Learning Skills from First Person Demonstrations

**Project Summary** Wearable (first person) cameras such as Snapchat Sunglasses and GoPro open up a new opportunity to closely record our candid and unscripted interactions with surroundings. In particular, as following the visual attention of the camera wearer, it can measure subtle eye-hand coordination, e.g., head location, gaze direction, and finger configuration. Such subtlety has been largely ignored in prior studies of object interaction due to limited resolution of third person cameras (surveillance camera or robot mounted camera). In this research proposal, the PI proposes to develop a theory and algorithm to learn skills such as playing tennis or violin from first person cameras. However, understanding first person videos is challenging because state-of-the-art computer vision systems built upon third person videos do not apply. The first person view is highly dynamic, local, and person-biased due to severe head movements, which produces larger variations of visual data. The PI will address this challenge by building a 3D reconstruction algorithm of interactions with surrounding objects, which enables effectively learning first person demonstrations using an imitation learning framework. This computational model of interactions with environments will offer a new opportunity to train and teach robots in an intuitive, cost effective, and scalable way. It will also enable quantitative skill assessment and training.

**Intellectual Merit** First person videos naturally tell us about the camera wearer in two ways. (1) First person camera motion follows visual attention of the wearer while interacting with surroundings. The PI will leverage the 3D camera motion to recognize the objects in action, e.g., tracking the tennis ball during forehand stroke, and study the relationship between the camera and motion, i.e., how an object’s movement influences the gaze direction in 3D. An algorithm will be developed to learn the dynamic relationship from first person videos. (2) A first person video records how hands interact with the object, e.g., how grasping a stuffed animal affects the deformation of the objects. The PI will develop a theory and algorithm to reconstruct object movement and deformation from first person perspective, which will be further used to learn eye-hand coordination conditioned on context of the object interactions. To the end, these two aspects of first person cameras will be tightly integrated to build a computational model that forecasts controls similar to human demonstrations.

**Broader Impacts** A unique characteristics of human intelligence is an ability to predict future outcome. Investigating the dynamics of eye-hand coordination will make strong impact on developing robot intelligence for human-robot collaborations. Also it will provide a quantitative assessment of collaborative teamwork, e.g., which surgical team performs better in the operating room, and longitudinal behavioral analysis, e.g., which social signals trigger autistic children and how they will respond. This research will be integrated into a sequence of the PI’s newly designed computer vision courses in the University of Minnesota and textbook currently being written by the PI. Outreach programs through the College of Science and Engineering will be utilized to ensure the dissemination of the results. Also the dataset and algorithm will be open source where the PI will organize tutorial and workshop at the top computer vision venues to share broader audiences.
Imagine you learning forehand strokes from a tennis coach. The coach will demonstrate the sequence of his swings: pose preparation (gripping racket and steps), back swing, forward swing while rotating the racket, and full body rotation for following swing. He may emphasize the attack angle of the racket, foot steps, elbow locations and show the head pose that guides precise gaze direction to track the ball (eye-hand coordination). First, you may begin with imitating his sequential poses and then, you will adapt the swing to your body anthropometric configuration. Now consider a robot designed for forehand strokes. Can the robot learn from the coach’s demonstrations? A camera mounted on the robot may observe his skeletal pose and recognize the racket. To some extent, it can predict the 3D body pose and head direction. A question is that “is the robot camera perception is sufficient to learn such complex sequence of actions?”

Unfortunately, its capability is fundamentally limited due to insufficient visibility of his actions as a third person observer, e.g., a few body parts (face, finger, and elbow) are occluded by others, their visual attention cannot be precisely localized, and the contact point of the ball on the racket is not estimated accurately because of viewing angle. Prior studies taught robot either by hard coding, i.e., providing exact coordinates to perform the task, which is not completely adaptable, or by optical human motion capture, i.e., transferring actions using marker-based tracking, which requires infrastructural instrumentation of optical tracker and has limited spatial resolution (e.g., a few markers represent all fingers).

