A First Person Perspective on Computational Vision

Kristen Grauman
Department of Computer Science
University of Texas at Austin

Recent advances in sensor miniaturization, low-power computing, and battery life have carved the path for the first generation of mainstream wearable cameras. Images and video captured by a first-person (wearable) camera differ in important ways from traditional third-person visual data. A traditional third-person camera passively watches the world, typically from a stationary position. In contrast, a first-person camera is inherently linked to the ongoing experiences of its wearer. It encounters the visual world in the context of the wearer's physical activity, behavior, and goals.

To grasp this difference concretely, imagine two ways you could observe a scene in a shopping mall: in the first, you watch a surveillance camera video and see shoppers occasionally pass by the field of view of the camera; in the second, you watch the video captured by a shopper's head-mounted camera as he actively navigates the mall, darting in and out of stores, touching certain objects, swinging his head around to read signs or look for a friend. While both cases represent similar situations---and indeed the same exact physical environment---the latter highlights the striking difference in capturing the visual experience from the point of view of the camera wearer.

This distinction has intriguing implications for computer vision research---the realm of artificial intelligence and machine learning that aims to automate *visual intelligence* so that computers can "understand" the semantics and geometry embedded in images and video. First-person computational vision is poised to enable a class of new applications. The applications are in domains ranging from augmented reality, behavior assessment, perceptual mobile robotics, video indexing for life-loggers or law enforcement, and even the quantitative study of infant motor and linguistic development. What's more, the

first-person perspective on computational vision has the potential to transform the basic research agenda of computer vision as a field: from one focused on "disembodied" static images, heavily supervised machine learning for closed-world tasks, and stationary testbeds---to one instead encompassing embodied learning procedures, unsupervised learning and open-world tasks, and dynamic testbeds that change as a function of the system's own actions and decisions.

My group's recent work explores first-person computational vision on two main fronts:

- **Embodied visual representation learning:** How do the visual observations from a first-person camera relate to its 3D ego-motion? What can a vision system learn simply by moving around and "looking", if it is cognizant of its own ego-motion? How should an agent choose to move, so as to most efficiently resolve ambiguity about a recognition task? These questions have interesting implications for modern visual recognition problems and representation learning challenges underlying many tasks in computer vision.
- **Egocentric summarization:** An always-on first-person camera is a double-edged sword: the entire visual experience is retained without any active control by the wearer, yet the entire visual experience is not substantive. How can a system automatically *summarize* a long egocentric video, pulling out the most important parts to construct a visual index of all significant events? What attention cues does a first-person video reveal, and when was the camera wearer engaged with the environment? Could an intelligent first-person camera predict when it is even a good moment to take photos or video? These questions lead to applications in personal video summarization, sharing first-person experiences, and in-situ attention analysis.

Throughout both research threads above, our work is driven by the notion that the camera wearer is an active participant in the visual observations received. We consider ego-centric or first-person cameras of varying sources---namely, those worn by people, autonomous vehicles, or mobile robots.

(a) (b)

Figure 1: (a) The status quo in computer vision is to learn object categories from massive collections of "disembodied" Web photos that have been labeled by human supervisors as to their contents. (b) In first-person vision, we have the opportunity to learn from embodied spatio-temporal observations, capturing not only what is seen but also how it relates to the movement and actions of the self (i.e., the egocentric camera) in the world. Right image is shared by user Daniel under the Creative Commons license.

Embodied visual learning: How does ego-motion shape visual learning and action?

First, I will show how to exploit ego-motion when learning image representations. Cognitive science tells us that proper development of visual perception requires internalizing the link between "how I move" and "what I see". For example, in their famous "kitten carousel" experiment, Held and Hein (1963) examined how the visual development of kittens is shaped by their self-awareness and active control (or lack thereof) of their own physical motion. However, today's best computer vision algorithms, particularly those tackling recognition tasks, are deprived of this link, learning solely from bags of images downloaded from the Web and labeled by human annotators. We argue that such "disembodied" image collections, though clearly valuable when collected at scale, deprive feature learning methods from the informative physical context of the original visual experience. See Figure 1.

