

FACE READERS: The Frontier of Computer Vision and Math Learning*

Beverly Woolf^{1,*}, Margrit Betke², Hao Yu², Sarah Adel Bargal³, Ivon Arroyo¹, John Magee⁴, Danielle Alessio¹ and William Rebelsky¹

¹University of Massachusetts-Amherst, Massachusetts, MA 01003, USA

²Boston University, Massachusetts, MA 02215, USA

³Georgetown University, Washington, D.C. 20057, USA

⁴Clark University, Worcester, MA 01610, USA

Abstract

The future of AI-assisted individualized learning includes computer vision to inform intelligent tutors and teachers about student affect, motivation and performance. Facial expression recognition is essential in recognizing subtle differences when students ask for hints or fail to solve problems. Facial features and classification labels enable intelligent tutors to predict students' performance and recommend activities. Videos can capture students' faces and model their effort and progress; machine learning classifiers can support intelligent tutors to provide interventions. One goal of this research is to support deep dives by teachers to identify students' individual needs through facial expression and to provide immediate feedback. Another goal is to develop data-directed education to gauge students' pre-existing knowledge and analyze real-time data that will engage both teachers and students in more individualized and precision teaching and learning. This paper identifies three phases in the process of recognizing and predicting student progress based on analyzing facial features: Phase I: Collecting datasets and identifying salient labels for facial features and student attention/engagement; Phase II: Building and training deep learning models of facial features; and Phase III: Predicting student problem-solving outcome.

Keywords

facial expression recognition, intelligent tutors, machine learning

1. Introduction

As students engage with online learning technologies, they experience a variety of emotions (confusion, excitement, frustration, anxiety) and various levels of engagement, depending on a combination of motivation, mood, and background knowledge. Students' affective states and engagement are tightly correlated with learning gains [1, 2]. Having affective and engagement information accessible to teachers (or digital tutors) can aid in understanding students' progress and suggest when and which students need further assistance.


AIED 2023 Workshop: Towards the Future of AI-Augmented Human Tutoring in Math Learning, July 07, 2023, Tokyo, Japan

*Corresponding author.

✉ bev@umass.edu (B. Woolf); betke@bu.edu (M. Betke); haoyu@bu.edu (H. Yu); sarah.bargal@georgetown.edu (S. A. Bargal); arroyo@umass.edu (I. Arroyo); jmagee@clarku.edu (J. Magee); alessio@umass.edu (D. Alessio); wrebelsky@umass.edu (W. Rebelsky)



© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

This paper describes the design and evaluation of a suite of tools for facial expression recognition called FaceReaders, or tools that detect users' faces and gestures for the purpose of identifying and predicting engagement, motivation, and future behavior. For example, if an intelligent tutor can predict that a student's future behavior will be to "Give up", the system might provide an intervention (example problem, formula, hints or easier problem). Ethnographic surveys are used to first identify activities and concrete questions that teachers ask in real-time: Who needs my help most right now? Who is wheel-spinning right now? Is the class ready to move to the next topic? How often is Arjun skipping, guessing or giving up? Answers to these questions help teachers strategize responses, adapt class pedagogy and provide interventions. This paper presents a survey of research activities addressed by our laboratories towards the future of computer vision-augmented tutoring in math learning. One goal is to design, develop and evaluate these tools. Specifically, we describe Phase I: Collecting datasets and identifying labels for faces and gestures; Phase II: Identifying students' attention in math learning, and Phase III: Predicting problem-solving outcome.

2. Related and Prior Work

Intelligent Tutoring Systems. Intelligent Tutoring Systems (ITS) produce learning gains with effects close to one letter grade improvement [4, 5]. Students using these tutors outperform students from conventional classes in 92 percent of the controlled evaluations with performance measured twice as high as for students using typical (non-intelligent systems) [6, 7, 8]. One meta-analysis of findings from controlled ITS evaluations shows that test scores increased by 0.66 standard deviations over conventional levels, or from the 50th to the 75th percentile [8]. In an emotion-sensitive ITS, student emotion is automatically detected through facial expressions, body posture and gestures, speech, text, or physiological metrics. Measuring physiological signals is the rarest metric as it requires an intrusive learning experience [9]. Strain and D'Mello [10] studied the role of emotion in ITS engagement, task persistence, and learning gain. D'Mello et al. used gaze prediction based on natural language dialogues. Their system responded to students' boredom and tried to engage students; boredom is one of the most frequent states a student experiences during learning and negatively correlates with learning gain.

