Student Engagement Dataset

Kevin Delgado
Boston University
kevry@bu.edu

Juan Manuel Origgi*
Boston University
jmoriggi@bu.edu

Tania Hasanpoor*
Boston University
taniahsp@bu.edu

Hao Yu
Boston University
haoyu@bu.edu

Danielle Allessio
University of Massachusetts Amherst
allessio@umass.edu

Ivon Arroyo
University of Massachusetts Amherst
ivon@cs.umass.edu

William Lee
University of Massachusetts Amherst
williamlee@cs.umass.edu

Margrit Betke
Boston University
betke@bu.edu

Beverly Woolf
University of Massachusetts Amherst
bev@umass.edu

Sarah Adel Bargal
Boston University
sbargal@bu.edu

Abstract

A major challenge for online learning is the inability of systems to support student emotion and to maintain student engagement. In response to this challenge, computer vision has become an embedded feature in some instructional applications. In this paper, we propose a video dataset of college students solving math problems on the educational platform MathSpring.org with a front facing camera collecting visual feedback of student gestures. The video dataset is annotated to indicate whether students' attention at specific frames is engaged or wandering. In addition, we train baselines for a computer vision module that determines the extent of student engagement during remote learning. Baselines include state-of-the-art deep learning image classifiers and traditional conditional and logistic regression for head pose estimation. We then incorporate a gaze baseline into the MathSpring learning platform, and we are evaluating its performance with the currently implemented approach.

1. Introduction

Online learning can be boring, impersonal and non-interactive. Currently few online systems account for the context-sensitive nature of learning, *i.e.*, motivation, social and emotional learning, and climate as well as com-

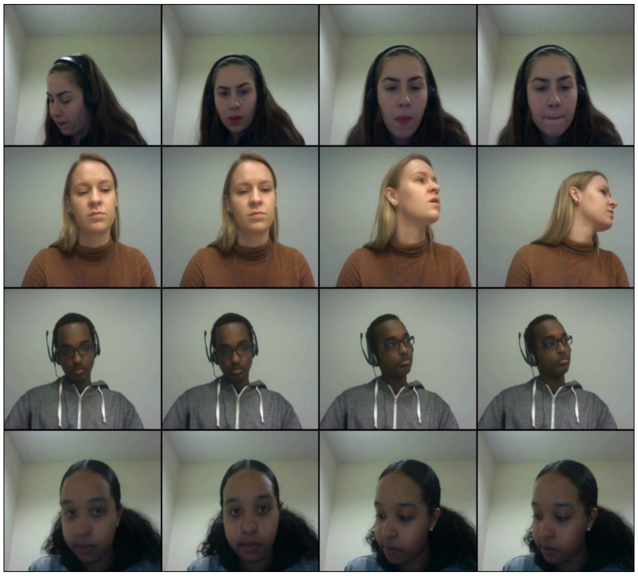


Figure 1. **Student Engagement Dataset.** Examples of video frames from our proposed Student Engagement Dataset for four sample students. The videos in the dataset capture student faces and gestures as students solve math problems. Video frames are then annotated to indicate whether a student is engaged ('looking at the screen' or 'looking at their paper') or wandering. Annotations are then used to train computer vision models that automatically detect wandering. A computer vision model is then integrated into MathSpring, an online tutor system, in order to trigger interventions whenever students are predicted to be wandering.

[*] Equal contribution.

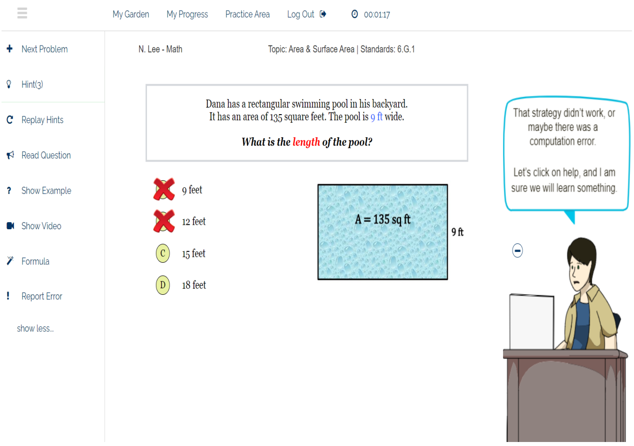
plex interactions among these factors [15, 26]. Student engagement and emotion are tightly correlated with learning gains [10, 13]; emotion drives attention and attention drives learning [19]. Establishing and maintaining student engagement is critical, especially as students increasingly move online for their education. A major challenge for online learning is the inability of systems to support students' emotion nor to maintain student engagement. When students are placed in such environments, many external distractions can lead to disengagement and wandering. These distractions lead to a decline in the learning process.

In response to this challenge, computer vision has become an embedded feature in some instructional applications. One goal of online intelligent instruction is to build systems that analyze student behavior in the wild and contribute to trust and understanding between teacher/students and computers. This paper describes development and early evaluation of technology that monitors student engagement in real-time, detects waning attention and distraction, and assesses which interventions lead to more productive learning. The system detects student disengagement through recognition of head orientation and gaze expression.

In order to develop such computer vision modules, they need to be trained on benchmark datasets that are curated to help machines solve such tasks. We collect videos of consenting students solving math problems on MathSpring [1], a game-like intelligent math tutor that offers a more personalized approach to online learning, as well as to track and respond to student performance and engagement. We then pre-process and crowdsource label frames of the videos to propose a publicly available dataset that aids research in automated student engagement prediction. Figure 1 presents sample video frames from our Student Engagement Dataset. The dataset will be made publicly available. We note that the dataset only includes individuals who have provided written consent that their data may be used publicly for research purposes.