In this research program, we propose to harness a first person camera such as Snapchat Sunglasses and GoPro cameras to learn the demonstrated actions. A first person camera is an ideal sensor to measure a sequence of actions associated with eye-hand coordination because it allows us to tap into what he sees. For instance, it naturally follows the visual attention of the camera wearer and records the interactions with hit hand and racket at proximal distance, which can capture subtle details of actions at high resolution, e.g., gripping, wrist angle, and racket rotation.

A first person camera can tell us about the camera wearer in two ways: 1) the camera records the hand interactions with surroundings at the ideal perspective as the wearer actively moves to see; 2) the 3D camera egomotion reflects the head movement of the wearer and his visual attention. These two signals are reciprocal to each other through eye-hand coordination. We will leverage these two signals to build a computational model to learn skills from first person demonstrations without hard coding or motion capture systems.

However, understanding first person videos is challenging because state-of-the-art computer vision systems built upon third person videos do not apply. The first person view is highly dynamic, local, and person-biased due to severe head movements, which produces larger variations of visual data. The PI will address this challenge through following research thrust:

1. **(Thrust 1) Representing First Person Interactions**: We will study a theory and algorithm to reconstruct the interactions between hands and surroundings in 3D using first person cameras. The algorithm will recognize the physical properties of objects in actions such as mass and appearance and compute the 3D nonrigid transformation of the objects with respect to the actions. Unlike previous reconstruction methods, we will represent objects in actions in egocentric coordinate, which will allow us to construct a computational model for the interactions.

2. **(Thrust 2) Learning First Person Dynamics and Controls**: We will learn the dynamics of interaction model in Thrust 1, i.e., how will an object deform if I grasp? In particular, we will study the first person visual dynamics that allows us to model the eye-hand coordination. This learned dynamics will be used to predict control policy of human demonstrations.
These two thrusts are synergetic as shown in Figure ??: the 3D egocentric representation of first person interactions (Thrust 1) will allow us to register diverse configurations of eye-hand coordination. We will cluster and classify such configurations to learn visual dynamics of the interactions (thrust 2). The overarching goal is closing the loop, i.e., the learned dynamics and controls are integrated into learning 3D representation.

1 PI’s Prior Contribution

The core objectives of this proposal is 1) to represent the dynamic interaction between objects and actions and 2) to learn the control of actions from first person demonstrations.

1.1 3D Representation of Object-Person Interactions

Building a 3D representation of surrounding objects from a first person video is a geometrically ill-posed problem, equivalent to estimating depth from a single pixel. In prior studies, many assumptions about the scene (piecewise rigidity, shape prior, and isometry) have been made to represent highly constrained objects such as face, car, and hands \[10, 14, 31, 32\]. However, such assumptions are often object and coordinate dependent, which is not suitable to describe a general dynamic scene captured by nonstationary first person cameras. The PI proposed a coordinate independent representation of objects using a generic trajectory basis (DCT) \[19, 21, 22\], which can express general motion from moving cameras. This generic representation is highly scalable: we have demonstrated 3D reconstruction of person-object interactions at unprecedented spatial resolution (5 mm reconstruction accuracy) using the largest markerless motion capture system (480 VGA cameras) \[13\] as shown in Figure 1.

A video from a first person camera can tell us about not only the scene in front of the camera but also the photographer standing behind the camera. For instance, the slope of horizon (vanishing lines) and unique visual landmarks (e.g., kitchen and desk) in the first person image reveals the head orientation and location. Therefore, such visual cues can be used to infer full 3D camera pose using structure from motion \[11\]. The PI leverages the first person camera pose as a proxy of gaze direction (visual attention) to estimate 3D joint attention—attention shared by a group \[5, 17, 18, 20\]. Figure 2 shows that the four 3D joint attentions are found where 12 people wearing the first person cameras interact with each other in a happy hour session. Since a group of people persistently engage joint attention, i.e., they are mentally connected, it can delineate the way they interact with each other from first person videos. For instance, in a basketball game, the group attention of players moves along with the one who possesses the ball. Inspired
Figure 3: (a) We design a geometrically persistent visual representation to learn pixel affordance, e.g., is this pixel walkable? We use this representation to predict future movement from a first person image. (b) A first person video encodes the intent of the actor. We decode the intent by retrieving the control input in a form of force and torque, e.g., thrust force through pedaling and braking.

