#2879 Atari-HEAD

Atari Human Eye-Tracking and Demonstration Dataset

Ruohan Zhang*, **Calen Walshe, Zhuode Liu, Lin Guan, Karl Muller, Jake Whritner, Luxin Zhang, Mary Hayhoe, Dana Ballard**

*zharu@utexas.edu

The University of Texas at Austin

Dataset Link

Abstract

Large-scale public datasets have been shown to benefit research in multiple areas of modern artificial intelligence. Many human decision-making tasks require visual attention to obtain high levels of performance. Therefore, measuring eye movements can provide a rich source of information about the strategies that humans use to solve decision-making tasks. We provide a large-scale, high-quality dataset of human actions with simultaneously recorded eye movements while humans play Atari video games. We introduce a novel form of gameplay, in which the human plays in a semi-frame-by-frame manner. This leads to near-optimal game decisions and game scores that are comparable or better than known human records. We demonstrate the usefulness of the dataset through two applications: predicting human gaze and imitating human demonstrated actions. The quality of the data leads to promising results in both tasks. Moreover, using a learned human gaze model to inform imitation learning leads to an 115% increase in game performance. We interpret these results as highlighting the importance of incorporating human visual attention in models of decision making and demonstrating the value of the current dataset to the research community. We hope this dataset can provide more opportunities to researchers in the areas of visual attention, imitation learning, and reinforcement learning.

Keywords: Visual Attention; Eye Tracking; Imitation Learning

Motivation and Overview

- Many successful stories of Atari gaming AIs and reinforcement learning algorithms in the past few years
- [AI] How can we collect demonstration data that better suited for training artificial learning agents?
- [Cognitive ergonomics] What is the level of human performance when the Atari gaming environment is made less cognitively demanding?
- [Visuomotor control[How do human players allocate their attentions in a variety of dynamic environments and make decisions?

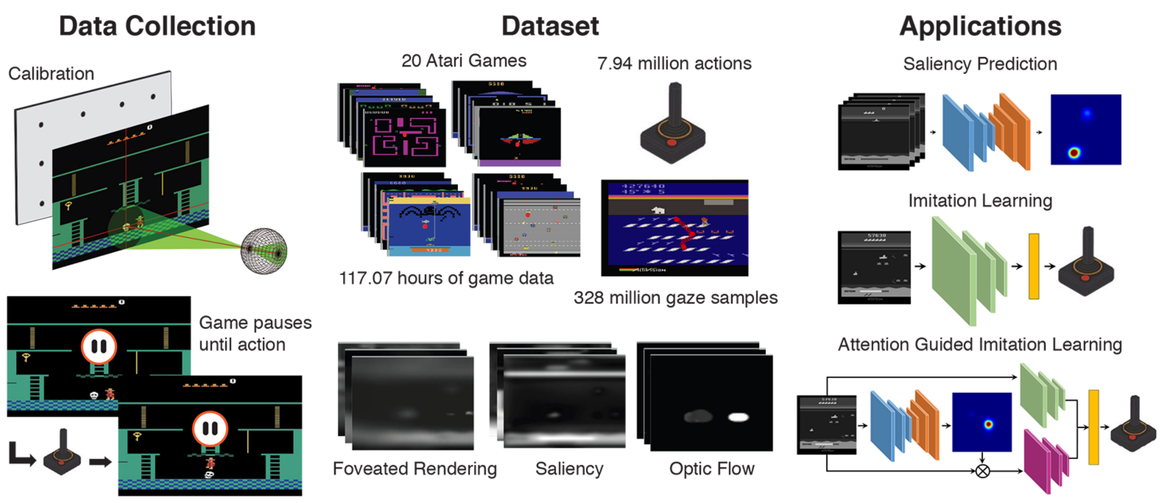

Figure 1: Project schematic of Atari-HEAD

Data Collection

- 20 Atari video games from the Arcade Learning Environment [Bellemare et al., 2012]
- Game image frames, human keystroke action, reaction time, gaze positions, and immediate reward received are recorded at every timestep
- Gaze positions are recorded using EyeLink 1000 eye tracker at 1000Hz; average gaze positional error: 0.40 visual degrees ($< 1\%$ of the image size)
- **Semi-frame-by-frame game mode**: Game pauses until action. Players can hold down a key and the game will run continuously at 20Hz
  - Eliminates errors due to human sensori-motor delays. Which is typically ∼250ms (∼15 frames at 60Hz game speed). Action a(t) could be intended for a state s(t') ∼250ms ago. Ensuring the action (label) matches the state (input) is important for supervised learning algorithms such as behavior cloning
  - Reduces inattentional blindness and allows sophisticated planning
- World-record level human performance
- Dataset link: QR code or https://zenodo.org/record/3451402