Deep learning has become state-of-the-art in many computer vision applications. Using our proposed dataset, collected from the University of Massachusetts Amherst, we developed a computer vision model that uses state-of-the-art technology to detect student engagement and compare its performance to traditional baselines. The developed models predict whether a student is 'looking at their paper', 'looking at their screen', or 'wandering' at any point in time.

We then incorporate one of our baselines in the MathSpring tutor. Figure 2 exhibits an example problem presented to students on MathSpring. The tutor targets sensing and interpreting facial signals relevant to student emotion, and provides students with real-time interventions that can aid their progress, suggesting when and who needs further assistance, and identifying which interventions are working. The system is used in real classrooms, interacts in a "human-centered" and engaging manner, and serves as digital assistants to students. We use the implemented computer vision module to alert wandering students to re-gain their attention.



Dana has a rectangular swimming pool in his backyard.
It has an area of 135 square feet. The pool is 9 ft wide.

What is the length of the pool?

9 feet
12 feet
C 15 feet
D 18 feet

That strategy didn't work, or maybe there was a computation error.

Let's click on help, and I am sure we will learn something.

Figure 2. **MathSpring Problem Example.** This figure presents a sample MathSpring problem. Hints, worked-out examples, tutorial videos, and formulas are available from the corresponding buttons on the left. A learning companion (right) talks to students and guides them when they make mistakes.

2. Related Work

In this section, we survey the datasets and computer vision approaches that have addressed engagement in online learning.

Datasets. Most datasets that involve facial expressions and head orientation, are not integrated with an online learning environment. Examples include the *Affective-MIT Facial Expression* [25] and *Aff-Wild* [23, 22, 35] datasets. The *Affective-MIT Facial Expression Dataset (AM-FED)* dataset contains frames from 242 videos of people watching commercials using their front-facing webcam. This dataset has frame-by-frame annotation however, there is not much variation visible in head pose. The *Aff-Wild* dataset is an in-the-wild dataset consisting of 500 videos taken from YouTube that exhibits emotions while watching videos, reacting to comedy and performing activities. Frame-by-frame annotation was also completed for valence and arousal.

Publicly available datasets that do consider a student environment include the *DAiSEE* dataset [16] and the *Kaur* dataset [21]. The *DAiSEE dataset* contains video recordings of students in an online learning environment. To imitate such environment, subjects were presented with two separate videos, one educational and one recreational to capture focused and relaxed settings. This allowed variation in user engagement levels. Frame-by-frame annotation was accomplished by crowdsource labels for frustration, engagement, confusion, boredom. The *Kaur* dataset contains 195 videos from 78 subjects in a simulated learning environment. These subjects were presented with purely educational videos, *Learn the Korean Language in 5 minutes*, *a pictorial video (Tips to learn faster)* and *(How to write a research paper)*. Each video was given a single label for the overall level of engagement. In contrast, our proposed dataset has been collected in a real online learning environment, has been labeled for engagement at frequently sampled points in time within each video, models students math skills, provides adaptive problems based on student performance, moves from topic to topic, and is suited for engagement and wandering detection.

Student Engagement Systems. Many studies have focused on different levels of student engagement, and defined the levels differently [12, 29]. During this study [12], three ways of engagement detection were proposed based on how involved the student was in reporting their involvement. The methods were manual, semi-automatic, and automatic. These methods are further broken down based on their data type such video, audio, text, etc. The automatic category uses computer vision techniques on facial expressions to detect engagement by an external observer. They used part-based methods for the face that focus on different parts of the face, and appearance-based methods which focuses on the face as a whole. They used posture and gesture techniques, and eye movement techniques as well, which is a popular method to detect engagement. Sharma *et al*. [29] focused on levels of engagement or concentration indexes as very engaged, nominally engaged, and not engaged at all. This work demonstrated that students with higher test scores, have higher concentration indexes.

Head pose, eye gaze, and facial expressions are mostly common used modalities in engagement detection techniques based on computer vision. Researchers believe that since humans are able to detect engagement using the cues above, machines can do it as well [31]. Altuwairqi *et al*. [4] focused on engagement detection using students' behavior such as their mouse movement, keyboard keystroke, in addition to facial emotions. Huang *et al*. [17] combined all the different features discussed above such as eye gaze direction, head pose and eye coordinates and achieved an accuracy that was higher than each one of them individually. To build an end-to-end system, Abedi and Khan [2] approached the problem as a spatio-temporal classification problem on the *DAiSEE* dataset using Residual Network (ResNet) and Temporal Convolutional Network (TCN) to detect the engagement level of students in videos. TCN was used to analyze the video frames and observe the temporal changes in them in order to detect student's engagement level.

One of the classifications we consider in this study is attention wandering. Eye tracking is one highly used method to detect wandering in terms of engagement for Massive Open Online Course (MOOCs) [18]. This work is focused on consumer-grade eye tracking, which requires a laboratory setting. Other studies that focus on the MOOC aspect of engagement detection include [18, 27, 28]. For example, a study in 2018 [28] hypothesizes that students are not aware of their 'learning behavior' while studying and that leads to most students not completing MOOCs successfully. This work uses eye and face trackers to detect whether a face is present or absent in order to detect whether the student is engaged or they have lost focus.

