Social and Physical Interactions from First Person Cameras

Hyun Soo Park
Third person camera
Internal state: intention, attention, and memory

Third person camera
Internal state:
intention, attention, and memory

Can we recover the internal states from a video?
First person video
First person video
First person video
First person video tells us about me.
What can first person video tell us?
What can first person video tell us about my physical interactions with scene?
What can first person video tell us about my physical interactions with scene?
What can first person video tell us about my physical interactions with scene?
Can we understand activity?
Can we understand detailed activity?
Can we understand **physics** of activity?

1. **Gravity**

   ![Gyroscope](image)
Can we understand **physics** of activity?
Can we understand physics of activity?
Can we understand physics of activity?
Time: 0.17 sec
Speed: 2.0 m/s
Air Drag: 1.77 N

\[ \text{Th: } -177 N \]
\[ \text{Ro: } 91 Nm \]
\[ \text{Lt: } 4 N \]
\[ \text{No: } -842 N \]
\[ \text{Yw: } 13 Nm \]
Time: 0.17 sec
Speed: 2.0 m/s
Air Drag: 1.77 N

Th: -177 N
Lt: 4 N
No: -842 N

Yw: 13 N

mg
Time: 0.17 sec
Speed: 2.0 m/s
Air Drag: 1.77 N

- Pi: -93 Nm
- Th: -177 N
- Lt: 4 N
- No: -842 N
- Ro: 91 Nm
- Yw: 13 Nm

Roll torque

mg
Time: 0.17sec
Speed: 2.0m/s
Air Drag: 1.77N

Thrust force: -177N
Rear: 91Nm
Left: 4N
No: -842N
Right: 13Nm
Time: 0.17 sec
Speed: 2.0 m/s
Air Drag: 1.77 N

Thrust force: -177 N
Ro: 91 Nm
Lt: 4 N
No: -842 N
Yw: 13 Nm
3D reconstruction
3D reconstruction

3D scene points

3D camera trajectory
Is geometry or kinematics enough to understand the biker’s behaviors?

*What causes motion?*
1. Compute Gravity
Prediction via CNN

Ground truth

Prediction error: 0.5 degree
Probability of gravity direction in 2D via CNN

\[ p(I \mid g) \]

Probability of gravity direction in 3D
Visual Gravity Prediction

\[ p(\hat{g} | I_1, \ldots, I_N) \propto p(\hat{g}) \prod_i^N p(I_i | \hat{g}) \]
2. Compute Speed
Camera trajectory $\mathcal{C}$

Velocity: $v = \frac{d\mathcal{C}}{dt}$
Camera trajectory $\alpha \mathbf{c}$

where $\alpha$ is arbitrary scale.

Velocity: $\mathbf{v} = \alpha \frac{d\mathbf{c}}{dt}$
Dynamic balance:

\[ mg \]

Scale factor:

\[ \alpha = \frac{g}{a_{\text{ctrpeta|l}}} \tan \theta_b \]

Measured by 3D reconstruction

Banked turn
3. Compute Control

Pedaling

Braking
Position: $c(t)$
Angle: $q(t)$

Linear force:
\[ m \frac{d^2 c}{dt^2} = F \]
Position: $c(t)$
Angle: $q(t)$

Air drag

Linear force:
$$m \frac{d^2 c}{dt^2} = F = \sum F_{\text{passive}} + \sum F_{\text{active}}$$

Forces:
- $F$ (total)
- $mg$ (gravitational force)
- $F_{\text{circular}}$ (circular force)
- $F_{\text{drag}}$ (air drag)
- $F_n$ (normal force)
- $F_F$ (friction force)
Position: $c(t)$
Angle: $q(t)$

Linear force:

$$m \frac{d^2 c}{dt^2} = F = \sum F_{\text{passive}} + \sum F_{\text{active}}$$
Position: $c(t)$
Angle: $q(t)$

Linear force:
$$m \frac{d^2 c}{dt^2} = F = \sum F_{\text{passive}} + \sum F_{\text{active}}$$

Angular force (torque):
$$J \frac{d^2 q}{dt^2} + \frac{dq}{dt} \times J \frac{dq}{dt} = \sum T_{\text{passive}} + \sum T_{\text{active}}$$

