Leveraging Affect Transfer Learning for Behavior Prediction in an Intelligent Tutoring System

Nataniel Ruiz[1], Hao Yu[1], Danielle A. Allessio[2], Mona Jalal[1], Ajjen Joshi[3], Thomas Murray[2], John J. Magee[4], Jacob R. Whitehill[5], Vitaly Ablavsky[6], Ivon Arroyo[5], Beverly P. Woolf[2], Stan Sclaroff[1], and Margrit Betke[1]

[1] Boston University [2] University of Massachusetts Amherst [3] Affectiva
[4] Clark University [5] Worcester Polytechnic Institute [6] University of Washington

2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021) | 978-1-6654-3176-7/21/$31.00 ©2021 IEEE | DOI: 10.1109/FG52635.2021.9667001

***Abstract*— In this work, we propose a video-based transfer learning approach for predicting problem outcomes of students working with an intelligent tutoring system (ITS). By analyzing a student's face and gestures, our method predicts the outcome of a student answering a problem in an ITS from a video feed. Our work is motivated by the reasoning that the ability to predict such outcomes enables tutoring systems to adjust interventions, such as hints and encouragement, and to ultimately yield improved student learning. We collected a large labeled dataset of student interactions with an intelligent online math tutor consisting of 68 sessions, where 54 individual students solved 2,749 problems. We will release this dataset publicly upon publication of this paper. It will be available at `https://www.cs.bu.edu/faculty/betke/research/learning/`. Working with this dataset, our transfer-learning challenge was to design a representation in the source domain of pictures obtained "in the wild" for the task of facial expression analysis, and transferring this learned representation to the task of human behavior prediction in the domain of webcam videos of students in a classroom environment. We developed a novel facial affect representation and a user-personalized training scheme that unlocks the potential of this representation. We designed several variants of a recurrent neural network that models the temporal structure of video sequences of students solving math problems. Our final model, named ATL-BP for *Affect Transfer Learning for Behavior Prediction*, achieves a relative increase in mean F-score of 50% over the state-of-the-art method on this new dataset.**

I. Introduction

Research on developing intelligent tutoring systems (ITS) is a promising avenue for improving learning and education [32], [4], [47]. Previous work has shown that real-time signals from students can be used to improve their learning [3], [16], [17]. Predicting whether students are having trouble with problems can allow an ITS to provide interventions, such as providing hints or encouragement, which could help the students understand or solve the problem, thus improving learning outcomes.

MathSpring [32] is a popular online browser-based ITS that uses multimedia to encourage and support students as they solve math problems. Using the MathSpring ITS, a dataset named MathSpringSP [22] was collected, which includes 1,596 segmented videos of study sessions of students interacting with the ITS. Each problem tackled by a student has an associated outcome label automatically annotated by the ITS. Some example labels are *skipped*, *solved on first try*, *solved with hint*, among others. In this work we address the problem of predicting the outcome label from a video feed of the student while they are solving the problem. Being able to have a model that can successfully predict outcomes while a student is completing a problem can help the ITS provide interventions such as hints or encouragement when the student is having difficulties.

978-1-6654-3176-7/21/$31.00 ©2021 IEEE

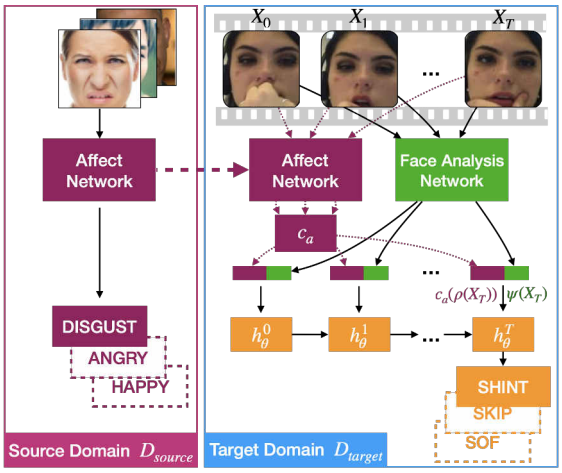

Fig. 1. Our proposed *Affect Transfer Learning for Behavior Prediction* (ATL-BP) model for predicting the behavior of students working with an intelligent tutoring system. The target-domain ATL-BP model consists of three components, an affect network trained for the source domain problem of affect recognition, a facial analysis network, and an LSTM.

