Social Saliency Prediction

Hyun Soo Park and Jianbo Shi

Penn
Nonverbal social signals

Hogan and Stubbs, *Can't Get Through 8 Barriers to Communication*, 2003.

Nonverbal social signals


Geometric Localization of Joint Attention

Gaze ray

Head location

Head orientation

Location of joint attention

Patron-Perez et al., Structured Learning of Human Interactions in TV Shows, PAMI, 2012
Fathi et al., Social Interactions: A First Person Perspective, CVPR, 2012
Park et al., 3D Social Saliency from Head-mounted Cameras, NIPS, 2012
Challenges in Social Scenes
Challenges in Social Scenes

Marin-Jimenez et al., *Detecting People Looking at Each Other in Videos*, IJCV 2014
Can we localize joint attention without measuring gaze directions?

True positive head detection

Challenges in Social Scenes

Marin-Jimenez et al., *Detecting People Looking at Each Other in Videos*, IJCV 2014
Structure from Motion
Social saliency: likelihood of joint attention

Output
Halloween show

Social saliency: likelihood of joint attention

Output
Distance between face and camera

Third person view

First person view

Cristani et al., BMVC 2011
Fathi et al., CVPR 2012
Park et al., NIPS 2012

Rodriguez et al., ICCV 2011
Lan et al., PAMI 2012
Chakraborty et al., CVPR 2013
Yang et al., CVPR 2011
Alahi et al., CVPR 2014
Choi et al., ECCV 2014

Li et al., ICCV 2013
Ryoo et al., CVPR 2013
Pusiol et al., CogSci 2014

Arev et al., SIGGRAPH 2014
Joint Attention from First Person Cameras

Croquet

Park et al., 3D Social Saliency from Head-mounted Cameras, NIPS, 2012.
Noninvasiveness

Measurement accuracy

Third person view

Distance between face and camera

First person view

3D estimation error < 10 cm

Cristani et al., BMVC 2011

Fathi et al., CVPR 2012

Park et al., NIPS 2012
3D estimation error <10cm

Measurement accuracy

Critani et al., BMVC 2011
Fathi et al., CVPR 2012
Park et al., NIPS 2012

Noninvasiveness
Prediction
Learning

Distance between face and camera
First person view
Third person view

3D estimation error <10cm
$g\left( \left\{ \begin{array}{c} +_i \\ i=1 \end{array} \right\}^N \right) = \text{Ground truth joint attention}$

where $N$ is the number of social members.
Geometric localization: triangulation

where $N$ is the number of social members.

Geometric localization: triangulation
\[ g \left( \{ \text{Head location}_i \}_{i=1}^{N} \right) = \text{Ground truth joint attention} \]

where \( N \) is the number of social members.
$g\left(\left\{+i\right\}_{i=1}^{N}\right) = -$ 

where $N$ is the number of social members.

ex. $g\left(\left\{+i\right\}_{i=1}^{N}\right) = \frac{1}{N} \sum_{i=1}^{N} +$ = $-$

Geometric localization: center of mass

: Ground truth joint attention

: Head location

: Center of mass
where $N$ is the number of social members.
$g\left(\{\text{Head location}_i\}_{i=1}^{N}\right) = \text{Ground truth joint attention}$

where $N$ is the number of social members.
First Person Social Interaction Data

<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
$g\left(\left\{\begin{array}{c} +_i \\ i \end{array}\right\}_{i=1}^{N}\right) = -$ 

Social formation

where $N$ is the number of social members.
\[ g \left( \left\{ \hat{s}_i \right\}_{i=1}^{N} \right) = - \]

where \( N \) is the number of social members.
$g\left(\left\{\left[\begin{array}{c}i \\
\end{array}\right]\right\}_{i=1}^{N}\right) = \text{Social formation}$

where $N$ is the number of social members.
Representation: Social Dipole Moment
Water molecule, $\text{H}_2\text{O}$
\[ q_e = \sum_{i=1}^{N} (\mathbf{e} - \mathbf{p}_i) \]

Electric dipole moment

Water molecule, \( \text{H}_2\text{O} \)

**Electric dipole moment**
Electric dipole moment

Water molecule, $\text{H}_2\text{O}$

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

Electric dipole moment
Electric dipole moment

Water molecule, $\text{H}_2\text{O}$

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

Electric dipole moment
Electric dipole moment

Water molecule, H₂O

$$\mathbf{q}_e = \sum_{i}^{N} (\mathbf{e} - \mathbf{p}_i)$$

Electric dipole moment
Social dipole moment

\[ \mathbf{p}_i \]

\[ \mathbf{q} \]

\[ \mathbf{s} \]

\[ \mathbf{c} = \frac{1}{N} \sum_{i}^{N} \mathbf{p}_i \]

