Force from Motion: Decoding Physical Sensation in a First Person Video

Hyun Soo Park Jyh-Jing Hwang Jianbo Shi
University of Pennsylvania
{hypar,jyh,jshi}@seas.upenn.edu

Abstract

A first-person video can generate powerful physical sensations of action in an observer. In this paper, we focus on a 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.

The sensation of gravity can be observed in a natural image. We 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. 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.

1. Introduction

A wingsuit BASE jumper, Jeb Corliss, dives from a cliff in Alps with his body-mounted GoPro camera[1] (Figure 1). This camera records a beautiful scenery that he had seen but also captures what he *experienced* and *controlled* via the camera egomotion. This egomotion is a resultant of physical interactions between passive forces from environments (e.g., gravity and air pressure) and active forces exerted by him to control his egomotion, e.g., angular momentum change along the roll axis to shift the heading direction. In this paper, we study a problem of *Force from Motion*—reconstructing force and torque from an egocentric video to revive the physical sensation.

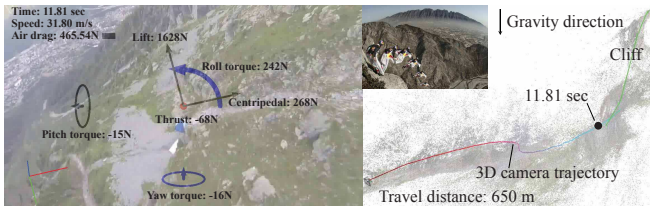


Figure 1. This paper presents *Force from Motion*—decoding the sensation of 1) passive forces such as the gravity, 2) the physical scale of the motion (speed) and space, and 3) active forces exerted by the observer. We model egomotion with rigid body dynamics integrated in a bundle adjustment that allows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

Extracting such forces requires to explicitly measure his muscle tension—the acceleration computed by a camera or inertial measurement unit (IMU) is not directly applicable to find active forces exerted by him because only net acceleration can be measured. Our key question is "can we extract his input in a form of active force and torque without measuring muscle tension from an egocentric video?"

We show that it is possible to estimate an active force and torque profile that generates the egomotion. This requires to overcome three fundamental challenges: a) limited observations of body parts (body pose is often not visible from an egocentric video); b) scale and orientation ambiguity inherent in structure from motion; c) scene and activity variability (different appearance, camera placement, and motion).

We address these challenges by modeling the observed camera egomotion with rigid body dynamics that integrates three key sensations: 1) gravity force; 2) physical scale of the world; and 3) input force and torque.

The gravity force sensation is captured in the visual image itself. The gravity affects how physical environment is formed, i.e. trees and buildings are usually vertical and horizon perpendicular to gravity direction. We learn such image

[1] https://www.youtube.com/watch?v=IM1vss7FXs8

cues to predict a 2D gravity direction in a 2D image using a convolutional neural network designed to recognize the orientation of the image. The prediction of multiple frames is consolidated using 3D reconstructed camera orientation to estimate the 3D gravity direction. Note only the camera orientation information is needed in this step, and we are still affected by the unknown scale factor.

The physical scale of the space is important sensation since it tells us how fast we are going exactly. The absolute scale of our motion is revealed to us when the body is in a dynamically balanced state. During a banked turn, the torques produced by centripetal force and gravity force are balanced with the body leaning angle. This physical constraint together with the known gravity constant, i.e., 9.81 m/s^2, allows us to compute the physical scale exactly.

The input force sensation includes 3D active force (thrust) and torque (roll and yaw). For each type of first person sport video, we construct a rigid body dynamics and model egomotion as a function of the input forces and gravity. Given the physical scale and gravity direction, we minimize the 2D geometrical reprojection error (in video) with respect to the unknown 3D world and egomotion governed by rigid body dynamics. The reconstructed camera egomotion that is corrected by physical scale and gravity direction is used for an initialization of the bundle adjustment for active force and torque estimation.

In total, our system takes an input, a first person sport video, and outputs active force and torque profile in metric scale as shown in Figure 1. We predict the 3D gravity direction by integrating 2D prediction by a convolutional neural network and recover physical scale using the roll torque equilibrium. These factors are embedded in the bundle adjustment that finds a plausible active force and torque profile that can simulate the camera egomotion via inverse dynamics while simultaneously minimizing reprojection error.

Why Egocentric Video? As a form factor of a video camera facilitates seamless integration into body, hundreds of thousands of egocentric videos are captured and shared via online video repositories such as YouTube, Vimeo, and Facebook. For instance, currently more than 6,000 GoPro videos are posted in YouTube in a day. Many of these videos capture speed sport activities such as downhill mountain biking (1-10 m/s), glade skiing (5-12 m/s), skydiving (60-80 m/s) from first person view. These videos excite visual motion stimuli that are strongly dominated by physical sensation. Decoding such physical sensation provides a new computational representation of such videos that can be not only applied to vision tasks such as activity recognition, video indexing, content generation for virtual reality [32] but also computational sport analytics [28], sensorimotor learning [39], and sport product design [7].

Contributions This paper includes three core technical contributions. (1) Force from motion: we integrate rigid body dynamics into a bundle adjustment to estimate active force and torque profile; (2) Gravity direction estimation: we learn image cues to predict gravity direction and upgrade to 3D by employing the reconstructed camera orientations; (3) physical scale recovery: we recover a scale factor from the roll torque equilibrium relationship. We quantitatively evaluate our method using a controlled experiment with inertial measurement units (IMU). 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.

2. Related Work

This paper studies physics based human behavior modeling via egocentric vision. In this section, we briefly review the most related work.

2.1. Human Behavior Modeling in 3rd Person View

Johansson's experiment [12] has shown that human motion can be perceived and predicted by a sparse representation with short duration of visual observation. However, enabling such perception for a machine is still challenging without prior knowledge due to a large degree of freedom of an articulated body structure. This requires a compact representation to describe human body motion. While a large body of literature have studied this problem based on geometry [2, 34, 42] and statistical model [33, 6, 35], we focus on physics based representation.

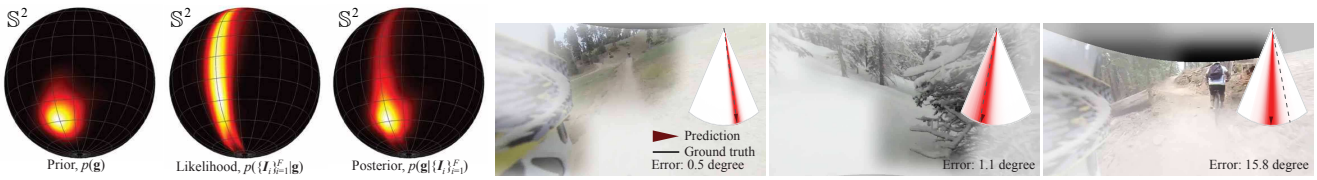
Markerless motion capture often benefits from physics based approaches[2]. Brubaker et al [4, 3] explicitly modeled the ground reaction force as an impulse function during bipedal walking. Wei and Chai [38] have shown a keyframe based human motion reconstruction where physics based simulation interpolates between keyframes. Vondrak et al. [37, 36] introduced a feedback control system based on multibody dynamics that provides a Bayesian prior to track human body motion.