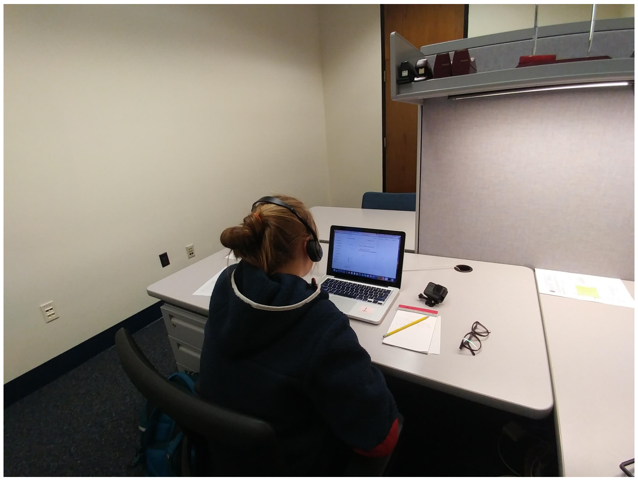


Figure 3. **Data Collection Setup.** This Figure presents the lab environment setup used to collect the data. Students worked on their math problems on a laptop with a webcam. Students had a notebook and pencil available, initially placed to the right of the laptop. The webcam captured their face as demonstrated in Figure 1.

Our approach differs from related work in that it: 1) explores different deep convolutional networks to predict student engagement using a single holistic visual cue; 2) uses crowdsourcing to identify what students are paying attention to, rather than only identifying student distraction; and 3) integrates video frame classification into an existing learning platform to re-orient students' attention towards learning. We are interested in knowing whether students are paying attention (looking at the screen or at the paper), or whether their attention is wandering (looking away).

3. Dataset

In this section we discuss the collection and annotation process for our dataset of video recordings of students solving math problems. We also discuss distribution of data samples among classes, and how training and testing splits are created.

Data Collection. Nineteen students participated in our dataset and each student was video recorded while solving math problems on MathSpring. The setup created for stu-

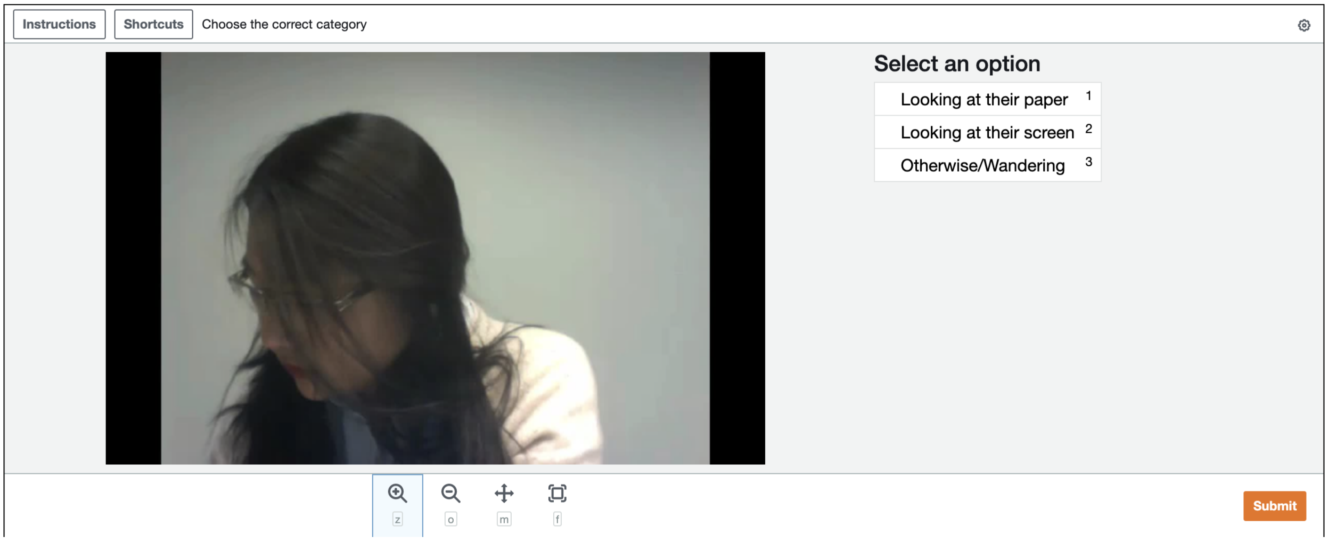


Figure 4. **Sample Annotation Sub-task.** This Figure demonstrates the interface presented to MTurk workers for labeling a video frame of a student working on MathSpring. The labels include: 1) Looking at their paper, 2) Looking at their screen, or 3) Otherwise/Wandering. Each crowd-worker is presented with ten such frames randomly selecting from sessions of different subjects.

dents to conduct their sessions is presented in Figure 3. The platform adapted problems for students, *i.e.*, less/more difficult problems were provided based on a model of student skills. Each student was provided with a piece of paper to the right of their computer to take notes. Videos were recorded using a front-facing webcam.

Dataset Details. All 19 students consented to have their data publicly available for research. Each student solved a different set of math problems. The intelligent tutor adapted problems based on student skill level. A single video corresponds to a single problem solved by a single student. We collected 400 videos from the 19 participants. For annotations, we sampled video frames at one frame-per-second (FPS) for annotation, resulting in a total of 18,721 frames. To reduce research bias, researchers were restricted from labeling the frames, and instead, an external process for annotating each frame was conducted, as described next.

Annotation Tool. We utilized Amazon Mechanical Turk (MTurk); a crowdsourcing marketplace for individuals to outsource their tasks to a distributed workforce, allowing them to perform jobs virtually. We specified in our task settings that only MTurk workers who had previously completed at least 1,000 tasks (a.k.a HITs) are eligible to take on our posted tasks. We also compensated the work of all crowd workers who participated in our labeling tasks.

