Delving into Egocentric Actions

Yin Li, Zhefan Ye, James M. Rehg
School of Interactive Computing, College of Computing, Georgia Institute of Technology

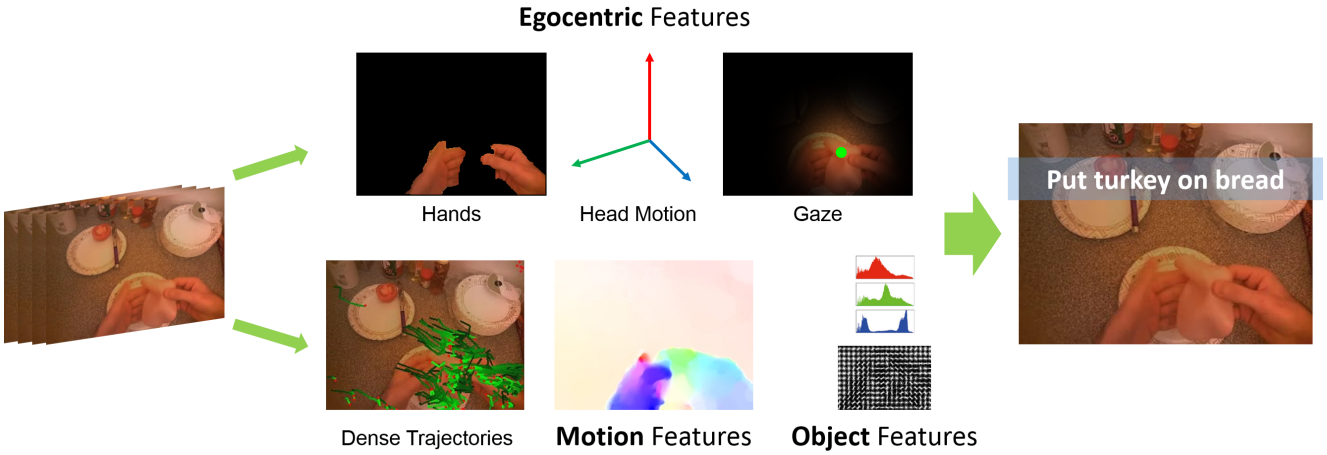

Figure 1: We address the challenging problem of recognizing the camera wearer's actions from videos captured by an egocentric camera. We propose to combine a novel set of mid-level egocentric cues with low-level object and motion cues for recognizing egocentric actions. Our *egocentric* features encode hand pose, head motion and gaze direction. Our *motion* and *object* features come from local descriptors in Dense Trajectories, with motion compensation using head motion. We design a systematic benchmark to evaluate how different types of features contribute to the final performance, and seek the best recipe using motion, object and egocentric cues. Our findings significantly advance the results in all major benchmarks. More details can be found on our project website www.cbi.gatech.edu/egocentric

Egocentric action recognition, which aims at analyzing the first person's behavior using egocentric videos, has received a growing interest in the computer vision community [1, 3, 5, 6, 7]. Despite many effort on action recognition in a surveillance setting, it is unclear that whether they can be successfully applied to egocentric actions. Frequent camera motion can hamper motion-based representations underlie many successful action recognition systems. Thus, previous egocentric action recognition methods rely mainly on an object-centric representation. A systematic evaluation of motion cues in egocentric action recognition remains missing.

In addition, egocentric videos encode a rich set of signals regarding the camera wearer, including head movement, hand pose and gaze information. We consider these signals regarding the first person as mid-level egocentric cues. They usually come from low-level appearance or motion cues, e.g. hand segmentation or motion estimation, and are complementary to traditional visual features. These mid-level egocentric cues reveal the underlying actions of the first person, yet have been largely ignored by previous methods of egocentric action recognition.

We provide the first systematic evaluation of motion, object and egocentric features for egocentric action recognition. Our egocentric features are derived from a novel set of mid-level cues, including the first person's hand, head and gaze movement. We set up our baseline using local descriptors from Dense Trajectories (DT) [8], a successful video representation for action recognition in a surveillance setting. We then systematically vary the method by adding motion compensation, object features and egocentric features on top of DT. Different feature channels are combined by stacking multiple Improved Fisher Vectors [4] for each channel. Figure 1 provides an overview of our approach.

Our extensive benchmark demonstrates how these choices contribute to the final performance. We identify a key set of practices that produce statistically significant improvement over the state-of-the-art methods. In particular, we find that simply extracting features around the first-person's attention point works surprisingly well. Our findings lead to a significant performance boost over state-of-the-art methods on major datasets. Our best method that supplement DT with proper object and egocentric features, largely improves the recognition accuracy by 27.0% in GTEA [1], 13.9% in GTEA Gaze [2] and 10.7% in GTEA Gaze+ [2], in comparison to the state-of-the-art methods.

Our findings, derived from this large set of experiments, can be summarized into three parts:

- Motion compensation is important for egocentric actions. It leads to more reliable motion features, as well as identifying foreground moving regions that can be used to extract object features.
- Foreground object cues are of crucial importance in egocentric actions. The information of what object is been used greatly helps the performance of action recognition
- Using an "attention" point (manipulation/gaze point) to guide feature encoding works surprisingly well. A manipulation point derived from hand shape serves as a good approximation to the actual gaze point for egocentric action recognition.

[1] Alireza Fathi, Ali Farhadi, and James M Rehg. Understanding egocentric activities. In *ICCV*, pages 407–414, 2011.

[2] Alireza Fathi, Yin Li, and James M Rehg. Learning to recognize daily actions using gaze. In *ECCV*. 2012.

[3] Kris Makoto Kitani, Takahiro Okabe, Yoichi Sato, and Akihiro Sugimoto. Fast unsupervised ego-action learning for first-person sports videos. In *CVPR*, 2011.

[4] Florent Perronnin, Jorge Sánchez, and Thomas Mensink. Improving the fisher kernel for large-scale image classification. In *ECCV*. 2010.

[5] Hamed Pirsiavash and Deva Ramanan. Detecting activities of daily living in first-person camera views. In *CVPR*, 2012.

[6] M. S. Ryoo and L. Matthies. First-person activity recognition: What are they doing to me? In *CVPR*, 2013.

[7] Ekaterina H. Spriggs, Fernando De la Torre Frade, and Martial Hebert. Temporal segmentation and activity classification from first-person sensing. In *IEEE Workshop on Egocentric Vision, CVPR 2009*, 2009.

[8] Heng Wang, Alexander Kläser, Cordelia Schmid, and Cheng-Lin Liu. Dense trajectories and motion boundary descriptors for action recognition. *International Journal of Computer Vision*, 103(1):60–79, 2013.

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.