2.2. Egocentric Perception

An egocentric camera is a powerful tool to understand human behaviors as it records what the camera wearer has experienced. Therefore, it is a viable solution for behavior science and quality of life technology [13, 27, 26], and this motivates many vision tasks such as understanding fixation point [18], identifying eye contact [43], and localizing joint attention [9, 24].

An egocentric video is biased by camera egomotion which is highly discriminative for activity recognition. Fathi et al. [8, 9] used gaze and object segmentation cues to classify activities. 2D motion features were exploited by Kitani et al. [14] to categorize and segment a first person sport video in a unsupervised manner. Coarse-to-fine

[2]Other applications of physics based approaches have been used to infer motion [20, 40].



(a) Gravity distribution (b) CNN Visualization

Figure 2. (a) We compute a maximum a posteriori estimate of the 3D gravity direction, $\hat{\mathbf{g}} \in \mathbb{S}^2$. We model the prior using a mixture of von Mises-Fisher distributions and learn a likelihood function using a convolutional neural network (CNN). (b) We show the likelihood given an image with the red heatmap. The dotted lines are the ground truth gravity direction. The per pixel evidence [19] is encoded as transparency, i.e., the stronger evidence, the more transparent. The CNN correctly predicts gravity direction while the last image produces 15 degree error due to the tilted bicycler.

motion models [30] and a pretrained convolutional neural network [31] provided a strong cue to recognize activities. Yonetani et al. [44] utilized a motion correlation between first and third person videos to recognize people's identity. Kopf et al. [15] stabilized first person footage via 3D reconstruction of camera egomotion. In a social setting, joint attention was estimated via triangulation of multiple camera optical rays [24] and the estimated joint attention was used to edit social video footage [1].

Another information that the egocentric camera captures is exomotion or scene motion. Pirsiavash and Ramanan [25] used an object centric representation and temporal correlation to recognize active/passive objects from a egocentric video, and Rogez et al. [29] leveraged a prior distribution of body and hand coordination to estimate poses from a chest mounted RGBD camera. Lee et al. [17] summarized a life-logging video by discovering important people and objects based on temporal correlation, and Xiong and Grauman [41] utilized a web image prior to select a set of good images from egocentric videos. Fathi et al. [9] used observed faces to identify social interactions and Pusiol et al. [26] learned a feature that indicates joint attention in child-caregiver interactions.

Our approach: To our best knowledge, this is the first paper that provides a computational framework to understand an egocentric video based on physical body dynamics. We leverage two motion cues: 1) 3D reconstruction from egomotion, and 2) gravity and scale recovery from exomotion. As an egocentric video has limited observation of body parts, estimating force and its control significantly differs from previous problems of physics based tracking and reconstruction. We introduce a novel *Force from Motion* method that computes the control input applied by the camera wearer. It also produces a scaled and oriented 3D reconstruction via dynamics.

3. Force from Motion

Gravity, scale, and active force are three key ingredients that generate physical sensation in movement. In this section, we estimate these physical quantities.

3.1. Gravity Direction

A natural image encodes gravity direction because it affects how physical environment is formed, i.e. trees and buildings are usually vertical and horizon perpendicular to gravity direction [23, 10]. We exploit such image cues learned by a convolutional neural network [16] to predict a gravity direction in a 2D image. This per image prediction is integrated over multiple frames by leveraging structure from motion.

We define a 3D unit gravity direction, $\hat{\mathbf{g}}(\theta, \phi) = \begin{bmatrix} \sin\theta\cos\phi & \sin\theta\sin\phi & \cos\theta \end{bmatrix}^\mathsf{T} \in \mathbb{S}^2$. We normalize the representation with respect to the instantaneous velocity direction such that $\hat{\mathbf{g}}(0,0) = \mathbf{v}/\|\mathbf{v}\|$ where $\mathbf{v}$ is the instantaneous velocity. This allows us to register different camera orientations in an unified coordinate system (with respect to the gravity).

We compute the maximum a posteriori (MAP) estimate of the gravity direction given a set of images, $\{\mathcal{I}_i\}_{i=1}^F$:

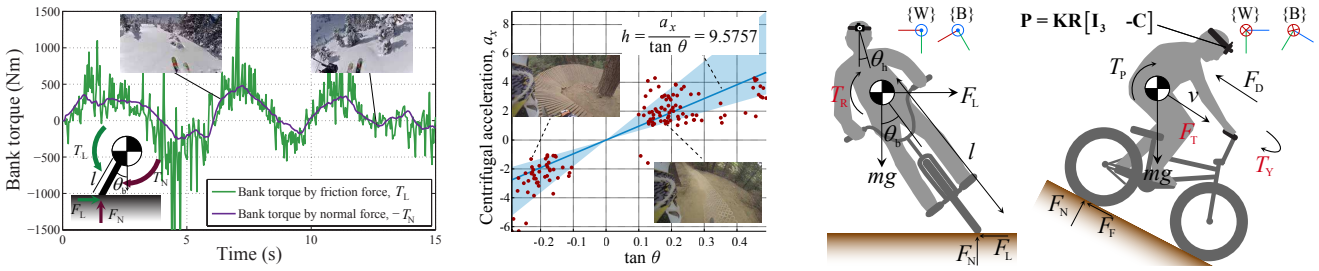
$$\begin{aligned} \hat{\mathbf{g}}^* &= \underset{\hat{\mathbf{g}}\in\mathbb{S}^2}{\text{argmax}}\ p(\hat{\mathbf{g}}|\mathcal{I}_1, \cdots, \mathcal{I}_F) \\ &= \underset{\hat{\mathbf{g}}\in\mathbb{S}^2}{\text{argmax}}\ p(\hat{\mathbf{g}}) \prod_{i=1}^{F} p(\mathcal{I}_i|\hat{\mathbf{g}}), \end{aligned} \tag{1}$$

where $p(\hat{\mathbf{g}})$ is a prior distribution of the gravity direction and a likelihood $p(\mathcal{I}_i|\hat{\mathbf{g}})$ measures how well the 3D gravity direction is aligned with image, $\mathcal{I}_i$.