We propose to develop *embodied visual representations* that explicitly link what is seen to how the sensor is moving. To this end, we present a deep feature learning approach that embeds information not only from the video stream the observer sees, but also the motor actions he simultaneously makes (Jayaraman &

Grauman, 2015). Specifically, we enforce that the features learned in a convolutional neural network exhibit *equivariance*, i.e., they respond predictably to transformations associated with distinct ego-motions. During training, the input image sequences are accompanied by a synchronized stream of ego-motor sensor readings; however, they need not possess any semantic labels. The ego-motor signal could correspond, for example, to the inertial sensor measurements received alongside video on a wearable or car-mounted camera. The objective is to learn a feature mapping from pixels in a video frame to a space that is equivariant to various motion classes. In other words, the learned features should change in predictable and systematic ways as a function of the transformation applied to the original input. See Figure 2. To exploit the features for recognition, we augment the neural network with a classification loss when class-labeled images are available. In this way, ego-motion serves as side information to regularize the features learned, which we show facilitates category learning when labeled examples are scarce. We demonstrate the impact for recognition, including a scenario where features learned from ego-video on an autonomous car substantially improve large-scale scene recognition.

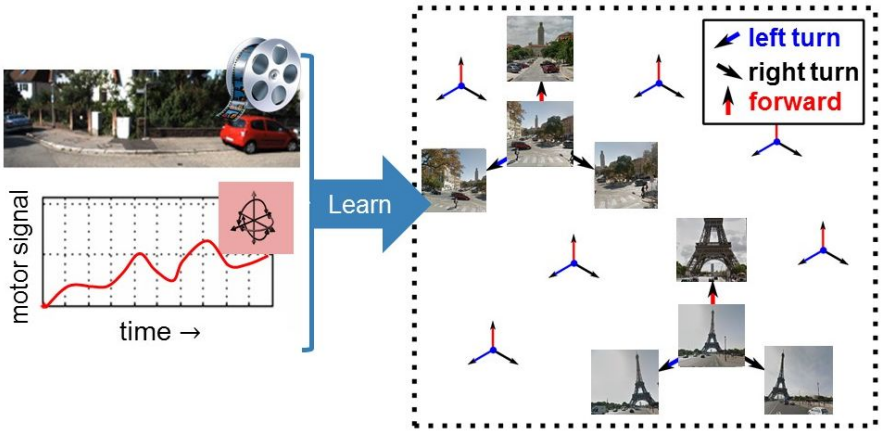

Figure 2: We propose to learn visual representations that are equivariant with respect to the camera's ego-motion. Given an unlabeled video accompanied by external measurements of the camera's motion (left), the approach optimizes an embedding that keeps pairs of views organized according to the ego-motion that separates them (right). In other words, the embedding requires that pairs of frames which share an ego-motion be related by the same transformation in the learned feature space. Such a learned representation injects the embodied knowledge of self-motion into the description of what is seen.

Building on this concept, we further explore how the system can actively choose *how to move* about a scene, or *how to manipulate* an object, so as to recognize its surroundings using the fewest possible observations (Jayaraman & Grauman, 2016). The goal is to learn how the system should move to improve its sequence of observations, and how a sequence of future observations is likely to change conditioned on its possible actions. We show how a recurrent neural network-based system may be trained to perform end-to-end learning of motion policies suited for this "active recognition" setting. In particular, the three functions of control, per-view recognition, and evidence fusion are simultaneously addressed in a single learning objective. Results so far show the impact on recognizing a scene by instructing the egocentric camera where to point next, and recognizing an object manipulated by a robot arm by determining how to turn the object in its grasp to get the sequence of most informative views. See Figure 3.