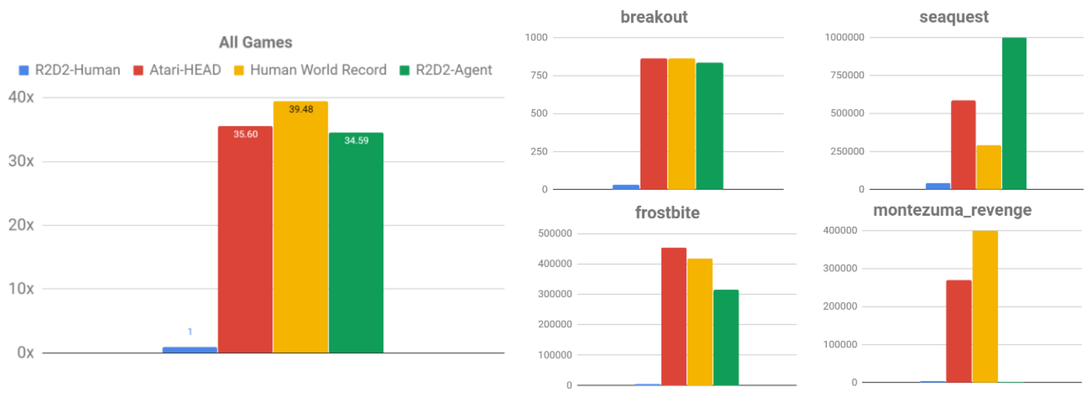

Figure 2: The way we collected data results significantly better human performance than previous human baselines [Kapturowski et al., 2018]. Human still outperforms one of the best reinforcement learning algorithm (R2D2) in many difficult games, when given enough time to make decisions.

- How do humans perceive game states?
  - Foveated representation: Humans have foveal vision with high acuity for only 1-2 visual degrees

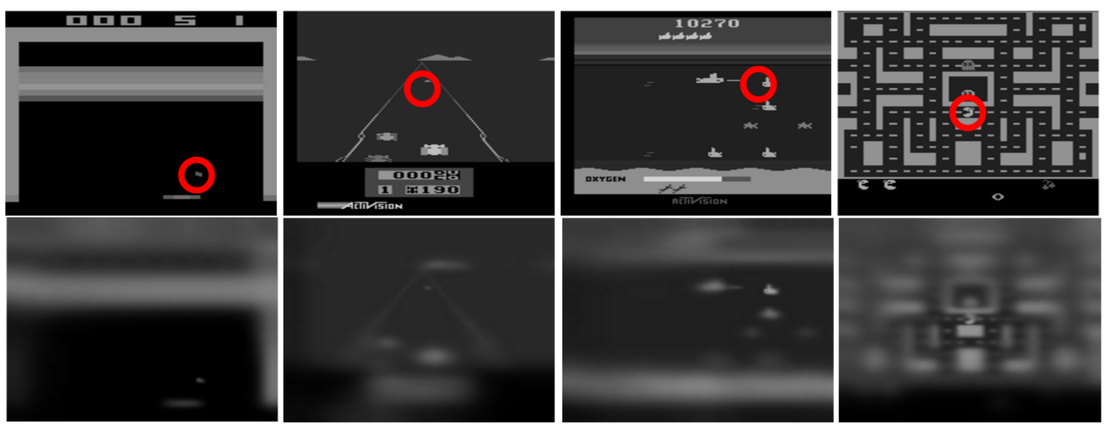

Figure 3: Foveal vision, generated using the foveated rendering algorithm [Perry and Geisler, 2002]

Learning to Predict Human Visual Attention and Actions

- Human gaze prediction as a standard saliency prediction problem
  - A 6-layer convolution-deconvolution network
  - Highly accurate, avg. AUC across 20 games = 0.97 (random = 0.5; max = 1)
  - Model captures predictive eye movements
  - Model identifies the target object from a set of visually identical objects
  - Model captures divided attention

Figure 4: Visualization of gaze prediction results for eight games. The solid red dot indicates the ground truth human gaze position. The heatmap shows the model's prediction as a saliency map, computed using the gaze network. Average area under the curve (AUC) score across 20 games on testing dataset: **0.97**.

