Joint Attention Estimator

Wallace Lawson*
Anthony M. Harrison*
LT Eric S. Vorm
J. Gregory Trafton
U.S. Naval Research Laboratory
[ed.lawson,anthony.harrison,eric.vorm,greg.trafton]@nrl.navy.mil

ABSTRACT

Joint attention has been identified as a critical component of successful human machine teams. Teaching robots to develop awareness of human cues is an important first step towards attaining and maintaining joint attention. We present a joint attention estimator that creates many possible candidates for joint attention and chooses the most likely object based on a human teammate's hand cues. Our system works within natural human interaction time (< 3 seconds) and above 80% accuracy. Our joint attention estimator provides a meaningful step towards ensuring robots enable human social skills for successful human machine teaming.

KEYWORDS

human robot interaction, object detection, novel object finding

ACM Reference Format:
Wallace Lawson, Anthony M. Harrison, LT Eric S. Vorm, and J. Gregory Trafton. 2020. Joint Attention Estimator. In *Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction (HRI '20 Companion), March 23–26, 2020,Cambridge, United Kingdom.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3371382.3378247

1 INTRODUCTION

Joint attention, or looking at the same object as another, is a critical component to human interaction [14] and develops relatively early in human infants [4]. Some researchers have suggested that joint attention is strongly related to the ability to infer others' mental states, and represents a component of how people develop a theory of mind [2]. Joint attention in humans typically occurs through gaze [4], gesture [9], language [5], and handling [16]. HRI researchers have shown that when robots have joint attention with a human collaborator, they are perceived as more competent and more socially interactive [7].

Most robotic systems, however, struggle having joint attention with a human partner, particularly recognizing which object is being referenced by the human. Some robots use an eye-tracker to track a human's gaze (e.g., [12, 18]) or head direction [7, 8, 17].

Another common approach to help a robot identify which object a human teammate is referring to is to provide a simple environment with very few target objects. For example, Trafton et al. (2009) modeled a developmental experiment and provided only two objects for the robot to find. Imai et al. could identify a limited number of objects a person was looking at. De Silva et al. used three possible objects available for joint attention. Performing joint attention in a cluttered environment is especially difficult.

A third method is to perform joint attention off-line. For example, Marintson [10] had a human pick up objects while a mobile sensor collected data; offline selection of possible bounding boxes then occurred. Similarly, Narayanan et al. [11] created bounding boxes of objects held by humans after the interaction had occurred. Azagra [1] performed automatic segmentation in a very cluttered environment, but reached an accuracy of less than 25%.

This brief summary highlights the difficulty that today's robots have in identifying an object that a human partner references. If the object is well known to the robot, well specified, and there are only a few possible objects in the environment to choose from, current systems seem to be able to successfully perform joint attention with a human partner. If offline training can occur, some of these constraints can be relaxed, but performing joint attention with a human partner in a resaonable amount of human-like interaction time (3-5 seconds) in a non-pristine environment is very difficult.

We describe here a system we call the "Joint Attention Estimator" which represents our initial attempt to solve some of these problems. Our specific goal is to create a system that attains joint attention with a human teammate using natural gestures, within a natural human time scale. In our system:

- The human must be able to identify a single object out of a group.
- The object may or may not be known to the system *a priori*.
- The system performance must not take longer than 5 seconds (we think of this as "interaction time" where the user will not get frustrated waiting for the system to respond).
- The system must be able to deal with a cluttered environment or other objects in the scene.

2 JOINT ATTENTION ESTIMATOR

In order to determine joint attention, the robot must have a set of candidate objects. A naive approach for this is to use sliding window(s) to select candidate objects. Although this detects known and unknown object candidates, it also adds a computational bottleneck as it results in a large number of potential object candidates. Processing this requires substantially longer than typical human interaction time. Instead, we used a learned approach to find potential