by this characteristics of the joint attention, we design an algorithm to learn the dynamics of group behaviors and further, predict their future movements [30].

Similarly, wearable cameras mounted on body parts can tell us about entire body movement. The PI has shown a 3D reconstruction algorithm for articulated human body using wearable cameras [27]. The wearable camera poses in 3D are linked through joint angle, e.g., the camera on the upper arm is connected to the one on the lower arm via elbow joint. This study eliminates the necessity of an infra-structured motion capture system to model object-person interactions.

1.2 Learning Interactions from First Person Videos

At every moment, we interact with surroundings and consciously or unconsciously make a decision. For instance, we actively steer the bike to maintain our balance and to travel to the desired future direction. First person videos encode such behavioral controls with respect to the surroundings, allowing us to understand what they intended and how they achieved it. The PI has developed a computational model to learn the way they interact with objects [7–9], scenes [15–16], and people [17,30] from their first person videos.

A key feature of first person interaction is that the visual scene is often deployed such that it affords taking actions. For instance, a first person street scene in Figure 3(a) depicts the 3D space that we can walk into (road and entrance) or cut through obstacles such as people. We design a visual representation that encodes semantics of pixels in terms of actions, e.g., is this pixel walkable? The representation (left in Figure 3(a)) is geometrically persistent to severe head motion, which enforces objects to appear similar locations and shapes. This makes learning walking actions from a large variation of first person visual data efficient and further, predicting on where to move in the future possible. Such visual affordance can be also used to identify objects important to the actor, e.g., a cup about to touch and a basketball about to throw [7,9]. Objects in action appear consistently in a first person image because the 3D pose of the objects with respect to the first person camera is highly limited due to anthropometric constraints such as arm length, sight, and personal handedness. We leverage the consistency of object appearance to recognize them in an unsupervised manner [8].

The learned visual representation of first person interactions provides a strong cue to infer the underlying intent of the action. In particular, when the actor undergoes physical motion such as mountain biking, skiing,
and skydiving, the intent/control of the motion can be retrieved, i.e., how much force is exerted by the actor to change her direction. The PI presented a method to decode the control input of the actor in a form of force and torque from first person videos as shown in Figure 3(b) [16]. The control input includes thrust force applied through pedaling and braking, steering torque to control the direction, and roll torque to balance the posture. We employed the optimal control theory with Newtonian mechanics to estimate the control input that generates the motion sensation of the video. Such control scheme can be also applied to stochastic dynamics such as social interactions where the dynamics can be learned from the first person videos [30].

2 Thrust 1: Representing First Person Interactions

Young infants between 2 and 6 month old develop their knowledge of physical world [6] based on what they see, e.g., a baby can identify that a thrown ball will eventually fall down and collide with the ground plane. This understanding of the physical world becomes more concrete as they learn to manipulate objects with their hands, i.e., relating what they see with what they touch. They build a visual representation of objects in terms of appearance (color and shape), geometrical attributes (distance and angle), and physics properties (weight, friction, and elasticity), which effects on the way they interact with the objects. This is not limited to knowledge of physics but implicit social cues where we learn through interacting with others.