The prior distribution encodes how the gravity is oriented with respect to the heading direction. Given a gravity direction in a training dataset[3], we model this prior distribution using a mixture of von Mises-Fisher distributions:

$$p(\hat{\mathbf{g}}) = \sum_{k=1}^{K} \frac{\kappa_k}{4\pi \sinh \kappa_k} \exp\left(\kappa_k \hat{\mathbf{g}}^\mathsf{T} \hat{\mathbf{m}}_k\right) \tag{2}$$

[3]Our training data consists of 32 Bike, 19 Ski, 23 Urban bike, 23 Jetski, 29 Wingsuit fly, and 30 Speed fly sequences and each sequence ranges between 1 mins to 38 mins. We annotate the 2D gravity direction of images in the training set and reconstruct it in 3D. This 3D reconstructed gravity allows us to propagate over 100 frames. Optionally, we also use IMU attached camera to automatically annotate the gravity. See the supplementary material for the detailed description of the training data.



(a) Bank torque (b) Scale factor (c) Geometry

Figure 3. (a) We recover the physical scale of a 3D reconstruction by exploiting the torque equilibrium at a banked turn where the torques generated by normal force and centripetal force $T_N + T_L = 0$ must be canceled to maintain the leaning angle, θ_b. (b) The scale factor can be estimated by the slope, $|\mathbf{a}_x|/\tan\theta_b$. (c) We model the egomotion of a camera wearer using single rigid body dynamics (6 degree of freedom). Force and torque are decomposed into passive components (gravity, $m\mathbf{g}$; centripetal force, F_L; normal force, F_N; friction force, F_F; air drag, F_D; pitch torque, T_P) and active components (thrust, F_T; roll torque, T_R; yaw torque, T_Y).

where $\{\hat{\mathbf{m}}_k, \kappa_k\}$ is a set of modes and concentration parameters that can be learned by an Expectation-Maximization algorithm as shown in Prior of Figure 2(a).

The image likelihood, $p(\mathcal{I}_i|\hat{\mathbf{g}})$ measures how well the projected 3D gravity direction onto the i^{th} image agrees with the image cues learned from the training data. By the projection, we measure the orientation of the image, $\xi = \text{atan2}(\mathbf{r}_2^\mathsf{T}\hat{\mathbf{g}}, \mathbf{r}_1^\mathsf{T}\hat{\mathbf{g}}) \in \mathbb{S}$ where $\mathbf{R}(t_i) = \begin{bmatrix} \mathbf{r}_1^\mathsf{T} & \mathbf{r}_2^\mathsf{T} & \mathbf{r}_3^\mathsf{T} \end{bmatrix}^\mathsf{T}$ and $\mathbf{R}(t) \in SO(3)$ is the camera orientation at the t^{th} time instant. We learn this likelihood function using the convolutional neural network (CNN) proposed by Krizhevsky et al. [16] with a few minor modifications. We correct the fisheye lens distortion and warp the image with a homography, $\mathbf{H} = \mathbf{K}\mathbf{R}_\mathbf{v}\mathbf{R}(\varphi_p)\mathbf{R}(\varphi_r)\mathbf{R}^\mathsf{T}\mathbf{K}^{-1}$ where $\mathbf{K}$ and $\mathbf{R}$ are the camera intrinsic parameter and orientation matrices, respectively. $\mathbf{R}_\mathbf{v}$ is the rotation matrix whose Z axis aligns with the instantaneous velocity, $\mathbf{v}$. The body coordinate system, $\{\mathbf{B}\}$ is defined in Figure 3(c), and $\mathbf{R}(\varphi_p)$ is the constant rotation about the pitch axis to minimize the area outside of the image. $\mathbf{R}(\varphi_r)$ is a rotation about the roll axis used for data augmentation. The warped image (1280×720) is resized to 320×180 as an input for the CNN. We train the network to predict a probability of the projected angle ξ discretized by 1 degree between $-30°$ and $30°$, i.e., $\xi = 0$ means the gravity direction is aligned with y axis of the image. We augment the data by rotating the image with $\mathbf{R}(\varphi_r)$ and its horizontal flip. Figure 2(b) illustrates the likelihood of the gravity directions learned by CNN as shown in the red heatmap and the ground truth gravity direction with dotted line.

Predictions on multiple images are consolidated by the 3D reconstructed camera orientations. Note that a single image cannot predict the 3D gravity direction due to 2D projection. Each image produces a streak in a likelihood distribution as shown in Likelihood of Figure 2(a)—any gravity direction along the streak is projected onto the same direction in 2D. The product of multiple image predictions in Equation (1) by leveraging the 3D reconstructed camera orientations can collapse the streak into a unimodal distribution[4].

3.2. Physical Scale

The leaning angle, θ_b, at a banked turn is formed to balance the roll torque at the center of mass. The normal force, F_N, produces a torque, $T_N = lF_N\cos\theta_b$ and the friction force, or centripetal force (no slip condition), F_L produces an opposite directional torque $T_L = lF_L\sin\theta_b$ with respect to the center of mass where l is the distance between the center of mass to the ground contact point as shown in Figure 3(a) and 3(c). These two torques must be balanced to maintain the leaning angle, i.e., the tangential velocity, $\mathbf{v}$, is defined by the leaning angle and the curvature of the turn.

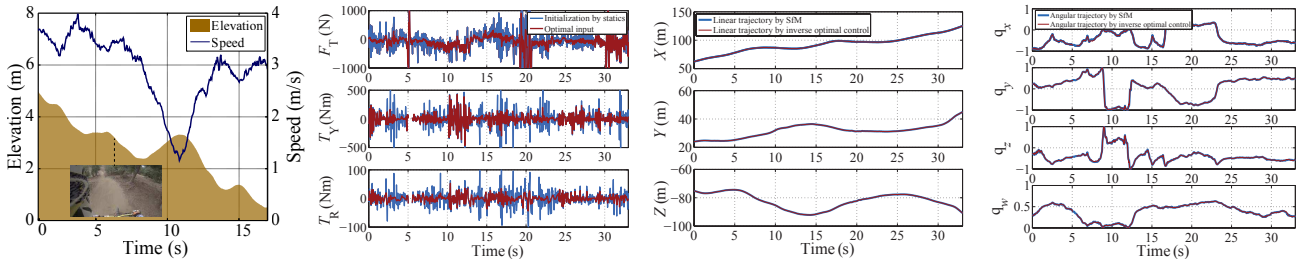
By equating these two torques, i.e., $T_L + T_N = 0$, we obtain the following relationship with gravity constant:

$$\|\mathbf{g}\| = 9.81\ \text{m/s}^2 = c\frac{|\hat{\mathbf{a}}_x|}{\tan\theta_b}, \tag{3}$$

where $\hat{\mathbf{a}}_x$ is the linear acceleration in the lateral direction, which is measured from the reconstructed 3D camera trajectory in $\{\mathbf{W}\}$ (Figure 3(c)) and c is a scale factor that maps from the 3D reconstruction to the physical world.

