Who are they looking at? Automatic Eye Gaze Following for Classroom Observation Video Analysis

Arkar Min Aung
Worcester Polytechnic Institute
aaung@wpi.edu

Anand Ramakrishnan
Worcester Polytechnic Institute
aramakrishnan@wpi.edu

Jacob R. Whitehill
Worcester Polytechnic Institute
jrwhitehill@wpi.edu

ABSTRACT

We develop an end-to-end neural network-based computer vision system to automatically identify *where* each person within a 2-D image of a school classroom is looking ("gaze following"), as well as *who* she/he is looking at. Automatic gaze following could help facilitate data-mining of large datasets of *classroom observation* videos that are collected routinely in schools around the world in order to understand social interactions between teachers and students. Our network is based on the architecture by [27] but is extended to predict whether each person is looking at a target inside or outside the image; and to predict not only where, but who the person is looking at. Moreover, since our focus is on classroom observation videos, we collected a dataset from scratch of publicly available classroom sessions from 70 YouTube videos and collected labels from 408 labelers who annotated a total of $17,758$ gazes in $2,263$ unique image frames. Results of our experiments indicate that the proposed neural network can estimate the gaze target – either the spatial location or the face of a person – with substantially higher accuracy compared to several baselines.

Keywords

Automatic Eye Gaze Following; Classroom Observation Videos; Deep Neural Networks

1. INTRODUCTION

The nature and quality of teacher-student interactions in school classrooms are predictive of learners' development. Numerous observational studies and several causal studies have demonstrated the link between emotional and instructional support in the classroom and children's cognitive, social, and emotional skills [18, 23]. In order to discover how classroom interactions are related to learning outcomes, educational researchers often conduct *classroom observation* sessions, whereby human coders score either live or video-recorded classroom observations (typically 1 hour long each) along different dimensions, such as positive climate, teacher sensitivity, language modeling, quality of feedback, etc [25]. The Gates Foundation Measures of Effective Teaching (MET) project [16], in particular, recorded tens of thousands of hours of classroom observations across the United States with the aim of discovering best practices for how to teach students most effectively.

This material is based upon work supported by the National Science Foundation under Grant No. #1551594 and Spencer Small Research Grand No. #201800131.

One of the major impediments to learning more from classroom observation video datasets is the difficulty and labor involved in coding them. Deep understanding of teacher-student interactions requires the coder to consider how the affective, linguistic, and pedagogical channels interact, and to interpret interactions within the context of classroom instruction. However, classroom observations contain multiple students and teachers interacting simultaneously in different parts of the classroom. It is easy for human coders to miss a subtle but important interaction. As a result, scores often can vary across coders, and multiple codes per video must be collected to obtain a reliable estimate. It would thus be invaluable to devise methods that could at least partially automate the process of classroom observation coding. Such a system could be useful not only for educational data-mining of large-scale classroom observation datasets, but also facilitate teachers' professional development by showing them video examples from their own classrooms in which they scored particularly high or low along different dimensions.

One important element of effective teacher-student interactions involves the students' and teachers' **eye gaze**: Does the teacher convey respect to his/her students by looking them in the eye when he/she is talking to them (*positive climate*)? Does the teacher notice when specific persons in the room are bored, confused, or even bullied (*teacher sensitivity*)? Tracking the eye gazes of students can also provide information on their thoughts and intentions [5] and may indirectly reveal how engaged they are in their learning.

In this paper, we take a tiny step towards creating an automated classroom observation scoring system. In particular, we build a prototype computer vision-based system for *automated eye gaze following* that estimates, for each person in the classroom, where she/he is looking. Such a system can be used to data-mine classroom observation video datasets. It could also facilitate "smart classrooms", which track gazes of both students and teachers, identify disengaged or distressed students, and help teachers to better recognize whether they are paying attention to the right thing or the right student in the classroom.

