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

Presenter: Jimmy Xin Lin

Prof. Kristen Grauman

Department of Computer Science

University of Texas at Austin

Outline

- Introduction
  - Target Problem, Essential Concepts, Motivations
  - A Visual Demo
  - Challenges, Related Work
- Framework: Force From Motion
  - Gravity Direction
  - Physical Scale: Speed and Terrain
  - Active Force and Torque
- Experimentation
  - Quantitative Evaluation
  - Qualitative Evaluation
- Conclusion and Discussion

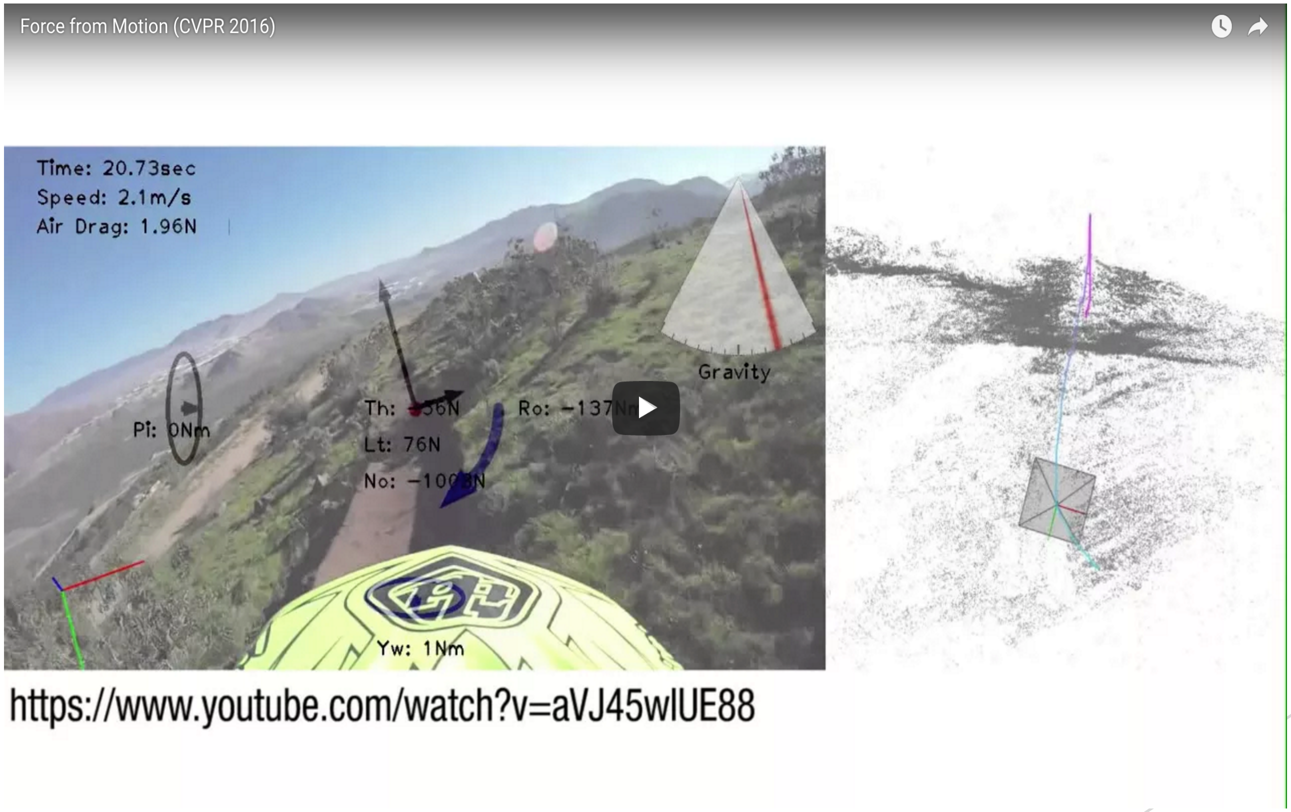
Introduction

Target Problem, Essential Concepts, Motivations

Introduction I: First Conceptual Touch

- Target problem: Model camera carrier’s physical sensation over the videos at his/her first-person perspective.
- This paper **initiate a computational framework** to evaluate the ego-motion from an egocentric video with the domain knowledge of **physical body dynamics**.
- What is the physical sensation?
  - Conceptually, analytic components of one’s physical motion.
  - Mechanically, three ingredients: gravity, physical scale, and active force and Torque.

Introduction II: A Visual Demo



Force from Motion (CVPR 2016)

https://www.youtube.com/watch?v=aVJ45wlUE88

Introduction III: More on Techniques

- Technical Challenges
  - Limited observations of one's body parts (body pose is not visible from the camera).
  - Scale and orientation are ambiguous from the motion.
  - Scene and activity vary case by case (environmental appearance, camera placement, and motion pattern).
- Applications:
  - computational sport analytics (mountain biking, urban bike racing, skiing and etc.).
  - activity recognition, video indexing, and content generation for virtual reality.

