Understanding Human Behaviors from Wearable Cameras

Hyun Soo Park
hypar@seas.upenn.edu

Figure 1(a) Joint attention

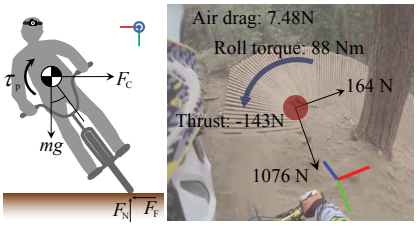
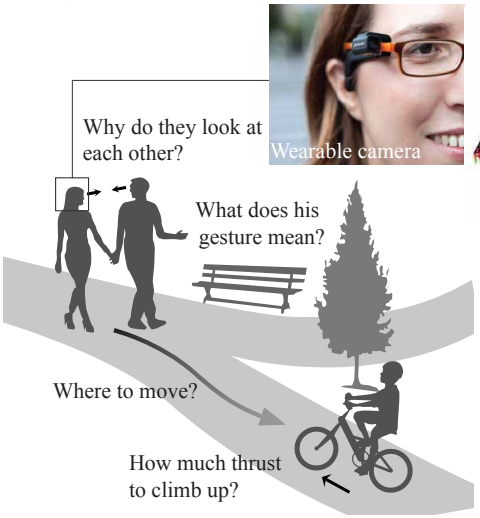


Figure 1(b) Physical sensation



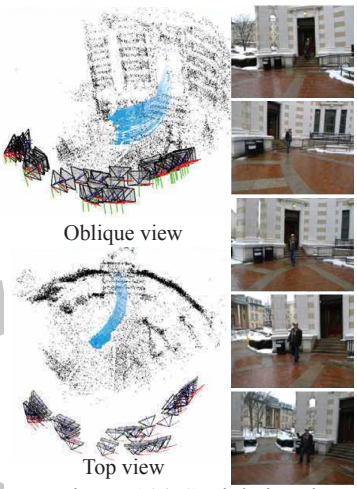


Figure 1(c) Social signal

Imagine a family going out for a walk. The mother reasons *beyond* what she is seeing: she may guess "why her husband suddenly looks at their son", "what his gesture means", and "where to move next". The son riding a bicycle may reason about "how hard I would pedal my way up the slope". My research goal is identifying such deep reasoning based on my conjecture—a wearable camera can tap into their experiences by *putting yourself in her/his shoes*.

Two key properties of a wearable camera support my conjecture. First, a wearable camera, notably a head-mounted camera, e.g., a Google Glass, records the *first person experiences* of the wearer. These experiences include physical and visual sensation, e.g., how visual stimuli change due to my speed, and social interactions, e.g., how often my partner makes eye-contact with me. This enables relating two distinctive events via similar sensation akin to Déjà vu experiences. Second, the wearable camera provides *in-situ* measurements of candid and unscripted behaviors, allowing a direct access to the true intent. This differs from third person photographs and videos that are often staged and controlled.

My research focuses on **designing machines that measure, decode, and learn the deeper meaning of human behaviors by exploiting the in-situ measurements of wearable cameras.** A simple usage of the wearable cameras does not solve behavioral understanding problems due to the first person biases such as camera placement, anthropometric configurations, and physical/social interactions. Therefore, representing human behaviors via first person perception is challenging, and I address this challenge through the following three ingredients:

- **Joint attention** To understand social behaviors, e.g., social formations, in a form of joint attention, or social saliency (Figure 1(a)) [1, 2, 3, 4].
- **Physical sensation** To predict one's behaviors by decoding first person sensation into physical quantities such as force, momentum, and energy (Figure 1(b)) [5, 6, 7].
- **Social signal** To recognize the meaning of human behaviors, e.g., social signals, by reconstructing their activities in 3D (Figure 1(c)) [8, 9, 10, 11, 12].