Deep learning for gaze following in classrooms: In this work, we explore a machine learning-based approach to automatic recognition of where a person in the image is looking. In particular, we build an end-to-end deep neural network that takes 2-D static images of multiple people as inputs and infers (x, y) coordinates of where *each* person

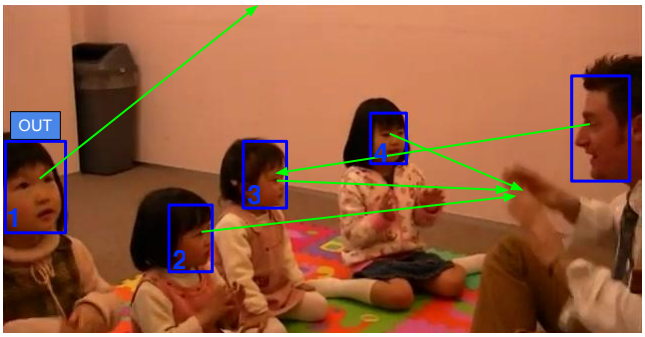

Figure 1: Eye gaze targets labeled by a human labeler for each person in the image. Labelers also indicate targets that are located outside the field-of-view (indicated by "OUT"). Can we build a computer vision system that can estimate *where* each person is looking? In this image, the man is looking at child #3. Can we identify automatically *who* each person is looking at? Image from `https://goo.gl/xUdYbC`

is looking at in the image as outputs. This computational problem is known as *gaze following* [10]. Gaze following from 2-D images is particularly challenging since 1) no additional information of the scene, such as depth information, is available and a person can be looking at any of the different planes of depth in the image, 2) people in the image can be looking at objects either inside the image or outside the image, 3) the eyes of some people may be blurred or partially invisible. Nonetheless, requiring only 2-D images is attractive because of the ubiquity and greater convenience of using commodity 2-D cameras. Our automated system is based on the architecture by [27], who tackled a similar problem for general images from the web. However, our approach differs from theirs in several ways, including the prediction outputs, deep neural network architectures, training techniques, dataset collection, and application focus.

Contributions: (1) We explore a deep learning-based architecture, based on related work by [27], for automatic eye-gaze following from 2-D images of classroom observation videos. (2) We extend the model of [27] to support gaze targets that can be *outside* the camera's field-of-view. Especially due to the lack of depth information, this is a highly challenging problem, both for human labelers and the machine. (3) Our application focus is on school classrooms, which contain many subjects (not just a few, as in [27]), who gaze not only inside but sometimes also outside the field-of-view. We thus collected and annotated (see Figure 1) a new dataset of images from classroom videos. (4) Since classroom observation analysis is largely about interaction between subjects, we explore the accuracy of our automatic gaze following system in identifying which *face* (not just object) each person is looking at. Detailed methodology and results for contribution (1), (2) and (3) are described in Section 3 and those of contribution (4) are described in Section 5.

2. RELATED WORK

Eye gaze following: Due to the importance of following gaze of others, which humans do naturally when communicating, collaborating and socializing, researchers in the field of robotics, computer vision and machine learning have recently started to formulate and tackle the problem of automatic gaze following within different contexts: In some settings [15, 3, 11], there is only a single person whose gaze is being followed, e.g., a student who is interacting with a mobile phone or a tablet [19] to play an educational game [31]. In other settings (such as ours), the camera examines an entire scene containing many people, and the gaze of *each* person in the scene is followed [24] [21] [28] [27]. While most of the prior work uses RGB data, some approaches also use depth information [24]. More recently, researchers have considered gaze following not only from static images but also how to harness temporal information from an entire video to better estimate the person's gaze target [28]. In this work, we only consider gaze following from static 2-D images extracted from classroom observation videos but future work can explore following gaze by using temporal information from a sequence of images.

Saliency modeling: Gaze following is related to saliency modeling, whereby image features of different levels of abstraction (low-, mid-, and high-level) are examined to consider the most likely locations in the image to which an observer would visually attend [15]. [3] made a connection between these two by stating that an observer looking at an image containing people may follow the gaze of people rather than actually fixating on salient objects in that image. Therefore, gaze following can play a complementary role in solving the problem of saliency model of attention. [7] explored the problem of predicting a driver's gaze behaviours and identifying the attention of a driver by detecting saliency in a complex driving environments.

