Who are they looking at?

Automatic Eye Gaze Following for Classroom Observation Video Analysis

Arkar Min Aung, Anand Ramakrishnan, and Jacob Whitehill

Worcester Polytechnic Institute (WPI)
Context
Classroom observation

• In the USA (and other countries), it is commonplace for administrators, researchers, and other teachers to make classroom observations:
  • Live
  • Video-based
Classroom observation

These observation sessions are used for:

- Professional development
- Accountability
- Educational research
Classroom observation protocols

• Classroom sessions are coded using one of several standard observation protocols to characterize different aspects of classroom instruction.

• One of the most commonly used protocols is the Classroom Assessment Scoring System (CLASS; Pianta, et al. 2008).
CLASS
Pianta, et al. (2008)

• An underlying assumption of the CLASS is that the quality of teacher-student interactions can be measured independently of the curriculum being taught.

• Significant evidence that CLASS scores predict children’s downstream academic, cognitive, and emotional outcomes, e.g.:
  • Reading achievement (Ponitz, et al. 2009)
  • Engagement (Curby, et al. 2014)
  • Executive functioning (Weiland, et al. 2013)
<table>
<thead>
<tr>
<th>Domain</th>
<th>Indicators</th>
</tr>
</thead>
<tbody>
<tr>
<td>Emotional support</td>
<td>Positive climate, Negative climate, Teacher sensitivity, Regard for child perspectives</td>
</tr>
<tr>
<td>Classroom organization</td>
<td>Behavioral management, Productivity, Instructional learning formats</td>
</tr>
<tr>
<td>Instructional support</td>
<td>Concept development, Quality of feedback, Language modeling, Literacy focus</td>
</tr>
</tbody>
</table>
# CLASS

Pianta, et al. (2008)

<table>
<thead>
<tr>
<th>Domain</th>
<th>Dimension</th>
</tr>
</thead>
<tbody>
<tr>
<td>Emotional support</td>
<td>Positive climate</td>
</tr>
<tr>
<td></td>
<td>Negative climate</td>
</tr>
<tr>
<td></td>
<td>Teacher sensitivity</td>
</tr>
<tr>
<td></td>
<td>Regard for child perspectives</td>
</tr>
<tr>
<td>Classroom organization</td>
<td>Behavioral management</td>
</tr>
<tr>
<td></td>
<td>Productivity</td>
</tr>
<tr>
<td></td>
<td>Instructional learning formats</td>
</tr>
<tr>
<td>Instructional support</td>
<td>Concept development</td>
</tr>
<tr>
<td></td>
<td>Quality of feedback</td>
</tr>
<tr>
<td></td>
<td>Language modeling</td>
</tr>
<tr>
<td></td>
<td>Literacy focus</td>
</tr>
</tbody>
</table>
## CLASS

*Pianta, et al. (2008)*

<table>
<thead>
<tr>
<th>Domain</th>
<th>Dimension</th>
<th>Indicators</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Emotional support</strong></td>
<td>Positive climate</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Negative climate</td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>Teacher sensitivity</strong></td>
<td><strong>Awareness</strong>&lt;br&gt;<strong>Responsiveness</strong>&lt;br&gt;<strong>Address problems</strong></td>
</tr>
<tr>
<td></td>
<td>Regard for child perspectives</td>
<td></td>
</tr>
<tr>
<td><strong>Classroom organization</strong></td>
<td>Behavioral management</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Productivity</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Instructional learning formats</td>
<td></td>
</tr>
<tr>
<td><strong>Instructional support</strong></td>
<td>Concept development</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Quality of feedback</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Language modeling</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Literacy focus</td>
<td></td>
</tr>
</tbody>
</table>
# CLASS

Pianta, et al. (2008)

<table>
<thead>
<tr>
<th>Domain</th>
<th>Dimension</th>
<th>Indicators</th>
<th>Behavioral markers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Emotional support</td>
<td>Positive climate</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Negative climate</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Teacher sensitivity</strong></td>
<td></td>
<td><strong>Awareness</strong></td>
<td><strong>Notices lack of understanding</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>Responsiveness</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>Address problems</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Regard for child perspectives</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Classroom organization</td>
<td>Behavioral management</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Productivity</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Instructional learning formats</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Instructional support</td>
<td>Concept development</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Quality of feedback</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Language modeling</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Literacy focus</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Manual classroom observation

• With the CLASS, human annotators assign one number (1-7) to each dimension once every 15 minutes.
  • Sparse
  • Expensive
  • Non-specific (difficult to label which children/teachers were most important)
Automated classroom observation

- It could be useful to (partially) automate this process:
  - More frequent and specific feedback to teachers
  - Improved lens to estimate impact of educational interventions
Automated classroom observation: feasibility

- Some dimensions are likely more automatable than others.