Force From Motion

Gravity, scale, and active force and torque are three key ingredients that evaluates the physical sensation of one's motion.

Force from Motion I: Gravity Direction

- Intuition: image cues (i.e. trees and buildings) imply the gravity direction.
- Approach: construct a convolutional neural network [16] to predict a gravity direction in a 2D image. This per image prediction is **integrated over multiple frames** $\{\mathcal{I}_i\}_{i=1}^F$ by leveraging structure from motion.
- Define a 3D unit gravity direction

$$\hat{\mathbf{g}}(\theta,\phi) \;=\; \left[\begin{array}{ccc}\sin\theta\cos\phi & \sin\theta\sin\phi & \cos\theta\end{array}\right]^{\mathsf{T}} \in \mathbb{S}^2$$

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

$$\hat{\mathbf{g}}^* = \underset{\hat{\mathbf{g}}\in\mathbb{S}^2}{\operatorname{argmax}}\; p(\hat{\mathbf{g}}|\mathcal{I}_1,\cdots,\mathcal{I}_F)$$

$$= \underset{\hat{\mathbf{g}}\in\mathbb{S}^2}{\operatorname{argmax}}\; p(\hat{\mathbf{g}})\prod_{i=1}^{F} p(\mathcal{I}_i|\hat{\mathbf{g}}),$$

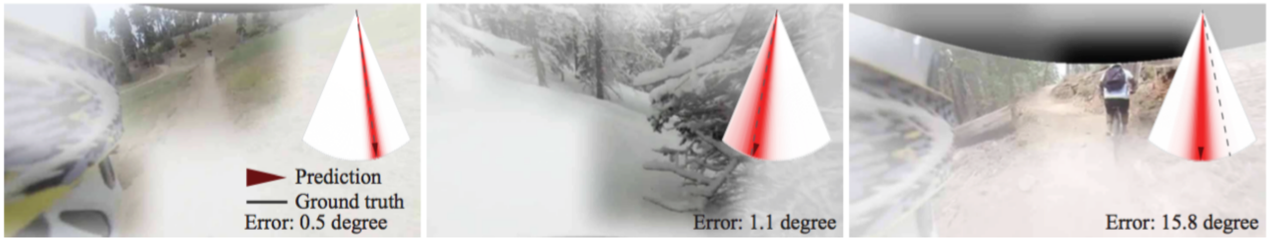
- Prior distribution encodes how the gravity is oriented with respect to the heading direction. Prior distribution using a mixture of von Mises-Fisher distributions

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

Gravity Direction (cont.)

- Image likelihood $p(\mathcal{I}_i|\hat{\mathbf{g}})$ measures how well the aligned 3D gravity direction is consistent with cues on the i-th image.

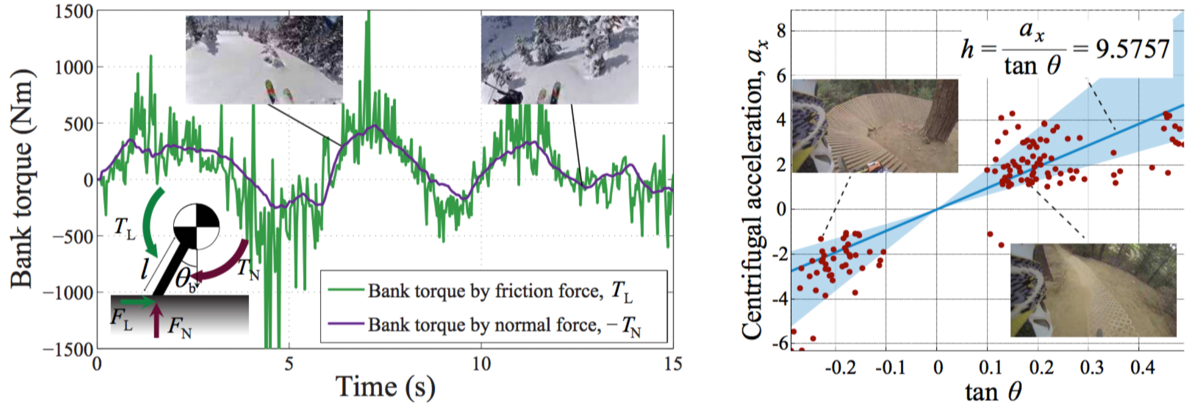
- Learn the image likelihood function using the convolutional neural network (CNN) proposed by Krizhevsky et al. [16] with a few minor modifications.
  - Resizing: warped images (1280 × 720) are resized to 320 × 180 as inputs for the CNN
  - Target Shrinking: train the network to predict a probability of the projected angle discretized by 1 degree between −30 and 30 degrees.