In Figure 3(b), we plot the scale factors measured from different time instances with their median and variance. The slope of the data points represents the scale factor of the reconstruction. We compute these data points along the video sequences that include a number of banked turns. Figure 3(a) shows the torques produced by the scale factor and two torques are roughly canceled out. Note that $-T_N$ is plotted for a direct comparison. This allows us to reconstruct physical dimension of the terrain and speed as shown in Figure 4(a). Note that the speed profile is physically meaningful, i.e., average speed of the mountain biking ranges between 1-6 m/s^2.

[4]If ones goes straight without changing camera orientation, the streak remains constant as shown in Likelihood of Figure 2(a).



(a) Elevation (b) Optimal active force (c) Linear trajectory (d) Angular trajectory

Figure 4. (a) The recovered gravity direction and scale allow us to identify physical dimension of elevation and speed. (b) We compute active force and torque by parametrizing them into a bundle adjustment in Equation (7). (b) The bundle adjustment in Equation (7) produces plausible active force and torque profile that produces a camera trajectory concerting with the video ((c) and (d)).

3.3. Physics of Rigid Body Dynamics

A single rigid body that undergoes motion as a resultant of forces and torque can written as:

$$m\mathbf{a} = \mathbf{F}_{\text{in}} + \mathbf{F}_{\text{ex}} \tag{4}$$

$$\mathcal{J}\boldsymbol{\alpha} + \boldsymbol{\omega} \times \mathcal{J}\boldsymbol{\omega} = \mathbf{T}_{\text{in}} + \mathbf{T}_{\text{ex}}, \tag{5}$$

where m is mass, $\mathbf{a} \in \mathbb{R}^3$ is linear acceleration, $\boldsymbol{\alpha} \in \mathbb{R}^3$ is angular acceleration, $\mathcal{J} \in \mathbb{R}^{3\times3}$ is moment of inertia, and $\boldsymbol{\omega} \in \mathbb{R}^3$ is angular velocity. We denote $\mathbf{F}_{\text{in}}$ and $\mathbf{T}_{\text{in}}$ as active force and torque that are applied by the camera wearer (input signal). $\mathbf{F}_{\text{ex}}$ and $\mathbf{T}_{\text{ex}}$ are passive force and torque that are applied by external sources such as gravitation force, centripetal force, and pitch moment created by an unbalance impact between two wheels in a bicycle as shown in Figure 3(c). Note that a biking is used for an illustrative purpose while this dynamics can be applied general activities such as skiing, jetskiing, speedflying, and wingsuit flying with a few minor modifications such as body mass, moment of inertia, and air lift instead of normal force for a flying activities[5].

We represent Equation (4) in the world coordinate system, $\{\mathbf{W}\}$, and Equation (5) in the body coordinate system, $\{\mathbf{B}\}$[6] as shown in Figure 3(c). The active force and torque are composed of thrust force F_{T}, roll torque T_{R}, and yaw (steering) torque T_{Y}:

$$\mathbf{F}_{\text{in}} = F_{\text{T}}\frac{\mathbf{v}}{\|\mathbf{v}\|}, \qquad \mathbf{T}_{\text{in}} = \begin{bmatrix} 0 & T_{\text{Y}} & T_{\text{R}} \end{bmatrix}^{\mathsf{T}},$$

where the thrust force is applied along the velocity direction, $\mathbf{v}$[7].

The passive force and torque are composed of the following components:

$$\mathbf{F}_{\text{ex}} = m\mathbf{g} + (F_{\text{D}} + F_{\text{F}})\frac{\mathbf{v}}{\|\mathbf{v}\|} + \begin{bmatrix} F_{\text{L}} & F_{\text{N}} & 0 \end{bmatrix}^{\mathsf{T}}$$

$$\mathbf{T}_{\text{ex}} = \begin{bmatrix} 0 & 0 & lF_{\text{N}}\sin\theta_{\text{b}} - lF_{\text{L}}\cos\theta_{\text{b}} \end{bmatrix}^{\mathsf{T}},$$

where $\mathbf{g} = \begin{bmatrix} 0 & 9.81 & 0 \end{bmatrix}^{\mathsf{T}}$ m/s^2 is the gravitational acceleration, $f_{\text{D}} = -0.5C_{\text{D}}\rho A\|\mathbf{v}\|^2$ is the air drag force where $C_D \approx 1.0$, $\rho = 1.23$ kg/m^3, and A are air drag coefficient, air density, and cross sectional area perpendicular to the velocity, respectively. $F_{\text{F}} \leq 0$ and F_{L} are frictions along velocity and lateral directions, respectively. l is the distance between the center of mass and the ground contact point, and θ_b is the body leaning angle.

Equation (4) and (5) can be together written as a compact form:

$$\mathcal{M}\ddot{\mathbf{q}} + \mathcal{C}(\dot{\mathbf{q}}) = \mathbf{J}\mathbf{u} + \mathbf{E}, \tag{6}$$

where $\mathcal{M}$ is the inertial matrix, $\mathcal{C}$ is the Coriolis matrix, $\mathbf{E}$ is the passive force and torque, and $\mathbf{u} = \begin{bmatrix} F_{\text{T}} & T_{\text{R}} & T_{\text{Y}} \end{bmatrix}$ is the active component. The state $\mathbf{q} = \begin{bmatrix} \mathbf{C}^{\mathsf{T}} & \boldsymbol{\Omega}^{\mathsf{T}} \end{bmatrix}^{\mathsf{T}}$ describes the camera egomotion where $\mathbf{C} \in \mathbb{R}^3$ is the camera center and $\boldsymbol{\Omega} \in \mathbb{R}^3$ is the axis-angle representation of camera rotation, i.e., $\exp\left([\boldsymbol{\Omega}]_{\times}\right) = \mathbf{R} \in SO(3)$ where $[\cdot]_{\times}$ is the skew symmetric representation of the cross product [21]. $\mathbf{J}$ is a workspace mapping matrix written as:

$$\mathbf{J} = \begin{bmatrix} \mathbf{v}^{\mathsf{T}}/\|\mathbf{v}\| & 0 & 0 & 0 \\ \mathbf{0} & 0 & 1 & 0 \\ \mathbf{0} & 0 & 0 & 1 \end{bmatrix}^{\mathsf{T}}.$$

Equation (6) describes motion in terms of active force and torque component, $\mathbf{u}$, which allows us to directly map between input and the resulting motion. Solving for $\mathbf{u}$ is inverse dynamics that is integrated in our bundle adjustment in Section 4.

