COMP 150 Project Proposal

Cassie Collins, Raina Galbiati, Doo-yun Her

Abstract. This research addresses the challenges of developing a robot attention model capable of incorporating human social cues. Humans, as incredibly social creatures, supplement their bottom-up and top-down attentional mechanisms with social understanding when judging the saliency of objects in their visual field. A crucial skill that contributes to this social understanding is joint attention, or the ability to direct one's attention towards the object of another's attention. Much of the previous work in the area of creating an attentional model that accounts for human-robot interactions has focused on implementing either a bottom-up attentional model or a top-down attentional model. We propose implementing a model that uses a top-down gaze-tracking paradigm supported by reinforcement learning in order to initially identify the observer's fixation points. Additionally, a bottom-up attentional model will supplement the top-down learning algorithm to help the robot accurately identify the most salient object as the one the observer is gazing at.

1 Introduction

In order to function in a world rich with information, humans selectively attend to only the stimuli they identify as most salient in their environments. Visual information is gated for saliency according to the dynamic interaction between two attentional mechanisms: stimulus-driven processing and goal-driven processing.[1] Stimulus driven processing is often referred to as "bottom-up" processing because it involves analyzing the low-level visual features of the scene such as color, edges, contrast, and motion. Bottom-up processing can result in rapid, involuntary shifts of attention to visually prominent features.[2] On the other hand, goal-driven processing, or "top-down" processing, involves a more conscious biasing of attention. Top-down processing relies on the observer's prior knowledge and expectations about the scene to direct attention.[12] For example, a human might be biased to attend to visual cues related to food when they are hungry.[2] In short, human bottom-up attention quickly and automatically identifies potentially visually salient objects in our environment while top-down attention helps us deliberately direct our attention to objects relevant to our current state.[2]

Moreover, humans do not only account for their own goals when directing their attention; they also react to social concerns by taking into account the attention and goals of other agents in their environment. Joint visual attention is the ability for two agents to coordinate their attention by directing it towards the same object. Humans uses non-verbal physical cues such as gaze to indicate the direction of their attention.[3]

Gaze following, or re-directing one's attention based on another's head or eye position, is crucial for the success of joint visual attention.[3] Research shows that gaze following develops in human infants as young as three months. Infants start by only re-directing their attention based on head direction cues but progress to considering eye direction cues when as well.[4]

A big problem still facing the field of developmental robotics is getting robots to direct their attention according to their ongoing, dynamic interactions with humans. If joint human-robot interactions are ever to appear natural and smooth, robots must account not only for bottom-down visual features and their own goals, but also use top-down processing to account for the goals of humans in the environment.[5] In other words, robots must correctly identify the most salient object in a scene as the one attended to by a human. Moreover, robots must complete all this serial processing and gating of incoming sensory information for saliency in a timely manner.[6]

Robots have been built with different kinds of gaze-tracking methods to try and help solve the problem of joint attention. Some methods use sophisticated combinations of tracking eye movements, head-orientation and facial features. However, the process of gaze fixation identification has been shown to be a simple yet effective method of gaze-tracking that utilizes only eye-tracking. Human eye movements consist of fixations, or times when the eyes pause over regions of interests, and saccades, or rapid eye movements between fixations. Gaze fixation identification involves separating fixations from saccades, removing the saccades, and then collapsing the fixations into a single representative sequence.[7]

We propose a scenario in which a robot simulator is presented as input a video of a scene of a human in a room with two objects. One object has low-level features that are more visually salient than the other object. Both objects are of equal relevance to the robot's own goals, but the human's gaze is directed towards one object, indicating it as more salient. A model of bottom-up attention processing alone would naturally direct the robot's attention to the more visually salient object, although a human would use their social understanding to direct their attention towards the object the observer's gaze was directed at.[3] Since robots lack this human social understanding, we propose the implementation of a gaze-following paradigm with Reinforcement Learning (RL) and the support of a bottom-up attentional model. RL has been proposed as a viable technique for learning top-down visual attention that promotes joint attention with humans.[6] The robot will first use the gaze-following paradigm suggested by Salvucci, Dario, and Goldberg (2000) in a top-down manner to identify points of human gaze fixation as potential regions of interest in the scene.[7] Then, it will further evaluate those regions of interest based on Itti, Koch, and Niebur's (1998) classic bottom-up attentional model.[8] Ultimately, we show that implementing a gaze-following RL paradigm in combination with a bottom-up attentional model can address the challenge of achieving joint human-robot attention

