some of the objects were swapped with other objects for each participant in the experiment.

We gathered 12 subjects (22 to 24 year-old students) for this experiment. Each subject put their chin on the chinrest of a gaze tracking system facing the table 1.5 m away from the nearest line of objects. The instructions were given as follows:

1) For the preparation, wait for a few moments with your chin on the chinrest. [no-task]
2) Find pairs of objects somewhat similar to each other, e.g., a canned drink and a cup are often used together, a book and toilet paper are both made of paper, etc. [matching-task]
3) Memorize each position of the objects [memorizing-task]
4) For the post-processing, wait for a few moments with your chin on the chinrest. [no-task]

This experiment consists of three tasks, including the no-task. After the experiment, we checked that each subject understood the tasks and performed them just as we intended by interview. The matching-task and memorizing-task continued until the subject’s gaze began to still. The average duration of those tasks for all subjects was around 50 s. In contrast, the periods of no-task lasted for 30 s each. The data of both no-tasks were brought together for analysis.

The data was analyzed as follows. For each task, we first found the attended positions where the subject stopped their gaze for more than 200 ms. As shown in Fig. 2, each gaze path between two attended positions was categorized into three types as follows:

- Movement to adjacent objects at the same depth. [depth-adjacent]
- Movement to adjacent objects in a similar 2D position. [2D-adjacent]
- Movement to objects not covered by the above two conditions. [others]

Figure 3 shows the occurrence distribution of the three types of gaze path in the three tasks. The results are averaged over all 12 subjects. Assuming that humans only randomly moved their gaze between objects, the proportion of depth-adjacent paths would be 26.7%, but all of the observed proportions of depth-adjacent were significantly higher than 26.7% (p < 0.05). This assumption may be extreme, but most existing visual attention models assume that all objects (regions) have the same chances of being gazed at if the
saliencies of the objects are same. The same arguments hold for the 2D-adjacent gaze paths. To address this issue, some existing models introduce a feature map that imitates a fovea central. This *fovea map*, however, causes the proportion of 2D-adjacent gaze paths to increase more than that of depth-adjacent gaze paths. The result of this experiment, which does not show such a distribution, suggests a need for a mechanism to increase the occurrences of depth-adjacent gaze paths.

We next check whether the occurrence distribution of the depth-adjacent gaze path varies according to task. As the result of multiple comparisons (Tukey’s Honesty Significant Difference test) shows, a significant difference was found in the proportions of the depth-adjacent gaze paths between the matching-task and memorizing-task ($p < 0.05$). This suggests that the mechanism for depth-adjacent gaze movement should have a top-down control interface to change its behavior.

### III. Proposed Method

Taking into account the results in Section 2, we propose a novel visual attention model that can simulate the depth of field effect of human eyes and recreate the occurrence distribution observed in the previous experiment. In this section, we briefly explain Itti’s model and Ozeki’s model, which are the basis of our model, and then describe the proposed method.

The visual attention of a human is divided into two types: the first is bottom-up attention, which is a stimulus-driven mechanism, while the other is top-down attention, which is a knowledge-driven mechanism that modulates the bottom-up attention. It is generally accepted that top-down attention is further divided into at least three categories: feature-based, spatial-based, and object-based.

One of the most well-known computational models of bottom-up attention is the saliency map model [20]. This model first calculates several feature maps based on the difference of an elementary feature (intensity, color, etc.) from its surroundings. The feature map is implemented as an image, each pixel of which represents the measure of conspicuity corresponding to the feature. A saliency map is then obtained by integrating the normalized feature maps into one global measure of conspicuity. The saliency map model implements a feature-based, top-down modulation by changing the weights of individual feature maps. Top-down modulation is an essential function to adapt the attentional behavior according to the given situation or task.

Ozeki et al. proposed a derived model from Itti’s model by adding a particle filter on the saliency map as the attentional distribution on the field of view [21]. The outline of this model is shown in Fig. 4. The particle filter provides a spatial-based, top-down modulation, such as a searchlight. The attended region has a flexible form and moves continuously over the field of view. By this characteristic, the point of gaze, which is obtained as the mode of the attentional distribution, tends to move the location of the (apparently) neighboring region in the field of view, compared with that of Itti’s model.

In the proposed method in this paper, two other feature maps are added to Ozeki’s model: the depth-of-field map and the fovea map. Each pixel of the depth-of-field map represents a measure of conspicuity defined by the reciprocal of its $z$-axis distance from the current point of gaze. Adding this map encourages the gaze to move to the regions that have a depth close to the current point of gaze. The depth-of-field map has a parameter, called depth-of-focus, for top-down attentional control. For instance, lower depth-of-focus decreases the saliencies of the regions that have a longer $z$-axis distance from the point of gaze. Figure 5 outlines the generation of the depth-of-field map.

To obtain the fovea map, the difference of the 2D distance from the point of gaze is used instead of the $z$-axis difference. The remaining procedure is similar to that of the depth-of-field map. Since the fovea map, which is used in some existing attention models, is not the main topic this paper, we do not mention the fovea map in detail in the remaining part.