4. Inverse Dynamics for Optimal Control

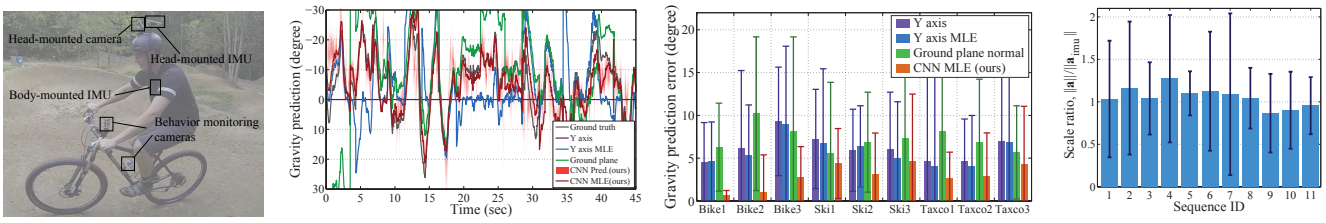
We integrate three ingredients for physical sensation, gravity direction, physical scale, and active force and torque into the following cost:

$$\begin{aligned} \underset{\mathbf{u}(t),\{\mathbf{X}_j\}}{\text{minimize}} \quad & \sum_{i,j}\mathcal{D}\left(\mathbf{P}(t_i)\mathbf{X}_j, \mathbf{x}_{ij}\right) + \lambda_{\mathcal{R}}\int_0^T \dot{\mathbf{u}}(t)^{\mathsf{T}}\dot{\mathbf{u}}(t)\mathrm{d}t \\ \text{subject to} \quad & \mathbf{P}(t_i) = \mathbf{K}\mathbf{R}(t_i)\begin{bmatrix} \mathbf{I}_3 & -\mathbf{C}(t_i) \end{bmatrix} \\ & \mathcal{M}\ddot{\mathbf{q}} + \mathcal{C}(\dot{\mathbf{q}}) = \mathbf{J}\mathbf{u} + \mathbf{E}, \end{aligned} \tag{7}$$

[5] See the supplementary material for activity dependent coefficients.

[6] Forces in world coordinate system are semantically meaningful as the Y axis aligns with the gravity direction while torques in the body coordinate system are more interpretable (roll, pitch, and yaw) [21].

[7] The choice of the input force and torque components depends on the constraints of motion while it has to satisfy the controllability criterion.



(a) Experiment setup (b) Gravity prediction (c) Gravity prediction comparison (d) Scale recovery

Figure 5. (a) For quantitative evaluation, we design a control experiment with an experienced mountain biker. (b) We compare our prediction with three baseline algorithms (see the description of the baseline algorithm in Section 5.1. The red heatmap indicates the likelihood at each time instant $p(\mathcal{I}|\hat{\mathbf{g}})$. Our predictor uses the image likelihood in conjunction with the reconstructed camera orientation. (c) We measure error across different scenes. (d) We recover physical scale and compare with IMU in terms of linear acceleration. Our method correctly estimate the scale (perfect recovery if 1; median 1.0287 with 0.6186 standard deviation).

where $\mathcal{D}$ measures reprojection error, i.e., $\mathcal{D}(\mathbf{x}_1, \mathbf{x}_2) = (x_1/z_1 - x_2/z_2)^2 + (y_1/z_1 - y_2/z_2)^2$ where $\mathbf{x} = \begin{bmatrix} x & y & z \end{bmatrix}^\mathsf{T}$. $\mathbf{P}(t_i) \in \mathbb{R}^{3\times4}$ is the camera projection matrix at time t_i instant, $\mathbf{X} \in \mathbb{P}^3$ is a 3D point, and $\mathbf{x}_{ij} \in \mathbb{P}^2$ is the j^{th} 2D point measurement at t_i time instant. The goal is to infer both the unknown 3D world structure $\mathbf{X}$, as well as control forces for the rigid body dynamics, $\mathbf{u}(t)$, assuming the gravity force and the scale of the space is given. The last term in the cost function regularizes active forces such that the resulting input profile over time is continuous. $\lambda_{\mathcal{R}}$ is a control weight for input regularization.

Equation (7) consolidates a bundle adjustment cost from structure from motion with the optimal control theory that finds the optimal control profile to generate the desired output trajectory. Equation (7) is highly nonlinear due to reprojection error and rigid body dynamics, which requires a good initialization. We reconstruct 3D discrete camera pose trajectory, $\{\mathbf{P}(t_i)\}$, and a set of 3D points, $\{\mathbf{X}_j\}$, using structure from motion. The acceleration and velocity of the camera pose are approximated by differentiating the discrete camera pose. This allows us to approximate active and passive components, $\mathbf{u}$ and $\mathbf{E}$, by solving statics, i.e., each time instant independently. Given this discrete input profile, we build a continuous piecewise linear function as an initialization of $\mathbf{u}(t)$.

We minimize the objective function using Levenberg-Marquardt algorithm [22] where the ordinary differential equations for rigid body dynamics are solved via the Runge-Kutta 4th-order method on $SE(3)$ [5] whenever evaluating the objective function. Figure 4(b) shows comparison between initialization and the refined active force and torque. The initialization contains implausible input profile due to noisy acceleration computation. The optimization allows us to find the smooth input profile that simultaneously minimizes reprojection error, which agrees with structure from motion result as shown in Figure 4(c) and 4(d).

5. Result

We evaluate our algorithm on real world data. For all sequences, a camera trajectory is reconstructed by structure from motion at 30 Hz. We assume all videos have the fixed resolution (1280 × 720) and intrinsic parameters (focal length, principle coordinates, and fisheye lens distortion) because 97% of first person sport videos are taken by the same mode of GoPro 2 Hero or GoPro 3 Hero. We use 29 Bike sequences ranging from 5 mins to 20 mins (about 1 million images[8]) to fine-tune the CNN pre-trained by [16] using Caffe [11]. For computational efficiency, we divide a video into a set of 10 second videos (300 frames) to optimize Equation (7).

5.1. Quantitative Evaluation

We quantitatively evaluate our algorithm with a controlled experiment conducted by an experienced mountain biker with head-mounted inertial measurement unit (IMU) as shown in Figure 5(a). Additional IMU was attached on his body to measure disparity between head and body motion[9]. Two cameras are also attached on the bike to monitor his behaviors such as pedaling and braking. Our evaluations are performed to verify our method in three criteria: gravity prediction, scale recovery, and active force and torque estimation.