Crowdsourcing Task. Crowd workers on MTurk were tasked to label a frame with one of the following labels: ‘looking at their paper’, ‘looking at their screen’, or ‘wandering’. Figure 4 presents the annotation interface with one sub-task, *i.e.* frame to be labeled, with which crowdsource workers were presented. Each sub-task had a total of three labels assigned by three different crowdworkers. Every one of the 18,721 frames was assigned to three different crowdworkers. A task consisted of ten sub-tasks, and was allotted a maximum of one hour to complete. Crowdworkers were paid $0.10 per task/assignment.

Crowdsourcing Task Results. We processed 56,163 (18,721 images * 3 votes per image) crowdsourced results to assign a winning label, ‘looking at paper’ or ‘looking at screen’, ‘wandering’ for each frame. It took each crowdworker an average of 46 seconds to complete a single task. We used majority voting to combine the three crowd-collected selections into a single vote for each frame.

Dataset Splits. The dataset consists of Ĩ9K frames for three classes and 19 different students. The resulting class split, Table 1 (a), is very imbalanced: the ‘screen’ class counts 14 times more samples than the ‘wander’ class and three times more samples than the ‘paper’ class. By analyzing the distributions of the different samples for each class, we notice that the ’paper’ and ’screen’ classes contain a large number of similar frames. We create a second smaller version of the original dataset by removing the similar samples for each class and balance the dataset. After selecting and removing the similar frames, we obtain a more equally distributed dataset, Table 1 (b), consisting of around 2,000 frame samples. Finally, we split the balanced dataset into a training and a testing set. 20% of the samples were selected for our test set, and the remaining for the training set. In order to test and train the model on samples coming from different students, we choose the test samples from only three of the original 19 students.

Time Series Data of Student Behavior. We developed

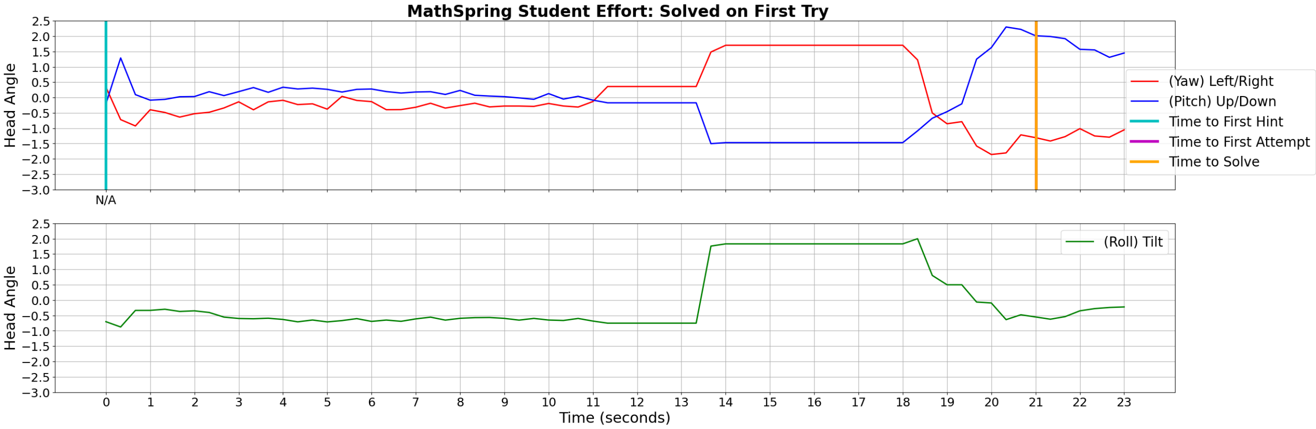


Figure 5. **Time Series Data of Student Behavior.** Students' head positions are graphed while they solve math problems. The horizontal axes reflect time to solve the problem and the vertical axis represents head position. The double graphs show the head position within the time frame of a single problem. The top chart displays two lines to indicate head position, left/right (red, yaw) and up/down (blue, pitch). The bottom chart shows student's head tilt (green); the position is neutral at 0 for head tilt (yaw/pitch-green). The head position is neutral around 0 of vertical axis, and is pronounced when outside of the interval [-1,1]. This student solved a problem on the first try/first attempt (orange vertical line; purple vertical line (time to first attempt) is not visible and is covered by orange line).

Class	Original dataset (a)	Balanced dataset (b)
Paper	4,655	638
Screen	13,483	826
Wander	583	509
Total	18,721	1,973

Table 1. **Dataset Samples Per Class.** Column (a) presents the samples distribution for each class in the real-world raw data. Column (b) presents the same distribution after down-sampling and balancing the original dataset. Both versions will be made publicly available for non-commercial research purposes.

time series graphs that reflect students' head positions while they solve mathematics problems and analyze data collected within the time frame of a single problem (see Figure 5). These graphs integrate head position and yaw/pitch information along with log data of a students' actions within in the tutor. Head position is neutral at 0 in the y-axis. The top graph displays two lines to indicate left/right (red, yaw) and up/down (blue, pitch) while the bottom chart shows student's head tilt (green). This student solved a problem on the first try (orange vertical line). The online tutor records information about the time required for students to request a hint, time until they first tried to solve the problem and the time recorded to solve the problem (Figure 5, top, vertical orange line). We found invaluable information about the juxtaposition of head position and problem-solving activity logs. For example, a 'head tilt' (green line in the bottom time graph) appears to occur when the student is also moving his/her head up and down; thus 'head tilt' may be a sign of concentration and cognitive engagement. This time series data of student behavior will also be made publicly available with the annotated dataset.

