Predicting When Teachers Look at Their Students in 1-on-1 Tutoring Sessions

Han Jiang
Worcester Polytechnic Institute
hjiang@wpi.edu

Karmen Dykstra
Aalto University
kldykstra@gmail.com

Jacob Whitehill
Worcester Polytechnic Institute
jrwhitehill@wpi.edu

***Abstract*—We propose and evaluate a neural network architecture for predicting when human teachers shift their eye-gaze to look at their students during 1-on-1 math tutoring sessions. Such models may be useful when developing affect-sensitive intelligent tutoring systems (ITS) because they can function as an *attention model* that informs the ITS when the student's face, body posture, and other visual cues are most important to observe. Our approach combines both feed-forward (FF) and recurrent (LSTM) components for predicting gaze shifts based on the history of tutoring actions (e.g., request assistance from the teacher, pose a new problem to the student, give a hint, etc.), as well as the teacher's prior gaze events. Despite the challenging nature of the task – we are asking the network to *predict* whether or not the teacher will shift her/his eye gaze during the next one-second time interval – the network achieves an AUC (averaged over 2 teachers) of** 0.75**. In addition, we identify some of the factors that the human teachers in our study used when making gaze decisions and show evidence that the two teachers' gaze patterns share common characteristics.**

***Index Terms*—eye gaze prediction**

I. Introduction

Since the early 2000s [1], [2], [3], [4], one of the chief goals of the intelligent tutoring systems (ITS) community has been to develop *affect-sensitive* ITS that can perceive and respond to their students' affective states, e.g., frustration, boredom, and engagement. Due to tremendous, contemporaneous progress in machine learning and computer vision research, the accuracy of automatic detectors of emotions from images and video, both in general (e.g., basic emotions) and educational settings (e.g., detection of student engagement [5], [6]), has increased to the point that they are becoming practical. However, much less research has been done on how automatic affective sensor measurements should be integrated into the ITS' decision-making process.

One key question is: During *which specific moments* of the tutoring session are the students' emotions most important to perceive and respond to? While it is sometimes feasible simply to run an array of detectors on every frame of the videostream (captured from one or even multiple cameras), there are reasons why this is not a good idea: (1) **Computational cost**: as of 2017, the most accurate object detection and recognition systems (e.g., [7], [8], [9]), based on deep convolutional neural networks, are computationally very intensive, more so than "previous generation" detectors such as the classic Viola-Jones [10] approach. In order to maintain real-time responsiveness and low energy cost (particularly relevant for ITS on mobile devices), it may be preferable to sacrifice temporal resolution (i.e., run the detectors less frequently) in exchange for higher recognition accuracy. (2) **Redundancy**: there is a strong correlation between emotion estimates over time. (3) **Data overload**: Estimating a variety of facial expression and emotional states in every video frame can result in a huge amount of data that the ITS must somehow analyze and use to teach more effectively. The magnitude of this data may increase the challenge of training of downstream systems – e.g., a control system that uses "engagement" estimates to adjust the difficulty of the curriculum. It may instead make more sense to attend only to specific moments; indeed, the trend in recent deep-learning research on image- and video-based event recognition is to deploy *neural attention models* [11], [12] that automatically select *dynamically* which parts of an image or video are most salient, based on information contained in the image/video itself. In particular, if the salient moments (when full analysis of all sensors is necessary) can be determined using just a few less computationally expensive, lower-bandwidth sensor readings – e.g., audio rather than video, or low-resolution peripheral vision [13] rather than high-resolution direct gaze – then it is possible that significant computation could be saved.