Visual Facial Action Units. A correlational analysis found evidence for relationships between visual facial Action Unit (AU) factors and self-reported traits such as academic effort, study habits, and interest in subjects [11]. Detected facial cues gave insight into the learner's

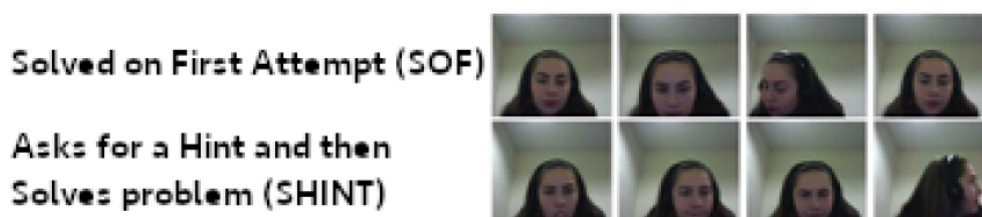


Figure 1: Images of a student solving a problem on the first attempt (SOF) (top row) and when she required hints to solve the problem (SHINT) (bottom row). Originally published in [3]

mental state, but potential cues to predict learning did not offer a consistent signal. Behavior prediction can support improved learning by tailoring the interventions of the ITS to the predicted actions of the student. Our work focuses on using predicted deep affect embeddings learned from a large facial affect dataset to improve behavior prediction in an ITS [12].

Transfer Learning in Facial Analysis. Prior research in transfer learning for facial analysis applications mostly focuses on transfer learning within the same application to improve results or bridge domain gaps, e.g., personalize a prediction system to specific individuals [13, 14, 12], fine-tune neural networks pre-trained on external datasets for a similar prediction task [15], or pre-training on a related facial analysis task [16]. In contrast, our work tackles transfer learning across domains and tasks, which is a form of transductive transfer learning [17]. We explore transfer learning from the facial analysis problem of in-the-wild affect recognition of affect to a webcam video behavior prediction problem. Work exists to explore transfer learning from facial analysis to behavior analysis such as AutoRate that uses VGGFace facial recognition embeddings to improve predictions of driver attention scores [18].

Interventions in Online Tutors. Affective messages delivered by avatars and empathetic messages respond to students' recent emotions [19]. Interventions in MathSpring ITS led to improved grades in state standardized exams [20] as well as influencing students' perceptions of themselves as learners [21]. Empathetic characters generate superior results to improve student interactions with the system and address negative student emotions [22, 23, 24].

MathSpring Intelligent Tutoring System. MathSpring.org, is a freely available game-like system [25]; students grow gardens that visually represent their mathematics progress. It is *multimedia* in that it provides both audio and visual support, *intelligent* in that it builds an internal model of students, as would a good teacher, and *personalized* in that it provides remedial tutoring when needed. For instance, MathSpring might be set to teach Grade 7 mathematics and will seamlessly move back to grade 6 and 5 material as needed, in a way that is unnoticeable to the student. Animated learning companions (LCs) provide emotional support and build students' socio-emotional skills, instilling a growth mindset [26], encouraging students to consider mistakes as a natural part of learning and stating that intelligence is malleable. LCs support students' learning processes and use well established instructional strategies. Students' emotions are assessed regularly and the tutor offers support when students become frustrated or anxious [25]. MathSpring provides a positive affective impact on students in the USA and Argentina. In controlled studies, students showed an increase in mathematics and reading comprehension. The combination of the character and role design of pedagogical agents makes a significant positive impact on student learning and behavior (e.g., [27, 28, 29, 30, 31]).

3. Phase I: Collecting Datasets and Identifying Labels for Face and Gestures

Students experience a variety of emotions while working online, e.g., boredom, frustration, interest, and surprise [32] and these displayed emotions correlate well with students' achievement in the learning task [33]. Equipping an intelligent tutor with the ability to interpret such affective signals could potentially enable it to monitor students' progress, provide timely interventions and present appropriate affective reactions via a virtual tutor. For example, machine learning

classifiers can be trained to recognize the subtle differences in facial behavior between when a student requires hints to solve a problem, see Figures 1- 2, so that the tutor can intervene accordingly.