4. Experiments

Given the collected, annotated, and balanced dataset of students solving mathematical problems, we now discuss the models used to predict student engagement. We consider state-of-the-art deep learning architectures that classify a student's gesture into 'looking at their screen', 'looking at their paper', or 'wandering'. We then compare these to baselines that rely on head pose estimation.

Deep Convolutional Networks. We explore different deep convolutional neural networks for the task of classifying the frames. The architectures we use as the backbone model are: MobileNet [5], VGG16 [30] and Xception [11]. We fine-tune models that are pre-trained on ImageNet [20]. On top of the pre-trained model, we add the following custom layers: one global average pooling 2D, one dense layer with 128 neurons and ReLU activation, and a final output layer with three neurons and softmax activation. To avoid overfitting, we use multiple data augmentation techniques at the input layer (gaussian noise, color channel changes, and cropping) and neurons drop-out at the head layers. To compare the performance of the different models we use the global and per-class accuracy score. After training with frozen weights for the backbone, we fine-tune the last layers of the backbone to reach better accuracy (the number of layers fine-tuned depends on the model complexity).

Head Pose Estimation. We estimate the head poses (*i.e.*, yaw, pitch and roll) of students using a deep neural network FSA-Net [34]. Given a facial image detected and cropped using MTCNN [33], a deep cascaded multi-task face detector, FSA-Net predicts the head pose based on feature aggregation and regression. It combines feature maps from different layers/stages by spatially grouping and aggregating features to harvest multi-scale information. The learned meaningful intermediate features are then used for performing soft stage-wise regression. The pose estimation model is pre-trained on the 300W-LP synthetic dataset [36] which contains 122,450 facial images with labelled head poses. The dataset synthesized faces across large poses (above 45°), ensuring that the trained model is robust to self-occlusion in our student dataset. In situations which MTCNN fails in detecting a face due to occlusion, we switch to using a single deep neural network Single Shot Detector(SSD) [24] and ResNet architecture for head pose estimation.

We focus on two approaches for baseline classifiers. Our first method is a conditional approach with yaw and pitch head angles as the features. Three "if" conditions were implemented to distinguish head poses as either 'looking at their screen', 'looking at their paper', or 'wandering'. When students look at their paper, visible positive spikes in the pitch angle and a negative spike in the yaw angle were observed. When students look at their screen, the yaw and pitch angles were neutral around 0. Both these observations can be seen in Figure 5. The conditions for the conditional classifier are as follows 1) if the yaw angle is negative and the pitch angle is positive, we classify the set of angles as 'looking at their paper'; 2) if the yaw and pitch poses are 0.0 ±0.05, we classify the set of angles as 'looking at their screen'; 3) if both conditions were not met, we classify the set of angles as 'wandering'. Our second approach uses the classical Logistic Regression to model the probability of a certain class. Each set of head angles (yaw and pitch of the student's head pose in a frame) correspond to a data point with each data point being annotated as one of the three labels. We train a 2-feature Logistic Regression classifier with a total of 1,589 sets of angles as training. Each class was weighted with respect to the class size for balancing the dataset. Cross-Entropy loss was used as the loss function and Stochastic Average Gradient Descent as the optimizer.

Results. Table 2 presents the accuracy of predicting student engagement for the different baselines. The deep learning models show different results depending on the model size and number of parameters. Important to notice is that all the pre-trained models reach similar performances when trained with frozen backbone weights (between 74-79% test accuracy), but they differ when we fine-tune the backbone model (results in Table 2). A smaller model such as MobileNet allows us to fine-tune more layers without overfitting, compared to deeper or larger models like VGG16 and Xception. This allows the MobileNet model to obtain a feature representation of the input images that is more relevant for this classification task, and by consequence this model reaches a higher final accuracy compared to the others. The results and training strategy may vary when we use different dataset configurations. We can also conclude from the table that all convolutional neural networks perform significantly better than the head pose estimation strategies. Figure 6 presents the per class accuracy for the best deep learning and head pose estimation models.

Method	Accuracy (%)
MobileNet (pretrained ImageNet)	94
Xception (pretrained ImageNet)	88
VGG16 (pretrained ImageNet)	85
Head pose Estimator (Logistic Reg.)	60
Head pose Estimator (Conditional)	55

Table 2. **Results.** Global accuracy score of predicting student engagement using different deep learning and head pose estimate approaches. From these results we can conclude that the deep learning models are more suitable for this type of classification task compared to the head pose estimation. Also, depending on the complexity of the deep learning model we reach different accuracy scores, with the best results obtained by the model with less complexity, MobileNet.

TestBed Intelligent Tutor. MathSpring is our testbed tutor. Built at UMass-Amherst [9, 7, 6, 8, 3], it incorporates 1,000 mathematics problems, has been used by more than 20,000 students, and covers topics in grades 5-12. Additionally, it provides detailed student- and class-level analytics data to help teachers inform adjustments to classroom instruction and pacing. MathSpring, represents 214 Common Core Standards/topics including geometry, equations, fractions, statistics and algebra. A cognitive model automatically assesses students' knowledge based on their behavior and adapts problem difficulty and feedback. The tutor uses an effort-based tutoring algorithm to select the next problem for each student, maintaining students within a zone of proximal development, by selecting problems that are not too easy nor too hard [32]. Rich multimedia help provides on-demand support and is offered when students make mistakes. Students accomplish goals within their comfort level supported by alternative presentation modes (*i.e.*, highlights and graphics) or easier word problems and remedial hints as needed. Teachers access reports about individual student progress. Like a human tutor, the tutor sustains engagement and provides practice required for students to become better learners, while facilitating logging, pre/post-testing and data collection.