Gravity prediction We compare our prediction using CNN and reconstructed camera orientation with three baseline methods: a) Y axis: prediction by the image Y axis as a camera is often oriented upright; b) Y axis MLE: prediction by a) consolidated by the reconstructed camera orientation; c) ground plane normal. The ground plane is estimated by fitting a plane with RANSAC on the sparse point cloud. Figure 5(b) shows a comparison with baseline algorithms where our method produces median error 2.7 degree with 3.64 standard deviation (mean: 4.40 degree). Note that we do not compare our final MAP estimate for fair comparison. We also test our method on manually annotated data in Figure 5(c) where our method consistently outperforms others significantly ($\times2 \sim \times10$). Note that only biking sequences

[8]Note that the scenes change rapidly due to fast egomotion and thus, the data capture variety of scene cues.

[9]A quantitative analysis on the relationship between body and gaze orientation is included in the supplementary material.

	Bike 1			Bike 2			Bike 3			Bike IMU			Ski 1			Ski 2			Ski 3			Taxco 1			Taxco 2			Taxco 3		
	Mean	Med.	Std.	Mean	Med.	Std.	Mean	Med.	Std.	Mean	Med.	Std.	Mean	Med.	Std.	Mean	Med.	Std.	Mean	Med.	Std.	Mean	Med.	Std.	Mean	Med.	Std.	Mean	Med.	Std.
Y axis	5.62	4.44	4.72	8.10	6.18	9.06	10.15	9.29	6.34	16.02	13.11	10.88	8.31	7.24	5.80	8.11	7.37	6.94	6.86	5.93	4.79	8.00	4.62	13.10	5.77	4.66	4.92	9.66	7.00	8.84
Y axis MLE	5.92	4.57	4.66	6.08	5.31	5.91	10.68	8.97	9.11	15.83	12.28	11.21	10.09	6.72	8.72	7.80	6.54	6.28	7.00	6.37	4.75	6.90	4.06	12.73	5.94	4.01	5.97	10.41	6.83	10.85
Ground plane	7.45	6.28	5.14	12.69	10.20	8.99	11.31	8.16	11.01	11.98	10.24	9.03	8.27	5.50	8.36	7.36	6.90	5.17	7.87	6.86	5.84	10.44	8.13	13.04	8.07	6.79	7.44	7.09	5.67	5.44
CNN MLE (ours)	**0.76**	**0.61**	0.60	**2.53**	**1.00**	4.38	**4.40**	**2.70**	3.64	**11.21**	**9.11**	8.18	**5.17**	**4.37**	4.08	**4.97**	**2.59**	11.17	**4.53**	**3.05**	4.88	**3.37**	**2.68**	3.02	**4.60**	**2.89**	5.06	**5.86**	**4.26**	6.80

Table 1. Gravity prediction error (degree). Med.: median, Std.: standard deviation

are used for the training data while Bike 1, 2, and 3 were not included in the training dataset. Table 1 summarizes the gravity prediction comparison.

Scale recovery We recover the scale factor and compare the magnitude of linear acceleration with IMU, i.e., $\|\mathbf{a}\|/\|\mathbf{a}_{\mathrm{m}}\|$ where $\mathbf{a}$ and $\mathbf{a}_{\mathrm{imu}}$ are acceleration of ours and IMU, respectively. Note that IMU data is noisier than our estimation but the ratio remains approximately 1 (head: 1.0278 median, 1.1626 mean, 0.6186 std.; body: 0.9999 median, 1.1600 mean, 0.7739 std.). We recover scale factors for 11 different sequences each ranges between 1 mins to 15 mins as shown in Figure 5(d). This results in overall 1.0188 median, 1.1613 mean, and 0.7003 std.

Active force estimation We identify the moment that thrust force (pedaling and braking) is applied[10]. We use a thresholding binary classifier, $\xi^{+}(t)$ and $\xi^{-}(t)$ to detect pedaling and braking, respectively: $\xi^{+}(t) = 1$ if $\int_{t-1}^{t} F_{\mathrm{T}}(t)\mathrm{d}t > \epsilon_T$, and 0 otherwise; $\xi^{-}(t) = 1$ if $\int_{t-1}^{t} F_{\mathrm{T}}(t)\mathrm{d}t < -\epsilon_T$, and 0 otherwise[11]. Figure 6(a) shows active force profile and ground truth manually annotated from the videos of behavior monitoring cameras as shown in Figure 5(a). Our active force profile accords with the ground truth, i.e., pedaling when $F_{\mathrm{T}} > 0$ and braking when $F_{\mathrm{T}} < 0$. In Figure 6(b) and 6(c), we compare our method with net acceleration measured by IMU and structure from motion. We also compare against optical flow to measure acceleration that is often use for egocentric activity recognition tasks [14, 30]. Also we compare with Pooled Motion Feature representation [31], which requires a pre-trained model. Our active force identification outperforms other baseline methods that do not take into account active force decomposition. This verifies that a trivial extension by attaching IMU on camera is not sufficient enough to estimate the active force applied by the camera wearer—the measured acceleration needs to be decomposed.

Active torque estimation We compare the estimated angular velocity with measurements from gyroscope in Figure 6(d). Note that the velocity computation by differentiating the reconstructed camera trajectory does not directly apply as different framerate between IMU and camera and noisy reconstruction. The optimally estimated active force and torque generate plausible angular velocity profile. Table 2 summarizes error of angular velocity measured by 11 different scenes. The correlation is also measured, which produces 0.87 mean correlation.

	1	2	3	4	5	6	7	8	9	10	11
Mean(rad/sec)	0.25	0.31	0.27	0.31	0.27	0.26	0.41	0.29	0.30	0.30	0.40
Med. (rad/sec)	0.18	0.30	0.17	0.27	0.26	0.22	0.36	0.23	0.22	0.24	0.36
Std. (rad/sec)	0.24	0.20	0.26	0.23	0.19	0.19	0.32	0.23	0.27	0.26	0.31
Corr.	0.91	0.94	0.90	0.88	0.88	0.61	0.82	0.83	0.90	0.86	0.86

Table 2. Angular velocity comparison with gyroscope. Med.: median, Std.: standard deviation, Corr: correlation (perfect if 1)

5.2. Qualitative Evaluation

We apply our method on real world data downloaded from YouTube. 5 different types of scenes are processed: 1) mountain biking (1-10 m/s); 2) Flying: wingsuit jump (25-50 m/s) and speedflying with parachute (9-40 m/s) (); 3) jetskiing at Canyon (4-20 m/s); 4) glade skiing (5-12 m/s); 5) Taxco urban downhill biking (5-15 m/s). Figure 1 and 7, estimated gravity direction, physical scale of force and velocity, and active force and torque. Also passive components such as air drag, pitch torque, and normal force are shown. Thrust force is applied when climbing up the hill in Biking or when accelerating in Jetskiing. For Skiing, periodic lateral forces and roll moments are observed as the camera wearer was banking frequently. For flying case[12], strong air drag force and lifting forces are observed. Also unstable angular momentum along the roll axis comparing to other axes is observed, which requires skillful body control to balance left and right wings.