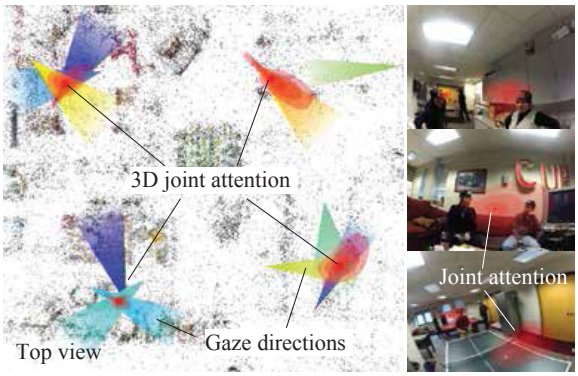
In my Ph.D. and postdoctoral research, I have demonstrated the validity of my representation through video editing [3], sport analytics [4], performance capture [12], and behavior prediction [1, 2, 6]. Research projects have been featured in major media including IEEE Spectrum, NBC News, Discovery News, and Wired, and I have co-organized a tutorial based on my thesis [13], "Group Behavior Analysis and Its Applications[1]" in conjunction with CVPR 2015.

[1] http://www.seas.upenn.edu/~hypar/GroupBehavior/cvpr15_tutorial_group_behavior.html

NOTE My work is best seen in videos and pdf. Please find a link in the footnote or a hyperlink in each image to access the corresponding video and project page.

Joint Attention

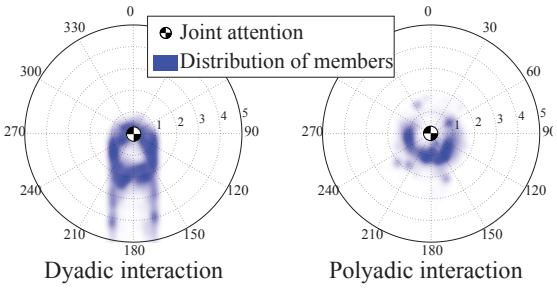
In Figure 1, the mother makes eye-contact with her husband signaling that she is listening to what he says. When her husband suddenly turns his gaze direction to his son, she immediately follows his gaze movement. They engage *joint attention*—the shared focus of multiple individuals. The joint attention plays a key role in a learning process at a developmental age, which is strongly related with social behaviors, e.g., we form a circular shape around a street busker (social formation) where the joint attention stays at the busker. My research aims **to localize joint attention and to understand its relationship with social behaviors from wearable cameras.** This research facilitates computational understanding of social communications, e.g., "what information is transferred from where to where, and how?", and decoding personal relationships, e.g., "why are they attracted to each other?"



First person cameras are used to estimate the 3D locations of joint attention by finding the intersection of gaze directions.

However, this is challenging because joint attention is *invisible*. We know that the joint attention would form at the son's location because his parents pay attention to him but how can we measure it? I approach this question by using head-mounted cameras [1][2]. The key insight is that the joint attention forms at the intersection of the gaze directions. We use the head orientation as a proxy for the gaze direction, measured by the head-mounted camera. In the left figure, the four locations of joint attention are found in 3D where 12 people wearing the head-mounted cameras interact with each other in a happy hour session. This method has been applied dynamic and chaotic social interactions such as children playing basketball. Further, the joint attention can be used to edit the multiple videos of a social event [3][3] and to detect a social anomaly, e.g., how my attention agrees with others [2][4].

Social behaviors are strongly influenced by joint attention. For instance, the parents may change their path because the son takes another path. The son's location (joint attention) is spatially and temporally related with the parents' moving direction (social behaviors). I have studied the spatial relationship through Kendon's F-formation theory that characterizes geometric social formations. The core idea is that the location of joint attention can be described as a function of people's locations, which I empirically prove using in-situ measurements of joint attention from wearable cameras [4][5]. The right figure illustrates the spatial distribution of people from top view where the origin is the location of joint attention. Based on this spatial relationship, we discovered "social hot spot" in a third person video (Figure 1(a)).