978-1-5386-2335-0/18/$31.00 ©2018 IEEE

Human visual attention in one-on-one tutoring: Even in one-on-one tutoring settings, the teacher does not look at her/his pupil during the entire session. In contexts where the student and teacher share a common workspace – e.g., a piece of paper on which to write – the teacher divides her/his attention between the student, workspace, and other objects around the room. The choice of where the teacher decides to look is motivated by several factors, including: (1) **Privacy**: it would likely be uncomfortable for the student to be stared at the entire time; (2) **Information transmission**: From the psychology literature, there is evidence that increased eye gaze by the teacher is associated with more efficient encoding and subsequent recall of information [14], [15], [16] by the student. (3) **Information gathering**: The teacher looks at the student at moments that she/he judges to be most informative for making tutoring decisions. As an example of how these factors can influence visual attention, the teacher might generally avoid looking at the student (to maintain privacy) but decide to "check in" if, after asking her/him to tackle a math problem, the student pauses for a long time without giving any cue that she/he is trying to solve it. This can both help the teacher to know whether the student is confused (information gathering), and it may also cue the student that the teacher is waiting for a response (information transmission).

When developing an ITS that *selectively* perceives its students' emotions, it is necessary to develop an algorithm that decides *when* to look. One approach might be based on reinforcement learning. However, tutoring sessions are relatively expensive and slow to conduct compared to the robotics settings in which reinforcement learning is usually used, likely rendering it impractical. An alternative paradigm, which we pursue in this paper, is to train a model of visual attention using supervised learning from one-on-one tutoring sessions collected from *human* tutors. To the extent that skilled human tutors employ sensible visual attention strategies, this approach could help an affect-sensitive ITS to look at the student during the most important moments.

Human tutors may decide how to shift their eye-gaze based on the high-level actions of the tutoring session – e.g., the student has asked the teacher to help her/him in solving a problem – as well as visual cues such as hand gestures, facial expressions, etc. Tutors' visual attention may also exhibit temporal patterns, e.g., if the teacher just ascertained that the student was "engaged" one second ago, then it might not be necessary to check again during the next second. To date, there has been scant research on how tutors decide when to look at their students (see Related Work); one of the goals of our paper is to start to fill this gap. In one sentence: **the purpose of our work is to explore the extent to which machine learning can be used to predict human tutors' future eye-gaze events, using high-level actions, behavioral cues, as well as the history of prior eye-gaze events, as predictors.**

We emphasize that we are *not* trying to estimate the tutor's *current* eye-gaze (i.e., gaze following [17]) by examining an image of the tutor's face or eye region – this is an interesting and important problem but arguably easier (most human observers can solve this problem easily) than ours. Instead, we are trying to *predict* whether the tutor will *change* her/his eye-gaze during the next time-step. In particular, we assume that the teacher has knowledge of the *high-level actions* (defined in Section II-A) of the session (e.g., give an explanation, request assistance, attempt a problem, etc.); such actions could be obtained, for example, by analyzing the measurements from low-bandwidth (compared to full video) sensors such as speech. We also assume that the teacher knows the history of gaze events she/he has executed so far. Our research harnesses a tutoring video dataset (described below) of two teachers, each of whom tutors 10 middle-school students in a math topic (for a total of 20 unique students), which has been densely annotated for the teachers' (as well as the students') eye-gaze. The focus of our work is on modeling the decision process of human tutors, as well as exploring computational architectures for deciding when to look.

A. Related Work

There is a large body of literature [18], [19], [20] on visual saliency and attention prediction. While much of this research focuses on predicting where subjects will look within a single image, there has also been significant prior work on predicting gaze *shifts* in interactive settings, e.g. an airplane flight simulator [21], multi-party conversations [22], and urban driving [23]. To date, there have only been a few studies on visual saliency within *educational* settings: Penaloza, et al. [24] built a model of the *student*'s visual attention to enable a robot to more accurately emulate the cognitive development of infants. We are aware of only 2 prior studies that explicitly model how the *teacher* attends to the student. One is by Dykstra, et al. [25]: on a dataset of 1 teacher with 10 students, they developed a logistic regression-based model that predicts eye gaze shift events (similar to our work) based on the joint actions taken by the tutor and student in one-on-one tutoring sessions. The other is a behavioral study by van den Bogert, et al.[26], who compare expert versus novice teacher's eye-gaze in traditional classrooms (not tutoring sessions).