- Predictions on multiple frames are consolidated to predict the 3D gravity direction by the reconstructed 3D camera orientations.

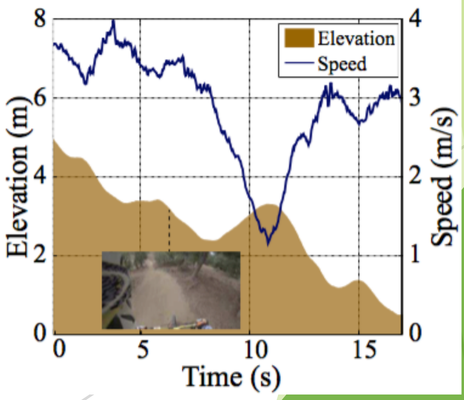
Force from Motion II: Physical Scale

- Two yielded torques must be balanced to maintain the leaning angle θ_{b}:
  - The normal force, F_{N}, produces a torque, $T_{\mathrm{N}} = lF_{\mathrm{N}}\cos\theta_{\mathrm{b}}$.
  - the friction force F_{L} produces an opposite directional torque $T_{\mathrm{L}} = lF_{\mathrm{L}}$
- By equating $T_{\mathrm{L}} + T_{\mathrm{N}} = 0$, we got $\|\mathbf{g}\| = 9.81\ \mathrm{m/s}^2 = c\dfrac{|\hat{\mathbf{a}}_{\mathrm{x}}|}{\tan\theta_{\mathrm{b}}}$,
- $\hat{\mathbf{a}}_{\mathrm{x}}$ is the linear acceleration in the lateral direction, which is measured from the reconstructed 3D camera trajectory, c is a scale factor that maps from the 3D reconstruction to the physical world.

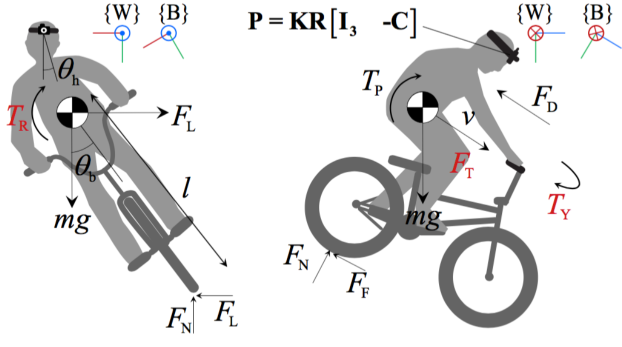


(a) Bank torque

(b) Scale factor



Force from Motion III: Active Forces and Torque



(c) Geometry

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

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

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

- Represent the first formula in world coordinate system {W} and the second in the body coordinate system {B}.
- The active force and torque are composed of thrust force F_{T}, roll torque T_{R}, and yaw (steering) torque T_{Y}

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

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

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

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

Active Forces and Torque (cont.)

- Compact form of motion description:

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

- Where
  - $\mathcal{M}$ is the inertial matrix, $\mathcal{C}$ is the Coriolis matrix.
  - $\mathbf{E}$ is the passive force and torque, and $\mathbf{u} = \begin{bmatrix} F_{\mathrm{T}} & T_{\mathrm{R}} & T_{\mathrm{Y}} \end{bmatrix}$ is the active component.
  - The state $\mathbf{q} = \begin{bmatrix} \mathbf{C}^{\mathsf{T}} & \mathbf{\Omega}^{\mathsf{T}} \end{bmatrix}^{\mathsf{T}}$ describes the camera ego-motion where $\mathbf{C} \in \mathbb{R}^3$ is the camera center and $\mathbf{\Omega} \in \mathbb{R}^3$ is the axis-angle representation of camera rotation.
- J is a workspace mapping matrix written as:

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

- This describes motion in terms of active force and torque component, $\mathbf{u}$, which allows us to directly map between input and the resulting motion.

Optimal Control: Inverse Dynamics

- Integration of three ingredients for physical sensation (gravity direction, physical scale, and active force and torque) into the following optimization problem:

