15th IAPR International Conference on Machine Vision Applications (MVA)
Nagoya University, Nagoya, Japan, May 8-12, 2017.

Connecting the Dots: Embodied Visual Perception from First-person Cameras

Jianbo Shi
School of Engineering and Applied Science
University of Pennsylvania, USA

1 Introduction

A computer has a complete photographical memory. It creates massive but isolated sensory moments. Unlike such fragmented photographic memory, human memories are highly connected through episodes that allow us to relate past experiences and predict future actions. How to computationally model a human-like episodic memory system that connects photographically accurate sensory moments? Our insight is that an active interaction is a key to link between episodes because sensory moments are fundamentally centered on an active person-self. Our experiences are created by and shared through our social and physical interactions, i.e., we connect episodes driven by similar actions and, in turn, recall these past connected episodes to take a future actions. Therefore, connecting the dotted moments to create an episodic memory requires understanding the purposeful interaction between human (person-self) and world.

Photographs are only half of our world experience: it records what are out there. What are in our head, our intention-attention-physiological states during the social and physical interactions, are missing from the memory recording. This needs creating an embodied memory link between our inner 'selves' with the external episode, and a first person camera is an ideal sensor to capture, model, and predict the embodied memory link because it encodes a complete visual audio sensation of the camera wearer's interaction with the world. We leverage purposeful actions measured by first person cameras to reveal the internal states of the camera wearer, and use the similar internal states to connect the wearer's episodic sensations of the world.

Here, we review several of our recent results on embodied visual perception from first-person cameras.

2 Decoding Physical Sensation

A first-person video can generate powerful physical sensations of action in an observer. Consider the problem of Force from Motion [**?**]: decoding the sensation of 1) passive forces such as the gravity, 2) the physical scale of the motion (speed) and space, and 3) active forces exerted by the observer such as pedaling a bike or banking on a ski turn (Fig. **??**).

The sensation of gravity can be observed in a natural image. We can learn this image cue for predicting a gravity direction in a 2D image and integrate the prediction across images to estimate the 3D gravity direction using structure from motion. The sense of physical scale is revealed to us when the body is in a dynamically balanced state. We compute the unknown physical scale of 3D reconstructed camera motion by leveraging the torque equilibrium at a banked turn that relates the centripetal force, gravity, and the body leaning angle.

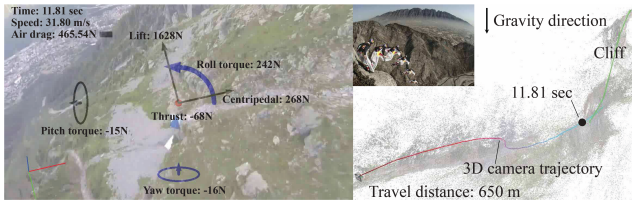

Figure 1. Force from Motion—decoding the sensation of 1) passive forces such as the gravity, 2) the physical scale of the motion (speed) and space, and 3) active forces exerted by the observer. We model egomotion with rigid body dynamics integrated in a bundle adjustment that allows us to recover the three sensations (left) via the physical scale and gravity aware reconstruction of the egomotion (right).

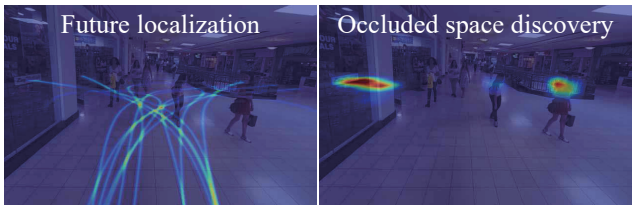

Figure 2. Where am I supposed to be after 5, 10, and 15 seconds? We predict a set of plausible future trajectories given a pair of egocentric stereo images. As a byproduct of the predicted trajectories, the occluded space by foreground objects such as the space inside of the shop or behind the ladies are discovered.