II. SDMATH Dataset

The San Diego Multimodal Adaptive Tutoring Human-to-human (which we call SDMATH) dataset consists of labeled video recordings of 20 one-on-one tutoring sessions. There are 2 tutors in the dataset, one female, one male, both of whom are accredited middle-school math teachers. Each tutor taught 10 students (5 male, 5 female each; no student was taught by both teachers), who were all 8th grade students of 13 years of age. There were 20 unique students in total. Before participating in the tutoring session, both the teachers and the students (and parents) gave informed consent/assent to participate, be videorecorded, and have their face images published in scientific publications (University of California, San Diego's IRB: 090920).

All sessions were captured using both frontal camera to capture student and teacher and an overhead camera to capture the scratch paper which both participants shared as a common workspace (see Fig. 4, right). Each tutoring session was approximately one hour in duration and consisted of a 10-minute pretest, 40-minute tutoring session, and finally a 10-minute posttest. The teachers were instructed to teach naturally in order to help each student to practice and learn the material as effectively as possible. The students were instructed simply to do the best they could. The "fundamentals of logarithms" were chosen as the topic of instruction. Logarithms were selected since they were expected to be challenging for the students (since they are typically taught to students in higher grade-levels than the participants in our study) but still learnable to significant degree within a 40-minute tutoring session.

A. Annotation

The SDMATH dataset was annotated for multiple channels (see Figure 1 for a schematic):

Actions: Based both on the teachers' and students' speech, head nods and shakes, as well as the content of what they

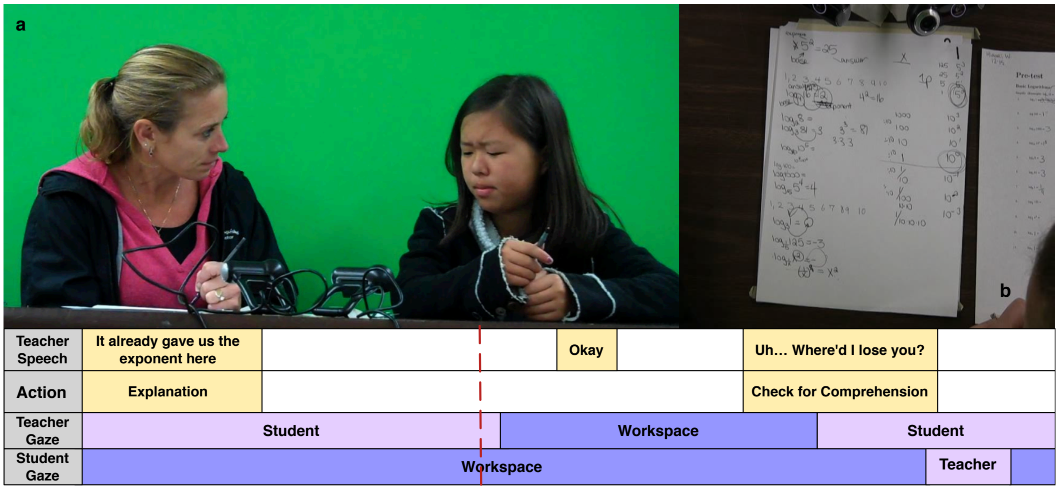

Fig. 1: A moment from the SDMATH dataset showing both (a) frontal and (b) overhead views. The labels underneath show the teacher's utterances, the corresponding teacher speech action labels, and teacher and student eye gaze labels. The dashed red line indicates the moment at which the image was extracted from the video.

wrote on the paper, each tutoring session was coded for the *actions* that were taken by each participant at each moment in time. There were 13 possible labels for the teachers' actions (explanation, present problem, solicit content, solicit explanation, solicit procedure, request for participation, provide hint, check for comprehension, direct negation, indirect negation, confirmation, encouragement, and socializing) and 7 for the student (correct attempt, incorrect attempt, incomplete attempt, request assistance, express lack of comprehension, socializing).