Modeling non-verbal cues of students and teachers: There has been substantial prior work on analyzing learners' affective states from video using computer vision [17, 12, 4, 30]. Much of this work has focused on intelligent tutoring systems. More recently, researchers in multi-modal machine learning and educational data mining have investigated how to characterize the dynamics of an entire classroom. For example, [9, 8] explored approaches for segmenting and recognizing students' and teachers' speech in unconstrained classrooms based on different configurations of Microsoft Kinect cameras. For automated classroom observation scoring (e.g., of CLASS [25]), we are only aware of one prior work: [26] developed a computer vision system, optimized within a multiple-instance learning framework [22], to estimate which 3-minute snippets of classroom videos were most relevant for CLASS coders to watch.

3. EXPERIMENT I: METHODOLOGY

3.1 Data collection

Since the application focus of our study is gaze following in *school classrooms*, we collected our own dataset of classroom observation sessions. In particular, we harvested 70 videos publicly available on YouTube of school classrooms. The study was approved under WPI IRB 18-0101. In contrast to publicly available annotated data on gaze following (the only such dataset of which we are aware is GazeFollow [27]), classroom observation videos often contain *many* people per image frame, and the kinds of background clutter differ significantly from that of GazeFollow, which largely consists of images used for more general object detection research.

From each video in our collection, we extracted 1 frame approximately every 10 seconds. After extracting frames from videos, we used Faster R-CNN for face detection [14] to obtain face bounding boxes (top left (x, y) coordinate, width and height) in extracted frames.

Annotation: Ground-truth gaze annotations from the image frames were collected using at least 3 labelers per image on Amazon Mechanical Turk (AMT). Labelers used an online annotation tool that we custom-built for this work, using JavaScript and HTML5, to annotate two main components of each subject in each scene. The first component is to identify the gaze target for each person (identified automatically by the face detector as described above) which is indicated by a line, starting between the eyes of a person and ending on an object or a person which the person is attending to. The second component is the indication of whether the person is looking at something inside or outside the image. We collected three gaze annotations each for $17,758$ faces in $2,263$ images, resulting a total of $48,907$ gaze annotations from 408 unique annotators.

3.2 Approach

Using the datasets annotated on AMT, our goal is to build a convolutional neural network (CNN) which takes in the whole image of the scene and predicts the gaze target of each person in the image along with the indication of whether that target is inside or outside the image. We have observed from our annotated datasets that predicting gaze can be ambiguous. If there are multiple people or several salient objects in the image, or the eyes of individuals in the image are not clearly visible, human labelers may disagree when predicting gaze locations. Due to this inherent uncertainty in the problem, we explore various options to design our model to support multimodal predictions.

We can formulate gaze following as either a regression or a classification task. **Regression**: the network regresses to (x, y) coordinates of the gaze target of each person in the image using the Euclidean distance between the predicted and ground-truth as the cost function. The disadvantage of using regression is that our predictions are constrained to be unimodal. Since each face in each image was labeled by multiple annotators, we can define the ground-truth by either (a) computing the mean (x, y) location over all labels per face, or (b) treating each location as a separate label. **Classification**: the gaze location is quantized into one cell on an $N \times N$ grid, and the network's job is to choose the correct cell for each person in the image. As the cost function, we can use cross-entropy loss. Classification naturally supports multimodal outputs since multiple gaze annotations at different cells can be treated as soft labels [1]. The disadvantage of this approach is that the choice of grid size can affect the precision of predictions (i.e. smaller numbers of grid cells N will result in poor precision). Another issue is that cross-entropy loss does not gradually penalize mistakes based on distance – misclassification which is off by one grid cell is penalized just as much as misclassification which is off by several cells on a grid.

3.3 Architecture

The deep learning architecture is based on the model by [27] and is depicted in Figure 2. The gaze target for each person