During Phase I, we collected and annotated databases of facial affect videos of students interacting with MathSpring, an intelligent tutor, Figure 3. Considering the dearth of large-scale, publicly available affect video datasets in learning and education settings, we made these datasets and annotations public [22, 3, 12]. The video datasets consist of college students solving math problems with a front facing camera collecting visual feedback of student gestures. Datasets consist of video clips which were obtained by trimming the raw videos based on problem start and end times recorded in MathSpring’s log file.

In the initial dataset collection [3], we collected 1596 video clips of 30 different students solving math problems. Each video clip is automatically annotated by MathSpring’s learning log data. The labels used to annotate the video clips are: ATT (student did not see any hints but solved the question after 1 incorrect attempt), GIVEUP (student performed some action but did not solve the problem at all), GUESS (student did not see hints, but solved the question after greater than 1 incorrect attempts), NOTR (student performed some action, but the first action was too rapid for him to have read the problem), SHINT (student eventually got the correct answer after seeing one or more hints), SKIP (student skipped problem with no action) and SOF (student answered correctly in first attempt, without seeing any hints). In the next iteration [12], we collected and annotated an extended version of [3], resulting in a dataset of 2749 video clips of math problems solved by 54 students. Next, we provide additional labels for video frames for a subset of 400 video clips of 19 different students. The labels used to annotate the extracted video frames are: “looking at their screen”, “looking at their paper”, or “wandering”. This resulted in 18,721 annotated frames. An interface was presented to MTurk workers for labeling to indicate whether students’ attention was engaged or wandering. Each of the 18,721 frames was assigned to three different crowdworkers and we processed 56,163 (18,721 * 3) results.

Our contributions are summarized as follows:

- Introduced a unique video dataset of 1596 student interactions labeled for problem outcome, extracted from more than 30 hours of raw video data [3];
- Augmented the dataset of [3] to include 2749 student interactions labeled for problem outcome [12];
- Provided annotations for a subset of 400 student interactions labeled for attention/engagement [22];
- Made these datasets publicly accessible to encourage and foster research in the intersection of the education and computer vision communities; and
- Provided a set of baseline results predicting student learning outcomes and attention/engagement solely from facial affect signals.

Analysis of the Outcome Classes. We provided an exploratory analysis of the different problem outcome classes that result when students interact with MathSpring, using typical facial action unit activations to analyze students’ faces. We developed baseline models to predict students’ problem outcome labels (e.g., ask for hints, solve problem) and discussed how early

problem outcome labels can be forecasted to provide possible interventions. Each data instance in the data set consists of a video clip of a student working on a problem and its corresponding label of the student’s problem-solving behavior. We researched baseline models to investigate the problem of directly predicting the learning outcome of students solely from affect signals. To visually illustrate the prediction of problem outcomes and to understand student behavior, we present visual examples of an eighth grade student using MathSpring, see Figure 2. The student used MathSpring for one session of around 20 minutes and consented to have his face and screen recorded. Figure 2 shows the evolution of student expressions and gestures, and their corresponding problem outcomes. When the student successfully solves the problem on the first attempt (SOF), we observe that he focused tightly on the problem during the period (first row). When he finally solved the problem correctly, he clenched his fist which may indicate his excitement and passion (second row). When asked for hints, the student looked confused.

To visually illustrate the prediction of problem outcomes and to understand student behavior,