2 Related Work

Dr. Yukie Nagai developed a model to enable robots to learn to recognize objects and their movement by using a bottom-Up approach for visual attention.[9] Her model utilized two main mechanisms which dictate what the robot attends to. First, she reduced the complexity of the peripheral through retinal filtering. This created an image with the most clarity at the fixation point and blurred acuity elsewhere. This technique was inspired by human vision which is most clear in the fovea, and much less clear in the peripheral. The model then used the retinal image to create a saliency map. In past research, this involved computing the difference in acuity of a given pixel and its surroundings. However, this paper proposed a different method which used color, intensity, orientation, and flicker to calculate saliency. The resulting saliency map was used for stochastic attention selection which found an attentional point. This model allowed the robot to focus on a particular location while also enabling it the flexibility to shift its attention to a new location. However, her model is limited in that it identifies a single salient object but does not take the attention of an outside actor into consideration. This means that the model works in controlled test environment however might miss information in a more complex real-world environment with many salient objects. Additionally, as Kim et. al. recognizes, the model in this study does not factor the caregivers field of vision into the calculation of salient objects and as a result may evaluate ambiguous situations incorrectly.[10]

Salvucci and Goldberg explored various fixation identification algorithms.[7] In their research, they developed a model to classify eye tracking algorithms based on how they

use spatial and temporal information. Finally, they use this model to qualitatively evaluate and compare different algorithms. In particular, this research is applicable in how a robot should recognize fixations versus seccades, an essential skill for following a gaze. Thus, while this research greatly adds to our understanding of eye-tracking, leaves ample space for further research and application as it pertains to attention and robotics.

Kim et. al. presented a model for the development of robotic gaze following.[10] This approach first taught the robot about various salient objects semi-autonomously by showing them to the robot head one at a time. Next, the robot learned to recognize faces, discriminate between poses, and estimate approximate distance of the face from the robot. The idea was then that the robot would learn gaze direction information by interacting with its "caregiver" through reinforcement. One limitation to this study is that the robot can only recognize salient objects that it has been taught ahead of time. Furthermore, Kim et. al. did not utilize the eye direction when calculating the fixation point which would decrease the accuracy of focus attentional point.

Thus, our research seeks to develop a model that utilizes the models of previous studies and solves the issues and challenges identified above. The rest of the paper will go into greater depth on our technical approach and expected results.

3 Problem Formulation and Technical Approach

This project aims to explore the possibilities of the question: can we build a model of robot attention that replicates human attention by prioritizing top-down processing of object saliency that is supported by bottom-up attention? In order to employ a potential solution, our main concern involves how to modulate existent attentional models of object saliency to incorporate human-based values of saliency using gaze estimation. To accomplish this, we will attempt to build three levels of attention: first, replication of a bottom-up attentional model supported in literature; next, construction of a top-down attentional process via gaze following of a human caregiver; finally, integration of the two separate components with prioritization of human gaze over object saliency. This third stage of implementation will represent the robot's ability to engage in joint attention with a human, which will eventually be refined through a reinforcement learning model to allow for adaptive use of this new skill.

3.1 Saliency Model of Attention

The bottom-up attentional model first presented by Itti et al. has been successfully applied by various researchers, including Nagai.[8] The source code for this model is available through the USC iLab CVS server and will be replicated as closely as possible, with room for modification when needed.

3.2 Gaze Following

Because of the limitations to this project given by constraints of time and resources, gaze tracking will be accomplished using the open-source software, OpenGazer.[11] OpenGazer will be used to estimate gaze in a set of training videos, where each video will show the human caregiver facing the camera (i.e., the robot) directly and shifting her gaze to one of six regions into which the video image is divided. The human will utilize dots drawn on a board as points to fixate on in order to maintain consistency of these regions. Position of head and eyes are used as calibration points to produce a vector which indicates direction of gaze, and from this, a 2-dimensional plot of many points representing location of gaze

Figure 1 Fig 1. Eye movement sampling from Privitera and Stark (2000).[12]

will be collected for each separate video. These data points can be overlaid with the image from the training video to visually depict where the human gaze is fixated, similar to Fig 1.[12]

However, a simple visualization of human gaze does not provide a discrete value of its true location, which consequently does not allow for mathematical comparison between human and robot attention. A dispersion-based algorithm called Dispersion-Threshold Identification (I-DT) will therefore be used to cluster the given data in order to identify points of fixation over n number of regions (n = 6 in this case), henceforth referred to as regions of interest (ROI). This algorithm utilizes the fact that fixation points, because of their low velocity, tend to cluster closely together. I-DT identifies fixations as groups of consecutive points within a particular dispersion, or maximum separation, and above a particular duration threshold. The eye track is scanned by a window of variable size that computes the dispersion of consecutive points, which is expanded as long as the dispersion of the points is lower than a certain parameter. This I-DT algorithm is based on Widdel's 1984 data reduction algorithm.[7]

3.3 Prioritizing Human Gaze

The two different attentional processes converge in our integrated method for identifying another human's object of interest, where the robot will be tasked with choosing which object of those given is most salient (i.e., of stronger interest to the human). Two objects will be placed between the robot and the human within separate regions of the screen space and are visible in the lower half of the video image. One object is large and a vivid red color, being the 'more salient' object; the other object is relatively smaller and a subdued gray, being the 'less salient' object. The human will shift her gaze from making 'eye-contact' with the robot (as in, looking directly at the camera) to then fixate on the less salient object. The robot should choose to first follow the gaze of the human to identify the ROI, then use the saliency map to identify the object that lies within that ROI as the object of interest.