- For some emotional support dimensions, the behavioral markers are related to:
  - Facial expression
Automated classroom observation: feasibility

- Some dimensions are likely more automatable than others.

- For some emotional support dimensions, the behavioral markers are related to:
  - Facial expression
  - Physical proximity
Automated classroom observation: feasibility

• Some dimensions are likely more automatable than others.

• For some emotional support dimensions, the behavioral markers are related to:
  • Facial expression
  • Physical proximity
  • Mutual eye-gaze between students and teachers.
Gaze following
Problem Statement

Given a classroom observation video, we want to know:

- Where and at whom each person in the classroom is looking (gaze-following)?
Gaze-following in 2-D Static Images

• Annotating gaze locations in 2-D images:
  – Can be ambiguous since 2-D images does not have depth information.
  – **Assumption:** Knowing gaze location in 2-D images can be informative for downstream processing.

• 2-D images are a lot easier to obtain than 3-D images (RGB-D images).
Classroom observation videos

Video taken from: https://www.youtube.com/watch?v=cjNv2dQCFEk&list=PLI4ATdTjDSeSpcj9YJie416kV5heHs-598&index=4&t=1084s
Classroom observation videos

- Multiple students and teachers
- Highly cluttered
- Significant occlusion
- Extreme head poses (with faces sometimes pointing away from camera)
Differences in Datasets

MS COCO, SUN, Actions, Places, PASCAL Datasets

Classroom Observation Video Images
Data Collection
Data Sourcing

• Use 70 classroom observation videos\textsuperscript{[1]} publicly available on YouTube.
• Extract 1 frame approximately every 10 seconds.
• Use Faster R-CNN for face detection\textsuperscript{[2]} to obtain face bounding boxes in extracted frames.
• 7.85 faces per image on average (for the whole dataset)

Data Annotation

- **Tool built with HTML5+Javascript** and deployed on Amazon Mechanical Turk (AMT).

- Collects gaze location as well as binary indication of whether the gaze ends inside or outside the image.
Data Annotation

- 3 labelers per image on average on AMT to annotate the gaze of each face.
- 408 unique annotators.
- Collected three gaze annotations each for 17,758 faces in 2,263 images.
- After cleaning data, obtained a total of 48,907 gaze annotations.
Dataset

- Training data is augmented by flipping images and gazes left to right.

- Data split
  - 70% Training
  - 15% Validation
  - 15% Testing

- Sets of people in training, validation, and test don’t overlap.

- No image from the same video occurs in more than one data split.
Sample Annotations (for 3 labelers)
Network Design
To regress or to classify?

- The task of following the gaze of a person can be formulated as either:
  - A classification task
  - A regression task
(x,y) coordinates and soft labels

256x256 pixel image

Regression

Classification

Separate training label

(140, 130)

(100, 200)

(90, 235)

8x8 Grid
Deep Learning Architecture

• Approach is inspired by Recasens, et al (2015) \[^1\].
• We use VGG16\[^2\] as the base architecture.
• We use different optimization techniques.
  – Transfer learning with fine tuning.
• Multiple-tasks
  – Predict the gaze location.
  – Predict whether the gaze ends inside or outside the image (In/Out gaze).

---

Take this image for an example
We want to know the gaze of this girl
Face-to-Gaze pathway

Only have access to close-up face image and head location

Intuition:
1) Infer gaze from head pose
Only have access to image of the scene without knowing anything about where the subject of interest is.

Intuition:
1) Learn to detect salient objects
Objects tend to emerge in the filter kernels of deep layers of CNNs [1].
Research Questions

1. How accurately can the Merged Model predict gaze locations?

2. Can our Merged Model predict whom the person is looking at?
Results
Regression Baselines

- **Random Gaze**: Random location over the whole image.

Regression Baselines

- **Random Gaze**: Random location over the whole image.
- **Center Region**: Random gaze constrained to center 10% of the image. Motivated by Judd, et al\(^3\).

Regression Baselines

- **Random Gaze**: Random location over the whole image.
- **Center Region**: Random gaze constrained to center 10% of the image. Motivated by Judd, et al[3].
- **Linear regression**: use shallow network to predict \((x,y)\) from close-up cropped face and head location.
Regression Baselines

- **Random Gaze**: Random location over the whole image.
- **Center Region**: Random gaze constrained to center 10% of the image. Motivated by Judd, et al\[3\].
- **Linear regression**: use shallow network to predict \((x,y)\) from close-up cropped face and head location.
- **Face-to-Gaze**: Left half of Merged Model. Only have access to close-up cropped face and head location.