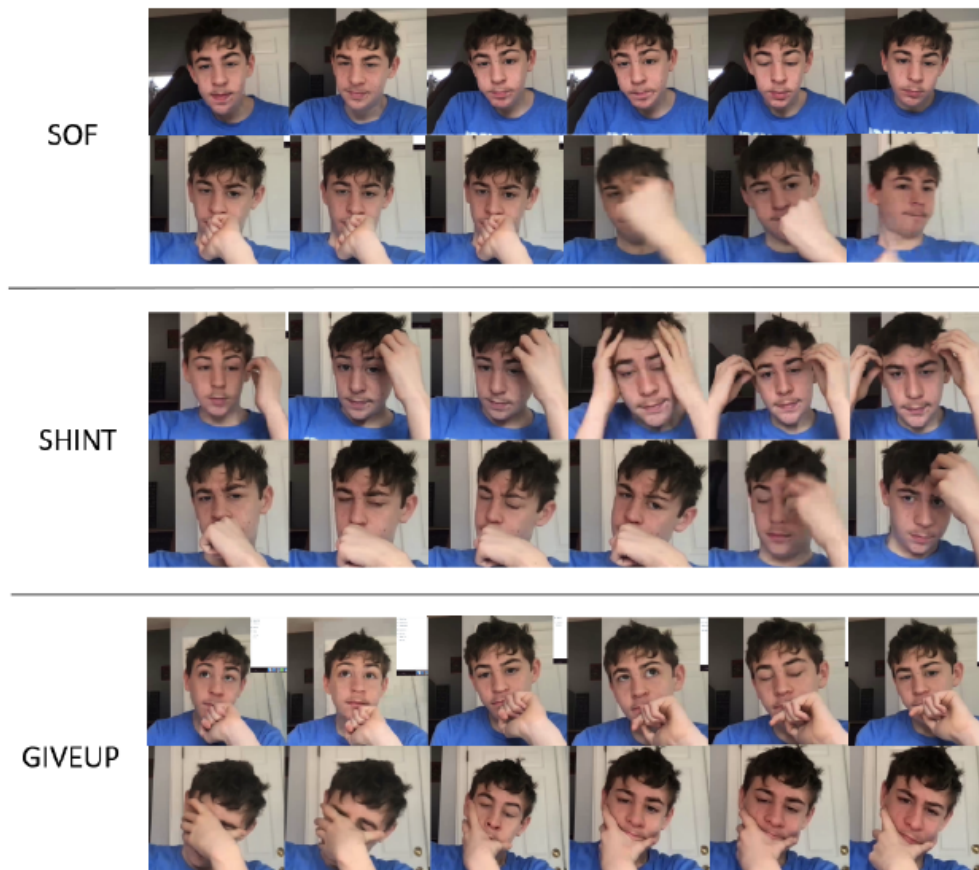


Figure 2: Example face-cropped images showing the evolution of student expressions and gestures, with the corresponding problem outcomes. In the top two rows the student solved the problem on the first attempt (SOF), 3-4th rows the student solved the problem with hints (SHINT), and in the bottom two rows the student tried but ultimately skipped the problem (GIVEUP). Originally published in [12]

we present visual examples of an eighth grade student using MathSpring, see Figure 2. The student used MathSpring for one session of around 20 minutes and consented to have his face and screen recorded. Figure ?? shows the evolution of student expressions and gestures, and their corresponding problem outcomes. When the student successfully solves the problem on the first attempt (SOF), we observe that he focused tightly on the problem during the period (first row). When he finally solved the problem correctly, he clenched his fist which may indicate his excitement and passion (second row). When asked for hints, the student looked confused scratching his head but still engaged and actively attempted to solve the problem (rows 3–4). For the last problem (GIVEUP), the student gradually became distracted and presented frustration and boredom (rows 5–6). These observations are consistent with our assumption that facial expressions and gestures provide important cues for inferring students’ learning outcomes.

Because of the interpretability of Facial Action Units (AUs), we visualized how often and with what intensity various action units occurred on average for the different effort classes of the entire dataset. For each data instance, we aggregated AU presence values weighted by their respective intensity values and normalized them by the total number of frames in which the face was detected. The input to our baseline models consists of variable-length webcam video clips of participants working on MathSpring problems. For each frame of all the videos in the dataset, 18 AU presence and 17 AU intensity values, along with head-pose and eye-gaze vectors, are extracted using OpenFace [34]. In order to compute an aggregate feature representation, we used statistics (mean, standard deviation, min and max) for each feature as well as statistics for their derivatives to produce a uniform length 376-dimensional feature representation. These

Figure 3: The MathSpring online tutor’s Practice Area interface provides “Hints” (textual and audio), worked-out examples, tutorial videos, and formulas. The companion encourages student perseverance and underscores the value of effort, especially when mistakes are made.

features are then used as input to machine learning models trained in Phase III to predict problem outcome labels. Other features that have been computed on our datasets are based on transfer learning and are described in phase III.

