Understanding egocentric imagery, for fun and science

David Crandall
School of Informatics and Computing
Indiana University

Joint work with: Denise Anthony (Dartmouth), Apu Kapadia, Chen Yu;
PhD Students: Sven Bambach, Roberto Hoyle, Mohammed Korayem, Stefan Lee, Robert Templeman; Undergrads: Steven Armes, Dennis Chen
- Google
- Narrative
- Samsung
- Autographer
- GoPro
Daguerreotypes, 1839

The Kodak, 1888

Polaroid Land Camera, 1948

Digital camera, 1975

J-phone, 2000
Daguerreotypes, 1839

The Kodak, 1888

Polaroid Land Camera, 1948

Digital camera, 1975

J-phone, 2000
What if your device were hacked?

Vision to the rescue!

• Could computer vision automatically help...
  – Organize and analysis egocentric image streams?
  – Find the great photos amongst all the bad ones?
  – Warn before sharing a photo with private data to Facebook?
  – Block or censor private photos from being taken by the device, and/or uploaded to the cloud?
What makes an image sensitive?

- 37 undergrads, wearing life-logging cameras for a week, each day reviewing images and labeling them in various ways.

Ethical, legal, IRB considerations

Photo Study Do's and Don'ts for Participants

During this study, you will be carrying a device on your neck that will be taking a picture of your surroundings every five minutes. It will also be taking samples of your movement, via an accelerometer, ambient sound levels, ambient light, and your location. There will be no audio recording of conversations. To make this experience pleasant for everyone involved, please observe a few guidelines while participating in the study.

Do's:
- DO keep the harness around your neck with the phone's screen facing you, and the rear camera of the phone facing out so that it is just peeking out of the harness.
- DO keep the harness outside of your regular clothing, so that it has an unblocked view of your surroundings.
- DO wear the harness so that the label is observable by those around you and in the phone's view.
- DO explain to persons who may be recorded the purpose of the study and the device.
- DO offer the informational cards provided, if you encounter anyone who requests more information about the study. The identifier on the card can be used to reference photos in an anonymous manner.
- DO put the device away if people around you express discomfort with the device or express their desire not to be recorded.
- DO pause the device whenever you want to stop recording, even if you need to put the device away. Pausing recording allows us to track how frequently people choose to avoid recording.
- DO contact us if you have any questions or concerns. We are reachable by email at [email protected].

Do nots:
- DO NOT wear the harness inside your clothing, so that it has a blocked view of your surroundings.
- DO NOT ever take photos of private areas.
- DO NOT have the device in your pocket or bag.
- DO NOT use the device in a manner that is disruptive or inappropriate.
- DO NOT use the device in a manner that would be considered unethical or illegal.

This is a study about how people use photologging applications, such as Google Glass and has been approved by Indiana University's Institutional Review Board and General Counsel. If you have any questions, please contact the researchers at photostudy@cs.soic.indiana.edu or view our FAQ at https://jhu.soic.indiana.edu/photoapp/faq.
Why not share the photo?

<table>
<thead>
<tr>
<th>Reason</th>
<th>Responses</th>
</tr>
</thead>
<tbody>
<tr>
<td>No good reason to share it</td>
<td>36.0%</td>
</tr>
<tr>
<td>Objects (other than people) in the photo</td>
<td>30.7%</td>
</tr>
<tr>
<td>Where this photo was taken</td>
<td>22.6%</td>
</tr>
<tr>
<td>People within the photo</td>
<td>18.4%</td>
</tr>
<tr>
<td>Participant was in the photo</td>
<td>11.5%</td>
</tr>
<tr>
<td>It had private information</td>
<td>11.5%</td>
</tr>
<tr>
<td>It would have been embarrassing to share it</td>
<td>5.4%</td>
</tr>
<tr>
<td>It would have violated someone else’s privacy</td>
<td>3.8%</td>
</tr>
<tr>
<td>It was a bad photo</td>
<td>1.5%</td>
</tr>
</tbody>
</table>
## Why not share the photo?