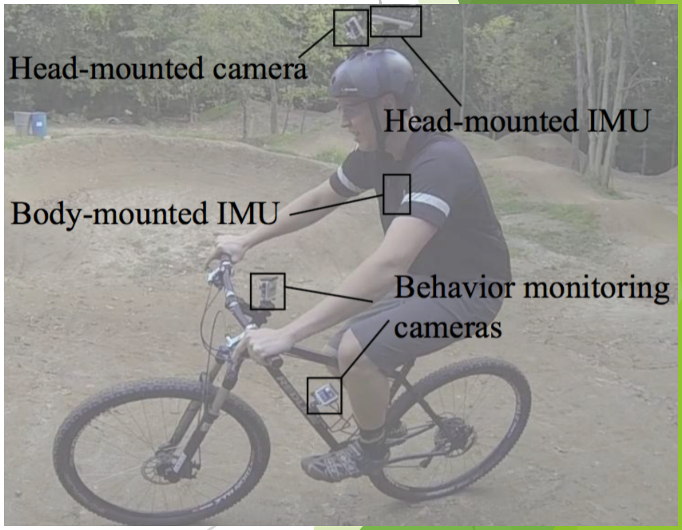
$$\begin{aligned}\underset{\mathbf{u}(t),\{\mathbf{X}_j\}}{\text{minimize}} \quad & \sum_{i,j} \mathcal{D}\left(\mathbf{P}(t_i)\mathbf{X}_j, \mathbf{x}_{ij}\right) + \lambda_{\mathcal{R}} \int_0^T \dot{\mathbf{u}}(t)^{\mathsf{T}} \dot{\mathbf{u}}(t) \mathrm{d}t \\ \text{subject to} \quad & \mathbf{P}(t_i)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

- Notations:
  - $\mathcal{D}$ measures reprojection error
  - the camera projection matrix at t_i time 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 the unknown 3D world structure X, and active component $\mathbf{u}(t)$ for the rigid body dynamics.
- Last term in the cost function regularizes active forces such that the resulting input profile over time is continuous.
- The above objective function can be solved using Levenberg-Marquardt algorithm [22].

Experimentation

Quantitative Evaluation

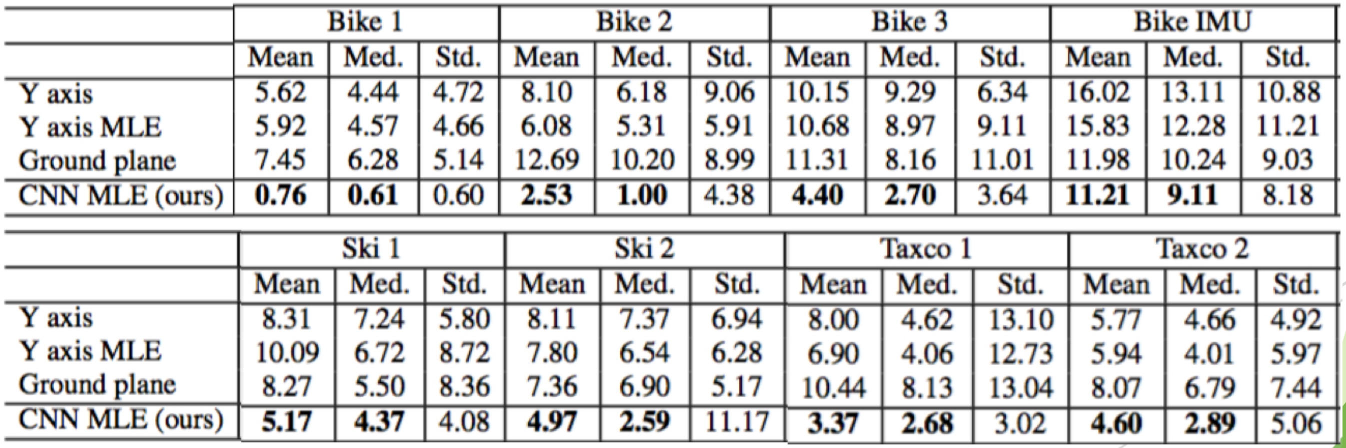
- Experimental setup:
  - Two Inertial Measurement Unit (IMU): one on head and the other on body.
  - Two Cameras: One on head and the other some place to monitor behaviors.



- Training Set: 29 Biking Sequences (10 secs with 300 frames).
- Operate quantitative evaluations in three criteria:
  - Gravity Prediction
  - Scale Recovery
  - Active Force and Torque Estimation

Gravity Prediction

- Compare our predictions 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.
- Test our method on manually annotated data