The active force and torque governs 3D egomotion through the physics of rigid body dynamics. Using an inverse dynamics optimization, we directly minimize 2D reprojection error (in video) with respect to 3D world structure, active forces, and additional passive forces such as air drag and friction force. We use structure from motion with the physical scale and gravity direction as an initialization of our bundle adjustment for force estimation. Our method shows quantitatively equivalent reconstruction comparing to IMU measurements in terms of gravity and scale recovery and outperforms method based on 2D optical flow for an active action recognition task. We apply our method to first person videos of mountain biking, urban bike racing, skiing, speedflying with parachute, and wingsuit flying where inertial measurements are not accessible.

3 Future Localization

With first-person cameras, we also address the problem of future localization [**?**]: to predict plausible future trajectories of ego-motion in egocentric stereo images (Fig. **??**). Our paths avoid obstacles, move between objects, even turn around a corner into space behind objects. As a byproduct

Figure 3. Predicting a group trajectory of basketball players from first person videos. The red is the ground truth and blue is the predicted trajectories with gaze direction.

of the predicted trajectories, we discover the empty space occluded by foreground objects.

One key innovation is the creation of an EgoRetinal map, akin to an illustrated tourist map, that 'rearranges' pixels taking into accounts depth information, the ground plane, and body motion direction, so that it allows motion planning and perception of objects on one image space. We learn to plan trajectories directly on this EgoRetinal map using first person experience of walking around in a variety of scenes. In a testing phase, given an novel scene, we find multiple hypotheses of future trajectories from the learned experience. We refine them by minimizing a cost function that describes compatibility between the obstacles in the EgoRetinal map and trajectories. We quantitatively evaluated our method to show predictive validity and apply to various real world daily activities including walking, shopping, and social interactions.

4 Social Behavior Prediction

Not only individuals, but first-person videos from a number of people can also be collectively exploited. For instance, we can predict future movements (location and gaze direction) of basketball players as a whole from their first person videos [**?**] (Fig. **??**). The predicted behaviors reflect an individual physical space that affords to take the next actions while conforming to social behaviors by engaging to joint attention.

The key innovation is to use the 3D reconstruction of multiple first person cameras to automatically annotate each other's the visual semantics of social configurations. We leverage two learning signals uniquely embedded in first person videos. Individually, a first person video records the visual semantics of a spatial and social layout around a person that allows associating with past similar situations.

Collectively, first person videos follow joint attention that can link the individuals to a group. We learn the egocentric visual semantics of group movements using a Siamese neural network to retrieve future trajectories. We consolidate the retrieved trajectories from all players by maximizing a measure of social compatibility–the gaze alignment towards joint attention predicted by their social formation, where the dynamics of joint attention is learned by a longterm recurrent convolutional network. This allows us to characterize which social configuration is more plausible and predict future group trajectories.

5 Unsupervised Learning of Important Objects

A first-person camera, placed at a person's head, captures, which objects are important to the camera wearer.

Figure 4. Given an *unlabeled* set of first-person images, our goal is to find all objects that are important to the camera wearer. Unlike most prior methods, we do so without using ground truth importance labels.

Most prior methods for this task learn to detect such important objects from the manually labeled first-person data in a supervised fashion. However, important objects are strongly related to the camera wearer's internal state such as his intentions and attention, and thus, only the person wearing the camera can provide the importance labels. Such a constraint makes the annotation process costly and limited in scalability.

Our recent work [**?**] makes it possible to detect important objects in first-person images without the supervision by the camera wearer or even third-person labelers (Fig. **??**). We formulate an important detection problem as an interplay between the 1) segmentation and 2) recognition agents. The segmentation agent first proposes a possible important object segmentation mask for each image, and then feeds it to the recognition agent, which learns to predict an important object mask using visual semantics and spatial features. We implement such an interplay between both agents via an alternating cross-pathway supervision scheme inside our proposed Visual-Spatial Network (VSN). Our VSN consists of spatial (?gwhere?h) and visual (?gwhat?h) pathways, one of which learns common visual semantics while the other focuses on the spatial location cues. Our unsupervised learning is accomplished via a cross-pathway supervision, where one pathway feeds its predictions to a segmentation agent, which proposes a candidate important object segmentation mask that is then used by the other pathway as a supervisory signal. Our method achieves similar or better results as the supervised methods.