A first person camera can directly measure two key visual signals regarding physical and social interactions: 1) what one sees and 2) how (s)he takes an action. First, the video naturally tracks the objects in action as following visual attention that often undergo the physical state transition in a form of motion and deformation. For instance, when passing a stuffed animal to a child, it undergoes structural deformation by grasping and translational and rotational motion by passing, which is closely recorded by the first person camera. Second, the first person camera also tells us about the actor as it partially sees body parts. This allows us to recognize what actions have been taken. In this research thrust, we will address the following two core research question to recover the two visual signals from first person videos:

- **Q1) How to represent the objects in action from a first person video?**: Objects in action undergo complex nonrigid transformations which are a function of intrinsic visual/physical properties such as appearance, shape, mass, and elasticity. We will design a theory and algorithm to reconstruct 3D object transformations by learning the intrinsic properties from first person videos.

- **Q2) How to represent the first person action from a first person video?**: A first person camera partially observes body parts associated with actions due to self-occlusion, e.g., the middle finger is often occluded by index finger when grasping. We will develop a 3D reconstruction algorithm for body parts by generating an egocentric panoramic image which can observe the interactions with surroundings from full 360 degrees.

2.1 How to Represent the Objects in Action from First Person Video?

A first person action can induce a complex kinematic transformation (rigid and nonrigid motion) of a 3D object. Representing such object transformation has been a central theme in computer vision, robotics, and graphics [30]. For instance, a three-way tensor can represent a collection of 3D points on a deformable object
\( \mathcal{X} \in \mathbb{R}^{3 \times T \times P} \) where \( T \) is the number of frame and \( P \) is the number of 3D points:

\[
\begin{align*}
\text{Shape : } \mathcal{X}^{i:t} & = \begin{bmatrix}
X_i^1 & \cdots & X_i^P \\
Y_i^1 & \cdots & Y_i^P \\
Z_i^1 & \cdots & Z_i^P
\end{bmatrix}, \\
\text{Trajectory : } \mathcal{X}^{i:p} & = \begin{bmatrix}
X_p^1 & \cdots & X_p^T \\
Y_p^1 & \cdots & Y_p^T \\
Z_p^1 & \cdots & Z_p^T
\end{bmatrix},
\end{align*}
\]

where \( \mathcal{X}^{i:p} \) is the \( p \)th frontal slice of the tensor that represents the \( p \)th 3D trajectory and \( \mathcal{X}^{i:t} \) is a lateral slice that represents the 3D shape at \( t \)th time instant as shown in Figure ??.

Reconstructing this tensor from a first person video is a geometrically ill-posed problem due to dimensional loss of image projection, i.e., at each time instant, an image observes the 2D projections of a frontal slice (different shape) and therefore, this is analogous to reconstructing 3D depth from a single image.

However, an 3D object exhibits two regularities. 1) Spatial regularity: the shape of the objects largely remains similar shape during the transformation and it is possible to compress the number of information that is needed to express with a few number of basis models, i.e., \( \mathcal{X}^{i:p} = \sum_{i=1}^{M} m_i S_i \) where \( S_i \) and \( m_i \) are spatial basis and its coefficient, respectively. 2) Temporal regularity: a point move smoothly over time and a few set of continuous functions can represent the point motion, i.e., \( \mathcal{X}^{i:t} = \sum_{j=1}^{N} n_j T_j \) where \( T_j \) and \( n_j \) are temporal basis and its coefficient, respectively. We have used analytic form of basis [22]. These regularities enable constraining 3D object transformations where 3D reconstruction is possible. Specifically, we will decompose the 3 way tensors into spatial and temporal basis:

\[
\mathcal{X} = S \times_3 C \times_3 T
\]

where \( S \in \mathbb{R}^{3 \times P \times M} \) is the shape basis tensor, i.e., the frontal slice corresponds to the shape basis, \( S_{i:m} = S_m \), \( C \in \mathbb{R}^{M \times N} \) is bilinear coefficients, and \( T \in \mathbb{R}^{N \times T} \) is the trajectory basis (can be prelearned). Prior studies in nonrigid structure from motion focused on learning a compact shape basis, \( S \), which is highly object dependent, e.g., face spatial basis cannot express the doll shape. The coefficient \( C \) is the control parameters that govern the dynamics of transformation.