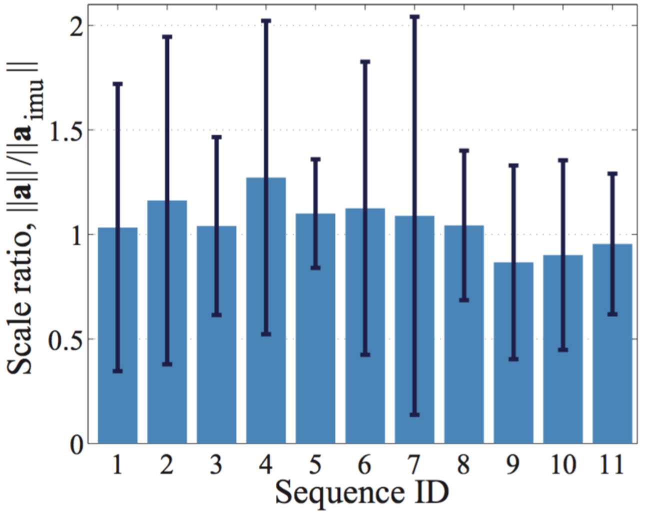


	Bike 1			Bike 2			Bike 3			Bike IMU		
	Mean	Med.	Std.	Mean	Med.	Std.	Mean	Med.	Std.	Mean	Med.	Std.
Y axis	5.62	4.44	4.72	8.10	6.18	9.06	10.15	9.29	6.34	16.02	13.11	10.88
Y axis MLE	5.92	4.57	4.66	6.08	5.31	5.91	10.68	8.97	9.11	15.83	12.28	11.21
Ground plane	7.45	6.28	5.14	12.69	10.20	8.99	11.31	8.16	11.01	11.98	10.24	9.03
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

	Ski 1			Ski 2			Taxco 1			Taxco 2		
	Mean	Med.	Std.	Mean	Med.	Std.	Mean	Med.	Std.	Mean	Med.	Std.
Y axis	8.31	7.24	5.80	8.11	7.37	6.94	8.00	4.62	13.10	5.77	4.66	4.92
Y axis MLE	10.09	6.72	8.72	7.80	6.54	6.28	6.90	4.06	12.73	5.94	4.01	5.97
Ground plane	8.27	5.50	8.36	7.36	6.90	5.17	10.44	8.13	13.04	8.07	6.79	7.44
CNN MLE (ours)	**5.17**	**4.37**	4.08	**4.97**	**2.59**	11.17	**3.37**	**2.68**	3.02	**4.60**	**2.89**	5.06

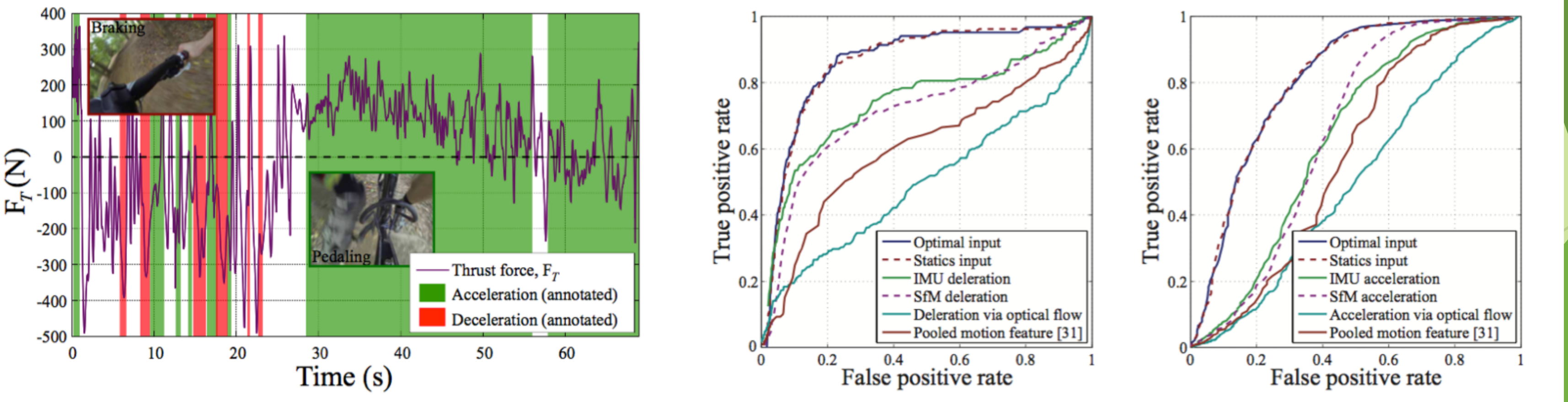
Scale Recovery

- Recover the scale factor and compare the magnitude of linear acceleration with IMU, $\|\mathbf{a}\|/\|\mathbf{a}_{\mathrm{m}}\|$.
- $\mathbf{a}$ is linear acceleration estimated by our method.
- $\mathbf{a}_{\mathrm{m}}$ is linear acceleration of IMU.
- The scale ratio $\|\mathbf{a}\|/\|\mathbf{a}_{\mathrm{m}}\|$ remains around 1.0 in training sequences:
  - head: 1.0278 median, 1.1626 mean, 0.6186 std.
  - body: 0.9999 median, 1.1600 mean, 0.7739 std
- Recover scale factors for 11 different sequences each ranges between 1 mins to 15 mins.
- The result is exciting:
  - overall 1.0188 median, 1.1613 mean, and 0.7003 std.