Gesture: Hand gestures were coded separately for the left hand and right hand of both the teacher and the student, for all 20 tutoring sessions. Hand gestures were labeled as one of four types (see [27]): *Deictic* (pointing) gestures are used to direct a listener's attention to a referent (e.g., writing on the paper). *Beat* gestures are small hand movements resembling flicks and occur with the rhythm of the speech, mostly placed on stressed syllables. *Iconic* gestures exhibit physical aspects of the scene described by speech. *Metaphoric* gestures are associated with abstract ideas and represent a metaphor of the speaker's idea or feeling about an object or concept.

Eye Gaze: The object of fixation of student and teacher eye gaze was labeled throughout each tutoring session. Distinctions were made between three mutually exclusive gaze fixations: (1) the paper workspace shared by the teacher and student, (2) the other tutoring session participant (teacher or student depending on the subject of labeling), and (3) elsewhere, defined as all eye gaze which does not fall into one of the first two categories. The median (over all 10 sessions per teacher) fractions of time that the teachers gazed at their students was 6% and 26% for Teachers 1 and 2, respectively.

III. Proposed Eye-Gaze Prediction Model

We developed a neural network (see Figure 2) to predict the binary outcome of *whether the teacher shifts her/his eye-gaze to look at the student during the next time-step*, based on the history of the student's and teacher's actions (e.g., hand gestures) as well as the prior eye-gaze events of both the student and teacher. In order to capture the *entire* history, we use an LSTM recurrent neural network (see Figure 3): the input $[x_t; f_t]$ consists of the *current* eye-gaze x_t at time t, along with the feature vector f_t describing the teacher's and student's actions; the output is the prediction $\hat{x}^{RNN}_{t+1}$ of what the teacher's eye-gaze x_{t+1} (at time $t+1$) will be, over all 3 eye-gaze targets (paper, student, elsewhere).

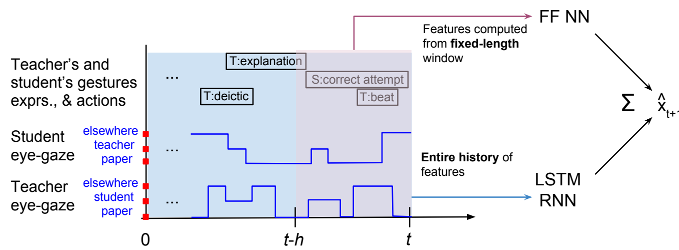

Fig. 2: Proposed neural network, consisting of both feed-forward (FF) and LSTM components, for predicting whether or not the teacher shifts her/his eye-gaze from {"paper","elsewhere"} to "student", at time $t+1$. The FF network analyzes features computed from a fixed-length window in the pink block; the LSTM analyzes the entire history of the teacher's prior eye-gaze events. The final prediction is the combination of the probabilities of FF NN and LSTM RNN.

In addition, since simple feed-forward (FF) neural networks are often easier to train (compared to LSTM) without overfitting, we also use a two-layer FF network to analyze the same set of features (student's and teacher's actions) from the *recent* history over a fixed time-window $[t-h, t]$. The output of the network is a softmax over 2 categories (shift to student, do not shift to student). This is equivalent to logistic regression

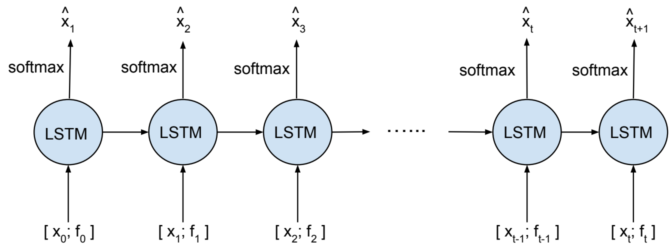

Fig. 3: LSTM subnetwork we used for eye-gaze prediction. During training, the target value at each timestep t is the ground-truth value of the next timestep $t+1$.

and is equivalent to the approach used by [25] (though with a different feature set).