$J$: Moment of inertia
Position: $c(t)$
Angle: $q(t)$

$$(c, q) = \text{ODE}(F_{\text{passive}}, F_{\text{active}}, T_{\text{passive}}, T_{\text{active}})$$
Position: $c(t)$
Angle: $q(t)$

$$(c, q) = \text{ODE}(F_{\text{passive}}, F_{\text{active}}, T_{\text{passive}}, T_{\text{active}})$$

$$\bigodot \approx (t, R)$$

where $P = K[R \ t]$
where $P = K[R \ t]$ and $(c, q) = \text{ODE}(F_{\text{passive}}, F_{\text{active}}, T_{\text{passive}}, T_{\text{active}})$ approximately equals $(t, R)$. The Correlation between body and head increases with speed.
Inverse control:

\[
\text{minimize } E_{\text{SfM}}^{F,T,X} \quad \text{Reprojection error}
\]

subject to \( (t,R) = \text{ODE}(F,T) \)

\[
(c,q) = \text{ODE}(F_{\text{passive}}, F_{\text{active}}, T_{\text{passive}}, T_{\text{active}})
\]

where \( P = K[R \ t] \)
Inverse control:

\[
\begin{align*}
\text{minimize} & \quad E_{\text{SfM}} + \lambda E_{\text{reg}}(F, T) \\
\text{subject to} & \quad (t, R) = \text{ODE}(F, T)
\end{align*}
\]

Temporal regularization

\[
(c, q) = \text{ODE}(F_{\text{passive}}, F_{\text{active}}, T_{\text{passive}}, T_{\text{active}})
\]

where \( P = K[R \ t] \)
Head-mounted camera
Body-mounted camera
Body-mounted IMU
Head-mounted IMU
Brake monitoring camera
Pedal monitoring camera

Evaluation
Mean: 12.42 degree
Mean: 1.76 degree
$m \frac{d^2 c}{dt^2} = F = \sum F_{\text{passive}} + \sum F_{\text{active}}$
Pedaling activity (acceleration)
Braking activity (deceleration)
Pedaling activity (acceleration)

Braking activity (deceleration)

Pedaling

Braking
Time: 19.97 sec
Speed: 6.8 m/s
Air Drag: 21.13 N

P1: -3 Nm

Th: 599 N
Rc: -60 Nm
Lt: 55 N
No: -223 N

Yw: 6 Nm

3D reconstruction

https://www.youtube.com/watch?v=aVJ45wIUE88
Time: 5.17 sec
Speed: 12.6 m/s
Air Drag: 73.60 N

PI: 2 Nm
Th: -699 N
Lt: -56 N
Ro: -3 Nm
No: -17 N

Yw: 12 Nm

Gravity

3D reconstruction

https://www.youtube.com/watch?v=rnvvsjstveM
What can first person vision tell us about my future motion?
Where will he go?
If I were him, how would I move into the scene?
What is he experiencing visually?
First person view
First person view
Occlusion

First person view
Future localization
Future localization
Why challenging?
Why challenging?

Training (past images)
Why challenging?

Training (past images)

Prediction (current image)
Why challenging?
Why challenging?

1. Geometric inconsistency
Why challenging?

1. Geometric inconsistency
Why challenging?

1. Geometric inconsistency

2. Semantic inconsistency

Looking down

Looking forward
Why challenging?

1. Geometric inconsistency
   → EgoRetinal representation
2. Semantic inconsistency
   → Preference learning
Prediction:
Configuration space (ground plane)
Ground plane

Image space \( \mathcal{I} \)

Predicted trajectory

\( (X_1, \cdots, X_F) \)

Prediction: Configuration space (ground plane)
Ground plane

Prediction:
Configuration space (ground plane)

Image space $\mathcal{I}$

Predicted trajectory

$(x_1, \cdots, x_F) = g(\mathcal{I})$
\[(x_1, \ldots, x_F) = g(f(I))\]

Projection to cfg. space

Prediction:
Configuration space (ground plane)
\((x_1, \cdots, x_F) = g(f(I))\)

**Prediction:**
Configuration space (ground plane)
Head orientation invariant

Image space \(I\)
Ground plane
\((x_1, \cdots, x_F) = g(f(I))\)

\(f(I)_{r, \theta}\)

\((r, \theta)\)
The image space \( \mathcal{I} \) is defined as:

\[
(X_1, \ldots, X_F) = g(f(\mathcal{I}))
\]

The projection function is given by:

\[
f(\mathcal{I})_{r,\theta} = \mathcal{I}_{\text{proj}(r,\theta)}
\]

This represents the projection of the image space onto the ground plane.