<table>
<thead>
<tr>
<th>Reason</th>
<th>Responses</th>
</tr>
</thead>
<tbody>
<tr>
<td>No good reason to share it</td>
<td>36.0%</td>
</tr>
<tr>
<td>Objects (other than people) in the photo</td>
<td>30.7%</td>
</tr>
<tr>
<td>Where this photo was taken</td>
<td>22.6%</td>
</tr>
<tr>
<td>People within the photo</td>
<td>18.4%</td>
</tr>
<tr>
<td>Participant was in the photo</td>
<td>11.5%</td>
</tr>
<tr>
<td>It had private information</td>
<td>11.5%</td>
</tr>
<tr>
<td>It would have been embarrassing to share it</td>
<td>5.4%</td>
</tr>
<tr>
<td>It would have violated someone else’s privacy</td>
<td>3.8%</td>
</tr>
<tr>
<td>It was a bad photo</td>
<td>1.5%</td>
</tr>
</tbody>
</table>
## Why not share the photo?

<table>
<thead>
<tr>
<th>Reason</th>
<th>Responses</th>
</tr>
</thead>
<tbody>
<tr>
<td>No good reason to share it</td>
<td>36.0%</td>
</tr>
<tr>
<td>Objects (other than people) in the photo</td>
<td>30.7%</td>
</tr>
<tr>
<td>Where this photo was taken</td>
<td>22.6%</td>
</tr>
<tr>
<td>People within the photo</td>
<td>18.4%</td>
</tr>
<tr>
<td>Participant was in the photo</td>
<td>11.5%</td>
</tr>
<tr>
<td>It had private information</td>
<td>11.5%</td>
</tr>
<tr>
<td>It would have been embarrassing to share it</td>
<td>5.4%</td>
</tr>
<tr>
<td>It would have violated someone else’s privacy</td>
<td>3.8%</td>
</tr>
<tr>
<td>It was a bad photo</td>
<td>1.5%</td>
</tr>
</tbody>
</table>

“This picture is of someone else's bedroom. It is private, and should not be shared without their permission.”
Why not share the photo?

<table>
<thead>
<tr>
<th>Reason</th>
<th>Responses</th>
</tr>
</thead>
<tbody>
<tr>
<td>No good reason to share it</td>
<td>36.0%</td>
</tr>
<tr>
<td>Objects (other than people) in the photo</td>
<td>30.7%</td>
</tr>
<tr>
<td>Where this photo was taken</td>
<td>22.6%</td>
</tr>
<tr>
<td>People within the photo</td>
<td>18.4%</td>
</tr>
<tr>
<td>Participant was in the photo</td>
<td>11.5%</td>
</tr>
<tr>
<td>It had private information</td>
<td>11.5%</td>
</tr>
<tr>
<td>It would have been embarrassing to share it</td>
<td>5.4%</td>
</tr>
<tr>
<td>It would have violated someone else’s privacy</td>
<td>3.8%</td>
</tr>
<tr>
<td>It was a bad photo</td>
<td>1.5%</td>
</tr>
</tbody>
</table>

“I wouldn't want the public to see my home for security reasons”
## Why not share the photo?

<table>
<thead>
<tr>
<th>Reason</th>
<th>Responses</th>
</tr>
</thead>
<tbody>
<tr>
<td>No good reason to share it</td>
<td>36.0%</td>
</tr>
<tr>
<td><strong>Objects (other than people) in the photo</strong></td>
<td>30.7%</td>
</tr>
<tr>
<td><strong>Where this photo was taken</strong></td>
<td>22.6%</td>
</tr>
<tr>
<td><strong>People within the photo</strong></td>
<td>18.4%</td>
</tr>
<tr>
<td>Participant was in the photo</td>
<td>11.5%</td>
</tr>
<tr>
<td>It had private information</td>
<td>11.5%</td>
</tr>
<tr>
<td>It would have been embarrassing to share it</td>
<td>5.4%</td>
</tr>
<tr>
<td>It would have violated someone else’s privacy</td>
<td>3.8%</td>
</tr>
<tr>
<td>It was a bad photo</td>
<td>1.5%</td>
</tr>
</tbody>
</table>
Place recognition in lifelogging images

• Given a stream of lifelogging photos, where was each photo taken?
  – In which specific room?
  – In which type of room?