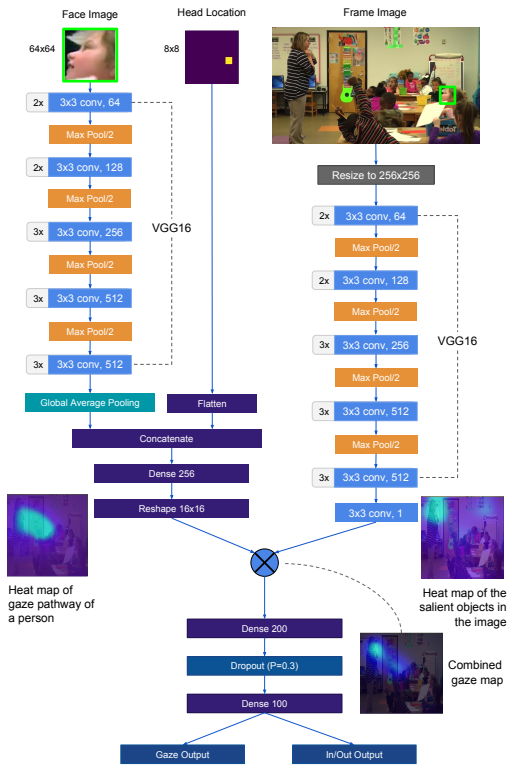

Figure 2: Deep neural network architecture, based on [27], for automatic eye-gaze following in school classrooms, consisting of two independent prediction pathways.

in the image is predicted independently based on two information sources: close-up information of the person's face (automatically detected by a separate face detection network [14]), and the whole image. Each information source is processed by a separate pathway consisting of a CNN, and the pathways' predictions about the person's gaze target are merged at the end. We call the combined architecture the *Merged Model*. In contrast to [27], we use the VGG16 [29] as the backbone of each CNN since we found empirically that it performed better than AlexNet [20]. Two other differences from [27] are the network optimization techniques and the use of multi-task learning (as described in Section 3.4). **Inputs**: The inputs of the Merged Model are a cropped, close-up face image (64×64 pixels); the $(r, c) \in N \times N$ location of the center of the person's head in the image; and the resized 256×256 pixels image of the whole frame. We chose $N = 8$ in our experiments. **Outputs**: For regression, the gaze target is represented as an (x, y) coordinate pair. For classification, the gaze target consists of a 1-hot vector indicating which of the $N \times N$ grid cells contains the gaze target. In addition (for both regression and classification), the network also contains an "in"/"out" binary prediction of whether the gaze target is inside or outside the image.

The intuition behind the Merged Model is that two CNNs are trained to solve two subproblems in a fully end-to-end fashion with only the gaze location and the "in"/"out" label as supervision to the model: (1) The close-up face CNN (left pathway in Figure 2) implicitly estimates the head pose and the direction of the gaze of the subject in order to produce

a heat map (shown as `Reshape 16×16` in Figure 2) of where the person is looking. In the figure, the heat map roughly shows a "cone" of possible gaze targets to the upper-left of the child's head. (2) The frame-image CNN (right pathway in Figure 2) identifies the salient objects in the image. This network has access to the entire original image but does not know the location of the subject. In the figure, the salient object heat map highlights the teacher in the upper-left of the image. In [32], the authors showed that objects tend to emerge in the filter kernels of the deep layers of CNNs; therefore, we take a filter kernel at the end of the right pathway (shown as 3×3 `conv, 1` in Figure 2). This produces the heat map of salient objects in the original image. Each heat map from each branch is combined by element-wise multiplication.

3.4 Training procedure

Data partitions: The 70 YouTube videos containing school classrooms were partitioned into training (12,430 gazes), validation (2,664 gazes), and testing (2,664 gazes) sets, such that none of the frames from any video was assigned to more than one set. The validation set was used for early stopping. The accuracy on the test set can be considered a performance estimate on faces that the network has never seen before.

Optimization: We used the following procedure for both the regression and classification formulations: We first performed transfer learning by initializing both CNNs with weights pre-trained on ImageNet [29]. We augmented the classroom images from our dataset by flipping the original images (frame image pathway) as well as the individually cropped face images, head locations and gaze locations (face pathway) left to right. We trained the final Merged Model first by freezing all the convolutional layers and training only the fully connected layers with RMSProp [13] (learning rate = 0.01, $\rho = 0.9$). Then all the previously frozen convolutional layers were unfrozen and the model was fine-tuned with SGD with momentum (learning rate=0.0001, momentum=0.9). The model was trained until there was no improvement in validation loss.