Long Term Goals to Identify Labels for Face and Gestures. One goal of Phase I is to improve the performance obtained by facial action units and baseline models. For example, a multi-modal model that utilizes signals from all streams of information in the dataset including the mouse movements and clicks, as well as the video stream of the screen activity will likely result in better predictive performance. Moreover, training models that explicitly utilize the temporal dynamics of how facial behavior evolves over the duration of the student’s interaction with the tutor could potentially yield further improvements in model performance. Finally, the biggest challenge in recognizing affect is to utilize affect-sensitive models to provide appropriate and effective interventions that quantifiably improve the learning experience. Some researchers ventured in this direction [22]. In future work, we plan to provide personalized interventions in MathSpring based on the proposed affect analysis models, and to conduct experiments to validate the effectiveness of the interventions. Lessons learnt from this initial analysis will also inform future data collection strategies. We intend to use richer data sets to investigate whether the system can predict changes in student learning behaviors and strategies.

4. Phase II: Identifying Students’ Attention in Math Learning

When students are bored or distracted online, they might disengage and wander, leading to a decline in the learning process. Currently, few online systems account for the context-sensitive nature of learning, i.e., motivation, social and emotional learning, and climate as well as complex interactions among these factors. In Phase II, we used computer vision to identify student engagement and emotion, which are correlated with learning gains [1, 2]; emotion drives attention and attention drives learning [35]. Computer vision-enhanced research can assist in supporting students’ emotion and maintaining their engagement by recognizing students’ head orientation and gaze expression.

We summarize our contributions as follows:

- Demonstrated the use of computer vision with live video data to infer affect as one indicator of students’ motivation;
- Developed a deep learning-based computer vision model to identify head pose as an indicator of engagement vs. distraction; and
- Communicated this information to teachers by showing facial expressions.

Research Approach and Results. Within the collected video dataset described in Phase I, annotations indicated whether students’ attention at specific frames was engaged or wandering [22], see Figure 2. In addition, we trained baselines for a computer vision module that determined 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 incorporated a gaze baseline into the MathSpring learning platform and evaluated its performance with the currently implemented approach.

Development and early evaluation of this technology monitored student engagement in real-time, detected waning attention and distraction, and assessed which interventions led to more productive learning. We used pre-process and crowdsourced label frames of the videos to propose a publicly available dataset that aids researchers in automated student engagement prediction [22].

The model was trained on benchmark datasets that were curated to help tutors solve such tasks. We incorporated one of our baselines in the MathSpring tutor. Figure 3 exhibits an example problem presented to students on MathSpring. The tutor targets sensing and interpreting facial signals relevant to student emotions and provides students with real-time classroom interventions that can aid their progress, suggesting when and who needs further assistance, and identifying which interventions are working. The implemented computer vision module alerts wandering students to regain their attention.

Given the collected, annotated, and balanced dataset of students solving mathematical problems, we considered state-of-the-art deep learning architectures that classified a student’s gesture into “looking at their screen”, “looking at their paper”, or “wandering”. We compared these to baselines that rely on head pose estimation.

We fine-tuned different convolutional architectures that are pre-trained on ImageNet [36] to classify video frames into the three classes. We also estimate head poses (i.e., yaw, pitch and roll) of students using a deep neural network FSA-Net [37]. The predicted head poses were used to classify video frames into the three aforementioned classes. We then compare the performance of the deep convolutional networks to the performance of the head pose estimator approach.

Pilot Study. We conducted a Pilot Study in which the head pose estimator was integrated into MathSpring. A student’s head pose was computed in real-time and used with real students during Summer 2021. The tutor detects whether a student is looking off-screen by analyzing the pose angle values and considers a student facing straight at the screen as being in a neutral state (i.e., the pose angle is 0°) and infers off-screen poses when the angle values exceed certain thresholds. Real-time interventions, e.g., showing a focus circle, an animated character, or a message, were delivered. 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 [38].

Results. All convolutional neural network architectures performed significantly better than the head pose estimation strategies. We presented the per-class accuracy for the best deep learning (94%) and head pose (60%) estimation models.

Long-Term Goals to Identify Students’ Attention. One long-term goal of Phase II is to evaluate students’ visual feedback in real classrooms through the head pose detector’s performance, which provides a coarse estimation of where students are looking. The intelligent tutor will acquire more information when students’ gaze direction can be detected and engagement is inferred. In this case, the head pose detector’s intervention (e.g., animated character, verbal message) is used as a learning companion for maintaining students’ level of engagement. During learning or problem-solving, it is quite common for students to keep relatively fixed head positions but the gaze direction moves frequently, which makes it insufficient to detect emotion from head poses only. Therefore, in Phase II we focused on methodology for inferring gaze direction while students interact in real-time.