Pilot Study. We conducted a Pilot Study where we integrate our head pose estimator into MathSpring. A student's head pose is computed in real-time, and is being used with real students during Summer 2021. The tutor detects whether a student is looking off-screen by analyzing the

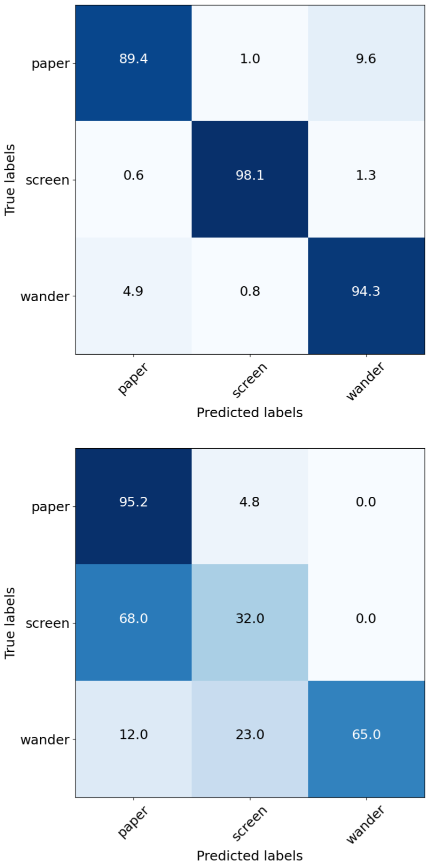
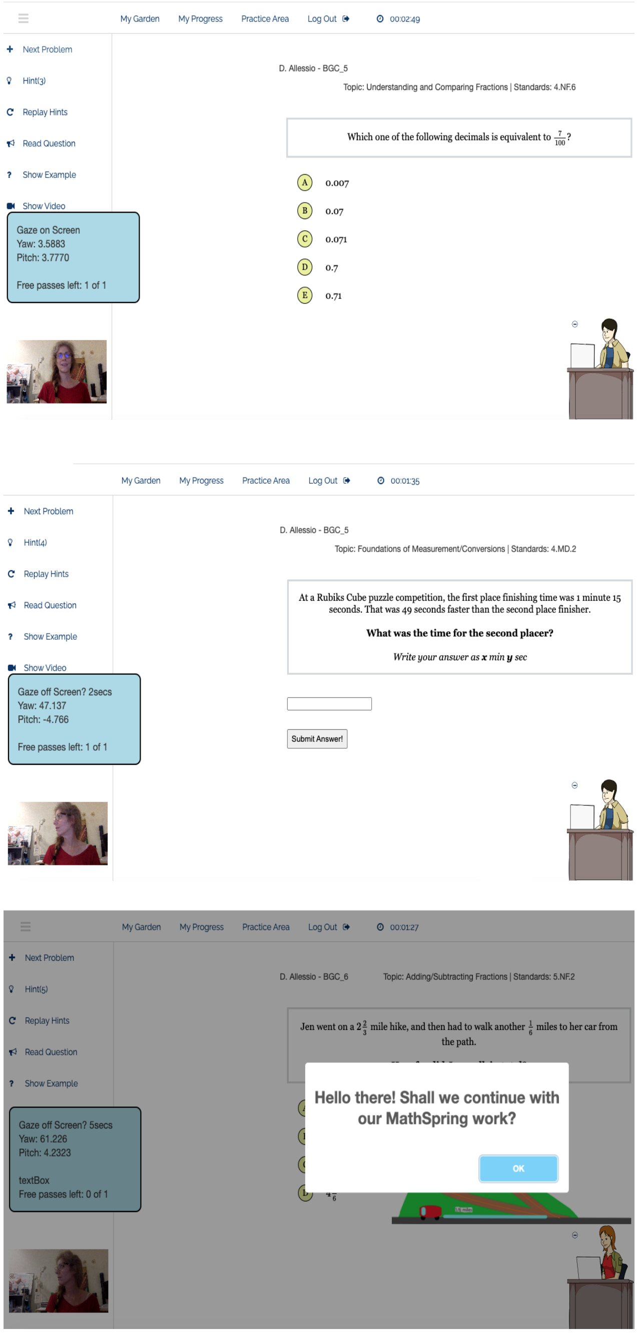


Figure 6. **Confusion Matrix Comparison.** The head pose estimation model (bottom) obtains a lower per-class accuracy score, compared to the deep learning model solution MobileNet (top), which not only reaches an overall higher accuracy but also consistently classifies different classes.

pose angle values as shown in Figure 4. Specifically, the tutor considers a student facing straight at the screen as a neutral state (*i.e.*, pose angle is 0°) and infer off-screen poses when the angle values exceed certain thresholds as previously detailed. The tutor is designed to deliver real-time interventions *e.g.*, showing a focus circle, an animated character, or a message as demonstrated in Figure 4. Such interventions target re-engaging a wandering student. The real-time detection and automatic responses help students sustain and effectively allocate attentional resources on learning tasks, which is critical for effective learning [14].