Multi-task learning: Since the Merged Model predicts the location of the gaze in the image as well as "in"/"out", it is performing multiple tasks, and we can use multi-task learning (MTL) [6] for training. Sharing the same hidden layers to solve several tasks forces the model to find representations which capture all of the tasks and thus reduce the risk of overfitting [2]. We found empirically that MTL helped to reduce overfitting and improve prediction accuracy. Table 1 compares the performance of the Merged Model with and without MTL. With MTL, the cross-entropy loss for both the grid output and the In/Out output is higher (worse) on the training set, but lower (better) on the testing set, compared to training two networks to handle each task separately. We thus adopted the MTL approach for training.

3.5 Accuracy measurement

Accuracy is measured for predicting the gaze target of each person (identified automatically by a face detector [14]) in each extracted frame from each of the YouTube videos (see Section 3.1). For **classification** of the gaze target among the $N \times N$ grid cells, we evaluated accuracy in terms of the cross-entropy (CE) loss w.r.t. the label distribution induced by the 3 annotators per example. For **regression** to an (x, y) location, we use mean absolute error (MAE), mean Euclidean distance and mean angular error (between the center of the person looking to their gaze target) in degrees, where the ground-truth is defined as the *average* annotation over all the annotators. In addition (for both regression and classification), we also used the Area Under the Receiver Operating Characteristics Curve (AUC) to evaluate the binary classification of whether the target is inside or outside the field-of-view.

Table 1: Effects of multi-task learning. CE Loss refers to Cross Entropy Loss and reported values are Cross Entropy Loss of Merged Model predicting gaze on 8×8 grid.

	Only grid output	Only In/Out output		Both grid output and In/Out output		
	CE Loss	CE Loss	AUC	CE Loss (Grid Output)	CE Loss (In/Out)	AUC (In/Out)
Training	3.27	0.32	0.63	3.39	0.33	0.60
Testing	3.59	0.46	0.59	3.58	0.43	0.62

3.6 Baseline comparison

When assessing the accuracy of any neural network, it is important to establish the relevant baselines for comparison. For classification, we use a uniform distribution over all $N \times N$ grid cells – in other words, a random guess in the whole image as to where the person is gazing. Alternatively, we can assume a center prior (motivated by [15]), consisting of the center 2×2 grid cells over the $N \times N$ grid. A variation on the center prior is to place a 2-D Gaussian – whose standard deviation σ is optimized directly on the *test set* for best possible accuracy – centered on the middle of the image, and assign probabilities to the $N \times N$ cells based on the Gaussian probability density function. For regression, we use a center prior corresponding to the midpoint in the image; we also compare to randomly selected points in the image.

As stronger baselines, we also consider linear regression to analyze the vectorized face pixels concatenated with head locations to predict (x, y) coordinates, as well as logistic regression to predict cells on a $N \times N$ grid. Finally, as a way of understanding which part of the Merged Model contains more information, we also compare to a Face-to-Gaze model consisting of a CNN that takes a cropped, close-up face image and location of head in the image as inputs, and predicts the location of the gaze in the image as well as "in"/"out" – this is the left pathway of Figure 2. Comparing with this baseline helps us understand how much the saliency pathway improves performance.

4. RESULTS I

Accuracy results on test images of the Merged Model compared to the baselines are shown in Table 2 (for regression) and Table 3 (for classification). Our Merged Model achieves mean Euclidean distance of 69.82 pixels on 256×256 pixel image (for regression) and cross entropy loss of 3.5855 on 8×8 grid (for classification) for gaze locations. These numbers are better than for the random gaze, center prior, center Gaussian, linear and logistic regression baselines. For comparison, human labelers exhibited a mean Euclidean distance of only 41.04 pixels on 256×256 pixel image, which

Table 2: Regression accuracy of the Merged Model for predicting the (x, y) location (within a 256×256 image) of where each person in each classroom image is looking. Accuracy is compared to human annotators and three baseline models.

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

Table 3: Classification results on 8×8 grid of the Merged Model compared to several baselines.

	Cross Entropy Loss (Grid Output)	AUC for In/Out
Center Gaze (Center 4 cells)	15.8047	-
Uniform Gaze	4.1589	-
Center Gaussian	4.0561	-
Logistic Regression	3.9997	-
Face-to-Gaze	3.7511	0.5459
Merged Model	**3.5855**	**0.6223**

is a bit more than half the error of the Merged Model, indicating that the machine's accuracy still has much room for improvement.