**Proposed Work**

What does 3D reconstruction of object transformations mean to understand first person interactions with surroundings? This question is still not addressed because the representation is solely based on the objects not including interactions with the actor. The dynamics of object transformation needs to be represented by a function of actions, i.e., “what cause the deformation and how will it evolve?”.

Based on PI’s work that demonstrated first person object recognition [7–9], we will develop an algorithm to learn the dynamics of 3D object transformation given actions from first person interactions. We will associate intrinsic visual and physical properties of objects such as appearance, topological structure, mass, and elasticity with the 3D reconstruction algorithm, which can be recognized by first person videos. For instance, the algorithm will predict the elasticity and deformability of a stuffed animal by recognizing the object class using visual semantics (appearance and shape of the doll), which allows anticipating the amount of deformation caused by grasping. Using these intrinsic parameters, we will learn nonlinear spatial and temporal basis with respect to actions, \( S(a) \) and \( T(a) \), respectively, where \( a \) is a sequence of first person actions.

**2.2 How to Represent First Person Action from First Person Video?**

Ironically, a first person video barely sees the camera wearer. A partial body part such as a stretched arms is visible: most body parts, e.g., face, shoulder, and legs often are located beyond the field of view. Therefore,
it is challenging to recover the first person action from first person video.

Two approaches have been introduced: data driven and geometry driven approaches. 1) Given partial observations of body parts, it is possible to retrieve most similar 3D body configuration from a large corpus of data. For instance, Rogez et al. [24, 25] have leveraged an incomplete first person depth image (3D occupancy grid) to recognize arm and hand configurations by fitting the human kinematic model and further, estimate contact force and grasp taxonomy. 2) Interestingly, the camera motion induced by attached body’s movement can tell us about the body part behind the camera. The PI has demonstrated recovering 3D body motion geometrically from wearable cameras [27] and visual attention from first person cameras [5, 17, 18, 20]. Recently, body facing head-mounted camera system has been studied for egocentric motion capture that particularly observes hands, elbows, shoulders, and feet [23].

Proposed Work

Similar to 3D object transformations in Section 2.1, prior work on first person actions mainly focused on reconstructing body pose in 3D accurately. Although precise reconstruction could produce accurate modeling of body parts, would it be expressive enough to describe the interactions between first person action and objects? We argue that the first person actions need to be integrated with the dynamics of surroundings which provide a contextual and semantic meaning of the actions. For instance, we stretch our arm in order to grasp the stuffed animal. If the stuffed animal does not exists, the action means something different, e.g., stretching.

In this thrust, we will develop an egocentric representation of semantic actions associated with object interactions. A key desired feature of the representation is the ability to learn the semantics of actions directly from the dynamics of surroundings without manual supervision. For instance, a pen and pencil are both involved in writing actions (similar sequence of 3D pose with respect to the objects). We will reconstruct 3D body pose and surroundings together using EgoPanorama—a panoramic image generated by multiple head-mounted cameras that observe 360 degree of egocentric surroundings. We will learn the interactions between first person action and associated objects using the EgoPanorama which will allows us to cluster and classify objects in similar actions.

3 Thrust 2: Learning First Person Dynamics and Controls

Although the existence of mirror neuron is still controversial, it is clear that humans are capable of putting themselves in other shoes. This ability is a key element in skill learning from others’ demonstrations, e.g., we mentally visualize what the tennis coach in YouTube would see to learn eye-hand coordination. Consider a robot trying to learn from human’s demonstrations. Can a robot to mentally visualize what the humans see? In fact, this is predicting an image from different perspective, which is extremely challenging because 1) estimating gaze direction in 3D from a robot’s perspective (third person view) is unstable when the orientation of head goes beyond 45 degrees and the angular resolution is not fine enough to model gaze behaviors [20]; and 2) dense 3D reconstruction is needed to visualize the scene from the actor’s point of view, which is not geometrically possible from a single third person image, e.g., holes. Then, how can we teach robot to learn human skills?