Facial and gesture analysis are valuable tools for predicting emotions, but the question of how to use them for predicting student performance with an ITS remains challenging since cues can be very subtle or ambiguous. A smile, for example, does not necessarily mean that the student is happily solving an exercise. Instead, it could indicate a student's embarrassment for not knowing the answer to a question. Moreover, in our experience, trying to obtain valid ground truth labels of the student videos from human annotators is a futile experimental task because humans have a very low accuracy rate when predicting problem outcomes from video. Just like automated facial analysis tools, human annotators struggle with interpreting the ambiguity in and limited amount of information given by student gestures.

Prior research in transfer learning for facial analysis tasks

mostly focuses on transfer learning for the same task in order to bridge domain gaps such as personalization of a prediction system to specific individuals [1], [7], [8], [43], [50], improving results on a benchmark by fine-tuning neural networks that are pre-trained on external datasets for a similar prediction task [25], or improving results by pre-training on a related facial analysis task [48]. In contrast, our work tackles the more challenging transfer learning across domains and tasks, which is a form of *transductive transfer learning* [39], Specifically, we tackle the problem of learning a representation in the source domain of in-the-wild pictures for the task of facial expression analysis and transferring this learned representation to the task of human behavior prediction in the domain of webcam videos in a controlled environment (Fig. 1). While prior work has explored transfer learning from facial analysis to behavior analysis, for example, using VGG-Face facial recognition embeddings to predict driver distraction [15], our work is, to the best of our knowledge, the first to propose leveraging an affect representation, learned using a deep neural network, for a behavior prediction task. Our learned affect representation is general and can be used not only for predicting problem outcomes on an ITS, but in any human behavior prediction problem where affect and expression are important cues.

The largest obstacle in training an end-to-end deep learning model for behavior analysis problems is the fact that data are relatively scarce, which increases the risk of overfitting. As a first step to alleviating the data problem, we present MathSpringSP+, an extended version of the MathSpringSP dataset, which is roughly double the size of the original dataset. Next, we propose a novel facial affect representation for behavior prediction problems that is learned from a large affect classification dataset. We show that, by incorporating this affect embedding, we can obtain improvements compared to more traditional deep face embeddings such as the VGG-Face facial recognition embedding [40]. We developed a two-layer *Long Short Term Memory* (LSTM) model [20] that takes into account the temporal structure of the problem and successfully leverages our affect embedding. We show that, by conducting user-personalized training where a small portion of a student's initial captured data is used to fine-tune the model, our method outperforms the previous state-of-the-art method [22] by 50%. Finally, we present a video dataset of problem-solving interactions of children and show that finetuning the ATL-BP affect network using children face images further improves the performance. We summarize our contributions as follows:

- We present MathSpringSP+: a large labeled dataset of student interactions with an intelligent online math tutor consisting of 68 sessions, where 54 individual students solved 2,749 problems. This dataset includes two views of students solving each problem as well as *problem outcome labels* that describe the performance of the students on each individual problem. We will release this dataset upon acceptance.
- We present a transfer learning facial affect representation that can be used for behavior prediction tasks. This representation is learned from a large facial affect dataset.
- We are the first to model the temporal structure of video sequences of students solving math problems using a recurrent neural network architecture, improving performance on existing datasets.
- Our proposed *Affect Transfer Learning for Behavior Prediction* (ATL-BP) model outperforms the previous state-of-the-art method by 50%.
- We present a dataset of children problem-solving interactions collected in the same manner as MathSpringSP+ and show that finetuning the ATL-BP affect network using children face images further improves the performance on this target domain.