For classifying whether the gazes end inside or outside the image, the Merged Model achieved an AUC of 0.62, whereas humans scored 0.70 on the same task. The relatively low human accuracy suggests that detecting whether a person is looking inside or outside the image is quite challenging in the classroom images.

Figure 3 shows qualitative results of some of the gaze predictions (represented by thick yellow arrows) by Merged Model. It can be seen that the model makes decent predictions on the general direction of gazes but sometimes misses the endpoints on salient objects in the scene. In Figure 3, three girls in the middle are looking at the man's hands but the gaze predictions end before the hand.

One notable fact is that the **Face-to-Gaze** model's performance is very similar to the Merged Model's performance. This suggests that our Merged Model is predicting gaze locations mainly by using the head pose and gaze pathway of the subject and less on the salient objects in the image. One possible explanation is that our dataset does not contain enough variety of classroom environments for the model to learn how to identify salient objects in classroom images.

5. EXPERIMENT II: WHO ARE THEY LOOKING AT?

We use the same neural network depicted in Figure 2 to predict *who* each person is looking at. This is especially useful in school classrooms, in which both students and teachers are often looking at other *people*, not just objects. Specifically, we use the *classification* approach to predict which of the $N \times N$ grid cells each person is gazing at. The face contained within that cell is then predicted to be target face of that person's gaze. We note that, depending on the grid size

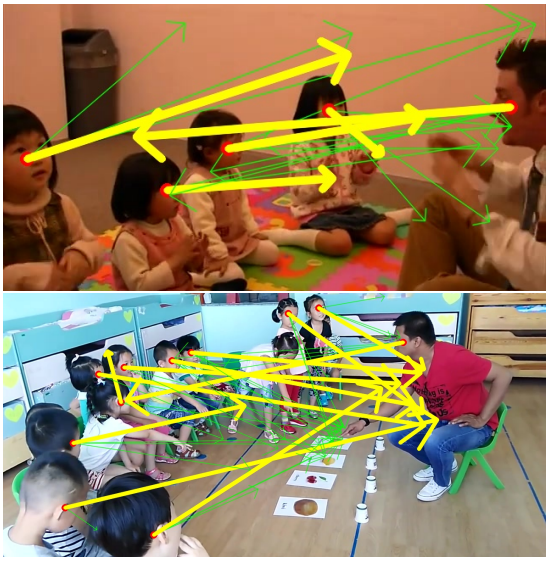

Figure 3: Qualitative results of gaze predictions by our Merged Model on the test set. Thin green arrows are ground truth annotations. Since there are multiple gaze annotations for each individual, there are multiple green arrows for each individual. Thick yellow arrows are predictions by Merged Model. Images (top to bottom) taken from: `https://goo.gl/xUdYbC`, `https://goo.gl/pcwQ5P`

and the specific image, multiple faces might appear within the same cell. A principled approach to handle to this issue would be to distribute the probability mass output by the neural network among all the faces within that cell in proportion to the size of each face. However, in this exploratory study, we simply assume that no grid cell contains more than 1 face.

5.1 Methodology

First, we computed the subset of all people in all image frames of our original YouTube dataset in which all annotators agreed that the person is looking at another *face* (not just another object somewhere in the image). Note that the labelers can still differ as to which particular face the person is looking at. By doing so, we obtained, 410 faces where all labelers agree that the person is looking at another face out of 17, 759 faces in our dataset. On the same data subset, we use the Merged Model to compute the softmax probabilities across all $N \times N$ grid cells of where each person was looking. From these probability outputs (for each person in each image), we remove every cell that does not contain any face (as determined by the face detector) and renormalize. We then choose the grid cell with the highest probability as the face that the person is most likely to be gazing at.

In order to evaluate how well our network is performing on determining which face a person is looking at, we took the top 1 face, top 2 faces, and top 3 faces. For the top-1 face, we choose the grid cell with the highest probability as the face that the person is most likely to be gazing at as predicted by the deep neural network. For top-2 and top-3 faces, if any of the top-2 and top-3 faces predicted by the network is the actual face which is agreed by the majority of human

labelers, the prediction is regarded as a correct prediction.