Active Force and Torque Estimation

- Active Force identification compete against
  - Net acceleration measured by IMU
  - Optical flow to measure acceleration (like in egocentric activity recognition tasks)
  - Pooled Motion Feature representation (requires a pre-trained model)



(a) Accel/decel detection

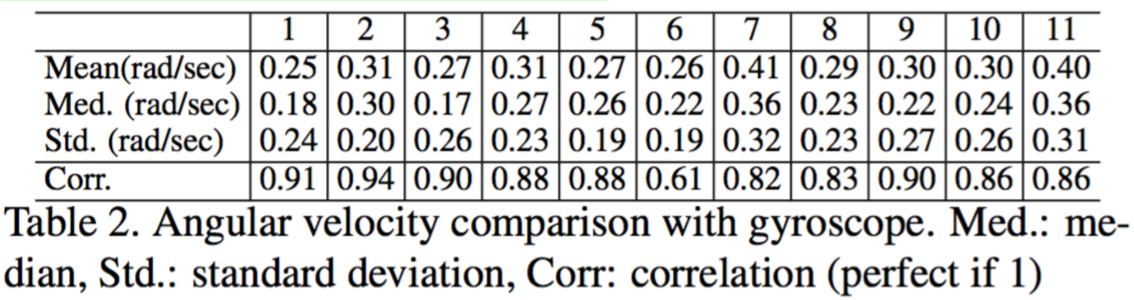
(b) Acceleration

(c) Deceleration

- Our active force identification outperforms other baseline methods that do not take into account active force decomposition.

Active Force and Torque Estimation

- Estimate angular velocity in 11 different scenes.
- Compare the estimated angular velocity with measurements of gyroscope.
- 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)

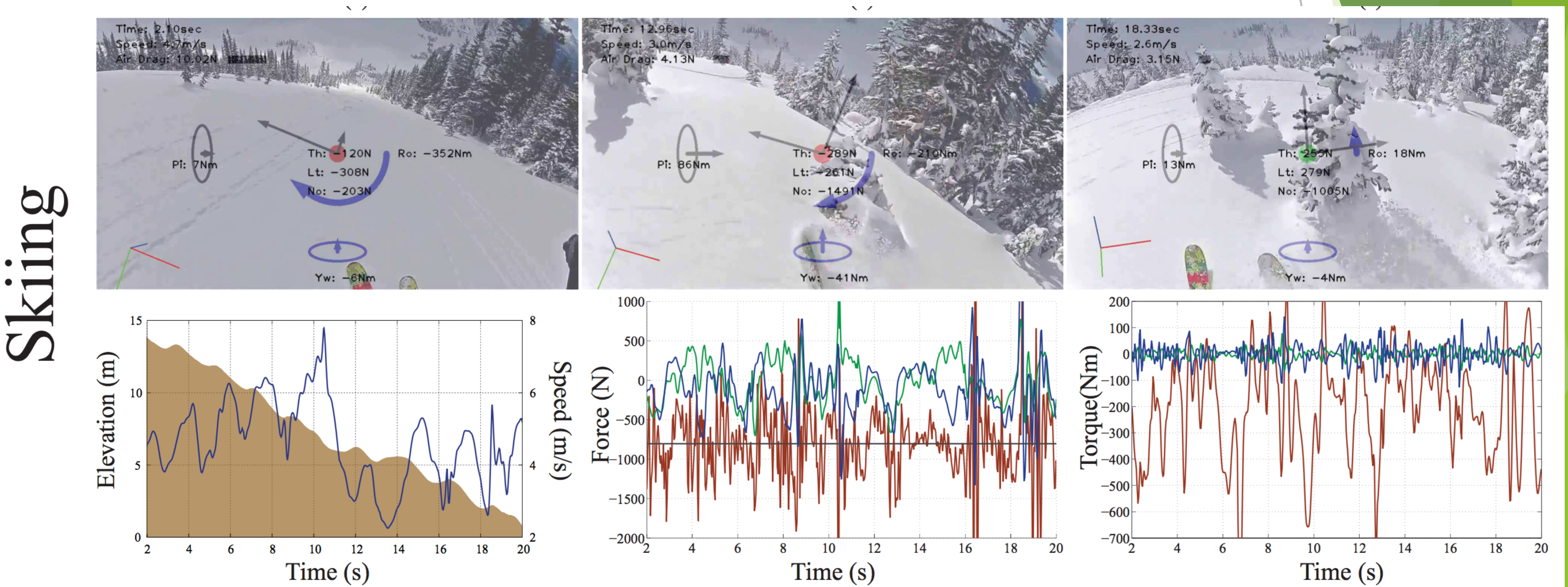
- Correlations in 11 different scenes are mostly close to 1.

Qualitative Evaluation

- Apply the framework on real world data downloaded from YouTube (5 categories)
  - 1) mountain biking (1-10 m/s)
  - 2) Flying: wingsuit jump (25-50 m/s) and speed flying 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)
- These Sports vary in
  - Appearance of the Environment
  - Speed Range of the Motion
  - Composition of Passive/Active Forces
- Sufficiently convincing to demonstrate the robustness of the proposed computational framework.