*Author order was determined by Rock Paper Scissors [15]. DISTRIBUTION A: Approved for public release, distribution is unlimited.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
HRI '20 Companion, March 23–26, 2020, Cambridge, United Kingdom
© 2020 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-7057-8/20/03.
https://doi.org/10.1145/3371382.3378247

objects. In the object detector, YOLO, [13]), the image is subdivided into grid cells. Each grid cell has a bounding box and probability of a previously trained set of objects of interest being present. However, because our goal is to have joint attention with both unknown and known objects, we were able to adjust the probability that an object of interest is present. By lowering this threshold to 1e-2, YOLO generates a very large number of possible object candidates. Figure 1 shows both the advantages of this approach (many candidate objects being identified) and the negatives of this approach (too many objects to easily find a unique object of joint attention from the human teammate). Another advantage of this approach is that YOLO was created to be very fast, so it can typically find objects on an image within interaction speed (see the results for the full data).

To deal with the negative component of having so many possibilities (bounding boxes), we can use joint attention from the human teammate to help select a reasonable bounding box. In this case, "reasonable" means capturing the vast majority of the object. We use OpenPose [3] to provide joint attention information to identify a unique bounding box. Specifically, we use OpenPose to track the wrists of the human which can define the maximum bounds of the joint object at the time of the verbal information from the human. These maximum bounds are then used by the robot to filter and merge the numerous elements in the environment taken in by the candidate object detector. This approach allows us to provide an estimation to the robot of a best guess of what the human wants the robot to attend to. This approach completely removes objects that are considered clutter by other systems and not between the person's hands. We combine the bounding boxes in between the human's hands using non-maximum suppression, a technique that filters proposals to find the most representative bounding box. This bounding box is very likely to contain the majority of the object for joint attention. An additional benefit to this approach is that we can identify an object for joint attention in multiple ways (e.g., the human can hold an object or put their hands on opposite sides of an object).

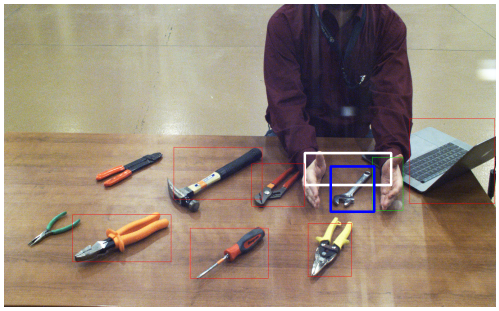

Figure 1: This figure shows what the robot sees when it is asked to perform joint attention when the human partner is surrounding a wrench. Notice the large number of candidate objects that are clearly not part of the human's joint attention. Each rectangle represents a candidate object, the best matching object selected through joint attention is shown in blue (a wrench), the region of interest shown in white.

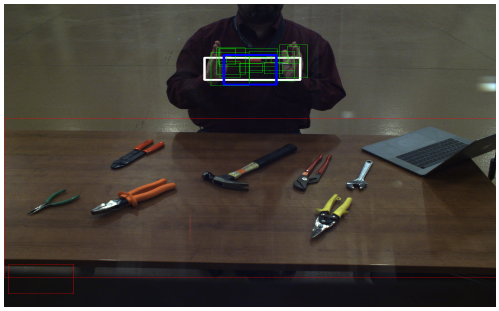

Figure 2: In this figure a human partner is holding a screwdriver. Green rectangles represent candidate objects, the best matching object selected through joint attention is shown in blue, the region of interest is shown in white. Also notice the screwdriver tip is not captured but neither are the hands, a reasonable trade-off in this case.

3 METHOD

Our testing environment consisted of a table with a cluttered background in a laboratory setting. Eight to twelve tools were scattered across the table during testing (see Figure 1). A human identified each of 10 objects to the robot in both of the following ways: (1) Holding a tool up for the robot or (2) providing a two handed gesture to the object. In both cases, the human labeled the object (i.e., "This is a screwdriver.") as the cue for the robot to engage in joint attention. Audio data was routed to a cloud speech recognition service provider. The robot was a DRC HUBO with a SCIPRR head [6] that contained a Point Grey Grasshopper 3. Processing is performed using an offboard Intel Core i7 with dual NVIDIA GPUs (for OpenPose and YOLO).