II. Related Work

a) ***Intelligent Tutoring Systems****:* Intelligent tutoring systems have been evaluated and shown to produce learning gains [14], [33], [34], [36], [45]. One meta-analysis shows test score improvements from the 50th to 75th percentile [29]. Some ITS have been shown to match the success of one-on-one human tutoring and students using these tutors outperform students from conventional classes in 92% of the controlled evaluations and perform twice as high as for students using typical (non-intelligent systems) [9], [18], [28].

There is a large amount of work that analyzes user affect, emotions and expressions from interactions with games or intelligent tutoring systems [2], [5], [10], [14], [13], [21], [23], [30], [37], [38], [41], [42], [46]. In certain cases the predicted affect information is used to improve learning. For example, Strain and D'Mello [44] have studied the role of emotion in ITS engagement, task persistence, and learning gain. Gaze prediction has also been used in an effort to respond to students' boredom and to perform interventions [14]. Further, relationships between visual facial Action Unit (AU) factors and self-reported traits such as academic effort, study habits, and interest in the subject have been studied [38].

In contrast to this body of work, our work focuses on using predicted deep affect embeddings that are learned from a large facial affect dataset to improve behavior prediction in an ITS. Behavior prediction can be useful in improving learning by tailoring the interventions of the ITS to the predicted actions of the student. To the best of our knowledge, our work is the first to use an affect embedding for behavior prediction in an ITS.

b) ***Interventions in an Online Tutor****:* Prior research has examined the impact of several interventions in ITS to improve student outcome and affect, specifically, affective messages delivered by avatars and empathetic messages that responded to students' recent emotions [47]. Interventions in the MathSpring ITS led to improved grades in state standardized exams [11] as well as influence students' perceptions of themselves as learners [24]. Empathetic characters which provide interventions generate superior results both to improve student interactions with the system, address negative

student emotions, and in the overall learning experience [27]. Predicting outcomes of problems for students is a valuable source of information for planning and executing ITS interventions for improving learning [12], [49]. For example the ITS could provide hints when the system predicts that the student will not be able to successfully complete the problem.

*c) **Predicting Exercise Outcome**:* Joshi et al. [22] presented a first attempt at tackling the problem of exercise outcome prediction. They did not explore deep learning representations but used traditional facial analysis features such as head pose, gaze and facial action units (AUs). They also did not attempt to model the temporal component of the videos, which is a rich source of information, and instead opted to summarize features from a video into one single feature vector. The method by Joshi et al. [22] can be considered the previous state of the art in student outcome prediction, and, thus, our experimental results include a performance comparison between this method and our models.

III. MATHSPRINGSP+ DATASET

In order to build an ITS capable of understanding student behavior and producing interventions, it is critical to build tailored datasets that allow development of behavior understanding techniques. To this end, in this work we expand the MathSpringSP dataset described by Joshi et al. [22], following the same data collection protocol. We name the extended dataset MathSpringSP+. MathSpringSP+ is roughly double the size of the original MathSpringSP dataset.

MathSpringSP+ consists of Webcam and GoPro videos that are recorded while college students solve math problems using the online tutor MathSpring [32] on a laptop. The webcam is positioned on the laptop and films the student at a frontal angle. The designated spot for the GoPro camera is above the notepad, to the front-right of the student. Figure 2 illustrates our data capturing setup from three viewpoints. Students work on solving math problems for 30–40 minutes or approximately 50 problems. The number of problems solved is variable between sessions depending on the rate at which each student solves problems. We divide each student's video session into shorter video segments, where each segment is associated with an individual math problem. Each math problem video clip has an associated problem outcome y, recorded in the log files of the ITS [22]. This problem outcome is automatically labeled by the software using a rule-based algorithm that chooses from the following seven possible student outcomes:

- ATT (attempted): Student did not see any hints and solved the problem after one incorrect attempt
- GIVEUP: Student tried to answer the problem or asked for a hint but ultimately skipped the problem
- GUESS: Student did not see hints, but solved the problem after more than one incorrect attempts
- NOTR (not read): Student performed some action, but the first action was too fast for the student to have read the problem
- SHINT (solved with hint): Student eventually submitted the correct answer after seeing one or more hints
- SKIP: Student skipped the problem without asking for a hint or attempting to answer the problem
- SOF (solved on the first attempt): Student answered correctly on the first attempt, without seeing any hints