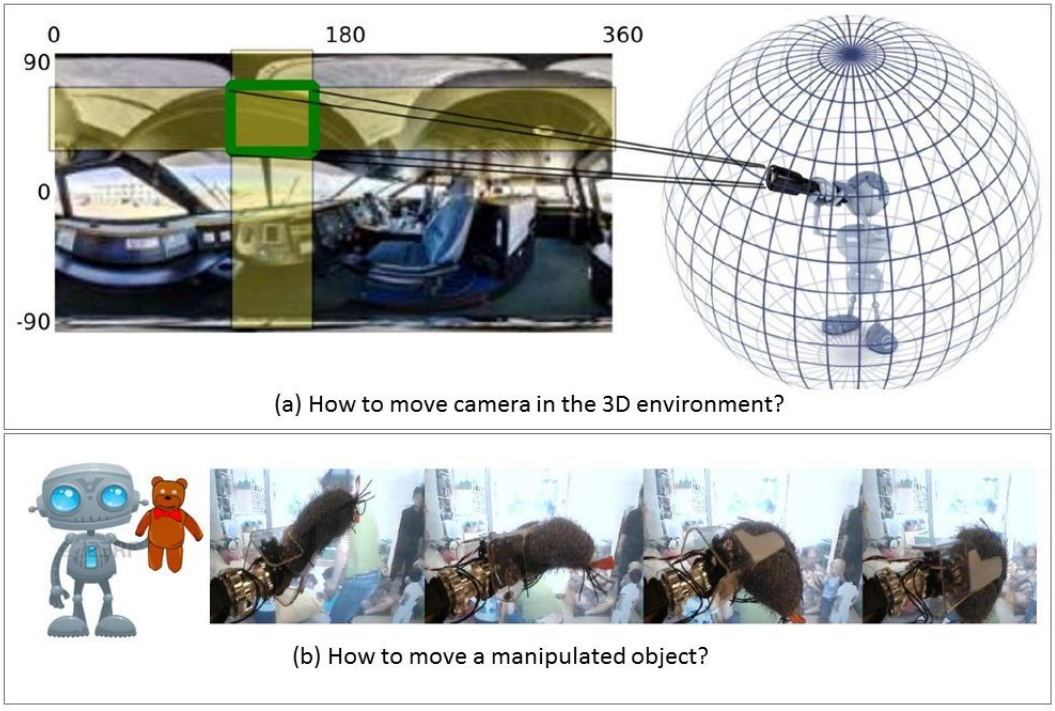

Figure 3: Active visual recognition requires learning how to move to reduce ambiguity in a task. A first-person vision system must learn (a) how to move its camera within the scene, or (b) how to manipulate an object with respect to itself, in order to produce more accurate recognition predictions more rapidly. Images in figure originally appear in (Jayaraman & Grauman, 2016), published by Springer.

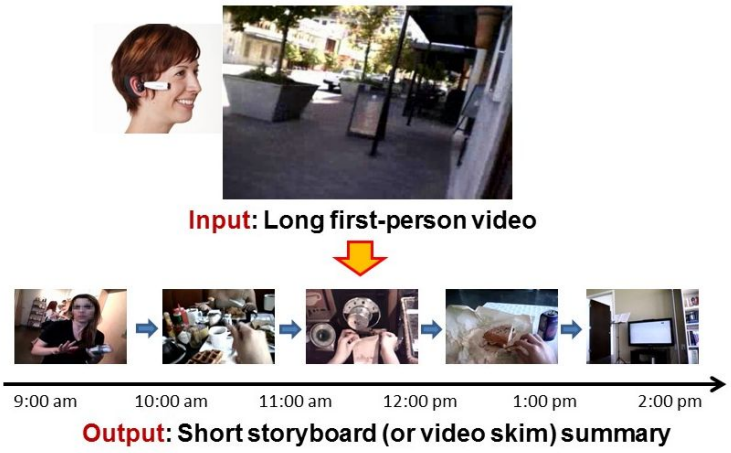

Figure 4: The goal in egocentric video summarization is to compress a long input video (here, depicting daily life activity) into a short human-watchable output that conveys all essential events, objects, and people to reconstruct the full story.