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

- : Ground truth joint attention
+ : Head location
○ : Center of mass
→ : Social dipole moment
Orientation Normalization

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

- **q**: Ground truth joint attention
- **s**: Head location
- **c**: Center of mass
- **\( p_i \)**: Social dipole moment

Social dipole moment
Scale Normalization

\[
\frac{1}{N} \sum_{i}^{N} \| p_i - c \| = 1
\]
Social Formation Feature

\[ \sum \]

\[ p_i \]

\[ q \]

\[ s \]

\[ c \]

\[ f^c \]

- Social formation feature
- Social dipole moment
Social Formation Feature
Learning Likelihood of Joint Attention

\[ \Phi(f^c, f^s; x = s) = 1 \]
Learning Likelihood of Joint Attention

\[ \Phi\left(f^c, f^s; x = s\right) = 1 \]
\[ \Phi\left(f^c, f^s; x \neq s\right) = 0 \]

AdaBoost binary classifier
Joint Attention Prediction

\[ \Phi \left( f^c, f^s ; x = s \right) = 1 \]

\[ \Phi \left( f^c, f^s ; x \neq s \right) = 0 \]

AdaBoost binary classifier
Joint Attention Prediction

Social saliency: Likelihood of joint attention

\[
\Phi \left( f^c, f^s; x = s \right) = 1 \\
\Phi \left( f^c, f^s; x \neq s \right) = 0
\]

AdaBoost binary classifier
Joint Attention Prediction

Social saliency: Likelihood of joint attention

\[ \Phi \left( f^c, f^s; x = s \right) = 1 \]
\[ \Phi \left( f^c, f^s; x \neq s \right) = 0 \]

AdaBoost binary classifier
Input video → Human detection
Input video

Human detection

Group detection

Learning via FPCs

Social saliency prediction
Result
Group meeting
Street performance
Class interactions

CF: Context feature (Lan et al., PAMI 2012)
Group meeting

Street performance

Class interactions

CF: Context feature (Lan et al., PAMI 2012)
Group meeting

Street performance

Class interactions

Social member location
Joint attention location
Center of mass (COM)
Center of circumcircle (CC)

SSF: Social formation feature
RF: Random forests

CF: Context feature (Lan et al., PAMI 2012)

<table>
<thead>
<tr>
<th>Scenes</th>
<th>SFF+Boosting</th>
<th>SFF+RF</th>
<th>CC</th>
<th>COM</th>
<th>CF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dance</td>
<td>0.2769</td>
<td>0.1381</td>
<td>0.3299</td>
<td>0.0419</td>
<td>0.0106</td>
</tr>
<tr>
<td>Meeting I</td>
<td>0.2941</td>
<td>0.3599</td>
<td>0.2418</td>
<td>0.2350</td>
<td>0.0649</td>
</tr>
<tr>
<td>B-boy I</td>
<td>0.7178</td>
<td>0.6907</td>
<td>0.2078</td>
<td>0.1232</td>
<td>0.1225</td>
</tr>
<tr>
<td>Class</td>
<td>0.7678</td>
<td>0.7386</td>
<td>0.1445</td>
<td>0.2757</td>
<td>0.1873</td>
</tr>
<tr>
<td>Busker</td>
<td>0.2919</td>
<td>0.2059</td>
<td>0.3432</td>
<td>0.1929</td>
<td>0.0103</td>
</tr>
<tr>
<td>Picnic</td>
<td>0.1364</td>
<td>0.1349</td>
<td>0.1115</td>
<td>0.1808</td>
<td>0.0244</td>
</tr>
<tr>
<td>Social game</td>
<td>0.5425</td>
<td>0.4419</td>
<td>0.3461</td>
<td>0.2463</td>
<td>0.0020</td>
</tr>
</tbody>
</table>

Mean average precision
Busker
Time Square

Source: https://www.youtube.com/watch?v=ezyrSKgcyJw
Basketball Scene Result
First Person Basketball Data

University team of Northwestern Polytechnical University (China)
Dynamic Joint Attention Prediction

\[ \Phi(f^c, f^s; x = s) = 1 \]
\[ \Phi(f^c, f^s; x \neq s) = 0 \]
Dynamic Joint Attention Prediction

\[ \Phi(f^c, f^s, v_{\text{com}}; x = s) = 1 \]
\[ \Phi(f^c, f^s, v_{\text{com}}; x = s) = 0 \]
Person detector:
Can we predict social saliency without measuring gaze directions?

Social formation ←→ Social saliency
Social Saliency Prediction

Hyun Soo Park and Jianbo Shi

Project website:
http://www.seas.upenn.edu/~hypar/socialsaliencyprediction.html

Poster #36