TABLE I
SIZE COMPARISON OF OUR EXTENDED MATHSPRINGSP+ DATASET COMPARED TO MATHSPRINGSP

	MathSpringSP	MathSpringSP+
Individual Students	30	54
Student Sessions	38	68
Problem Samples	1,596	2,749

Examples of the variation in student facial expression throughout the process of answering problems in the math tutor are shown in Figure 3. We note that expressions can be very subtle. Expressions can also be ambiguous: a frown can mean that the student is very focused and will solve the problem correctly or that they are having difficulties with the problem. Expression intensities and variance depend on the individual, and it is challenging to generalize to different identities. Finally, our method has to deal with hand gestures, face occlusions and extreme pose changes, some of which are shown in Figure 3. A total of 24 students participated in the extended study, compared to 30 in the original study. 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. Several students participated in multiple sessions. Each session lasted approximately one hour. In total, 30 student sessions were recorded, which yielded 1,153 problem samples. Thus, the extended MathSpringSP+ dataset contains videos of a total of 54 unique students, 68 student sessions and 2,749 problem samples. This amount of data almost doubles the original MathSpringSP dataset, which contains 38 student sessions and 1,596 problem samples. A detailed breakdown of the relative sizes of MathSpringSP and MathSpringSP+ are shown in Table I.

IV. METHOD

The dataset consists of labeled video pairs $(\boldsymbol{X}, \boldsymbol{y})$, where the video $\boldsymbol{X}$ is a time series of RGB frames $\boldsymbol{X} = \{\boldsymbol{X}_t \mid t = 1..T\}$ of a student solving a problem, and the scalar label $\boldsymbol{y}$ indicates the outcome class for that problem. The task at hand is a 7-label classification problem, i.e., $y \in \{1, ..., C\}$, for $C = 7$.

Our challenge was to work out how to leverage state-of-the-art affect recognition techniques to compute an output label y from the input video $\boldsymbol{X}$. Affect recognition models provide affect estimates from images of faces that typically show strong emotions, e.g., the disgust expressed in the women's face on the left in Fig.1. We decided to use a ResNet-50 network [19] and the AffectNet dataset [35], which contains more than one million facial images collected "in the wild" from the Internet, to solve the source domain problem of predicting eight emotions, neutral, happiness, sadness, surprise, fear, disgust, anger, contempt, plus the

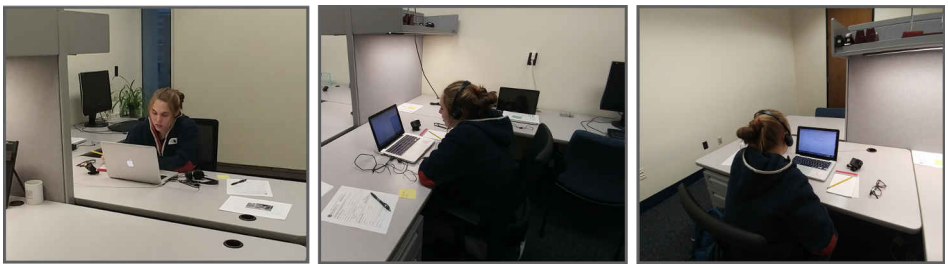

Fig. 2. Data capture setup for the MathSpringSP+ dataset from three views (front, side and back). The student completes problems on a laptop. The laptop webcam and Go-Pro camera on the right side of the student are used to capture the student's upper body and face during the completion of problems.