4 RESULTS

The system was able to correctly create a bounding box for 85% of the objects that the human identified for joint attention. Critically, identifying each object was very fast and clearly within interaction speed: Across all 20 trials, it took an average of 1.02 seconds to verbally request joint attention, 1.7 seconds for the speech processing to return from the cloud server, and .25 seconds to create and return the best bounding box. Thus, it took less than 3 seconds to identify a joint attention object with 85% accuracy in a cluttered environment.

Currently, the Joint Attention Estimator can only identify joint objects when the human uses two hands. We are currently working on other methods (e.g., gaze, pointing, single hand holding) to expand our interaction options.

ACKNOWLEDGMENTS

This work was supported by the Office of Naval Research (GT, WL). The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the U.S. Navy.

REFERENCES

[1] Pablo Azagra, Ana Cristina Murillo, Manuel Lopes, and Javier Civera. [n. d.]. Incremental Object Model Learning from Multimodal Human-Robot Interactions. ([n. d.]).

[2] Simon Baron-Cohen, Alan M Leslie, and Uta Frith. 1985. Does the autistic child have a "theory of mind"? *Cognition* 21, 1 (1985), 37–46.

[3] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2018. OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields. In *arXiv preprint arXiv:1812.08008*.

[4] Valerie Corkum and Chris Moore. 1995. Development of joint visual attention in infants. (1995).

[5] Peter F Dominey and Christelle Dodane. 2004. Indeterminacy in language acquisition: the role of child directed speech and joint attention. *Journal of Neurolinguistics* 17, 2-3 (2004), 121–145.

[6] Anthony M. Harrison, Wendy M. Xu, and J. Gregory Trafton. 2018. User-Centered Robot Head Design: A Sensing Computing Interaction Platform for Robotics Research (SCIPRR). In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction (HRI '18)*. ACM, New York, NY, USA, 215–223. https://doi.org/10.1145/3171221.3171283

[7] Chien-Ming Huang and Andrea L Thomaz. 2011. Effects of responding to, initiating and ensuring joint attention in human-robot interaction. In *2011 Ro-Man*. IEEE, 65–71.

[8] Michita Imai, Tetsuo Ono, and Hiroshi Ishiguro. 2003. Physical relation and expression: Joint attention for human-robot interaction. *IEEE Transactions on Industrial Electronics* 50, 4 (2003), 636–643.

[9] Connie Kasari, Marian Sigman, Peter Mundy, and Nurit Yirmiya. 1990. Affective sharing in the context of joint attention interactions of normal, autistic, and mentally retarded children. *Journal of autism and developmental disorders* 20, 1 (1990), 87–100.

[10] Eric Martinson. 2018. Interactive Training of Object Detection without ImageNet. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 1–9.

[11] Priya Narayanan, Magdalena D Bugajska, Wallace Lawson, and J Gregory Trafton. 2017. Impact of embodied training on object recognition. In *2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 1338–1343.

[12] P Ravindra, S De Silva, Katsunori Tadano, Stephen G Lambacher, Susantha Herath, and Masatake Higashi. 2009. Unsupervised approach to acquire robot joint attention. In *2009 4th International Conference on Autonomous Robots and Agents*. IEEE, 601–606.

[13] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. 2016. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 779–788.

[14] Michael Scaife and Jerome S Bruner. 1975. The capacity for joint visual attention in the infant. *Nature* 253, 5489 (1975), 265–266.

[15] Elaine Short, Justin Hart, Michelle Vu, and Brian Scassellati. 2010. No fair!! an interaction with a cheating robot. In *2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 219–226.

[16] Michael Tomasello et al. 1995. Joint attention as social cognition. *Joint attention: Its origins and role in development* 103130 (1995).

[17] J Gregory Trafton, Benjamin Fransen, Anthony M Harrison, and Magdalena Bugajska. 2009. *An embodied model of infant gaze-following*. Technical Report. NAVAL RESEARCH LAB WASHINGTON DC.

[18] Tomoko Yonezawa, Hirotake Yamazoe, Akira Utsumi, and Shinji Abe. 2007. Gaze-communicative behavior of stuffed-toy robot with joint attention and eye contact based on ambient gaze-tracking. In *Proceedings of the 9th international conference on Multimodal interfaces*. ACM, 140–145.