Estimating Visual Focus of Attention in Dyadic Human-Robot Interaction for Planar Tasks

Caitlyn Clabaugh, Tejas Ram, and Maja Matarić

University of Southern California

Abstract. Visual focus of attention (VFOA) has been shown to be a strong indicator of human intent and engagement, and thus, an important feature to human-robot interaction (HRI). As a mature research problem in computer vision, VFOA recognition has been implemented in environments as complex as group meetings. In this paper, we define and leverage the relatively simple and highly-constrained environments of recent dyadic HRI for planar tasks, and validate the application of simple statistical methods for VFOA classification. We discuss the potential of VFOA in emerging HRI research and platforms.

1 Introduction

Humans use head pose and gaze direction as nonverbal cues to communicate intent, as well as to intentionally or circumstantially express attention and engagement. Visual focus of attention (VFOA) has been described and investigated by literature across many fields, cultivating a strong base of psychological, philosophical, computational, and multidisciplined perspectives. We investigate and apply the substantial research on VFOA recognition to an emerging trend of human-robot interaction (HRI) research. We leverage the strong constraints of dyadic HRI for tasks performed in a planar task space, an increasingly common structure in HRI research, such that high-level, literature-eulogized features of VFOA may be extracted automatically using simple computational methods and openly available tools.

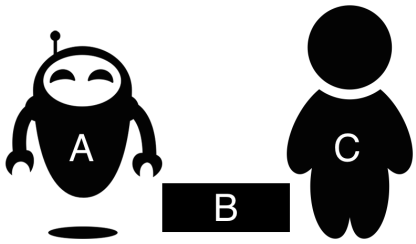

Fig. 1. Target dyadic, task-centered HRI application of VFOA classification: **(A)** robot, **(B)** planar task space, **(C)** human.

Langton et al. show that eye gaze direction [10], and particularly a combination of eye gaze and head pose [9], are important in determining the visual attention of an interaction partner. Stiefelhagen found head pose alone could accurately predict VFOA during group meetings [15]. Similar research efforts

have been spent to extend VFOA recognition to more complex, dynamic, and unstructured domains. While many real-world settings require the recognition of VFOA to be robust to multi-party and unstructured interactions, there are also a plethora of real-world problems in domains that are solely dyadic and environmentally well-structured that may benefit from VFOA analysis, especially in current HRI.

This paper considers VFOA in dyadic interactions, specifically HRI, given a front-facing camera and a planar task space (see Figure 1). We are inspired to explore this domain given the amounting research in dyadic, task-focused human-agent interaction. In this paradigm, the agent is positioned across from the participant, the task is performed in some planar space perpendicular and beneath both parties, and the surrounding environment is neutral and inconsequential to the interaction. Thus, VFOA may be classified as: at the interaction partner (i.e., a table-top robot), at the task space (i.e., a touch tablet), and away (i.e., any other target). We argue that this approach can be applied to other dyadic interactions given an equally static environment with few and sufficiently separated environmental targets of VFOA (e.g., a table and a projection screen).

This research takes an initial step toward VFOA classification in dyadic HRI for planar tasks, using simple and transparent methods to demonstrate growth potential. The main thrust of this paper is to define a novel, timely, and impactful application of VFOA in HRI research and discuss its potential. We support this overall goal with the following contributions: (1) an initial validation of simple VFOA classification in target domain, (2) a comparison of automated classification versus human annotation of VFOA, and (3) considerations for future inclusion and analysis of VFOA in target domain.

2 Related Work

VFOA is a key feature used in assessing and analyzing engagement and attention across psychology, HCI, and HRI literature. Intuitively, one may hypothesize that individuals focus their gaze upon stimuli with the most importance to them, and pose their heads accordingly. This intuition is backed by a plethora of research in psychology – Frischen et al. give an extensive review of research concerned with gaze behavior and its correlation with attention, emotional and mental states, and effects on the observer [5].

In HRI, it has been shown that robot speech and gaze modulates human gaze and comprehension [14] and that proper robot eye contact and attention develops joint attention important to human-robot collaboration [8]. We are motivated to explore automated VFOA recognition in human-robot collaborative or teaching tasks as this research shows that proper VFOA is both important to understanding humans' emotional and mental states as well as improving robot communication.