The final prediction of the network is the average of the two networks' predictions ($\hat{x}^{FF}_{t+1}, \hat{x}^{RNN}_{t+1}$).

A. Training

FF: We used as positive examples every time-point at which the teacher shifted her/his eye-gaze from *not* looking at the student (i.e., looking either at the paper or "elsewhere"), to looking at the student. A set of negative examples was created by sampling random timepoints when the teacher was likewise *not* looking at the student and also *did not immediately shift* her/his gaze to the student, subject to the constraint that every such negative example was at least 1 second before the onset and 1 second after the end of every time period during which the teacher gazed at the student. Based on this procedure, there were a total (over all 20 tutoring sessions) of 1836 and 3292 positive examples, and 3652 and 6584 negative examples, for Teacher 1 and Teacher 2, respectively. The value of h was optimized for each teacher to maximize prediction accuracy; this resulted in $h = 0.3$sec for Teacher 1 and $h = 0.2$sec for Teacher 2. The weights of the FF network were also regularized with a ridge term of strength 0.001.

LSTM: In SDMATH, eye-gaze labels are annotated using a *real-valued* clock (e.g., the teacher shifts her/his gaze at time 3.25sec from the paper to "elsewhere"). However, the LSTM recurrent neural network in our design uses a *discrete* clock (each t corresponds to 1 second of wall-clock time). When training the LSTM, we thus set the ground-truth label x_{t+1} that the network is trying to predict at time t to be the proportion of time, within the time interval $[t, t+1)$, that the teacher gazed at each of the 3 targets. At test time, the outputs $\hat{x}^{RNN}_{t+1}$ were converted (to match the format of $\hat{x}^{FF}_{t+1}$) into a probability vector over just 2 categories by summing the probabilities of "paper" and "elsewhere"; the result was then added to $\hat{x}^{FF}_{t+1}$ to produce the network's final eye-gaze estimate of whether or not the teacher gazes at the student. We trained the LSTM using the Adam optimizer (learning rate was 0.01) over 40 epochs. To optimize the number of hidden units in the LSTM layer (over the set $\{2,4,8,16,32\}$), we used subject-independent double cross-validation; the optimal number was 16.

Teachers' eye-gaze prediction accuracy (AUC)

	FF	LSTM	Combination
Teacher 1	0.77	0.76	**0.79**
Teacher 2	0.68	0.67	**0.70**

TABLE I: Eye-gaze prediction performance on the SDMATH dataset, using either the FF, LSTM, or combined networks. Results for each teacher are averaged over his/her 10 students.

IV. Results

We used SDMATH to estimate the accuracy of the network described above, for each teacher separately, using leave-one-session-out cross-validation. We measured accuracy separately for the FF and LSTM components, as well as of the overall network (combined predictions). To enable a fair comparison between the FF (real-valued clock) and LSTM (discrete clock) approaches, we tested the network at all timepoints t such that the time interval $[t, t+1)$ contained one of the positive or negative examples used for training+evaluating the FF network. Accuracy was measured using the Area Under the receiver operating characteristics Curve (AUC). Recall that the AUC of a classifier that guesses is 0.5, no matter what the prior class probabilities are.