In this research thrust, we will design an algorithm to learn skills from first person demonstrations using first person videos. A key insight is that a first person video can precisely encode the eye-hand coordination as it taps into what the person sees. For instance, a first person camera of an expert tennis player records a sequence of her gaze directions during forehand stroke. This requires not only perfect eye-hand coordination
but also predictability: how to re-design eye-hand coordination when the surroundings are changed (ball flies from different direction) and when the desired outcome are changed (sending ball to the other direction). Therefore, eye-hand coordination is not simple recording the location and orientation of hand and first person videos but learning flexible visual sensorimotor skill that can adapt to surroundings and controls. Therefore, what we have represented in the previous section is not enough.

We will study the problem by answering following two questions:

- **Q1) How to predict object interactions given first person actions?:** No interaction occurs the same situation. Even given the same task, the surroundings are different and depending on the context of the situation, our eye-hand coordination changes. We will develop an invariant algorithm that can handle various situation and re-design the coordination based on that.

- **Q2) How to learn the control policy?:** Not only surroundings but also the task can change. The desired outcome could be different from demonstrations. We will build an imitation learning framework that allows us to generalize the behaviors according to different tasks.

### 3.1 How to Predict Scene Dynamics from First Person Videos?

A quintessential property of human intelligence is the ability to predict what will happen next given an action. For instance, when opening the refrigerator’s door, we predict the pulling force that creates enough torque around the hinge to open the door. We can further predict the trajectory of the door to avoid a collision with our body. The action (pulling force) induces a state transition of the interacting object (refrigerator: closed $\rightarrow$ open) in a form of motion and deformation. Such predictions are not limited to object-person interactions but social (person-person) interactions where one’s social signals such as gaze direction and gestures trigger other person’s behaviors, e.g., the gaze directions of students follows a pointing gesture of a teacher.

The dynamics of scene interactions captured in first person videos can be expressed as follow:

$$ s_{t+1} = f_{\text{dyn}}(s_t, \cdots, s_1 | I_t, \cdots, I_1, a_t, \cdots, a_1), $$ (3)

where $I_t$ is the first person image at the $t^{th}$ time instant and $s_t$ is the state of the scenes/objects/others (3D transformation, $\mathcal{X}$ in Section 2.1 can be the state). $a$ is the controlled actions taken by the person and $f_{\text{dyn}}$ is the dynamics function that can predict the future state conditioned on actions and first person images, i.e., the first person image tells us about the contextual information about the scene interactions. For instance, $a$ can be grasping of the stuffed animal and the physical state of the object $s$ undergoes squeezed transformation.

Newton’s rules of motion is a typical example of $f_{\text{dyn}}$ without visual data. In our prior work [16], we modeled the dynamics of the actor given a sequence of first person videos, i.e., $f_{\text{dyn}} : F = ma$ where $F$ is force to accelerate the motion, $a$. We use the optimal control scheme to recognize the sequence of actions and then, predict the next physical state.

The dynamics is rather stochastic. In RL literature, this problem has been studied by learning value function, which can predict next physical state. Kalman filtering is one of such kind. However, applying this to the first person vision is challenging because 1) no global state representation and 2) it is impossible to interact with environment similar to Atari. We only have access to demonstrated activity.