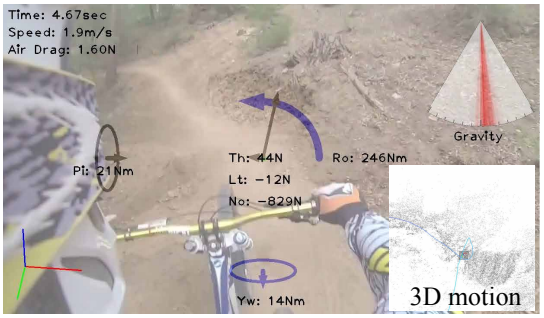


We prove the F-formation theory by characterizing social formations w.r.t. joint attention.

[2]http://www.cs.cmu.edu/~hyunsoop/gaze_concurrence.html
[3]http://graphics.cs.cmu.edu/projects/social-cameras/
[4]http://www.cs.cmu.edu/~hyunsoop/socialcharge.html
[5]http://www.seas.upenn.edu/~hypar/socialsaliencyprediction.html

Physical Sensation

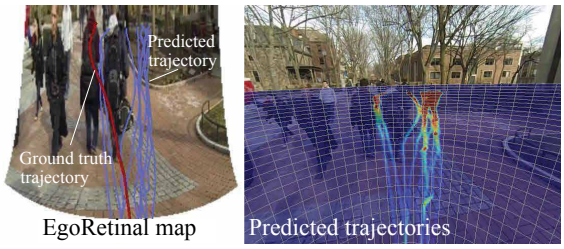
In the son's perspective (Figure 1), he sees the slope ahead, and he may pedal his bike hard to gain enough speed to climb up the slope. When he encounters a curve, he may slow down and lean against the turning direction as shown in Figure 1(b). He experiences force, momentum, and kinetic energy while interacting with the physical environments. Would it be possible to 're-experience' his physical sensation so that we can tap into his true intent? My research aims **to understand/predict human behaviors by decoding first person sensation into physical quantities such as force.** This research thrust enables not only physics based motion capture without wired strain gauge sensors but also in-situ behavioral imaging for sensorimotor skill learning, e.g., how does a child learn riding a bicycle in two weeks without understanding the internal dynamics? I have exploited my previous robotics research on controls, motion planning, and multibody dynamics for this research thrust [14, 15, 16, 17].



We have presented 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.*

How physical sensation is encoded in the video of a wearable camera? My hypothesis is that the dominant factors that constitute the physical sensation are the gravity direction, speed (momentum), and applied force/torque (kinetic energy). First, our activities are always predicated upon the gravity direction, e.g., most sports are a process of fighting against the gravity. Second, the speed determines the bandwidth of the visual perception, i.e., the visual receptive field for fast motion such as wingsuit flying (50-100m/s) is highly limited. Third, the egomotion is produced by the physical interactions between passive forces such as air drag and active forces to *control* one's motion, e.g., angular momentum change along the roll axis to shift the heading direction. I have studied these factors from first person sport videos such as biking, flying, and skiing [5][6] as shown in the left figure: the gravity direction learned from the image cues such as tree orientation and horizon (predicted gravity distribution on top right); the speed estimated by motion cue and force equilibrium (speed: 1.9 m/s); active/passive force computed by inverse dynamics with bundle adjustment (44 N of the thrust force).

How would one's physical sensation (perception) change future behaviors? Answering this question is challenging because it requires to infer his planning strategy for the long-term behaviors in relation to the physical sensation. I addressed this question by studying *future localization* [6][7]: to predict a trajectory of future locations, i.e., where am I supposed to be after 5, 10, and 15 seconds? (right figure). The intuition is that past movement experiences can be effectively associated with a novel scene via physical sensation akin to Déjà vu experiences. The physical sensation is encoded in *EgoRetinal Map*—a trajectory configuration space constructed by 're-arranging' pixels of the first person image, which allows us to retrieve the past memories. The predicted trajectories circumvent static obstacles (buildings), pass through a gap between moving people, and discover the partially occluded space (the space behind the people).