- Human action prediction as an imitation learning (behavior cloning) problem
- Incorporating human gaze as an additional channel of information in the behavior cloning network
  - Attention-guided imitation learning (AGIL) [Zhang et al., 2018]

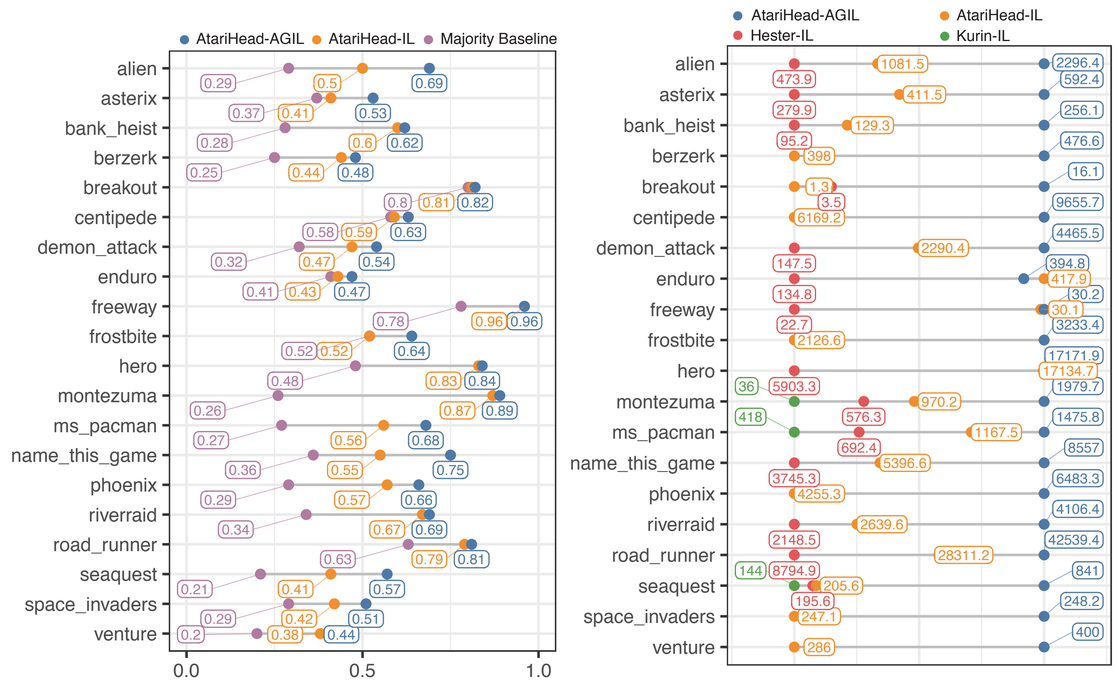

Figure 5: Human action prediction accuracy and game scores across 20 games. Visual attention helps the imitator better learn from human demonstration (**accuracy +0.07, game scores +115.3%**). Comparing to previous datasets [Kurin et al., 2017, Hester et al., 2018], the quality of this dataset leads to better performance.

- Why does human attention help? Hypothetically, visual attention simplifies the learning problem by indicating the object of interest for the current decision

Summary and Potential Future Research Directions

- [Cognitive ergonomics] A new human performance baseline
- [Vision science] A dataset for task-driven saliency prediction
- [AI] A high-quality dataset that is more suited for training learning agents
- [AI] Human attention-guided learning algorithms
- Human vs. machine attention: Do human players and gaming agents (e.g., reinforcement learning agents) pay attention to the same object given a state?
- Human decision time is an useful source of information as well, how do we use it for imitation learning?
- There is still a large performance gap between the human player and the learning agent in many difficult games, how can we do better?
- Visual attention + reinforcement learning: Applying learned attention model to speedup the learning process of reinforcement learning agents.

Acknowledgement

We would like to thank Wilson Geisler's lab at UT-Austin for providing data collection tools. The work is supported by NIH Grant EY05729/T32 EY21462-6, NSF Grant CNS-1624378, and Google AR/VR Research Award.

References

[Bellemare et al., 2012] Bellemare, M. G., Naddaf, Y., Veness, J., and Bowling, M. (2012). The arcade learning environment: An evaluation platform for general agents. *Journal of Artificial Intelligence Research*.

[Hester et al., 2018] Hester, T., Vecerik, M., Pietquin, O., Lanctot, M., Schaul, T., Piot, B., Horgan, D., Quan, J., Sendonaris, A., Osband, I., et al. (2018). Deep q-learning from demonstrations. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

[Kapturowski et al., 2018] Kapturowski, S., Ostrovski, G., Quan, J., Munos, R., and Dabney, W. (2018). Recurrent experience replay in distributed reinforcement learning.

[Kurin et al., 2017] Kurin, V., Nowozin, S., Hofmann, K., Beyer, L., and Leibe, B. (2017). The atari grand challenge dataset. *arXiv preprint arXiv:1705.10998*.

[Perry and Geisler, 2002] Perry, J. S. and Geisler, W. S. (2002). Gaze-contingent real-time simulation of arbitrary visual fields. In *Electronic Imaging 2002*, pages 57–69. International Society for Optics and Photonics.

[Zhang et al., 2018] Zhang, R., Liu, Z., Zhang, L., Whritner, J. A., Muller, K. S., Hayhoe, M. M., and Ballard, D. H. (2018). Agil: Learning attention from human for visuomotor tasks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 663–679.