As baselines, we can consider that the average number of faces (detected by the face detector [14]) per image was 6.87 on test set; hence, the baseline guess rate is $1/6.87 \approx 0.15$ for the test set. Moreover, we can estimate human accuracy in a leave-one-labeler-out fashion: for each unique labeler, in the subset of the dataset where all labelers agree that a person being annotated is looking at another face, we compare the face that the current labeler chooses with the face which the majority of other labelers agree on. In this fashion, we compute the accuracy (% correct) of the l^{th} labeler w.r.t. the other $l-1$ labelers. We then average across all labelers in our dataset. By doing so, we achieve the human level performance on determining whom the person is looking at in the classroom given that the person is looking at *a face*.

In order to make equal comparison with Merged Model's predictions, which is done on 8×8 grid, human annotations are quantized to cells on 8×8 grid and probability of one labeler agreeing with the rest of the labelers that a person being annotated is looking at a *specific* face (last row of Table 4).

6. RESULTS II

The results on test images, shown in Table 4, indicate that the Merged Model can predict the face target of people's eye gazes with substantially higher accuracy than just randomly guessing among all grid cells (8×8 grid) in the image containing faces. To put these results in context: if each classroom image contains 6.87 faces on average (as reported above), then the probability of 0.79 for $k = 3$ suggests that an automated gaze following system can usually determine at least which *group* of students a teacher is looking at. Interestingly, the accuracy of the Merged Model is close to that of human labelers when top 3 predicted faces are considered but still have room for improvement when only top 1 face is chosen.

7. CONCLUSION AND FUTURE WORK

The results in this paper indicate that an automatic neural network, based on the approach by [27] that analyzes 2-D images of school classrooms can estimate the gaze target location of each person in the image with accuracy substantially higher than chance and better than several other baselines as well. Moreover, the same architecture can be used to identify *who* each person is looking at more accurately than random guessing.

Future work: The most critical next steps are to (1) improve accuracy by collecting more training data and improving the accuracy of the annotations. (2) Given an improved eye gaze following system, we can begin to explore how automatic gaze estimates can be used to predict specific aspects of classroom observation protocols; for instance, the *positive climate* dimension of the CLASS is based explicitly (in part) on whether the teacher looks at his/her students [25]. Finally, (3) since multiple people often look at the same person (e.g., the teacher) in school classrooms, we will also investigate whether accuracy can be improved by estimating the gaze targets of all classroom participants *jointly* rather than separately.

Table 4: Probability of the Merged Model correctly identifying which face a person is looking at on 8×8 grid.

Top k faces	$k=1$	$k=2$	$k=3$
Random Face	0.15	0.30	0.45
Merged Model	**0.47**	**0.65**	**0.79**
Human	0.82		

8. REFERENCES

[1] Aung, A. M., and Whitehill, J. R. Harnessing label uncertainty to improve modeling: An application to student engagement recognition. In *IEEE Automatic Face & Gesture Recognition* (2018).

[2] Baxter, J. A bayesian/information theoretic model of learning to learn via multiple task sampling. *Machine learning 28*, 1 (1997).

[3] Borji, A., Parks, D., and Itti, L. Complementary effects of gaze direction and early saliency in guiding fixations during free viewing. *Journal of vision 14*, 13 (2014).

[4] Bosch, N., D'Mello, S., Baker, R., Ocumpaugh, J., Shute, V., Ventura, M., Wang, L., and Zhao, W. Automatic detection of learning-centered affective states in the wild. In *International conference on intelligent user interfaces* (2015).

[5] Brooks, R., and Meltzoff, A. N. Gaze following: A mechanism for building social connections between infants and adults. *In Mechanisms of social connection: from brain to group* (2014).

[6] Caruana, R. Multitask learning: A knowledge-based source of inductive bias. In *International Conference on Machine Learning* (1993).

[7] Deng, T., Yang, K., Li, Y., and Yan, H. Where does the driver look? top-down-based saliency detection in a traffic driving environment. *IEEE Transactions on Intelligent Transportation Systems 17*, 7 (2016).

[8] D'Mello, S. K., Olney, A. M., Blanchard, N., Samei, B., Sun, X., Ward, B., and Kelly, S. Multimodal capture of teacher-student interactions for automated dialogic analysis in live classrooms. In *ACM international conference on multimodal interaction* (2015).