Future research will provide evidence about whether head pose interventions are successful

in reorienting student attention towards learning and which deep learning models demonstrated superior classification performance. We also seek to determine which interventions are most effective in promoting learning gains compared to the non-pose-reactive tutor. Also of interest is whether individual student differences (e.g., in prior knowledge, aptitude, affective predispositions) moderate the effects of computer vision-enhanced interventions (for the teacher or student).

5. Phase III: Predicting Problem-Solving Outcome

In Phase III, we propose deep learning models that predict problem-solving outcomes for the video clips collected and annotated during Phase I. Predicting these outcomes allows tutoring systems to adapt interventions to enhance student learning. We first trained a classifier using traditional facial analysis features such as head pose, gaze and facial action units (AUs) to predict the exercise outcome. The multi-class model achieved a mean accuracy of 0.54 and a mean F-score of 0.27 for predicting one of seven possible outcome classes [3]. To improve prediction performance, we further developed a video-based transfer learning approach to predict problem outcomes of students by analyzing their facial expressions and gestures [12]. Our transfer learning challenge involved designing a representation for facial expression analysis using images from the Internet and transferring this knowledge to predict student behavior in webcam videos of students in a classroom setting. We introduced a novel facial affect representation and a user-personalized training scheme to harness the potential of this representation. Additionally, we developed various recurrent neural network variants that model the temporal structure of video sequences. Our final model, named ATL-BP for “Affect Transfer Learning for Behavior Prediction,” outperformed the previous work on the dataset, achieving a 50% relative increase in the mean F-score as well as an absolute 11 percentage point increase in accuracy.

Ideally, an affect-sensitive model should be able to accurately predict the effort label of the user as early as possible, in order to enable quick and effective interventions by the teacher or tutor. Therefore, our team is currently working on predicting the outcome of student performance using early visual and tabular cues demonstrating the efficacy of our approach and the potential impact of early outcome prediction for the development of better intelligent tutors. We will evaluate our classification models when only a fraction of the data is observed during test time. We are also incorporating tabular cues, e.g., timestamps of students performing specific actions. Again we are using a video-based transfer learning approach for predicting problem outcomes by analyzing students’ faces and gestures, and combining them with tabular data. The transfer-learning challenge is to design a representation in the source domain of images obtained from the internet for facial expression analysis and transfer this learned representation for human behavior prediction in the domain of webcam videos of students in a classroom environment.

6. Discussion and Conclusion

This paper presented a survey of research activities and challenges for the future of computer vision-augmented tutoring in math learning. The suite of computer vision tools that we

developed, called FaceReaders, uses facial expression recognition to identify and predict student engagement, motivation, affect, and future behavior early in students' interaction with online learning, specifically while students spend a brief time working on an exercise. We trained classifiers to directly predict the success or failure of a student's attempt to answer questions, based on features extracted from video streams. We extracted timing information from student log data, which includes the exact time students take for actions, e.g., asking for a hint or attempting to answer the exercise. Such information provides complementary insights into students' learning process and can be used to better understand their behavior and affective states.

To the best of our knowledge, no prior research combines visual affective analysis with student log data in the context of predicting student learning outcomes. One goal is to create and evaluate facial expression recognition tools with intelligent tutors.

Real-time teachers need answers for many questions, e.g., Who needs my help most right now? Is the class ready to move to the next topic? Answers to these questions will help teachers strategize responses, adapt class pedagogy and provide interventions. We expect this research to have a significant impact on development of better intelligent tutors. It should improve the diagnostic and predictive power of online learning by accurately predicting student exercise outcomes in the early stages.