• Use SVMs with standard image features
  – HOG, GIST, LBP, SIFT, ...

# Classifying photo streams with HMMs

<table>
<thead>
<tr>
<th></th>
<th>Bathroom</th>
<th>Bedroom</th>
<th>Garage</th>
<th>Living</th>
<th>Office</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bathroom</td>
<td>0.931</td>
<td>0.023</td>
<td>0.002</td>
<td>0.007</td>
<td>0.009</td>
</tr>
<tr>
<td>Bedroom</td>
<td>0.006</td>
<td>0.734</td>
<td>0.461</td>
<td>0.120</td>
<td>0.082</td>
</tr>
<tr>
<td>Garage</td>
<td>✓ 0.96</td>
<td>× 0.12</td>
<td>× 0.75</td>
<td>× 0.14</td>
<td>× 0.86</td>
</tr>
<tr>
<td>Living</td>
<td>0.014</td>
<td>0.020</td>
<td>0.420</td>
<td>0.127</td>
<td>0.162</td>
</tr>
<tr>
<td>Office</td>
<td>0.042</td>
<td>0.031</td>
<td>0.000</td>
<td>0.001</td>
<td>0.001</td>
</tr>
</tbody>
</table>

**Probabilities with individual photo classifiers:**

<table>
<thead>
<tr>
<th></th>
<th>Bathroom</th>
<th>Bedroom</th>
<th>Garage</th>
<th>Living</th>
<th>Office</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bathroom</td>
<td>0.896</td>
<td>0.436</td>
<td>0.060</td>
<td>0.015</td>
<td>0.010</td>
</tr>
<tr>
<td>Bedroom</td>
<td>0.010</td>
<td>0.052</td>
<td>0.026</td>
<td>0.004</td>
<td>0.002</td>
</tr>
<tr>
<td>Garage</td>
<td>✓ 0.99</td>
<td>✓ 0.93</td>
<td>✓ 0.74</td>
<td>✓ 0.84</td>
<td>✓ 0.83</td>
</tr>
<tr>
<td>Living</td>
<td>0.079</td>
<td>0.441</td>
<td>0.881</td>
<td>0.968</td>
<td>0.975</td>
</tr>
<tr>
<td>Office</td>
<td>0.006</td>
<td>0.027</td>
<td>0.009</td>
<td>0.009</td>
<td>0.012</td>
</tr>
</tbody>
</table>

**Probabilities after applying HMM:**

<table>
<thead>
<tr>
<th></th>
<th>Bathroom</th>
<th>Bedroom</th>
<th>Garage</th>
<th>Living</th>
<th>Office</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bathroom</td>
<td>0.896</td>
<td>0.436</td>
<td>0.060</td>
<td>0.015</td>
<td>0.010</td>
</tr>
<tr>
<td>Bedroom</td>
<td>0.010</td>
<td>0.052</td>
<td>0.026</td>
<td>0.004</td>
<td>0.002</td>
</tr>
<tr>
<td>Garage</td>
<td>✓ 0.99</td>
<td>✓ 0.93</td>
<td>✓ 0.74</td>
<td>✓ 0.84</td>
<td>✓ 0.83</td>
</tr>
<tr>
<td>Living</td>
<td>0.079</td>
<td>0.441</td>
<td>0.881</td>
<td>0.968</td>
<td>0.975</td>
</tr>
<tr>
<td>Office</td>
<td>0.006</td>
<td>0.027</td>
<td>0.009</td>
<td>0.009</td>
<td>0.012</td>
</tr>
</tbody>
</table>