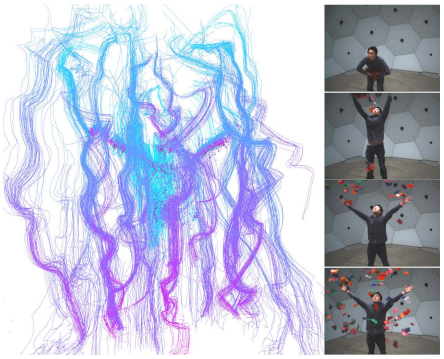


Future paths are predicted by leveraging past memory using the EgoRetinal map.

[6] https://youtu.be/LKjOvKCoO6Q
[7] https://youtu.be/yN0Gn7wSELk

Social Signal

The husband stresses his point with the hand gestures while talking to his wife as shown in Figure 1, and the son may wave his hand to attract the attention of his parents. They sends social signals in a form of body gestures, head movements, and facial expressions. My research aims **to recognize the meaning of social signals via 3D reconstruction of human activities**. This research allows performance capture, behavior monitoring for the elderly, and sport analytics. Notably this also has a broader impact on behavioral imaging, which will enable a computational analysis using a large scale social interaction data. In my prior study, I have primarily focused on reconstructing social signals in 3D using wearable cameras, which can be readily integrated in a recognition system.



The trajectories of human body and confetti are reconstructed in 3D at a unprecedented spatial resolution.

Why 3D? The video of a wearable camera inherits the first person biases produced by camera placements, anthopometric configurations, and target activities, which fundamentally limits a large scale learning capability. My hypothesis is that a 3D representation persistent to view points allows learning social signals across different people and scenes. A challenge of reconstructing the social signals in 3D is their *subtlety*, e.g., a small nod to signal my agreement, which requires a fine scale reconstruction algorithm. I have experimented with a large scale markerless motion capture system (480 VGA cameras)[8] where we have demonstrated 3D reconstruction of human body motion represented by a dense trajectories at a unprecedented spatial resolution (17,934 trajectories for 6.4 seconds of an activity) [12] (left figure). Also, the resulting algorithm can handle topological changes: the confetti (left figure) are dispersed in the middle of the sequence. This is ideal for social signals frequently involved with topological changes, e.g., hand-shake.

Would it be possible to reconstruct the husband's hand gestures in 3D from the wife's head-mounted camera? This is more challenging because the wearable camera introduces its camera egomotion induced by the attached body motion, i.e., her head motion, which needs to be compensated. Applying a prior knowledge about the scene such as a spatial or temporal motion model is a way to resolve this challenge. I have developed a reconstruction algorithm using a wearable camera by leveraging a temporal regularity, i.e., a trajectory is represented by a linear combination of analytic smooth functions [8, 10] (Figure 1(c))[9]. Further, an additional spatial constraint is studied [9][10] using body articulation, i.e., the 3D distance between two points on a limb remains constant.

21 body-mounted cameras are used to reconstruct human body motion in 3D by combining 3D reconstruction of camera motion and articulation constraints.

I also have exploited the egomotion of wearable cameras attached all limbs to reconstruct human body motion in 3D (right figure) [11][11]. 21 wearable cameras are jointly reconstructed to recover the underlying motion. This enables outdoor motion capture without a specialized instrumentation of the scene in contrast to existing optic based motion capture systems such as Vicon, and produces minimal drift while translating large distance, e.g., less than 2 cm drift for over 20 m translation, unlike IMU based motion capture systems.

[8] The Panoptic Studio at CMU, http://www.cs.cmu.edu/~hanbyulj/14/visibility.html
[9] http://www.cs.cmu.edu/~hyunsoop/trajectory_reconstruction.html
[10] http://www.cs.cmu.edu/~hyunsoop/articulated_trajectory.html
[11] http://drp.disneyresearch.com/projects/mocap/

Conclusion

Humans are complicated and unpredictable. Many factors collectively affect human behaviors and therefore, understanding the behaviors through an individual research goal in isolation is fundamentally limited: it is necessary to build a comprehensive representation that consolidates geometric, physical, and social factors. My research is oriented towards integrating such factors in collaborations with psychology and neuroscience to facilitate deeper understanding of the human behaviors: *reasoning beyond what we are seeing*.