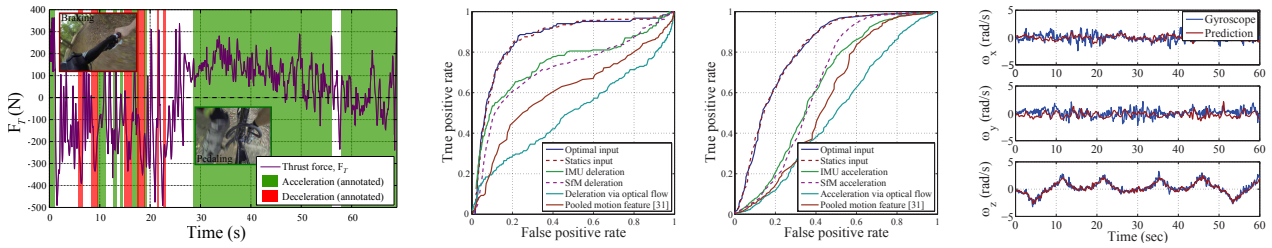
6. Discussion

In this paper, we present a method to reconstruct physical sensation of a first person video. We recover three ingredients for the physical sensations: gravity direction, physical scale, and active force and torque. The gravity direction is computed by leveraging a convolutional neural network integrated with the reconstructed 3D camera orientations. We recover the physical scale by using a torque equilibrium relationship along the roll axis at a bank turn. Active and passive components are modeled using rigid body dynamics which is integrated into a bundle adjustment that finds active force and torque profile concerting with the video. We quantitatively evaluate our method with controlled experiments where our method outperforms other baseline algorithms with a large margin ($\times 2 \sim \times 10$) and apply our method on real world data of various activities such as biking, skiing, flying, jetskiing, and urban bike racing.

[10] Active force and torque are difficult to directly measure using IMU because the measured acceleration is due to net force and torque not input. This requires special force/torque sensors attached human bodies that measures muscle tension.

[11] A sophisticated classifier such as recurrent neural networks can be a complementary approach when supervision is available.

[12] Unfortunately, the gravity direction cannot properly estimated as it was even challenging to a human annotator. Instead, we manually find frames that contain the horizon to estimate the gravity direction.



(a) Accel/decel detection (b) Acceleration (c) Deceleration (d) Angular velocity comparison

Figure 6. (a) We identify active forces by manually annotating frames when pedaling or braking. (b) and (c) Our method outperforms optical flow based representation including [31] with a large margin. (d) We compare our estimation with a gyroscope attached to the camera. Our estimation via active force and torque produces plausible angular velocity profile that accords with the gyroscope measurements.

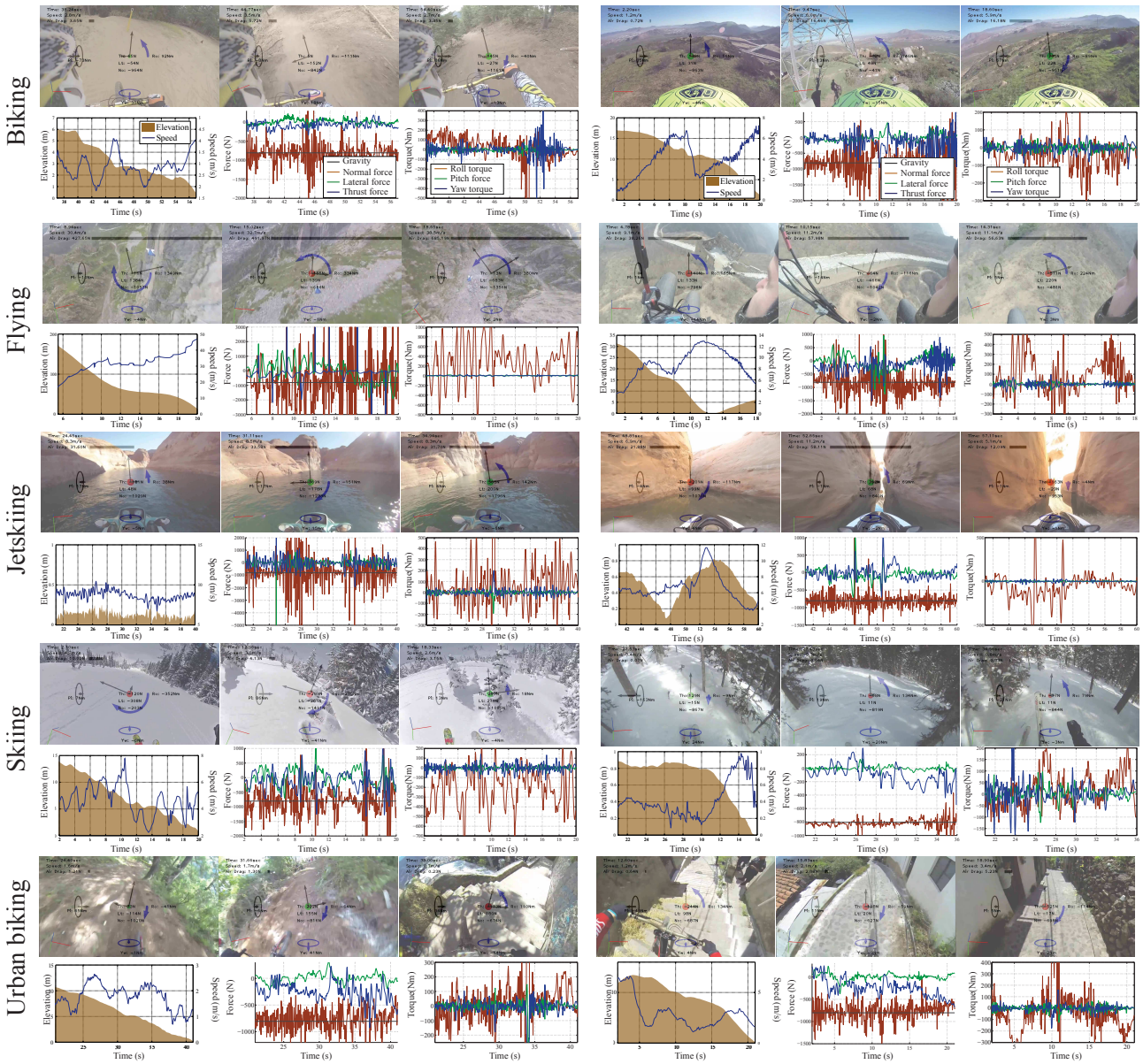


Figure 7. We compute gravity direction, physical scale factor, and active force and torque from a first person video. For each sequence, the top row shows image superimposed with speed, gravity, forces, and torque. Full trajectories of such physcial quantities are illustrated in the next row.