Future Work. Our research questions include: Are head pose interventions successful in reorienting student attention towards learning? How about deep learning models that have demonstrated superior classification performance in this work? Which interventions (focus circle, animated character, message) are most effective in promoting learning gains compared to the non-pose-reactive tutor? Do individual student differences in prior knowledge, aptitude, af-



D. Allessio - BGC_5

Topic: Understanding and Comparing Fractions | Standards: 4.NF.6

Which one of the following decimals is equivalent to $\frac{7}{100}$?

A 0.007

B 0.07

C 0.071

D 0.7

E 0.71

Gaze on Screen
Yaw: 3.5883
Pitch: 3.7770

Free passes left: 1 of 1

D. Allessio - BGC_5

Topic: Foundations of Measurement/Conversions | Standards: 4.MD.2

At a Rubiks Cube puzzle competition, the first place finishing time was 1 minute 15 seconds. That was 49 seconds faster than the second place finisher.

What was the time for the second placer?

Write your answer as ***x*** *min* ***y*** *sec*

Submit Answer!

Gaze off Screen? 2secs
Yaw: 47.137
Pitch: -4.766

Free passes left: 1 of 1

D. Allessio - BGC_6 Topic: Adding/Subtracting Fractions | Standards: 5.NF.2

Jen went on a $2\frac{2}{3}$ mile hike, and then had to walk another $\frac{1}{6}$ miles to her car from the path.

Hello there! Shall we continue with our MathSpring work?

OK

Gaze off Screen? 5secs
Yaw: 61.226
Pitch: 4.2323

textBox
Free passes left: 0 of 1

Figure 7. **MathSpring Gaze Tutor and Sample Intervention.** The Gaze Tutor in MathSpring detects a student's head pose including yaw and pitch (top) and predicts that attention has wandered if the head position is off center for a certain length of time (middle). The tutor then selects an intervention, *e.g.*, a message, visual cue or a pedagogical agent to bring the student back to task. We show a sample intervention of displaying a message after the system has predicted 'wandering' (bottom). The student's video and information about yaw and pitch (blue) in the lower left are not shown to the student.

fective predispositions towards learning mathematics moderate the effects of computer vision interventions (for the teacher and for the student) in learning and motivation?

5. Conclusion

This work proposes a student engagement video dataset of students in an online learning setting solving math problems. The dataset includes engagement annotation and time series information about head pose. The dataset is split in a way to ensure that no video frames of the same individual appear in both training and testing splits. We then train deep convolutional networks and head pose classifiers to provides baselines for predicting student engagement from snapshots of student status. Finally, we integrate one of our proposed baselines on the platform MathSpring such that whenever wandering is detected, a student is presented with an intervention that would help in re-gaining attention.

Acknowledgments

This work was supported in part by a National Science Foundation (NSF) grant #1551594, and gifts from Nvidia. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of NSF or the U.S. Government.

References

[1] Mathspring. http://ckc.mathspring.org/welcome.jsp? 2

[2] Ali Abedi and Shehroz S Khan. Improving state-of-the-art in detecting student engagement with resnet and tcn hybrid network. *arXiv preprint arXiv:2104.10122*, 2021. 3

[3] Danielle Allessio, Beverly Woolf, Naomi Wixon, Florence R Sullivan, Minghui Tai, and Ivon Arroyo. Ella me ayudó (she helped me): Supporting hispanic and english language learners in a math its. In *International Conference on Artificial Intelligence in Education*, pages 26–30. Springer, 2018. 6

[4] Khawlah Altuwairqi, Salma Kammoun Jarraya, Arwa Allinjawi, and Mohamed Hammami. Student behavior analysis to measure engagement levels in online learning environments. *Signal, Image and Video Processing*, pages 1–9, 2021. 3

[5] Menglong Zhu Andrew G. Howard, Dmitry Kalenichenko Bo Chen, Tobias Weyand Weijun Wang, and Hartwig Adam Marco Andreetto. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017. 5

[6] Ivon Arroyo, Winslow Burleson, Minghui Tai, Kasia Muldner, and Beverly Park Woolf. Gender differences in the use and benefit of advanced learning technologies for mathematics. *Journal of Educational Psychology*, 105(4):957, 2013. 6

[7] I Arroyo, D Shanabrook, BP Woolf, and W Burleson. Analyzing affective constructs: emotions, motivation and attitudes. In *International Conference on Intelligent Tutoring Systems*, 2012. 6

[8] Ivon Arroyo, Beverly Park Woolf, Winslow Burelson, Kasia Muldner, Dovan Rai, and Minghui Tai. A multimedia adaptive tutoring system for mathematics that addresses cognition, metacognition and affect. *International Journal of Artificial Intelligence in Education*, 24(4):387–426, 2014. 6

[9] Ivon Arroyo, Beverly P Woolf, David G Cooper, Winslow Burleson, and Kasia Muldner. The impact of animated pedagogical agents on girls' and boys' emotions, attitudes, behaviors and learning. In *2011 IEEE 11th International Conference on Advanced Learning Technologies*, pages 506–510. IEEE, 2011. 6

[10] Ryan SJd Baker, Sidney K D'Mello, Ma Mercedes T Rodrigo, and Arthur C Graesser. Better to be frustrated than bored: The incidence, persistence, and impact of learners' cognitive–affective states during interactions with three different computer-based learning environments. *International Journal of Human-Computer Studies*, 68(4):223–241, 2010. 2

[11] Francois Chollet. Xception: Deep learning with depthwise separable convolutions. *arXiv preprint arXiv:1610.02357*, 2017. 5