I envision that robots will be more closely integrated in our life: a flying drone capturing a romantic moment of the couple, a teaching robot helping the son to learn riding a bicycle, and a virtual reality displaying a future path to move. My research will provide a computational basis to design the artificial intelligence for such robots that respect the human behaviors in a socially acceptable way. Also it will have a significant impact on computational behavioral science, providing automatic recognition systems for the behavioral disease detection at an early age, such as autism, and an empirical foundation for sensorimotor skill learning.

Selected References

[1] H. S. Park, E. Jain, and Y. Sheikh, "3D social saliency from head-mounted cameras," in *Advanced in Neural Information Processing Systems*, 2012.

[2] H. S. Park, E. Jain, and Y. Sheikh, "Predicting gaze behavior using social saliency fields," in *International Conference on Computer Vision*, 2013.

[3] I. Arev*, H. S. Park*, Y. Sheikh, J. Hodgins, and A. Shimir, "Automatic editing of footage from multiple social cameras," *Transactions on Graphics (SIGGRAPH) (* indicates joint first authors)*, 2014.

[4] H. S. Park and J. Shi, "Social saliency prediction," in *Computer Vision and Pattern Recognition*.

[5] H. S. Park, J.-J. Hwang, and J. Shi, "Force from motion: Decoding physical sensation in a first person video," in *Computer Vision and Pattern Recognition (under review)*, 2016.

[6] H. S. Park, J.-J. Hwang, Y. Niu, and J. Shi, "Egocentric future localization," in *Computer Vision and Pattern Recognition (under review)*, 2016.

[7] G. Bertasius, H. S. Park, and J. Shi, "Exploiting egocentric object prior for 3D saliency detection," in *Computer Vision and Pattern Recognition (under review)*, 2016.

[8] H. S. Park, T. Shiratori, I. Matthews, and Y. Sheikh, "3D reconstruction of a moving point from a series of 2D projections," in *European Conference on Computer Vision*, 2010.

[9] H. S. Park and Y. Sheikh, "3D reconstruction of a smooth articulated trajectory from a monocular image sequence," in *International Conference on Computer Vision*, 2011.

[10] H. S. Park, T. Shiratori, I. Matthews, and Y. Sheikh, "Trajectory reconstruction under perspective projection," *International Journal of Computer Vision*.

[11] T. Shiratori, H. S. Park, L. Sigal, Y. Sheikh, and J. Hodgins, "Motion capture from body-mounted cameras," *Transactions on Graphics (SIGGRAPH)*, 2011.

[12] H. Joo, H. S. Park, and Y. Sheikh, "MAP visibility estimation for large-scale dynamic 3D reconstruction," in *Computer Vision and Pattern Recognition*, 2014.

[13] H. S. Park, "Social scene understanding from social cameras," *Carnegie Mellon University*, 2015.

[14] H. S. Park, S. Floyd, and M. Sitti, "Dynamic modeling of a basilisk lizard inspired quadruped robot running on water," in *IEEE/RSJ International Conference on Intelligent Robots and System*, 2008.

[15] H. S. Park and M. Sitti, "Compliant footpad design analysis for a bio-inspired quadruped amphibious robot," in *IEEE/RSJ International Conference on Intelligent Robots and System*, 2009.

[16] H. S. Park, S. Floyd, and M. Sitti, "Dynamic modeling and analysis of pitch motion of a basilisk lizard inspired quadruped robot running on water," in *International Conference on Robotics and Automation*, 2009.

[17] H. S. Park, S. Floyd, and M. Sitti, "Roll and pitch motion analysis of a biologically inspired quadruped water runner robot," *International Journal of Robotics Research*, 2010.