An additional challenge can be implemented by placing both objects in the same ROI, therefore prompting the robot to use more precise methods of gaze estimation. This can be accomplished by utilization of 'sub-classifiers' trained within regions to allow tighter dispersions of fixation while also improving avoidance of misclassification between neighboring regions.

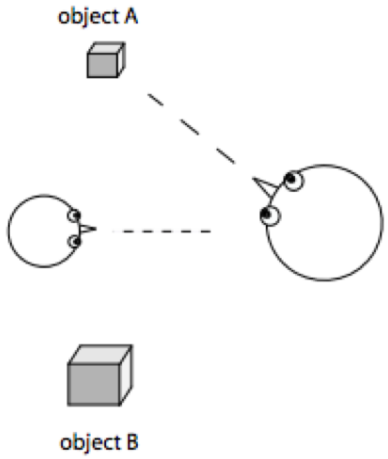

Figure 2 Fig 2. The setup of a similar prioritization task from Jasso et al. (2006).[3]

4 Expected Results and Experimental Validation

In the first stage of the project, using the bottom-up attentional model, we expect to obtain saliency maps of visual scenery in videos presented to the robot. Success is indicated by accurate object identification; i.e., the objects in the scene are depicted in the map as most salient on a density map.

Success is measured at the gaze-following level first by the robot's ability to identify the six ROIs presented by the human caregiver in the training videos, and subsequently by its ability to correctly identify the ROI of a particular interaction. The expected results are accuracy statistics which can be used to extend training and improve the system in later work, both in clustering methods as well as identification of a specific ROI.

Finally, prioritization is considered successful if the robot is able to ignore the more salient object to instead look to where the human is fixated. As the value of bottom-up cues are learned, the robot preferentially looks at object B, which is more salient. But as the robot learns to follow gaze, it starts to look more at object A, which is less salient but being looked at by the experimenter. An illustrated example of this is shown in Fig. 2 from H Jasso.[3] This system selection process would be modulated by a decision-making mechanism we will construct, which then prompts feedback/reward to the system for learning. A wide variety of this type of agent-critic structure has already been demonstrated by related research, and the exact methods we will use will be considered once clearer details of reinforcement learning parameters in this context are elucidated.

5 Timeline

- Oct 29: Begin replicating bottom-up attentional model
- Nov 11: Begin constructing gaze following mechanism
- Nov 19: Begin implementation of integrated system
- Dec 1: Finish up project
- Dec 5/7: Final presentation

References

1 H. E. Egeth and S. Yantis, "Visual attention: Control, representation, and time course," *Annual review of psychology* **48**(1), 269–297 (1997).

2 C. E. Connor, H. E. Egeth, and S. Yantis, "Visual attention: bottom-up versus top-down," *Current Biology* **14**(19), R850–R852 (2004).

3 H. Jasso, *A reinforcement learning model of gaze following*. PhD thesis, University of California, San Diego (2007).

4 M. Scaife and J. S. Bruner, "The capacity for joint visual attention in the infant," *Nature* **253**(5489), 265–266 (1975).

5 J.-D. Boucher, U. Pattacini, A. Lelong, *et al.*, "I reach faster when i see you look: gaze effects in human–human and human–robot face-to-face cooperation," *Frontiers in neurorobotics* **6** (2012).

6 A. Borji, M. N. Ahmadabadi, B. N. Araabi, *et al.*, "Online learning of task-driven object-based visual attention control," *Image and Vision Computing* **28**(7), 1130–1145 (2010).

7 D. D. Salvucci and J. H. Goldberg, "Identifying fixations and saccades in eye-tracking protocols," in *Proceedings of the 2000 symposium on Eye tracking research & applications*, 71–78, ACM (2000).

8 L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on pattern analysis and machine intelligence* **20**(11), 1254–1259 (1998).

9 Y. Nagai, "From bottom-up visual attention to robot action learning," in *Development and Learning, 2009. ICDL 2009. IEEE 8th International Conference on*, 1–6, IEEE (2009).

10 H. Kim, H. Jasso, G. Deák, *et al.*, "A robotic model of the development of gaze following," in *Development and Learning, 2008. ICDL 2008. 7th IEEE International Conference on*, 238–243, IEEE (2008).

11 P. Zielinski, "Opengazer: open-source gaze tracker for ordinary webcams." https://github.com/opengazer/OpenGazer (2008).

12 C. M. Privitera and L. W. Stark, "Algorithms for defining visual regions-of-interest: Comparison with eye fixations," *IEEE Transactions on pattern analysis and machine intelligence* **22**(9), 970–982 (2000).