For the experiment in Section 4, we implemented the
proposed system using a Microsoft Kinect for Windows. The concrete procedure to obtain the depth-of-field map is as follows:

1) Obtain a point of gaze from the color image using Ozeki’s method.
2) Calculate the z-axis difference between the point of gaze and every pixel of the depth image. The pixels that have z-axis difference beyond the range of -2600 to 2600 (corresponding to approximate -2.6 m to 2.6 m in real space) are set to zero saliency.
3) Scale the differences of depth by a factor of 3/2600 to convert the depth difference to a measure of conspicuity.
4) Determine the saliency of each pixel on the depth-of-field map as shown in Fig. 6.

The parameter, depth-of-focus, represents the variance of a normal distribution (Fig. 6). For example, a smaller depth-of-focus narrows the base of the normal distribution. This makes the robot concentrate its search in regions with depth similar to the current point of gaze. In this way, the depth-of-focus functions similar to the depth of focus of a photographic camera.

IV. Experiment

A. Demonstration

Figure 7 shows example outputs of the proposed method and a conventional method (a saliency map only with a fovea map). Figure 7(a) is an input image from the Kinect. In this scene, a person holds a picture, the saliency of which is less than that of the two pictures behind the person. The rectangle mark on the person’s face represents the current point of gaze.

Figure 7(b) is the output of a conventional method using only a saliency map with the fovea map. The saliency map is calculated regardless of the depth of the current point of gaze and the most salient regions are the pictures behind the person. In this case, the point of gaze will move to one of the background pictures next. In the case of a human viewer, however, the background pictures would have defocus blur because they are separated from the current point of gaze by depth. The previous method cannot take into account this issue as long as the background pictures show up as clearly in the input image as shown in Fig. 7(a).

Figure 7(c) is the output of the proposed method. We can see that saliency of the person’s body is almost equal to that of
the background pictures. The saliency of the section of picture held by the person is higher than the background pictures. In this case, the next position of the point of gaze will move to this picture. The gaze movement can be controlled so that it either continues to center on the face or moves to the background pictures by adjusting the depth-of-focus parameter. This result suggests that the depth-of-field map can simulate human eyes in terms of a depth of field effect using an ordinary camera.

B. Comparison with the Human Gaze Movements

Finally, we confirm that the proposed model can recreate the occurrence distribution (proportion of three types of gaze path) of the human gaze movement in Section 2. The input datasets consisted of six sets of a color video and depth video, captured by a Kinect located 2.0 m away from the nearest line of objects\(^3\). The duration of the input video was 60 s. For every dataset, each gaze path was categorized into one of the three types (depth-adjacent, 2D-adjacent, and others) in the same way as the human gaze movements. This process was repeated several times with various depth-of-focus settings until the closest results to those of humans were found. Meanwhile, the other parameters were fixed in the experiment.

Figure 8 shows the occurrence distribution of the three types of gaze path for three depth-of-focus values. Each graph represents an average for the six datasets. The line next to each graph represents the corresponding result of the human gaze movement shown in Fig. 3. We can see that the results obtained by the proposed model fit well with the results of human. Although we gave top-down control to the depth-of-field map as opposed to the fovea map, the proportions of the three types of gaze path are similar to those of humans.

As previously mentioned, smaller depth-of-focus leads the gaze to the regions with closer depth to the current point of gaze. In the case of this experiment, smaller depth-of-focus produces a gaze behavior similar to a sideways scan of the objects. Adopting this analogy to the experimental results of human, we can guess that the subjects took the strategy to scan the object sideways in the memorizing-task while they took a strategy to track objects that are three-dimensionally closer to each other in the matching-task.

\(^3\)The difference from the location of the chinrest derives from the difference of the angle of view.

V. CONCLUSION

In this paper, we propose a novel visual attention model with the depth-of-field map that enhances the saliencies of regions similar in depth to the current point of gaze and can reduce gaze movements between regions widely separated from each other. The experimental results show that the proposed model simulates the depth of field effect of human eyes and recreates the occurrence distribution observed in investigation of the human gaze movements. We verified the proposed method in just three tasks that do not always correspond to actual robot’s ones. However, the aim of the experiment is to show the need for top-down modulation on the gaze movement in the direction of depth, and our model has this function.

The proposed model is not biologically faithful, but it can give a depth of field effect to a robot’s eye through a conventional computational model for visual attention. This allows developers to use a reasonable sensor such as Kinect instead of an expensive camera with accurate and quick automatic-focus control. This approach also provides a comprehensible control interface for the range of the depth of field in much the same way as other top-down attentional controls.

We anticipate that our method will help to realize a robot capable of paying attention to objects similar to humans. For this purpose, the depth-of-focus must be changed according to the given situation or task. As with another parameters of existing models, an appropriate level of the depth-of-focus is
usually determined by the robot developer by experience or machine learning. We are now developing a robot who helps human in a desktop manipulation such as cooking, DIY, etc. In this case, when the robot guides the user the depth-of-focus is set for 0.2 and when the robot are looking around for some users the parameter is set for 0.3.

REFERENCES