---

Classifying photo streams with HMMs involves using Hidden Markov Models (HMMs) to analyze sequences of observations and infer the underlying states or categories. In this example, we classify photos into categories such as Bathroom, Bedroom, Garage, Living, and Office. The probabilities are calculated both with individual photo classifiers and after applying HMMs, which can improve accuracy by considering the sequence of photos rather than each one independently.
**Evaluation**

- Tested on 5 realistic lifelogging datasets

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Baseline</th>
<th>Local features + HMM</th>
<th>Global features + HMM</th>
<th>Local+global features + HMM</th>
<th>Local+global + human + HMM</th>
</tr>
</thead>
<tbody>
<tr>
<td>House 1</td>
<td>29.8%</td>
<td>89.2%</td>
<td>64.0%</td>
<td>89.2%</td>
<td>95.0%</td>
</tr>
<tr>
<td>House 2</td>
<td>31.0%</td>
<td>55.0%</td>
<td>56.4%</td>
<td>74.6%</td>
<td>76.8%</td>
</tr>
<tr>
<td>House 3</td>
<td>20.9%</td>
<td>97.4%</td>
<td>86.9%</td>
<td>98.7%</td>
<td>99.8%</td>
</tr>
<tr>
<td>Workplace 1</td>
<td>32.1%</td>
<td>75.5%</td>
<td>89.2%</td>
<td>87.7%</td>
<td>91.0%</td>
</tr>
<tr>
<td>Workplace 2</td>
<td>28.9%</td>
<td>92.3%</td>
<td>81.2%</td>
<td>98.7%</td>
<td>100.0%</td>
</tr>
<tr>
<td>Average</td>
<td>28.5%</td>
<td>81.9%</td>
<td>74.8%</td>
<td>89.8%</td>
<td>92.5%</td>
</tr>
</tbody>
</table>
Sample results
Detecting screens

CNNs, plus temporal smoothing, gives ~95% 2-way (monitor vs no-monitor) performance on real lifelogging data (vs ~73% baseline).

AS WE MAY THINK
A TOP U.S. SCIENTIST FORESEES A POSSIBLE FUTURE WORLD
IN WHICH MACHINES HELP US THINK

Vannevar Bush, *The Atlantic*, 1945
(Very broad) Research questions

• How do children coordinate their hand movement, eye gaze, and head movements?
  – How do these develop and change over time?
  – Are certain patterns predictive of later deficiencies?
  – ...

• How do children and parents interact, and how does this differ across subjects?
  – How do they jointly coordinate attention?
  – Which interaction patterns are most successful for learning?
  – ...

Typical experiments

• **13 child-parent dyads** (childrens’ mean age = 13 months, \(\sigma = 3.2\) months)
• Parents told to engage child with toys, naturally, e.g.:
  – Exchanging toys back and forth
  – Joint actions with toys
• \(~5\) minutes of video per trial
• In lab setting, more recently in naturalistic environments
Analysis

Video data processing:
- Pixel-level visual saliency estimation
  - Saliency map models e.g. of Itti et al
- Object segmentation and recognition
  - Positions of toys, hands, and faces
- Object holding activities

Motion data processing:
- Optical flow
- Head, hand inertial sensors
Eye gaze within the visual field

Head-eye coordination

Moving Head

Children

Parents

Stationary Head

\( \mu = (333, 253) \)
\( \sigma_x = 80, \sigma_y = 70 \)
\( N = 15,626 \)

\( \mu = (335, 226) \)
\( \sigma_x = 84, \sigma_y = 63 \)
\( N = 71,388 \)

\( \mu = (351, 229) \)
\( \sigma_x = 57, \sigma_y = 62 \)
\( N = 63,035 \)