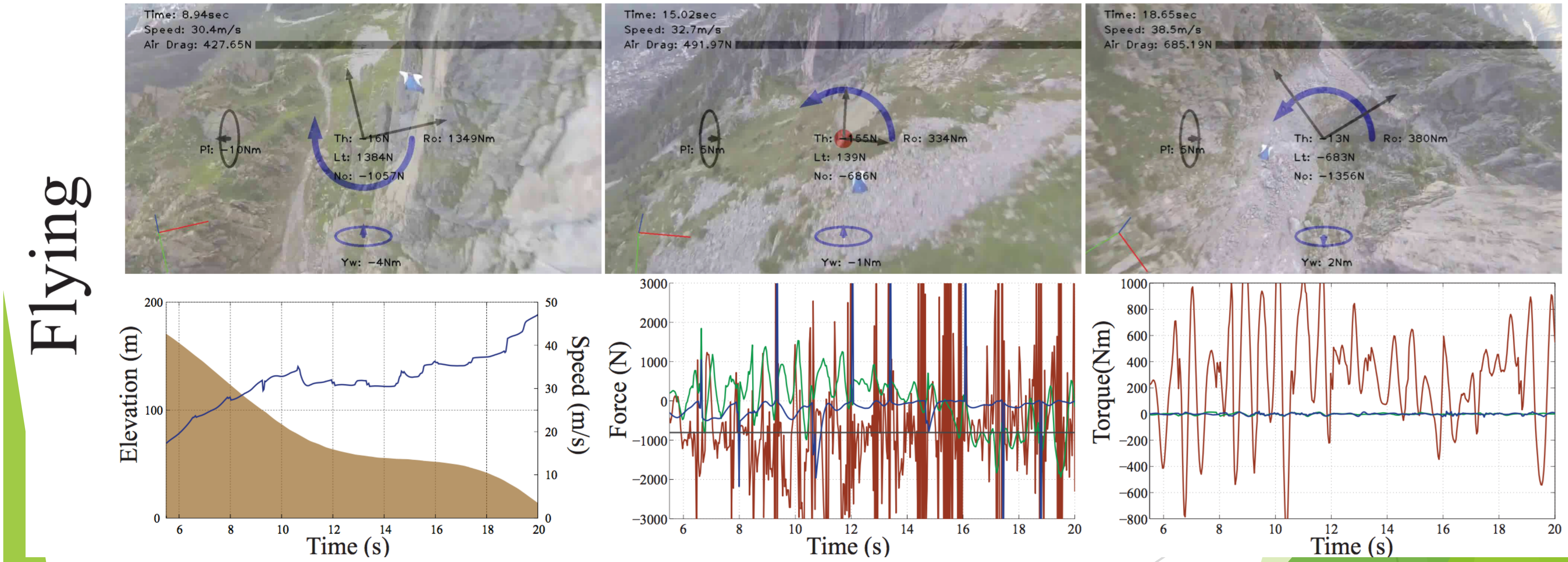
Qualitative Evaluation

- glade skiing (5-12 m/s);



Qualitative Evaluation

- Flying: wingsuit jump (25-50 m/s)