A. *Results: Predicting* teachers*' eye-gaze shifts*

Results (averaged over all 10 students of each teacher) are shown in Table I. The FF network was more accurate than the LSTM network, suggesting that – possibly due to the simplicity of the 2-layer FF network architecture – the short-term history of students' and teachers' actions is more easily capturable using the FF approach than the LSTM approach. However, we did observe evidence that the long-term history of events, as captured by the LSTM, can be helpful: the combined network (FF+LSTM) was statistically significant more accurate (0.79 versus 0.77 AUC for teacher 1, $t(9) = 3.949, p = 0.0036$; 0.70 versus 0.68 AUC for teacher 2, $t(9) = 2.4512, p = 0.03668$) than just the FF network by itself (i.e., the approach used in [25]), suggesting that long temporal windows can be useful for modeling human eye gaze and developing attention models for ITS. Using the combined network, the average AUC over both teachers was 0.75. Clearly, this would not be a high value for an *object recognition* problem such as *gaze following*. However, our problem is about *prediction* and is arguably more challenging.

B. *Results: Predicting* students*' eye-gaze shifts*

In addition to modeling *teachers*' eye-gaze, we also "reverse" the prediction problem and train models to predict when the *student* shifts her/his gaze to the teacher. This allows us to train predictive models for not just 2 teachers but also on 20 students, and to gain greater confidence in the ability of our model to generalize to new subjects. Using just the LSTM network (not the FF component, for simplicity), and using the same subject-independent cross-validation scheme (separately for each teacher), we trained predictive models of a student not

Fig. 4: **Left,middle**: Teacher 2 before/after shifting eye gaze to student. **Right**: Teacher 2's deictic hand gesture (pointing to an equation on the paper) before shifting eye gaze.

seen during training. The AUC for predicting students' eye-gaze, averaged over all 10 students of teacher 1, was 0.83; the average AUC over all 10 students of teacher 2 was 0.80. These numbers are consistent with the accuracies of predicting teachers' eye-gaze.

V. Identifying the most predictive features

What particular semantic and behavioral features did the teachers in SDMATH respond to when making decisions of where to look? To answer this question, we trained the FF neural network we used sequential additive logistic regression (similar to the FF network described above): For each teacher, we started with an empty feature set and iteratively added the feature (from the pool of 83 features) that maximized the increase in training accuracy, conditional on the already selected features. Selection was repeated for 10 iterations.

Results: The top 10 most predictive features of gaze-to-student events are shown in the tables below, along with the associated logistic regression coefficient:

Teacher 1

#	Person	Feature	Coef.	Cumulative AUC
1	**Teacher**	**deictic gesture (left)**	$+.26$	0.6231
2	**Teacher**	**explanation**	$+.24$	0.6745
3	**Teacher**	**prompting**	$+.11$	0.6917
4	**Teacher**	**check for comprehension**	$+.14$	0.7113
5	**Teacher**	**beat gesture (left)**	$+.13$	0.7194
6	Teacher	iconic gesture (left)	$+.12$	0.7256
7	**Teacher**	**present problem**	$-.11$	0.7320
8	**Teacher**	**iconic gesture (both)**	$+.11$	0.7369
9	Teacher	deictic gesture (both)	$+.10$	0.7430
10	Student	correct attempt	$+.07$	0.7471

Teacher 2

#	Person	Feature	Coef.	Cumulative AUC
1	**Teacher**	**present problem**	$-.25$	0.5739
2	**Teacher**	**explanation**	$+.13$	0.6025
3	**Teacher**	**prompting**	$+.17$	0.6318
4	Teacher	request for participation	$-.08$	0.6398
5	**Teacher**	**check for comprehension**	$+.12$	0.6473
6	**Teacher**	**beat gesture (left)**	$+.13$	0.6543
7	Student	eye gaze to paper	$-.05$	0.6600
8	**Teacher**	**deictic gesture (left)**	$+.10$	0.6646
9	**Teacher**	**iconic gesture (both)**	$+.14$	0.6694
10	Teacher	beat gesture (both)	$+.08$	0.6739

Seven out of the 10 features (shown in bold) overlap for the two teachers. The last column shows, for each selected feature, the cumulative accuracy on *training* data. Over both teachers, most of the top 10 features were positively correlated with gaze-to-student, meaning the presence of the feature increased the probability of the teacher shifting his/her gaze to the student. For example, the teachers were more likely to shift their gaze to the student after having started an *explanation*; this is intuitive since the teacher would likely want to sense the student's reaction to what he/she is saying. Similarly, there is a increased probability of gaze-to-student when the teacher *prompts* the student to answer a question, possibly because the teacher is now waiting for the student to deliver a response.