References

- [1] R. S. Baker, S. K. D'Mello, M. M. T. Rodrigo, A. 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 (2010) 223–241.
- [2] S. D'Mello, B. Lehman, R. Pekrun, A. Graesser, Confusion can be beneficial for learning, *Learning and Instruction* 29 (2014) 153–170.
- [3] A. Joshi, D. Alessio, J. Magee, J. Whitehill, I. Arroyo, B. Woolf, S. Sclaroff, M. Betke, Affect-driven learning outcomes prediction in intelligent tutoring systems, in: 2019 14th IEEE international conference on automatic face & gesture recognition (FG 2019), IEEE, 2019, pp. 1–5.
- [4] S. D'Mello, A. Olney, C. Williams, P. Hays, Gaze tutor: A gaze-reactive intelligent tutoring system, *International Journal of human-computer studies* 70 (2012) 377–398.
- [5] K. VanLehn, The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems, *Educational psychologist* 46 (2011) 197–221.
- [6] A. T. Corbett, J. R. Anderson, Student modeling and mastery learning in a computer-based programming tutor, in: *Intelligent Tutoring Systems: Second International Conference, ITS'92 Montréal, Canada, June 10–12 1992 Proceedings 2*, Springer, 1992, pp. 413–420.
- [7] A. C. Graesser, K. VanLehn, C. P. Rosé, P. W. Jordan, D. Harter, Intelligent tutoring systems with conversational dialogue, *AI magazine* 22 (2001) 39–39.
- [8] J. A. Kulik, J. Fletcher, Effectiveness of intelligent tutoring systems: a meta-analytic review, *Review of educational research* 86 (2016) 42–78.
- [9] M. S. Hussain, O. AlZoubi, R. A. Calvo, S. K. D'Mello, Affect detection from multichannel

- physiology during learning sessions with autotutor, in: *Artificial Intelligence in Education: 15th International Conference, AIED 2011, Auckland, New Zealand, June 28–July 2011 15*, Springer, 2011, pp. 131–138.
- [10] A. C. Strain, S. K. D 'Mello, Emotion regulation during learning, in: *Artificial Intelligence in Education: 15th International Conference, AIED 2011, Auckland, New Zealand, June 28–July 2011 15*, Springer, 2011, pp. 566–568.
- [11] B. D. Nye, S. Karumbaiah, S. T. Tokel, M. G. Core, G. Stratou, D. Auerbach, K. Georgila, Engaging with the scenario: Affect and facial patterns from a scenario-based intelligent tutoring system, in: *Artificial Intelligence in Education: 19th International Conference, AIED 2018, London, UK, June 27–30, 2018, Proceedings, Part I 19*, Springer, 2018, pp. 352–366.
- [12] N. Ruiz, H. Yu, D. A. Alessio, M. Jalal, A. Joshi, T. Murray, J. J. Magee, K. M. Delgado, V. Ablavsky, S. Sclaroff, et al., Atl-bp: a student engagement dataset and model for affect transfer learning for behavior prediction, *IEEE Transactions on Biometrics, Behavior, and Identity Science* (2022).
- [13] J. Chen, X. Liu, P. Tu, A. Aragonés, Person-specific expression recognition with transfer learning, in: *2012 19th IEEE International Conference on Image Processing, IEEE, 2012*, pp. 2621–2624.
- [14] J. Chen, X. Liu, P. Tu, A. Aragonés, Learning person-specific models for facial expression and action unit recognition, *Pattern Recognition Letters* 34 (2013) 1964–1970.
- [15] H. Kaya, F. Gürpınar, A. A. Salah, Video-based emotion recognition in the wild using deep transfer learning and score fusion, *Image and Vision Computing* 65 (2017) 66–75.
- [16] M. Xu, W. Cheng, Q. Zhao, L. Ma, F. Xu, Facial expression recognition based on transfer learning from deep convolutional networks, in: *2015 11th International Conference on Natural Computation (ICNC), IEEE, 2015*, pp. 702–708.
- [17] S. J. Pan, Q. Yang, A survey on transfer learning. *IEEE transactions on knowledge and data engineering*, 22 (10) 1345 (????).
- [18] I. Dua, A. U. Nambi, C. Jawahar, V. Padmanabhan, Autorate: How attentive is the driver?, in: *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019), IEEE, 2019*, pp. 1–8.
- [19] B. P. Woolf, I. Arroyo, K. Muldner, W. Burleson, D. G. Cooper, R. Dolan, R. M. Christopher, The effect of motivational learning companions on low achieving students and students with disabilities, in: *Intelligent Tutoring Systems: 10th International Conference, ITS 2010, Pittsburgh, PA, USA, June 14-18, 2010, Proceedings, Part I 10*, Springer, 2010, pp. 327–337.
- [20] S. Karumbaiah, R. Lizarralde, D. Alessio, B. Woolf, I. Arroyo, N. Wixon, Addressing student behavior and affect with empathy and growth mindset., *International Educational Data Mining Society* (2017).
- [21] Y. Kim, Empathetic virtual peers enhanced learner interest and self-efficacy, in: *Workshop on Motivation and Affect in Educational Software, in conjunction with the 12th International Conference on Artificial Intelligence in Education, 2005*, pp. 9–16.
- [22] K. Delgado, J. M. Origi, T. Hasanpoor, H. Yu, D. Alessio, I. Arroyo, W. Lee, M. Betke, B. Woolf, S. A. Bargal, Student engagement dataset, in: *Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021*, pp. 3628–3636.