[9] Donnelly, P. J., Blanchard, N., Samei, B., Olney, A. M., Sun, X., Ward, B., Kelly, S., Nystrand, M., and D'Mello, S. K. Multi-sensor modeling of teacher instructional segments in live classrooms. In *ACM international conference on multimodal interaction* (2016).

[10] Emery, N. J. The eyes have it: the neuroethology, function and evolution of social gaze. *Neuroscience & Biobehavioral Reviews 24*, 6 (2000).

[11] Fathi, A., Hodgins, J. K., and Rehg, J. M. Social interactions: A first-person perspective. In *Computer Vision and Pattern Recognition* (2012).

[12] Grafsgaard, J., Wiggins, J. B., Boyer, K. E., Wiebe, E. N., and Lester, J. Automatically recognizing facial expression: Predicting engagement and frustration. In *Educational Data Mining* (2013).

[13] Hinton, G. Rmsprop: Divide the gradient by a running average of its recent magnitude.

[14] Jiang, H., and Learned-Miller, E. Face detection with the faster r-cnn. In *IEEE Automatic Face & Gesture Recognition* (2017).

[15] Judd, T., Ehinger, K., Durand, F., and Torralba, A. Learning to predict where humans look. In *International Conference on Computer Vision* (2009).

[16] Kane, T. J., McCaffrey, D. F., Miller, T., and Staiger, D. O. Have we identified effective teachers? validating measures of effective teaching using random assignment. In *Research Paper. MET Project. Bill & Melinda Gates Foundation* (2013).

[17] Kapoor, A., Burleson, W., and Picard, R. W. Automatic prediction of frustration. *International journal of human-computer studies 65*, 8 (2007).

[18] Kontos, S., and Wilcox-Herzog, A. Teachers' interactions with children: Why are they so important? research in review. *Young Children 52*, 2 (1997).

[19] Krafka, K., Khosla, A., Kellnhofer, P., Kannan, H., Bhandarkar, S., Matusik, W., and Torralba, A. Eye tracking for everyone. In *Computer Vision and Pattern Recognition* (2016).

[20] Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (2012).

[21] Marín-Jiménez, M. J., Zisserman, A., Eichner, M., and Ferrari, V. Detecting people looking at each other in videos. *International Journal of Computer Vision 106*, 3 (2014).

[22] Maron, O., and Lozano-Pérez, T. A framework for multiple-instance learning. In *Advances in neural information processing systems* (1998).

[23] Mashburn, A. J., Pianta, R. C., Hamre, B. K., Downer, J. T., Barbarin, O. A., Bryant, D., Burchinal, M., Early, D. M., and Howes, C. Measures of classroom quality in prekindergarten and children's development of academic, language, and social skills. *Child development 79*, 3 (2008).

[24] Mukherjee, S. S., and Robertson, N. M. Deep head pose: Gaze-direction estimation in multimodal video. *IEEE Transactions on Multimedia 17*, 11 (2015).

[25] Pianta, R. C., La Paro, K. M., and Hamre, B. K. *Classroom Assessment Scoring System^{TM}: Manual K-3*. Paul H Brookes Publishing, 2008.

[26] Qiao, Q., and Beling, P. A. Classroom video assessment and retrieval via multiple instance learning. In *International Conference on Artificial Intelligence in Education* (2011).

[27] Recasens, A., Khosla, A., Vondrick, C., and Torralba, A. Where are they looking? In *Advances in Neural Information Processing Systems* (2015).

[28] Recasens, A., Vondrick, C., Khosla, A., and Torralba, A. Following gaze in video. In *Computer Vision and Pattern Recognition* (2017).

[29] Simonyan, K., and Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

[30] Wang, Z., Pan, X., Miller, K. F., and Cortina, K. S. Automatic classification of activities in classroom discourse. *Computers & Education 78* (2014).

[31] Zain, N. H. M., Razak, F. H. A., Jaafar, A., and Zulkipli, M. F. Eye tracking in educational games environment: evaluating user interface design through eye tracking patterns. In *International Visual Informatics Conference* (2011).

[32] Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., and Torralba, A. Object detectors emerge in deep scene cnns. *arXiv preprint arXiv:1412.6856* (2014).