More interesting is that *deictic hand gestures* were positively correlated with the teacher shifting his/her eye gaze to the student. In Figure 4, Teacher 2 is shown just before and just after she shifts her eye gaze from the paper to the student, along with the overhead view of the paper just before she shifts her gaze. At this moment, the teacher is making a deictic gesture with her left hand to point to the number 10 on the paper. One interpretation is that the teacher needs to gaze at the student to ascertain whether the student is attending to where the teacher had pointed. This suggests that it may be beneficial for an ITS, when pointing out a particular mistake that the student had made in a math derivation, to verify that the student is in fact attending to the tutor's explanation.

VI. Conclusion

We have proposed a neural network, combining both LSTM and FF components, for predicting whether the teacher in one-on-one tutoring sessions will shift her/his eye gaze to look at the student during the next timestep. This is a challenging problem that requires the model to predict future human behavior. The model was trained and evaluated on a dataset of 20 one-on-one math tutoring sessions from 2 human teachers and exhibited an overall accuracy (averaged over the two teachers) of 0.75 – this corresponds to a reduction in prediction error of about 50% (relative to the baseline guess AUC of 0.5). The accuracy of the overall neural network, comprising both an FF and LSTM component, was statistically significantly higher than just the FF subnetwork, suggesting that long-range temporal dependencies can be useful to capture for predicting eye-gaze events. In addition, we have identified particular high-level semantic actions and behavioral features that the teachers (implicitly) used to make their visual attention decisions. In **future work** it would be interesting to integrate into an affect-sensitive ITS the kind of neural attention model we have developed, and to explore what level of attention prediction accuracy is necessary for the ITS to teach effectively.

Acknowledgement

This material is based upon work supported by the National Science Foundation under Grant No. #1551594.

References

[1] B. Kort and R. Reilly, "An affective module for an intelligent tutoring system," in *Intelligent Tutoring Systems*. Springer, 2002, pp. 955–962.

[2] B. Kort, R. Reilly, and R. W. Picard, "An affective model of interplay between emotions and learning: Reengineering educational pedagogy-building a learning companion," in *2001 Proc. IEEE Int. Conf. Advanced Learning Technologies*, 2001, pp. 43–46.

[3] S. D'Mello, T. Jackson, S. Craig, B. Morgan, P. Chipman, H. White, N. Person, B. Kort, R. el Kaliouby, R. Picard *et al.*, "Autotutor detects and responds to learners affective and cognitive states," in *Proc. Emotional and Cognitive Issues Workshop at Int. Conf. Intelligent Tutoring Systems*, 2008.

[4] B. Woolf, W. Burleson, I. Arroyo, T. Dragon, D. Cooper, and R. Picard, "Affect-aware tutors: recognising and responding to student affect," *Int. Journal of Learning Technology*, vol. 4, no. 3, pp. 129–164, 2009.

[5] J. Whitehill, Z. Serpell, Y.-C. Lin, A. Foster, and J. R. Movellan, "The faces of engagement: Automatic recognition of student engagementfrom facial expressions," *Affective Computing, IEEE Transactions on*, vol. 5, no. 1, pp. 86–98, 2014.

[6] N. Bosch, S. K. D'mello, J. Ocumpaugh, R. S. Baker, and V. Shute, "Using video to automatically detect learner affect in computer-enabled classrooms," *ACM Transactions on Interactive Intelligent Systems (TiiS)*, vol. 6, no. 2, p. 17, 2016.