\( \mu = (374, 210) \)
\( \sigma_x = 47, \sigma_y = 56 \)
\( N = 33,981 \)
Eye-hand coordination

Empty Hands

Children

µ = (332, 231)
σ_x = 84, σ_y = 68
N = 55,704

Holding Toy

Parents

µ = (301, 242)
σ_x = 85, σ_y = 62
N = 10,937

µ = (336, 244)
σ_x = 71, σ_y = 70
N = 27,062

µ = (333, 270)
σ_x = 63, σ_y = 73
N = 11,239
Saliency in first-person views

Comparison of average saliency (based on 148,000 frames each, using Itti et al)

→ No significant difference in child and parent views
Saliency in first-person views

Comparison of average saliency within hotspot around gaze location

→ Gaze predictiveness differs significantly
Longer-term question

Can we jointly model head and hand pose, eye gaze, saliency, and activity, both to better perform egocentric computer vision, and to help explain human vision?
Why start with hands?

• Hands are in nearly every frame of egocentric video
• Hand configuration reflects what we are doing and what we are paying attention to
• Detecting hands is a fundamental problem for both computers and people

“The feeling of ownership of our limbs is a fundamental aspect of self-consciousness.”

[Ehrsson 2004]


Hand detection and disambiguation

• In egocentric video of interacting people, locate:
  – The observer’s hands (my left, my right)
  – The other person’s hands (your left, your right)
  – The other person’s head (your head)

• Our approaches so far:
  – Strong temporal models, weak spatial/appearance models

  – Strong spatial/temporal models, weak appearance models, explicit camera (head) motion model
    S. Lee, S. Bambach, C. Yu, D. Crandall. This hand is my hand: A probabilistic approach to hand disambiguation in egocentric video, *CVPR Egovision*, 2014.

  – Strong appearance models, weak spatial/temporal models
Strong appearance models: CNNs (of course)

- Mostly off-the-shelf Caffe, fine-tuned from ImageNet
- Generate candidates using domain-specific information
  - Sample from distribution over size, position, shape of hand regions in training data, biased by skin color detector
Hand detection & disambiguation results

- 4 actors x 4 activities x 3 locations = 48 unique videos
- 15,053 pixel-level ground truth segmentations
Hand detection & segmentation
Does hand pose alone reveal first person activities?

• Different interactions afford different hand grasps [Napier 1965]
• Train and test CNNs for 4 activities (puzzle, jenga, cards, chess) on masked out images
  – Single frame accuracy: 66.4% with GT masks, 50.9% with automatic masks (vs 25% baseline)
  – With temporal cues (50 frames): 92.9% with GT masks, 73.4% with automatic masks
Hand Positions during Sustained Attention, within child's field of view

- Parent’s Right Hand
- Parent’s Left Hand
- Child’s Left Hand
- Child’s Right Hand

6 child-parent dyads (children’s mean age = 19 months, $\sigma = 2.56$ months)
67,913 frames (~38 minutes) of child-view video
Daguerreotypes, 1839
The Kodak, 1888
Polaroid Land Camera, 1948
Digital camera, 1975
J-phone, 2000
Future work

• Detecting and disambiguating hands
  – Generalize hand-based activity recognition to more activities, finer-grained actions
  – More challenging social situations (e.g. more than two interacting people, moving camera wearer)

• Applications for egocentric video data
For more information:  http://vision.soic.indiana.edu/

Based on joint work with:

- **Faculty**: Denise Anthony, Apu Kapadia, Chen Yu
- **PhD Students**: Sven Bambach, Roberto Hoyle, Mohammed Korayem, Stefan Lee, Robert Templeman
- **Undergrad students**: Steven Armes, Dennis Chen

Sponsors:

- NSF (III CAREER, SaTC, EAGER, DIBBs), Intelligence Advanced Research Projects Activity (IARPA), Air Force Office of Scientific Research (AFOSR), Google, Nvidia, IU FRSP, IUCRG, IU D2I Center, Lily Endowment