Ground plane

Image space \( \mathcal{I} \)
\[(x_1, \ldots, x_F) = g(f(I))\]

\[f(I)_{r, \theta} = (I_{\text{proj}(r, \theta)}, u^T n)\]

Height of occluding object

Image space \( I \)

Ground plane
\[ \Delta r \propto \log \frac{1}{D} \]

where \( D \) is depth.
\[ \Delta r \propto \log \frac{1}{D} \] where \( D \) is depth.

Cf) Proxemics

Retinal representation
Persistent to 2D and 3D distance

Ground plane
\[ \Delta r \propto \log \frac{1}{D} \] where $D$ is depth.

Retinal representation
Persistent to 2D and 3D distance

Ground plane
Retinal representation
Persistent to 2D and 3D distance

\[ \Delta r \propto \log \frac{1}{D} \]

where $D$ is depth.
EgoRetinal image projection

EgoRetinal RGB
EgoRetinal image projection

EgoRetinal RGB

EgoRetinal depth

P1: Head orientation invariant

P2: 2D and 3D persistent

P3: Occlusion reasoning
Cartesian in image (LeCun et al.)

Configuration space
Cartesian in image (LeCun et al.)

Configuration space
Cartesian in image (LeCun et al.)

Cartesian in ground plane

Configuration space
Cartesian in image (LeCun et al.)

Cartesian in ground plane

Configuration space
Cartesian in image (LeCun et al.)

Cartesian in ground plane

Vanishing point

Configuration space
Cartesian in image (LeCun et al.)
Cartesian in ground plane
Vanishing point
EgoRetinal space
Configuration space
EgoMotion Dataset (outdoor)
<table>
<thead>
<tr>
<th>Scene</th>
<th>Frames</th>
<th>Duration</th>
</tr>
</thead>
<tbody>
<tr>
<td>IKEA</td>
<td>966</td>
<td>08:03</td>
</tr>
<tr>
<td>Costco</td>
<td>577</td>
<td>04:49</td>
</tr>
<tr>
<td>Mall</td>
<td>2683</td>
<td>22:22</td>
</tr>
<tr>
<td>Park</td>
<td>3088</td>
<td>25:44</td>
</tr>
<tr>
<td>School1/2</td>
<td>3754/3736</td>
<td>31:17/31:08</td>
</tr>
<tr>
<td>Downtown1/2</td>
<td>2856/3405</td>
<td>23:48/28:23</td>
</tr>
<tr>
<td>Grocery1/2/3</td>
<td>2858/2892/2834</td>
<td>23:49/24:06/23:37</td>
</tr>
<tr>
<td>Bus1/2</td>
<td>2292/1850</td>
<td>19:06/15:25</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Scene</th>
<th>Frames</th>
<th>Duration</th>
</tr>
</thead>
<tbody>
<tr>
<td>Campus1/2/3</td>
<td>2607/1884/1975</td>
<td>21:44/15:42/16:28</td>
</tr>
<tr>
<td>CVS1/2</td>
<td>2359/3337</td>
<td>19:40/27:49</td>
</tr>
<tr>
<td>Train Sta.1/2</td>
<td>4034/2568</td>
<td>33:37/21:24</td>
</tr>
<tr>
<td>River1/2</td>
<td>3378/2250</td>
<td>28:09/18:45</td>
</tr>
<tr>
<td>Dep. store</td>
<td>2250</td>
<td>13:20</td>
</tr>
<tr>
<td>Library</td>
<td>1255</td>
<td>10:30</td>
</tr>
<tr>
<td>Apartment</td>
<td>2050</td>
<td>17:05</td>
</tr>
<tr>
<td>Caffe</td>
<td>1550</td>
<td>13:00</td>
</tr>
</tbody>
</table>