In the PI’s prior work, a variant of Equation (3) has been used to learn the complex dynamics of basketball players, e.g., where they will move and look, using their first person videos [30]. A collective movements of the basketball players are modeled as state and their joint attention has been used to model the actions. We also demonstrate learning the vehicle dynamics for autonomous driving [12] where the visual data is used to recognize traffic situation.
Proposed Work

We learn skills from demonstrations by mentally visualizing their perspective based on our past experiences. Such mental visualization/simulation enables anticipating the results of our actions, which has been used for training professional athletes, i.e., imagining a sequence of actions in various situations helps them to build up highly responsive behaviors. To enable the mental visualization for a robot, the dynamics of scene states in Equation (3) needs to expand to predict next images, i.e.,

\[
(s_{t+1}, I_{t+1}) = f_{dyn}(s_t, \ldots, s_1 | I_t, \ldots, I_1, a_t, \ldots, a_1).
\]

Although this expansion seems trivial, predicting new image is a challenging task because an image is high dimensional vector (1280×960 for HD resolution), and pixels in the future image are temporally and spatially correlated with prior images which are a function of actions and depth.

Inspired by recent success in generative adversarial networks (GAN), we will design a recurrent neural network that learns the dynamics of images decoded by the GAN. In particular, we will learn visual scene dynamics of eye-hand coordination tasks captured by first person cameras. A key visual signal the network will learn is “how my first person image changes according to my actions with objects”. For proof of concept, we will plan to collect first person video data of playing table tennis and playing a musical instrument such as violin. We will associate the first person video with 3D body representation discussed in Section 2.2 to learn the visual dynamics of eye-hand coordination.

3.2 How to Learn First Person Control Policy?

Consider a pick and place task. You grab a laptop at your office and travel to your labspace by passing the corridor and finally drop the laptop on the lab desk. How do your actions change if you replace the laptop with a full cup of coffee? Your tasks are the same but the way you take the actions will be completely different: you may hold the cup higher with two hands with full attention, take a step by step extremely slowly, and look carefully the surface of the lab desk before placing. We achieve our goal by adapting our actions according to the visual scenes. The objective of this research thrust is to learn such adaptive control scheme from first person demonstrations.

Learning from first person demonstrations can be formulated as a cost minimization where the cost is expressed by difference from demonstrations:

\[
\begin{align*}
\text{minimize} & \quad D(a_1, \ldots, a_T, s_1, \ldots, s_T, I_1, \ldots, I_T) \\
\text{subject to} & \quad (s_{t+1}, I_{t+1}) = f_{dyn}(s_t, \ldots, s_1 | I_t, \ldots, I_1, a_t, \ldots, a_1)
\end{align*}
\]

where \( D \) is the cost that measure how different from demonstrations. The dynamics of states and first person images are embedded in Equation (5), i.e., the best sequence of actions are chosen to produce the desired states and images reflective of the demonstrations. If \( D \) can be defined by a quadratic cost, it is possible to compute the optimal control policy (linear quadratic regulator). The PI demonstrated that it is possible to compute physical force and torque via the optimal control theory from a first person sport videos [16] where \( f_{dyn} \) is defined by Newton’s second rule of motion, \( F = ma \). We also show that when the dynamics is purely modeled by images, the future trajectories of first person’s movement can be predicted in a urban scene [15] and social interactions [30].
Proposed Work

A key task for learning first person control policy is to design the cost function $D$ in Equation (5) and efficient optimization algorithm. In our preliminary study, we show that a convolutional generative adversarial imitation learning (cGAIL) is efficient to learn the complex visual dynamics of autonomous vehicle where the algorithm predicts a sequence of controlled actions (acceleration, brake, and steering) accurately [12]. This deep neural network is tuned to generate actions that produce trajectories similar to demonstrated trajectories given the visual traffic situation.

We will further leverage the cGAIL algorithm to understand the eye-hand coordination with respect to the 3D representation of interactions discussed in Section 2.1. The first person action data (table tennis and violin) will be used for evaluation, e.g., forecasting future gaze and hand movements.

4 Evaluation

5 Broader Impacts

5.1 Societal Impact

Time is ripe: social robots are poised to enter our social space, and a small form factor of a video camera accelerates seamless integration of first person (wearable) cameras into our bodies such as Snapchat Sunglasses [33], GoPro Hero, FirstVision [1], and Sony SmartEyeGlass [2]. Our proposed system will integrate these two emerging technologies to capture our daily lives that has societal impact in computational behavioral science and human-robot collaboration.