6 Assessing Player's Performance

We also devised a method to assess a basketball player's performance from his/her first-person video [**?**] (Fig. **??**). A key challenge lies in the fact that the evaluation metric is highly subjective and specific to a particular evaluator. We leverage the first-person camera to address this challenge. The spatiotemporal visual semantics provided by a first-person view allows us to reason about the camera wearer's actions while he/she is participating in an unscripted basketball game. Our method takes a player's first-person video and provides a player's performance measure that is specific to an evaluator's preference. To achieve this

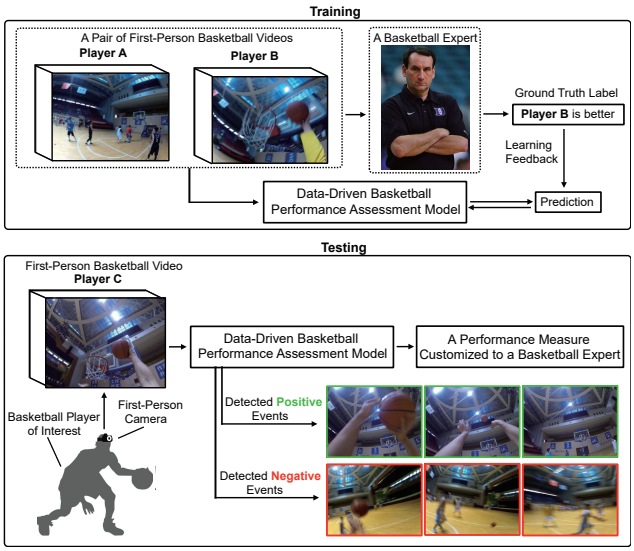

Figure 5. Assessing Player's Performance. Our goal is to assess a basketball player's performance based on an evaluator's criterion from an unscripted his/her first-person basketball video. During training, we learn such a model from the pairs of weakly labeled first-person basketball videos. During testing, our model predicts a performance measure customized to a particular evaluator from a first-person basketball video. Additionally, our model can also discover basketball events that contribute positively and negatively to a player's performance.

goal, we first use a convolutional LSTM network to detect atomic basketball events from first-person videos. Our network's ability to zoom-in to the salient regions addresses the issue of a severe camera wearer's head movement in first-person videos. The detected atomic events are then passed through the Gaussian mixtures to construct a highly non-linear visual spatiotemporal basketball assessment feature. Finally, we use this feature to learn a basketball assessment model from pairs of labeled first-person basketball videos, for which a basketball expert indicates, which of the two players is better. We demonstrate that despite not knowing the basketball evaluator's criterion, our model learns to accurately assess the players in real-world games. Furthermore, our model can also discover basketball events that contribute positively and negatively to a player's performance.

7 Conclusion

We have presented our recent results on embodied visual perception from first-person cameras. To computationally model a human-like episodic memory system that connects photographically accurate sensory moments, an active interaction is a key to link between episodes because sensory moments are fundamentally centered on an active person-self. Connecting the dotted moments to create an episodic memory requires understanding the purposeful interaction between human (person-self) and world. We leverage purposeful actions measured by first person cameras to reveal the internal states of the camera wearer, and use the similar internal states to connect the wearer's episodic sensations of the world.

References

[1] Hyun Soo Park, Jyh-Jing Hwang, and Jianbo Shi "Force from Motion: Decoding Physical Sensation in a First Person Video," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3834–3842, 2016.

[2] Hyun Soo Park, Jyh-Jing Hwang, Yedong Niu, and Jianbo Shi "Egocentric Future Localization," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4697–4705, 2016.

[3] Shan Su, Jung Pyo Hong, Jianbo Shi, and Hyun Soo Park "Predicting Behaviors of Basketball Players from First Person Videos," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[4] Gedas Bertasius, Stella X Yu, and Jianbo Shi "Exploiting Visual-Spatial First-Person Co-Occurrence for Action-Object Detection without Labels," arXiv 1611.05335, 2016.

[5] Gedas Bertasius, Stella X Yu, Hyun Soo Park, and Jianbo Shi "Am I a Baller? Basketball Skill Assessment using First-Person Cameras," arXiv 1611.05365, 2016.