**Dataset summary**

- 1280x960 stereo (100mm baseline, ~15m depth resolution)
- 26 scenes (13 indoor, 13 outdoor)
- 65.5k frames (9.1 hours)
$X_t = (r_t, \theta_t, \omega_t)$
Ground plane

\[ X_t = (r_t, \theta_t, \omega_t) \]

Trajectory topology
Ground plane

\[ \mathbf{x}_t = (r_t, \theta_t, \omega_t) \]

Trajectory topology

Traj A

Traj B
Ground plane

Trajectory topology

$X_t = (r_t, \theta_t, \omega_t)$
Walkable pixel \( X_t = (r_t, \theta_t, \omega_t) \)
Walkable height

\[ x_t = (r_t, \theta_t, \omega_t) \]
Testing image

Trajectory retrieval

\[
\min_{\mathcal{X}} E_D + E_{\text{RGB}} + \lambda \left\| \mathcal{X} - \mathcal{X}^* \right\|^2
\]

\( \mathcal{X}^* \): retrieved trajectory
Test image Trajectory retrieval Depth cost

\[
\min_{X} E_{D} + E_{RGB} + \lambda \left\| X - X^{*} \right\|^2
\]

Depth walking preference

\( X^{*} \): retrieved trajectory
minimize $E_D + E_{RGB} + \lambda \|X - X^*\|^2$

Walking preference

$X^*$: retrieved trajectory
Future localization
Outdoor scene
Indoor scene

Error rate (m/s)
0.8
0.6
0.4
0.2
0

Image+CNN+Image
Image+GIST+Image
Image+CNN+Image
Image+GIST+Image

Cartesian in image (LeCun et al.)
Cartesian on ground plane
Cartesian on ground plane

Short range error

Error rate (m/s)

Outdoor scene

Indoor scene

Short range
Ours

Short range error

Outdoor scene

Indoor scene

Error rate (m/s)
What can first person vision tell us?
What can first person vision tell us about our social interactions?
Gaze direction

Camera direction (frustum)
Gaze direction
Camera direction (frustum)
Autism spectrum disorder (ASD)  
Attention deficit hyperactivity disorder (ADHD) 

Joint Attention
Gaze direction
Camera direction (frustum)
Joint attention
Input Video: Meeting Scene

1x speed
3D Camera Pose Estimation
(Structure from motion)

Two groups

6x speed
Can we predict **without** first person cameras?
True positive head detection
True positive gaze detection
Halloween show

Social saliency: likelihood of joint attention
Halloween show

Social saliency

Children

Top view
$g\left(\left\{+_{i}\right\}_{i=1}^{N}\right) = -$  

where $N$ is the number of social members.
where $N$ is the number of social members.

Geometric localization: deterministic
$g\left(\left\{\text{\ding{125}}_i\right\}^{N}_{i=1}\right) =$  

Social formation

where $N$ is the number of social members.

Geometric localization: deterministic
where $N$ is the number of social members.
$g\left( \left\{ \begin{array}{c} i \end{array} \right\}^N_{i=1} \right) = -$

Social formation

where $N$ is the number of social members.

Scale variation
Orientation variation

- : Ground truth joint attention
+ : Head location
Representation: Social Dipole Moment
Electric dipole moment

Water molecule, H$_2$O

$q_e = \sum_{i}^{N} (e - p_i)$

Electric dipole moment
Water molecule, H₂O

\[ q_e = \sum_{i=1}^{N} (e - p_i) \]

Electric dipole moment
Water molecule, $\text{H}_2\text{O}$
Water molecule, \( \text{H}_2\text{O} \)
Water molecule, $\text{H}_2\text{O}$
Social dipole moment

\[ q = s - \frac{1}{N} \sum_{i}^{N} p_i = s - c \]

Electric dipole moment

\[ q_e = \sum_{i}^{N} (e - p_i) \]
Electric dipole moment

$$q_e = \sum_{i} (e - p_i)$$

Social dipole moment

$$q = s - \frac{1}{N} \sum_{i}^{N} p_i = s - c$$

Social member
Joint attention
Center of mass
Social dipole moment
Electric dipole moment

Social dipole moment

\[ q = s - \frac{1}{N} \sum_{i}^{N} p_i = s - c \]

Electric dipole moment

\[ q_e = \sum_{i}^{N} (e - p_i) \]
Electric dipole moment

\[ q_{e} = \sum_{i} (e - p_{i}) \]