It is common that VFOA is analyzed post-study and often annotated by hand. However, the expansive literature on VFOA in computer vision and HCI

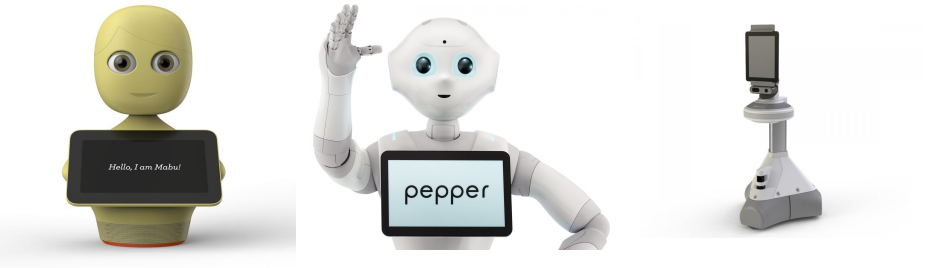

Fig. 2. HRI platforms with planar task spaces to benefit from VFOA classification, from left to right: Catalia Health's Mabu, Aldebaran's Pepper, and iRobot's Ava.

indicates that this expensive process may be unnecessary in common highly-constrained domains of HRI research. Automated recognition of VFOA is a mature topic in HCI research, as detailed in Murphy-Chutorian and Trivedi's review [12]. We show that methods that are relatively simple compared to the latest research in HCI are adequate for performing VFOA classification on highly-structured, dyadic HRI research.

A common structure of domain of research is currently emerging in HRI, defined by three key features: they are dyadic (i.e., one robot or agent and one human participant), focused on learning or collaborating to improve some skill or achieve some task, and take place in a planar task space. Additionally, touch tablets or computers are becoming an increasingly common feature of these interactions, enabling full-observability in task space. Figure 2 shows some recent platforms fitting of this domain: Catalia Health's Mabu [11], Aldebaran's Pepper [13], and iRobot's Ava [1]. Han gives an extensive summary of robot-aided learning research that includes many of these features and would benefit from VFOA analysis [7]. For this reason, we find that our proposed interaction and environmental constraints that simplify VFOA classification are reasonable to the field of HRI at present.

3 Approach

In this section, we describe our initial approach to VFOA classification for a specific case of the general domain of dyadic HRI for planar tasks. We also outline requirements and potential that are general to this research problem.

3.1 Study Design

As a general structure, we are inspired to look at VFOA classification in study designs containing two key features, illustrated in Figure 1, dyadic and with a planar task space. As explained in Related Work, these constraints are reasonable given the current direction of HRI research and beneficial to minimizing the complexity of VFOA classification.

Fig. 3. VFOA of child participant: at robot, at tablet, away; and the HRI setup.

We specifically consider a dataset of dyadic, human-child interactions focused on learning preschool mathematics [4]. The child participants, aged 4, interact with a socially assistive robot peer for about 20 minutes. This dataset contains features that match our problem's constraints: it is a dyadic interaction between a robot and a child with a planar task space and observable learning and performance. A high-definition camera faces the child participant (see Figure 3) such that estimates of eye gaze and head pose may be extracted using existing libraries (see Section 3.2). However, it is important to observe that in this dataset, the planar task space is not ideally perpendicular to the participant and robot, making this a challenging use case; in our discussion, we note the adverse effects of this on our classification and possible avoidance methods in future work.

3.2 Head Pose Extraction

For each participant video, we extract head pose as a key feature indicative of VFOA, as shown by HRI the research described in both the Introduction and Related Work. We used the Cambridge Face Tracker (CLM-framework) [3], a well-referenced and openly available library that implements head pose extraction. We use the automatically extracted roll, pitch, yaw of the child participants' heads in our final classification. We deemed it important to use tools accessible to other researchers to demonstrate feasibility and encourage pursuit of this research problem.

3.3 VFOA Classification

We frame this initial approach to VFOA as a supervised multi-class classification problem (see Tsoumakas and Katakis for an overview [16]). We use the human annotations (defined as at robot, at tablet, or away) as the labels of the three possible classes of each observation. The annotations were made by two trained human annotators using ELAN [17] with an inter-rater agreement of $k = 0.62$. For features of attributed, we consider head pose (defined as roll, pitch, yaw) coupled with the participants' interaction with the touch tablet. We limit our observations to each second of interaction, taking the average of each feature and label for all frames within that second.

Dataset	Percent Correct	Mean Absolute Error	Area Under ROC
Participant 1	60.78	0.39	0.74
Participant 2	71.37	0.38	0.60
Participant 3	67.90	0.38	0.67
All Participants	66.90	0.39	0.61

Fig. 4. Results of VFOA classification using multi-class logistics classifier on each participants' and participants combine datasets.

To further demonstrate the growth potential of this key research problem and encourage researchers unfamiliar with complex supervised machine learning, we limited our search for supervised classification algorithms to those well-established in the field and accessible through open source libraries. We implemented a simple logistic regression multi-class classifier [2]. This method is accessible through easy-to-use research tools such as Weka [6].

