Egocentric Vision: Integrating Human and Computer Vision

In today's digital age, our whole world is increasingly driven by data. A fascinating example of this is the commercial success of wearable devices which are filled with sensors that collect personalized behavioral data. Smartwatches and smartphones routinely include multiple cameras, heart rate sensors, altimeters, GPS, and much more. This wealth of noninvasive sensors is also shaping new research paradigms across many disciplines, as almost all of science is becoming more and more computational.

My research explores the potential of advanced machine learning to help make sense of vast amounts of behavioral data. As a computer vision scientist, I am particularly excited to harness the potential of wearable camera systems. I believe that such cameras provide a uniquely naturalistic insight into how a person interacts with the world and, since they can now be implemented with lightweight hardware, wearable cameras are becoming interesting tools across many areas in academia (e.g. research in human behavior, psychology, vision) and industry (e.g. smart glasses, police body cameras). Furthermore, as a cognitive scientist, I am very interested in exploring how to apply insights on human learning to machine learning and vice versa. I believe collecting dense and naturalistic behavioral data (e.g. the field of view of a toddler exploring new toys) and analyzing it with novel computational models (e.g. deep neural networks) is a promising new research direction.

These two principles – utilizing wearable sensors and exploring the mutual benefits of human and machine learning – underlie my previous work (as highlighted in the following examples) and motivate my future research plans.

Example 1: Egocentric Hand and Activity Recognition

Automatically analyzing first-person photo or video data is becoming a new and exciting area of computer vision research. One line of work [3, 4, 6] in this area that I have been investigating aims at analyzing hands. Your own hands are omnipresent in your field of view such that automatically recognizing hands can be beneficial in numerous ways. For example, in order to recognize a large number of daily activities, the shape and motion patterns of hands and fingers alone are surprisingly informative. Imagine your own field of view as you are brushing your teeth, typing on your computer, or perhaps playing with a deck of cards. Now imagine seeing only your hands, with no other context given. Each action is still clearly recognizable. Hands can convey emotions (think of someone giving a passionate speech) and hand gestures by themselves provide a rich means of communication. Utilizing wearable cameras to interpret hands will open new doors for human computer interaction (HCI), such as gesture control and even real-time sign language interpretation.

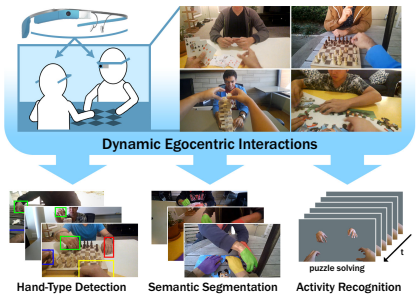

Figure 1: Examples of analyzing hands and activities based on video data of first-person interactions [4].

Towards this goal, my contributions include collecting the first densely-annotated, large-scale dataset [4] of hands recorded with first-person cameras, which is now widely used by the community. I have also proposed different methods to automatically detect and segments hands, as well as a probabilistic graphical model framework to semantically distinguish your own hands from others [6]. Finally, I have explored the potential of deep learning models to recognize different activities (e.g. playing chess) based on extracting and analyzing only the hands in view [3].

Example 2: Using Machine Learning to Study Object Recognition in Toddlers

Deep learning has transformed much of artificial intelligence research and driven major breakthroughs, including computers beating human experts at the board game *Go* for the first time. Computer vision research on convolutional neural networks has played a key role in this success, enabling computer models to learn to make sense of complex visual data by passively observing large amounts of images.

Meanwhile, developmental cognitive scientists are interested in understanding how humans learn to make

sense of the visual world based on the data they perceive each and every day. I believe that wearable camera systems allow new and potentially revolutionary research paradigms to explore the interdependencies between human and machine learning. Imagine recording the actual field of view of toddlers as they actively explore the world, resulting in massive visual datasets that accurately represent the real-world data humans use to train their cognitive visual systems. Deep learning-based computer vision models can help us analyze these statistics and gain insight into the ease and efficiency with which toddlers learn to recognize new concepts, which conversely may help us improve our A.I. models.

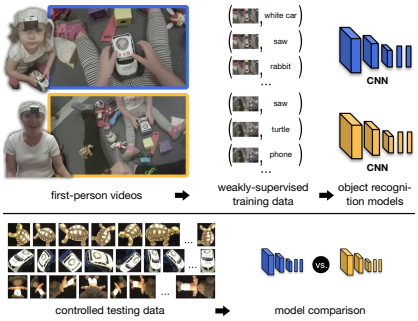

Figure 2: Using deep neural networks (CNNs) to analyze data in the field of view of toddlers and adults [2].

I have established an ongoing collaboration with leading developmental psychologists who study visual development in infants and toddlers. Successful pilot studies [1, 2, 5] include experiments where children and parents jointly play with a set of toys in a naturalistic and unconstrained environment while wearing head-mounted cameras. By using the recorded video data to train deep neural networks, we demonstrated that toddlers naturally collect high-quality training data of objects by creating visually diverse [1] and referentially unambiguous [2] object views.

Future Work

I intend to further establish my novel research direction of using advanced machine learning and vision models (deep neural networks) as tools to analyze large-scale behavioral/developmental data, especially visual data collected with egocentric camera systems. I believe that converging the study of learning in children and the study of machine learning will be beneficial for both areas. My ongoing collaboration with developmental psychologists will help me advance towards this direction as we are continuously collecting new and rich wearable sensor data with infants and toddlers.

Moreover, I intend to pursue core computer vision research in the area of first-person vision. New and exciting wearable camera devices continue to be developed and improving the state-of-the-art of vision algorithms (e.g. for gesture or action recognition) is a crucial component to ensure that wearable technology can become an increasingly practical tool that improves our lives. Finally, I believe computer vision, in particular combined with wearable technology, provides a great opportunity to engage students in research as my experience has shown that students are naturally drawn to exploring intriguing new gadgets.

References

[1] Bambach, S., Crandall, D. J., Smith, L. B., and Yu, C. (2016). Active viewing in toddlers facilitates visual object learning: An egocentric vision approach. In *Proceedings of the Cognitive Science Society*, volume 38.

[2] Bambach, S., Crandall, D. J., Smith, L. B., and Yu, C. (2017a). An egocentric perspective on active vision and visual object learning in toddlers. In *2017 Joint IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob)*.

[3] Bambach, S., Crandall, D. J., and Yu, C. (2015a). Viewpoint integration for hand-based recognition of social interactions from a first-person view. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, pages 351–354.

[4] Bambach, S., Lee, S., Crandall, D. J., and Yu, C. (2015b). Lending a hand: Detecting hands and recognizing activities in complex egocentric interactions. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1949–1957.

[5] Bambach, S., Zhang, Z., Crandall, D. J., and Yu, C. (2017b). Exploring inter-observer differences in first-person object views using deep learning models. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV) Workshops*.

[6] Lee, S., Bambach, S., Crandall, D. J., Franchak, J. M., and Yu, C. (2014). This hand is my hand: A probabilistic approach to hand disambiguation in egocentric video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 543–550.