Computational Behavioral Science Two factors fundamentally limits the efficacy of early detection of behavioral diseases such as autism spectrum disorder. 1) Subjective bias: unlike physiological signals such as heart beat, no quantitative measurement unit exists for behavioral analysis. Therapists and doctors rely on their visual inspections through a few session interaction or videos to assess the children’s behaviors, and therefore, it is not scalable and requires a consensus of multiple decisions. 2) Sporadic and short term measurements: behavioral diseases develop progressively, and therefore, longitudinal behavioral measurement is key for early detection. However, current behavioral studies focus on a short duration of a therapy session (at most 1-2 hours per month) that significantly prevents from understanding the progression of autistic behaviors. Our social event capture system will use multiple robots to continually record children’s behaviors over long duration of time, which will enable longitudinal behavioral study. In conjunction with a computational encoding of social behaviors using 3D reconstruction [13][19][22][27], our proposed work will enable a quantitative identification of the behavioral diseases. This can also apply to monitoring the physical condition of the elderly.

Human-robot Collaboration Close collaboration requires “putting yourself in her shoes”. We forecast others’ behaviors from their perspectives and adjust our behaviors accordingly. Our work will allow us to teach human skills to a robot without hard coding or marker based motion capture which will be readily applicable to work together as a team member. Also the output of this research can influence the development of personal assistant robots.

5.2 Outreach and Curriculum Development

The PI is leading the development of a new series of computer vision courses at the University of Minnesota (UMN). The sequence includes a general introductory course (CSCI 5561) followed by two advanced
courses focused on geometry (CSCI 5563) and statistical learning (CSCI 5565). Smaller problems from the proposed work will be given as course projects to ensure broader participation.

PI Park has given multiple workshop and tutorials in top tier computer vision conferences related to the proposed program: ECCV 2014 Workshop on “Human Behavior Understanding”, CVPR 2015 Tutorial on “Group Behavior Analysis and Its Application”, and CVPR 2016 Tutorial on “First Person Vision”. PI Park will give a CVPR 2017 Tutorial on “DYI Multicamera System: Panoptic Studio Teardown” which will study the visual representation of social interactions. Also the PI and co-PI will actively bridge the gap between robotics and computer vision by organizing workshop and challenges on social capture in CVPR and ICRA. In particular, co-PI Isler co-organized ICRA 2015 workshop on “Scaling Up Active Perception”. This proposed research will produce an ideal data and materials for the PI’s lectures, “CSCI 5980/8980: Multiview 3D Geometry in Computer Vision” and “CSCI 5563: 3D Geometry in Computer Vision”.

We will involve undergraduates in this research both through NSF’s REU supplement mechanism and the Undergraduate Research Opportunities Program (UROP) at the UMN. We will use programs in place at UMN to facilitate hiring of students from under-represented groups. For example, we will advertise research opportunities to Puckett Scholars (the program gives preference to students of color who demonstrate strong academic achievement, leadership and/or community involvement, financial need, and who have excellent potential to succeed at the college level) and the Jackie Robinson Foundation Scholars (the program provides scholarships of up to $7,500 annually to minority high school students showing leadership potential and demonstrating financial need to attend an accredited 4-year college or university of their choice).

This research will also be integrated into the Summer Technology Camp which is run by the College of Science and Engineering. The program offers hands-on experience with robots, programming, user interfaces, and computer vision to more than 25 middle-schoolers from underrepresented groups. We believe that such experiences, where students participate in the development of cutting-edge technology, will encourage them to pursue future careers as scientists and engineers.

6 Results from Prior NSF Support

PI Park has recently started his first tenure-track appointment and has not yet received any support from NSF.
References Cited