[12] M Ali Akber Dewan, Mahbub Murshed, and Fuhua Lin. Engagement detection in online learning: a review. *Smart Learning Environments*, 6(1):1–20, 2019. 3

[13] Sidney D'Mello, Blair Lehman, Reinhard Pekrun, and Art Graesser. Confusion can be beneficial for learning. *Learning and Instruction*, 29:153–170, 2014. 2

[14] Sidney K D'Mello. Gaze-based attention-aware cyberlearning technologies. In *Mind, Brain and Technology*, pages 87–105. Springer, 2019. 7

[15] Arthur C Graesser. Deeper learning with advances in discourse science and technology. *Policy Insights from the Behavioral and Brain Sciences*, 2(1):42–50, 2015. 2

[16] Abhay Gupta, Arjun D'Cunha, Kamal Awasthi, and Vineeth Balasubramanian. Daisee: Towards user engagement recognition in the wild. *arXiv preprint arXiv:1609.01885*, 2016. 2

[17] Tao Huang, Yunshan Mei, Hao Zhang, Sanya Liu, and Huali Yang. Fine-grained engagement recognition in online learning environment. In *2019 IEEE 9th international conference on electronics information and emergency communication (ICEIEC)*, pages 338–341. IEEE, 2019. 3

[18] Stephen Hutt, Jessica Hardey, Robert Bixler, Angela Stewart, Evan Risko, and Sidney K D'Mello. Gaze-based detection of mind wandering during lecture viewing. *International Educational Data Mining Society*, 2017. 3

[19] Robert J Jagers, Deborah Rivas-Drake, and Brittney Williams. Transformative social and emotional learning (sel): Toward sel in service of educational equity and excellence. *Educational Psychologist*, 54(3):162–184, 2019. 2

[20] Wei Dong Jia Deng, Li-Jia Li Richard Socher, and Fei-Fei Li Kai Li. Imagenet: A large-scale hierarchical image database.

In *2009 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2009. 5

[21] Amanjot Kaur, Aamir Mustafa, Love Mehta, and Abhinav Dhall. Prediction and localization of student engagement in the wild. In *2018 Digital Image Computing: Techniques and Applications (DICTA)*, pages 1–8. IEEE, 2018. 2

[22] Dimitrios Kollias, Mihalis A Nicolaou, Irene Kotsia, Guoying Zhao, and Stefanos Zafeiriou. Recognition of affect in the wild using deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 26–33, 2017. 2

[23] Dimitrios Kollias, Panagiotis Tzirakis, Mihalis A Nicolaou, Athanasios Papaioannou, Guoying Zhao, Björn Schuller, Irene Kotsia, and Stefanos Zafeiriou. Deep affect prediction in-the-wild: Aff-wild database and challenge, deep architectures, and beyond. *International Journal of Computer Vision*, 127(6):907–929, 2019. 2

[24] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016. 6

[25] Daniel McDuff, Rana Kaliouby, Thibaud Senechal, May Amr, Jeffrey Cohn, and Rosalind Picard. Affectiva-mit facial expression dataset (am-fed): Naturalistic and spontaneous facial expressions collected. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 881–888, 2013. 2

[26] Danielle S McNamara, Eileen Kintsch, Nancy Butler Songer, and Walter Kintsch. Are good texts always better? interactions of text coherence, background knowledge, and levels of understanding in learning from text. *Cognition and instruction*, 14(1):1–43, 1996. 2

[27] Tarmo Robal, Yue Zhao, Christoph Lofi, and Claudia Hauff. Intellieye: Enhancing mooc learners' video watching experience through real-time attention tracking. In *Proceedings of the 29th on Hypertext and Social Media*, pages 106–114. ACM, 2018. 3

[28] Tarmo Robal, Yue Zhao, Christoph Lofi, and Claudia Hauff. Webcam-based attention tracking in online learning: A feasibility study. In *23rd International Conference on Intelligent User Interfaces*, pages 189–197, 2018. 3

[29] Prabin Sharma, Shubham Joshi, Subash Gautam, Sneha Maharjan, Vitor Filipe, and Manuel JCS Reis. Student engagement detection using emotion analysis, eye tracking and head movement with machine learning. *arXiv preprint arXiv:1909.12913*, 2019. 3

[30] Karen Simonyan and Andrew Zisserman. Very deep convolutional network for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2015. 5

[31] Chinchu Thomas and Dinesh Babu Jayagopi. Predicting student engagement in classrooms using facial behavioral cues. In *Proceedings of the 1st ACM SIGCHI international workshop on multimodal interaction for education*, pages 33–40, 2017. 3

[32] Lev Semenovich Vygotsky. *Mind in society: The development of higher psychological processes*. Harvard university press, 1980. 6

[33] Jia Xiang and Gengming Zhu. Joint face detection and facial expression recognition with mtcnn. In *2017 4th international conference on information science and control engineering (ICISCE)*, pages 424–427. IEEE, 2017. 6

[34] Tsun-Yi Yang, Yi-Ting Chen, Yen-Yu Lin, and Yung-Yu Chuang. Fsa-net: Learning fine-grained structure aggregation for head pose estimation from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1087–1096, 2019. 5

[35] Stefanos Zafeiriou, Dimitrios Kollias, Mihalis A Nicolaou, Athanasios Papaioannou, Guoying Zhao, and Irene Kotsia. Aff-wild: valence and arousal'in-the-wild'challenge. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 34–41, 2017. 2

[36] Xiangyu Zhu, Zhen Lei, Xiaoming Liu, Hailin Shi, and Stan Z Li. Face alignment across large poses: A 3d solution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 146–155, 2016. 6