- [23] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.
- [24] H. Yu, A. Gupta, W. Lee, I. Arroyo, M. Betke, D. Allesio, T. Murray, J. Magee, B. P. Woolf, Measuring and integrating facial expressions and head pose as indicators of engagement and affect in tutoring systems, in: International Conference on Human-Computer Interaction, Springer, 2021, pp. 219–233.
- [25] I. Arroyo, B. P. Woolf, W. Bureson, K. Muldner, D. Rai, M. Tai, A multimedia adaptive tutoring system for mathematics that addresses cognition, metacognition and affect, *International Journal of Artificial Intelligence in Education* 24 (2014) 387–426.
- [26] C. S. Dweck, Messages that motivate: How praise molds students’ beliefs, motivation, and performance (in surprising ways), in: *Improving academic achievement*, Elsevier, 2002, pp. 37–60.
- [27] I. Arroyo, B. P. Woolf, D. G. Cooper, W. Bureson, K. 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, IEEE, 2011, pp. 506–510.
- [28] R. Azevedo, S. A. Martin, M. Taub, N. V. Mudrick, G. C. Millar, J. F. Grafsgaard, Are pedagogical agents’ external regulation effective in fostering learning with intelligent tutoring systems?, in: *Intelligent Tutoring Systems: 13th International Conference, ITS 2016, Zagreb, Croatia, June 7-10, 2016. Proceedings 13*, Springer, 2016, pp. 197–207.
- [29] A. L. Baylor, S. Kim, Designing nonverbal communication for pedagogical agents: When less is more, *Computers in Human Behavior* 25 (2009) 450–457.
- [30] M. C. Duffy, R. Azevedo, Motivation matters: Interactions between achievement goals and agent scaffolding for self-regulated learning within an intelligent tutoring system, *Computers in Human Behavior* 52 (2015) 338–348.
- [31] S. Lallé, N. V. Mudrick, M. Taub, J. F. Grafsgaard, C. Conati, R. Azevedo, Impact of individual differences on affective reactions to pedagogical agents scaffolding, in: *Intelligent Virtual Agents: 16th International Conference, IVA 2016, Los Angeles, CA, USA, September 20–23, 2016, Proceedings 16*, Springer, 2016, pp. 269–282.
- [32] S. D’Mello, R. W. Picard, A. Graesser, Toward an affect-sensitive autotutor, *IEEE Intelligent Systems* 22 (2007) 53–61.
- [33] R. Pekrun, T. Goetz, L. M. Daniels, R. H. Stupnisky, R. P. Perry, Boredom in achievement settings: Exploring control–value antecedents and performance outcomes of a neglected emotion., *Journal of educational psychology* 102 (2010) 531.
- [34] T. Baltrušaitis, P. Robinson, L.-P. Morency, Openface: an open source facial behavior analysis toolkit, in: 2016 IEEE winter conference on applications of computer vision (WACV), IEEE, 2016, pp. 1–10.
- [35] R. J. Jagers, D. Rivas-Drake, B. Williams, Transformative social and emotional learning (sel): Toward sel in service of educational equity and excellence, *Educational Psychologist* 54 (2019) 162–184.
- [36] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in: 2009 IEEE conference on computer vision and pattern recognition, Ieee, 2009, pp. 248–255.

- [37] T.-Y. Yang, Y.-T. Chen, Y.-Y. Lin, Y.-Y. Chuang, Fsa-net: Learning fine-grained structure aggregation for head pose estimation from a single image, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 1087–1096.
- [38] S. K. D’Mello, Gaze-based attention-aware cyberlearning technologies, *Mind, Brain and Technology: Learning in the Age of Emerging Technologies* (2019) 87–105.