## Regression Results

**Regression results (within 256x256 pixel image)**

<table>
<thead>
<tr>
<th></th>
<th>MAE*</th>
<th>Mean Euclidean Distance*</th>
<th>Mean Absolute Angular Error</th>
<th>AUC for In/Out</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random Gaze</td>
<td>79.74</td>
<td>124.15</td>
<td>67.24°</td>
<td>-</td>
</tr>
<tr>
<td>Center Region</td>
<td>52.76</td>
<td>82.11</td>
<td>48.36°</td>
<td>-</td>
</tr>
<tr>
<td>Linear Regression</td>
<td>49.63</td>
<td>77.34</td>
<td>55.21°</td>
<td>-</td>
</tr>
<tr>
<td>Face-to-Gaze</td>
<td>45.74</td>
<td>71.53</td>
<td>39.91°</td>
<td>0.54</td>
</tr>
<tr>
<td>Merged Model</td>
<td>44.49</td>
<td><strong>69.82</strong></td>
<td><strong>38.30°</strong></td>
<td><strong>0.62</strong></td>
</tr>
<tr>
<td>Human</td>
<td>25.91</td>
<td>41.04</td>
<td>18.38°</td>
<td>0.70</td>
</tr>
</tbody>
</table>

*Distance in pixels*
Qualitative Results (Regression)
Qualitative Results (Regression)

- The merged model sometimes accurately estimates the direction, but not the distance, of the gaze.
- E.g., the girl in red box is looking at teacher’s hands but the gaze endpoint stops before getting to the hands.
Who are they looking at?
Who are they looking at?

- Analyze subset of faces s.t. all annotators agree he/she is looking at another face (not just any other object).
- Prediction task: given that the person is looking at a face, whose face is he/she looking at?
Merged Model Predictions on faces

- Start with the network’s predictions on 8x8 grid.
- Remove any cells containing no faces.
- Find top $k=1$ cells with highest predicted gaze probability.
- Predict the face contained within that cell.
Merged Model Predictions on faces

- Start with the network’s predictions on 8x8 grid.
- Remove any cells containing no faces.
- Find top $k=1$ cells with highest predicted gaze probability.
- Predict the face contained within that cell.
- Can also consider top $k=1,2,3$ faces (c.f. object detection literature).

Face cells on 8x8 grid

Merged model predictions in color
(Top 1 face – 3
Top 2 face – 2 or 3
Top 3 faces – 1, 2 or 3)
Results for “Who are they looking at?”

Probability of correctly identifying which face a person is looking at on 8 × 8 grid.

<table>
<thead>
<tr>
<th>Top k faces</th>
<th>k = 1</th>
<th>k = 2</th>
<th>k = 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random Face</td>
<td>0.15</td>
<td>0.30</td>
<td>0.45</td>
</tr>
<tr>
<td>Merged Model</td>
<td>0.47</td>
<td>0.65</td>
<td>0.79</td>
</tr>
<tr>
<td>Human</td>
<td></td>
<td></td>
<td>0.82</td>
</tr>
</tbody>
</table>

- 6.87 faces per image on average (for test set)
Results for “Who are they looking at?”

*Probability of correctly identifying which face a person is looking at on an $8 \times 8$ grid.*

<table>
<thead>
<tr>
<th>Top $k$ faces</th>
<th>$k = 1$</th>
<th>$k = 2$</th>
<th>$k = 3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random Face</td>
<td>0.15</td>
<td>0.30</td>
<td>0.45</td>
</tr>
<tr>
<td>Merged Model</td>
<td>0.47</td>
<td>0.65</td>
<td>0.79</td>
</tr>
<tr>
<td>Human</td>
<td></td>
<td></td>
<td>0.82</td>
</tr>
</tbody>
</table>

- 6.87 faces per image on average (for test set)

- 79% of the time, NN can correctly “narrow down” the gazed-at face to a set of 3 people.
Summary
Summary

• With a modest-sized (70 classroom observation videos) dataset, we can train a NN to predict eye gaze (where & whom) from 2-D images.

• **Whom**: 79% of the time, NN can correctly “narrow down” the possible gaze targets to < 1/2 the number of classroom participants.
Summary

• With a modest-sized (70 classroom observation videos) dataset, we can train a NN to predict eye gaze (where & whom) from 2-D images.

• **Whom**: 79% of the time, NN can correctly “narrow down” the possible gaze targets to < 1/2 the number of classroom participants.

• Eye gaze is just one of many behavioral markers that could be useful for classroom observation.
Summary

• With a modest-sized (70 classroom observation videos) dataset, we can train a NN to predict eye gaze (where & whom) from 2-D images.

• **Whom:** 79% of the time, NN can correctly “narrow down” the possible gaze targets to < 1/2 the number of classroom participants.

• Eye gaze is just one of many behavioral markers that could be useful for classroom observation.

• Long-term goal is to integrate many (noisy) predictors into an automated — or hybrid — classroom observation system.
End