Egocentric summarization: What is important in a long first-person video?

The other major thread of interest in our first-person vision research deals with *video summarization* from the first person perspective. Given hours of first-person video, the goal is to produce a compact storyboard or a condensed video that retains all the important people, objects, and events from the original source video. If the summary is done well, it would serve as a good proxy for the original in the eyes of a human viewer. In other words, long video in; short video out. See Figure 4.

While summarization is valuable in many domains where video must be more accessible for searching and browsing, it is particularly compelling in the first-person setting due to 1) the inherently long-running nature of video generated from an "always on" egocentric camera, and 2) the inherent *storyline* embedded in the unedited video captured from a first person perspective. Our work is inspired by the potential application of aiding a person suffering memory loss, who by recounting their visual experience in brief could have improved recall (Hodges et al. 2011). Other applications include facilitating transparency and memory for

law enforcement officers wearing bodycams, or allowing a robot exploring uncharted territory to return with an executive visual summary of everything it saw.

To this end, we are developing methods to generate visual synopses from egocentric video. Leveraging cues about ego-attention and interactions to infer a storyline, the proposed methods automatically detect the highlights in long source videos. The main contributions so far entail learning to predict when an observed object/person is important given the context of the video (Lee & Grauman, 2015), inferring the influence between sub-events in order to produce smooth, coherent summaries (Lu & Grauman, 2013), predicting which egocentric video frames look as if they could be intentionally taken photographs (Xiong & Grauman, 2015), and detecting temporal intervals where the camera wearer's engagement with the environment is heightened (Su & Grauman, 2016). With experiments processing dozens of hours of unconstrained video of daily life activity, we show that long first-person videos can be distilled to succinct visual storyboards that are understandable in just moments.

Conclusion

Overall, the first-person setting offers exciting new opportunities for large-scale visual learning. The work described above offers a starting point towards the greater goals of embodied representation learning, first-person recognition, and discovering storylines in first-person observations. Future directions include expanding sensing to multiple modalities (audio, depth), giving an agent volition about its motions during training time as well as at the time of inference, investigating the most effective means to convey a visual or visual-linguistic summary, and scaling algorithms to cope with large-scale streaming video while making such complex decisions.

References

D. Jayaraman and K. Grauman. Look-Ahead Before You Leap: End-to-End Active Recognition by Forecasting the Effect of Motion. Proceedings of the European Conference on Computer Vision (ECCV), Amsterdam, October 2016.

Y-C. Su and K. Grauman. Detecting Engagement in Egocentric Video. Proceedings of the European Conference on Computer Vision (ECCV), Amsterdam, October 2016.

Z. Lu and K. Grauman. Story-Driven Summarization for Egocentric Video. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Portland, OR, June 2013.

D. Jayaraman and K. Grauman. Learning Image Representations Tied to Ego-Motion. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, Dec 2015.

Y J. Lee and K. Grauman. Predicting Important Objects for Egocentric Video Summarization. International Journal on Computer Vision, Volume 114, Issue 1, pp. 38-55, August 2015.

B. Xiong and K. Grauman. Intentional Photos from an Unintentional Photographer: Detecting Snap Points in Egocentric Video with a Web Photo Prior. Invited chapter. In *Mobile Cloud Visual Media Computing*. Springer International Publishing. Editors: G. Hua and X.-S. Hua. pp 85-111. November 2015.

R. Held and A. Hein. Movement-produced stimulation in the development of visually guided behavior. Journal of comparative and physiological psychology, 1963.

S. Hodges, E. Berry, and K. Wood. Sensecam: A Wearable Camera which Stimulates and Rehabilitates Autobiographical Memory. *Memory*, 2011.