Social dipole moment

\[ q = s - \frac{1}{N} \sum_{i}^{N} p_{i} = s - c \]
<table>
<thead>
<tr>
<th>Scene</th>
<th>N</th>
<th>T(sec)</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>B-boy I</td>
<td>18</td>
<td>105</td>
<td>317</td>
</tr>
<tr>
<td>B-boy II</td>
<td>18</td>
<td>450</td>
<td>1351</td>
</tr>
<tr>
<td>B-boy III</td>
<td>18</td>
<td>160</td>
<td>528</td>
</tr>
<tr>
<td>B-boy IV</td>
<td>18</td>
<td>50</td>
<td>180</td>
</tr>
<tr>
<td>Surprise party</td>
<td>11</td>
<td>120</td>
<td>2227</td>
</tr>
<tr>
<td>Class</td>
<td>11</td>
<td>360</td>
<td>3590</td>
</tr>
<tr>
<td>Croquet</td>
<td>6</td>
<td>300</td>
<td>6000</td>
</tr>
<tr>
<td>Busker I</td>
<td>6</td>
<td>120</td>
<td>3566</td>
</tr>
<tr>
<td>Busker II</td>
<td>6</td>
<td>180</td>
<td>5394</td>
</tr>
<tr>
<td>Card game</td>
<td>3</td>
<td>180</td>
<td>768</td>
</tr>
<tr>
<td>Hide and seek</td>
<td>3</td>
<td>180</td>
<td>214</td>
</tr>
<tr>
<td>Block building</td>
<td>3</td>
<td>700</td>
<td>2702</td>
</tr>
<tr>
<td>Social game</td>
<td>8</td>
<td>450</td>
<td>2086</td>
</tr>
<tr>
<td>Meeting I</td>
<td>11</td>
<td>120</td>
<td>832</td>
</tr>
<tr>
<td>Meeting II</td>
<td>5</td>
<td>440</td>
<td>1120</td>
</tr>
<tr>
<td>Picnic</td>
<td>6</td>
<td>60</td>
<td>965</td>
</tr>
<tr>
<td>Musical</td>
<td>7</td>
<td>180</td>
<td>2184</td>
</tr>
<tr>
<td>Dance</td>
<td>6</td>
<td>180</td>
<td>5301</td>
</tr>
<tr>
<td>4 way party</td>
<td>11</td>
<td>180</td>
<td>1909</td>
</tr>
<tr>
<td>Snowman</td>
<td>4</td>
<td>753</td>
<td>8256</td>
</tr>
</tbody>
</table>

Total 49,490 social formations
Dyadic interaction

Social formation theory (F-formation)*

Dyadic interaction

Center of mass with std.

Distribution of members
Social Group Detection
\[ \Phi(f^c, f^s) = \begin{cases} 1 & \text{if } x = s \\ \end{cases} \]
\[ \Phi(f^c, f^s) = \begin{cases} 1 & \text{if } x = s \\ 0 & \text{if } x \neq s \end{cases} \]
\[ \Phi(f^c, f^s) = \begin{cases} 
1 & \text{if } x = s \\
0 & \text{if } x \neq s 
\end{cases} \]
Social saliency: Likelihood of joint attention

\[ \Phi(f^c, f^s) = \begin{cases} 
1 & \text{if } x = s \\
0 & \text{if } x \neq s
\end{cases} \]
Social saliency: Likelihood of joint attention

\[ \Phi(f^c, f^s) = \begin{cases} 1 & \text{if } x = s \\ 0 & \text{if } x \neq s \end{cases} \]
First Person Basketball Data
What can first person vision tell us about?
What can first person vision tell us about?

Force from Motion
CVPR 2016
What can first person vision tell us about?

Force from Motion
CVPR 2016

Future Localization
CVPR 2016
What can first person vision tell us about?

- Force from Motion
  - CVPR 2016

- Future Localization
  - CVPR 2016

- Social Saliency Prediction
  - CVPR 2015
First person video tells us about me.
How do we control our vehicle w.r.t. visuals?
How do we control our vehicle w.r.t. visuals?
RPM: 24.56 %  ACC: 68.00 %  Handle: R 0.54 %  Speed: 1.00 %  BRK: 0.00 %