4 Evaluation

We trained our simple multi-class classifier using 10-fold cross-validation. Figure 4 gives the results of VFOA classification for each participant and all participants combined. We include the percentage correct, mean absolute error, as well as the area under the ROC curve for each set of data. We observe that our simple approach performs modestly on all datasets; however, we note that it performs moderately well on Participant 1 with 0.74 area under the ROC curve (where 0.5 is chance), indicating that personalized classification of VFOA may be more robust. Additionally, we note that the training and classification of VFOA using our approach takes two minutes per hour of data as compared to the 20 minutes per minute of data of human-annotation time.

5 Discussion & Future Work

The results from our initial and limited approach shows promise of VFOA classification to which the structure of current dyadic HRI research lends itself. Head pose as well as other features of interaction, such as tablet touch in our case, can be combined to predict VFOA without the intrusion of eye trackers or expense of human data annotation. We observe that personalized VFOA classification may perform better. Additionally, classification may perform better if there were a greater difference in angle of the task space and angle to the robot, unlike in the challenging dataset we present in this paper.

In future work, we will expand this analysis to include more participants. We are also interested in correlating VFOA classification with other features of interaction such as performance. It may also be of interest to analyze VFOA in relation to semi-static features of participants' personalities and learning styles. Another expansion of this work will be the application of VFOA to other datasets,

particularly datasets with more perpendicular planar task spaces. Lastly, the seminal goal of this research is to implement a open-sourced end-to-end VFOA extraction library and outline the optimal parameters for dyadic, task-focused HRI setups interested in using VFOA. We plan to do this through a combination of real-world dataset validations, like the one presented in this paper, and generative geometric models of human head pose relative to possible HRI setups.

References

1. Ava. http://www.irobot.com/For-Business/Platform-Opportunities.aspx, 2015.
2. Dankmar Böhning. Multinomial logistic regression algorithm. *Annals of the Institute of Statistical Mathematics*, 44(1):197–200, 1992.
3. P.W.D. Charles. Cambridge face tracker (clm-framework). https://github.com/TadasBaltrusaitis/CLM-framework, 2015.
4. Caitlyn Clabaugh, Gisele Ragusa, Fei Sha, and Maja Matari'c. Designing a socially assistive robot for personalized number concepts learning in preschool children. In *Fifth International Conference on Development and Learning and on Epigenetic Robotics (ICDL-EPIROB*, Providence, RI, August 2015.
5. Alexandra Frischen, Andrew P Bayliss, and Steven P Tipper. Gaze cueing of attention: visual attention, social cognition, and individual differences. *Psychological bulletin*, 133(4):694, 2007.
6. Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten. The weka data mining software: an update. *ACM SIGKDD explorations newsletter*, 11(1):10–18, 2009.
7. Jeonghye Han. *Robot-aided learning and r-learning services*. INTECH Open Access Publisher, 2010.
8. Michita Imai, Tetsuo Ono, and Hiroshi Ishiguro. Physical relation and expression: Joint attention for human-robot interaction. *Industrial Electronics, IEEE Transactions on*, 50(4):636–643, 2003.
9. Stephen RH Langton, Helen Honeyman, and Emma Tessler. The influence of head contour and nose angle on the perception of eye-gaze direction. *Perception & psychophysics*, 66(5):752–771, 2004.
10. Stephen RH Langton, Roger J Watt, and Vicki Bruce. Do the eyes have it? cues to the direction of social attention. *Trends in cognitive sciences*, 4(2):50–59, 2000.
11. Mabu. http://www.cataliahealth.com/about-us/, 2015.
12. Erik Murphy-Chutorian and Mohan Manubhai Trivedi. Head pose estimation in computer vision: A survey. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(4):607–626, 2009.
13. Pepper. https://www.aldebaran.com/en/a-robots/who-is-pepper, 2015.
14. Maria Staudte and Matthew W Crocker. Visual attention in spoken human-robot interaction. In *Proceedings of the 4th ACM/IEEE international conference on Human robot interaction*, pages 77–84. ACM, 2009.
15. Rainer Stiefelhagen and Jie Zhu. Head orientation and gaze direction in meetings. In *CHI'02 Extended Abstracts on Human Factors in Computing Systems*, pages 858–859. ACM, 2002.
16. Grigorios Tsoumakas and Ioannis Katakis. Multi-label classification: An overview. *Dept. of Informatics, Aristotle University of Thessaloniki, Greece*, 2006.
17. Peter Wittenburg, Hennie Brugman, Albert Russel, Alex Klassmann, and Han Sloetjes. Elan: a professional framework for multimodality research. In *Proceedings of LREC*, volume 2006, page 5th, 2006.