References

[1] I. Arev, H. S. Park, Y. Sheikh, J. K. Hodgins, and A. Shamir. Automatic editing of footage from multiple social cameras. *SIGGRAPH*, 2014. 3

[2] C. Bregler, A. Hertzmann, and H. Biermann. Recovering non-rigid 3d shape from image streams. In *CVPR*, 2000. 2

[3] M. A. Brubaker and D. J. Fleet. The kneed walker for human pose tracking. In *CVPR*, 2008. 2

[4] M. A. Brubaker, D. J. Fleet, and A. Hertzmann. Physics-based person tracking using simplified lower-body dynamics. In *CVPR*, 2007. 2

[5] J. C. Butcher. *Numerical Methods for Ordinary Differential Equations*. John Wiley and Sons, second edition, 2008. 6

[6] K. Choo and D. J. Fleety. People tracking using hybrid monte carlo filtering. In *ICCV*, 2001. 2

[7] A. Dal Monte, L. M. Leonardi, C. Menchinelli, and C. Marini. A new bicycle design based on biomechanics and advanced technology. *International Journal of Sport Biomechanics*, 1987. 2

[8] A. Fathi, A. Farhadi, and J. M. Rehg. Understanding egocentric activities. In *ICCV*, 2011. 2

[9] A. Fathi, Y. Li, and J. M. Rehg. Learning to recognize daily actions using gaze. In *ECCV*, 2012. 2, 3

[10] M. R. Greene and A. Oliva. Recognition of natural scenes from global properties: Seeing the forest without representing the trees. *Cognitive Psychology*, 2009. 3

[11] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014. 6

[12] G. Johansson. Visual perception of biological motion and a model for its analysis. *Perception and Psychophyics*, 1973. 2

[13] T. Kanade and M. Hebert. First person vision. In *IEEE*, 2012. 2

[14] K. M. Kitani, T. Okabe, Y. Sato, and A. Sugimoto. Fast unsupervised ego-action learning for first-person sports videos. In *CVPR*, 2011. 2, 7

[15] J. Kopf, M. Cohen, and R. Szeliski. First person hyperlapse videos. *SIGGRAPH*, 2014. 3

[16] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012. 3, 4, 6

[17] Y. J. Lee, J. Ghosh, and K. Grauman. Discovering important people and objects for egocentric video summarization. In *CVPR*, 2012. 3

[18] Y. Li, A. Fathi, and J. M. Rehg. Learning to predict gaze in egocentric video. In *ICCV*, 2013. 2

[19] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015. 3

[20] D. Metaxas and D. Terzopoulos. Shape and nonrigid motion estimation through physics-based synthesis. *TPAMI*, 1993. 2

[21] R. M. Murray, Z. Li, and S. S. Sastry. *A Mathematical Introduction to Robotic Manipulation*. CRC Press, 1994. 5

[22] J. Nocedal and S. J. Wright. *Numerical Optimization*. Springer, 2006. 6

[23] A. Oliva and A. Torralba. Modeling the shape of the scene: a holistic representation of the spatial envelope. *IJCV*, 2001. 3

[24] H. S. Park, E. Jain, and Y. Shiekh. 3D social saliency from head-mounted cameras. In *NIPS*, 2012. 2, 3

[25] H. Pirsiavash and D. Ramanan. Recognizing activities of daily living in first-person camera views. In *CVPR*, 2012. 3

[26] G. Pusiol, L. Soriano, L. Fei-Fei, and M. C. Frank. Discovering the signatures of joint attention in child-caregiver interaction. In *CogSci*, 2014. 2, 3

[27] J. M. Rehg, G. D. Abowd, A. Rozga, M. Romero, M. A. Clements, S. Sclaroff, I. Essa, O. Y. Ousley, Y. Li, C. Kim, H. Rao, J. C. Kim, L. L. Presti, J. Zhang, D. Lantsman, J. Bidwell, and Z. Ye. Decoding childrens social behavior. In *CVPR*, 2013. 2

[28] G. Robson and R. D'Andrea. Longitudinal stability analysis of a jet-powered wingsuit. In *AIAA Atmospheric Flight Mechanics Conference*, 2010. 2

[29] G. Rogez, J. Supancic, and D. Ramanan. First-person pose recognition using egocentric workspaces. In *CVPR*, 2015. 3

[30] M. S. Ryoo and L. Matthies. First-person activity recognition: What are they doing to me. In *CVPR*, 2013. 3, 7

[31] M. S. Ryoo, B. Rothrock, and L. Matthies. Pooled motion features for first-person videos. In *CVPR*, 2015. 3, 7, 8

[32] S. Shin, Y. Ahn, J. Choi, and S. Han. Design of a framework for interoperable motion effects for 4d theaters using human-centered motion data. In *International Conference on Advances in Computer Entertainment Technology*, 2010. 2

[33] H. Sidenbladh, M. J. Black, and D. J. Fleet. Stochastic tracking of 3d human figures using 2d image motion. In *ECCV*, 2000. 2

[34] L. Torresani, A. Hertzmann, and C. Bregler. Nonrigid structure-from-motion: Estimating shape and motion with hierarchical priors. *TPAMI*, 2008. 2

[35] R. Urtasun, D. Fleet, and P. Fua. 3d people tracking with gaussian process dynamical models. In *CVPR*, 2006. 2

[36] M. Vondrak, L. Sigal, J. K. Hodgins, and O. Jenkins. Video-based 3d motion capture through biped control. *SIGGRAPH*, 2012. 2

[37] M. Vondrak, L. Sigal, and O. Jenkins. Physical simulation for probabilistic motion tracking. In *CVPR*, 2008. 2

[38] X. Wei and J. Chai. Videomocap: Modeling physically realistic human motion from monocular video sequences. *SIGGRAPH*, 2010. 2

[39] D. M. Wolpert, J. Diedrichsen, and J. R. Flanagan. Principles of sensorimotor learning. *Nature Reviews Neuroscience*, 2011. 2

[40] C. R. Wren and A. Pentland. Dynamic models of human motion. In *IEEE Face and Gesture*, 1998. 2

[41] B. Xiong and K. Grauman. Detecting snap points in egocentric video with a web photo prior. In *ECCV*, 2014. 3

[42] J. Yan and M. Pollefey. A factorization-based approach for articulated nonrigid shape, motion and kinematic chain recovery from video. *TPAMI*, 2008. 2

[43] Z. Ye, Y. Li, Y. Liu, C. Bridges, A. Rozga, and J. M. Rehg. Detecting bids for eye contact using a wearable camera. In *FG*, 2015. 2

[44] R. Yonetani, K. M. Kitani, and Y. Sato. Ego-surfing first person videos. In *CVPR*, 2015. 3