[7] H. Jiang and E. Learned-Miller, "Face detection with the faster r-cnn," in *Automatic Face & Gesture Recognition (FG 2017), 2017 12th IEEE International Conference on*. IEEE, 2017, pp. 650–657.

[8] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 6, pp. 1137–1149, 2017.

[9] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *European conference on computer vision*. Springer, 2016, pp. 21–37.

[10] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, vol. 1. IEEE, 2001, pp. I–I.

[11] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *International Conference on Machine Learning*, 2015, pp. 2048–2057.

[12] V. Ramanathan, J. Huang, S. Abu-El-Haija, A. Gorban, K. Murphy, and L. Fei-Fei, "Detecting events and key actors in multi-person videos," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 3043–3053.

[13] S. Gould, J. Arfvidsson, A. Kaehler, B. Sapp, M. Messner, G. R. Bradski, P. Baumstarck, S. Chung, A. Y. Ng *et al.*, "Peripheral-foveal vision for real-time object recognition and tracking in video." in *IJCAI*, vol. 7, 2007, pp. 2115–2121.

[14] J. P. Otteson and C. R. Otteson, "Effect of teacher's gaze on children's story recall," *Perceptual and Motor Skills*, vol. 50, no. 1, pp. 35–42, 1980.

[15] R. Fry and G. F. Smith, "The effects of feedback and eye contact on performance of a digit-coding task," *The Journal of Social Psychology*, vol. 96, no. 1, pp. 145–146, 1975.

[16] J. V. Sherwood, "Facilitative effects of gaze upon learning," *Perceptual and Motor Skills*, vol. 64, no. 3c, pp. 1275–1278, 1987.

[17] A. Recasens, A. Khosla, C. Vondrick, and A. Torralba, "Where are they looking?" in *Advances in Neural Information Processing Systems*, 2015, pp. 199–207.

[18] A. Borji, H. R. Tavakoli, D. N. Sihite, and L. Itti, "Analysis of scores, datasets, and models in visual saliency prediction," in *Proceedings of the IEEE international conference on computer vision*, 2013, pp. 921–928.

[19] A. Borji and L. Itti, "State-of-the-art in visual attention modeling," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 1, pp. 185–207, 2013.

[20] S. Frintrop, E. Rome, and H. I. Christensen, "Computational visual attention systems and their cognitive foundations: A survey," *ACM Transactions on Applied Perception (TAP)*, vol. 7, no. 1, p. 6, 2010.

[21] S. M. Doane and Y. W. Sohn, "Adapt: A predictive cognitive model of user visual attention and action planning," *User Modeling and User-Adapted Interaction*, vol. 10, no. 1, pp. 1–45, 2000.

[22] E. Gu and N. Badler, "Visual attention and eye gaze during multiparty conversations with distractions," in *Intelligent Virtual Agents*. Springer, 2006, pp. 193–204.

[23] A. Borji, D. N. Sihite, and L. Itti, "What/where to look next? modeling top-down visual attention in complex interactive environments," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 44, no. 5, pp. 523–538, 2014.

[24] C. I. Penaloza, Y. Mae, K. Ohara, and T. Arai, "Using depth to increase robot visual attention accuracy during tutoring," in *IEEE International Conference on Humanoid Robots - Workshop of Developmental Robotics*, 2012.

[25] K. Dykstra, J. Whitehill, L. Salamanca, M. Lee, A. Carini, J. Reilly, and M. Bartlett, "Modeling one-on-one tutoring sessions," in *2012 Proc. IEEE Int. Conf. Development and Learning and Epigenetic Robotics*. IEEE, 2012, pp. 1–2.

[26] N. van den Bogert, J. van Bruggen, D. Kostons, and W. Jochems, "First steps into understanding teachers' visual perception of classroom events," *Teaching and Teacher Education*, vol. 37, pp. 208–216, 2014.

[27] D. McNeill, *Hand and mind: What gestures reveal about thought*. University of Chicago Press, 1992.