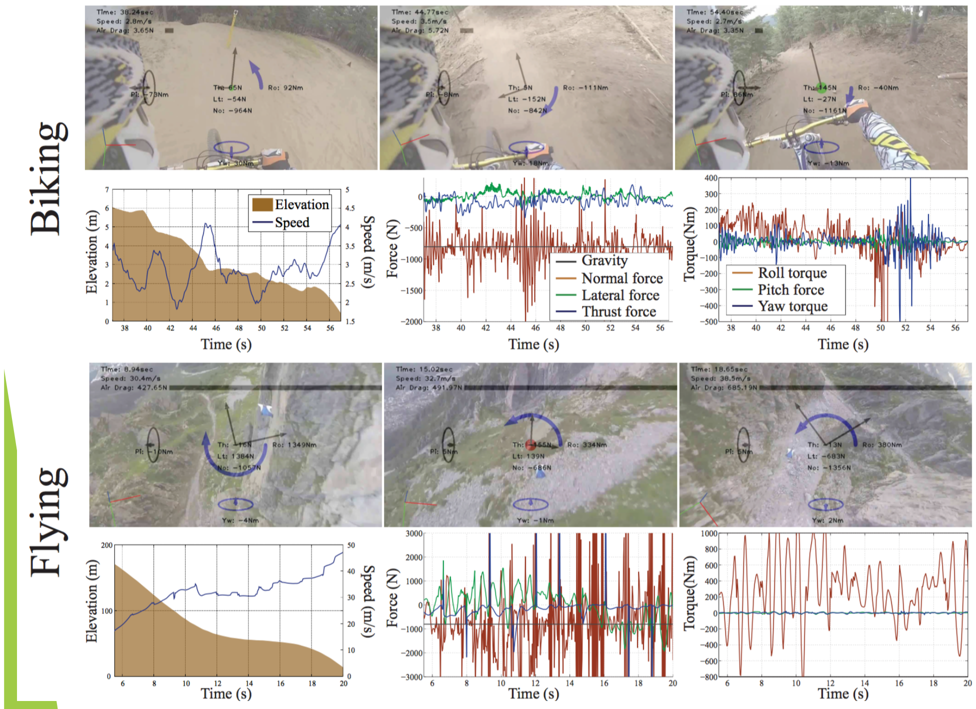


Conclusion & Discussion

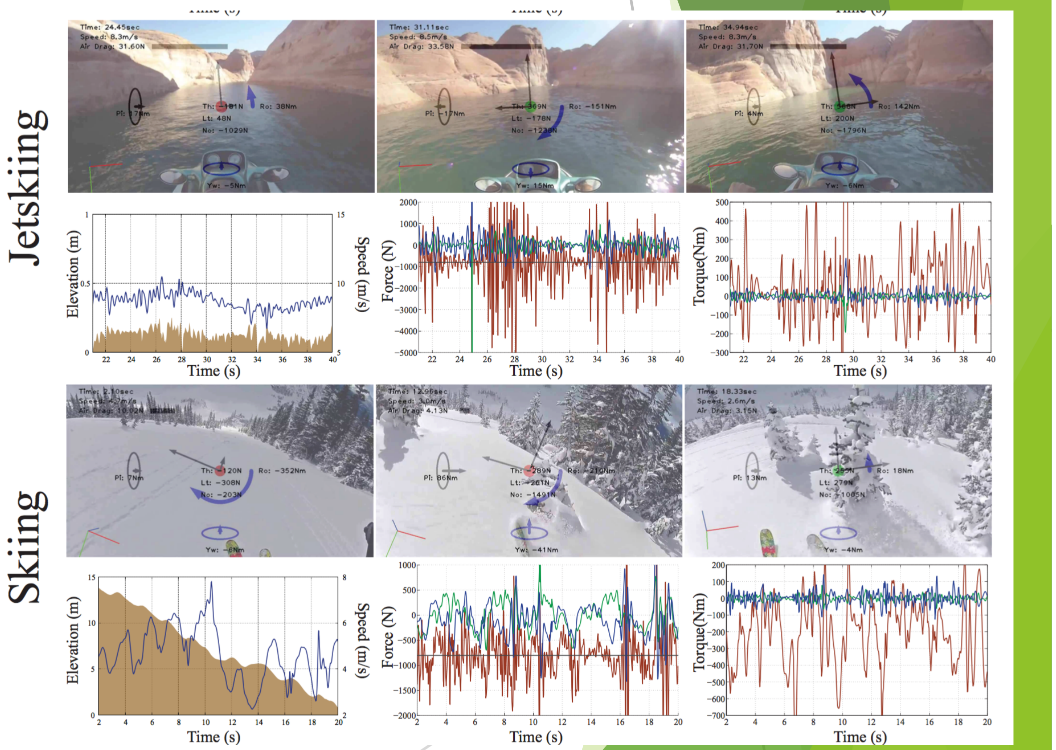
Conclusion

- This paper propose a new computational framework that evaluates camera wearer's physical sensation.
  - Gravity DirectionPrediction: through CNN + MLE (3D reconstruction of camera orientation)
  - Physical Scale (speed and terrain): through the 3D trajectory reconstruction
  - Active Force and Torque: through an optimization problem based on dynamics
- Quantitative experiments are operated on each individual estimation component and demonstrate the efficacy of these components.
- Qualitative experiments show that Force From Motion is decently applicable to a number of other sports (not shown in training set).

Questions?



Biking
Flying
Elevation
Speed
Elevation (m)
Speed (m/s)
Force (N)
Torque(Nm)
Time (s)
Gravity
Normal force
Lateral force
Thrust force
Roll torque
Pitch force
Yaw torque



Jetskiing
Skiing
Elevation (m)
Speed (m/s)
Force (N)
Torque(Nm)
Time (s)

Thanks!

References

[1] Force from Motion: Decoding Physical Sensation from a First Person Video. H.S. Park, J-J. Hwang and J. Shi. CVPR 2016.

- Visual Demo: http://www-users.cs.umn.edu/~hspark/ffm.html
- Gravity Prediction on CNN Models: https://github.com/jyhjinghwang/Force_from_Motion_Gravity_Models

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

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

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

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

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

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

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

[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.

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

References

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

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

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

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

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

[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.

[33] H.Sidenbladh,M.J.Black,andD.J.Fleet.Stochastictrack- ing 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.

[38] X. Wei and J. Chai. Videomocap: Modeling physically real- istic human motion from monocular video sequences. *SIG- GRAPH*, 2010.

[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.

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