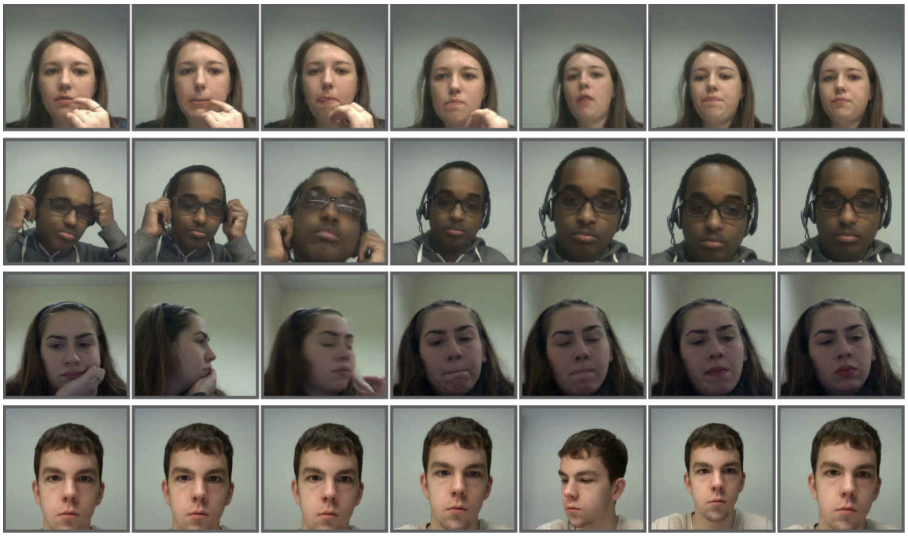

Fig. 3. Example face-cropped images from the MathSpringSP+ dataset showing the evolution of student expressions. In particular we notice changes in head pose, hand gestures, face occlusion and facial gestures throughout the videos. Expressions in videos can be very subtle, as well as ambiguous, making the prediction problem challenging.

two classes, uncertain, and non-face. We employ this trained affect network to solve the target-domain problem of student outcome prediction.

The proposed ATL-BP model consists of three main components, see Figure 1, the affect network, a facial analysis network, and an LSTM. We also study variants of our model by either removing the affect network or replacing it with a face recognition network.

First, from the last layer of the trained affect network, ATL-BP extracts a fixed-size embedding of size 8,192, computed for each frame $\boldsymbol{X}_t$, and compresses it into a lower-dimensional vector $\rho(\boldsymbol{X}_t)$ by learning the weights for a fully-connected neural network layer c_a (Fig. 1, magenta).

Second, ATL-BP uses a facial analysis model to extract facial Action Unit (AU) presence and intensity, gaze direction, and head pose for each frame $\boldsymbol{X}_t$. We note these traditional facial analysis features as $\psi(\boldsymbol{X}_t)$ (Fig. 1, green). We chose the OpenFace 2.0 model [6] to compute student head position, head pose, gaze, facial AU presence, and facial AU intensity from individual frames in each video segment.

For our main ATL-BP model we devised a feature representation that is based on concatenating the outputs of our proposed affect representation and the facial analysis components:

$$\phi(\boldsymbol{X}_t) = c_a(\rho(\boldsymbol{X}_t)) \oplus \psi(\boldsymbol{X}_t),$$

where $\oplus$ is the concatenation operation for vectors. The compressed embedding $c_a(\rho(\boldsymbol{X}_t))$ is 100-dimensional. The full feature vector $\phi(\boldsymbol{X}_t)$ has dimension 149 for every frame $video_t$.

For our model variants, we replace the affect network by a face recognition model in order to extract face related features. We selected the pre-trained VGG-Face network [40], which computes an embedding ξ of dimension 2,622. ATL-BP compresses the feature representation $\xi(\boldsymbol{X}_t)$, computed by this network for each video frame $\boldsymbol{X}_t$, using another fully-connected layer c_v, into $c_v(\xi(\boldsymbol{X}_t))$.

Finally, in order to model the temporal nature of the videos, we designed a unidirectional 2-layer LSTM classifier h_θ with 200 hidden units that processes the feature vector $\phi(\boldsymbol{X}_t)$ frame by frame and produces the final estimate of student outcome y (Fig. 1, orange).

V. Experiments

We present experiments on problem outcome prediction on the MathSpringSP+ dataset. These experiments study our contributions, which include incorporating temporal information from video streams by using an LSTM and using our affect transfer learning representation. The experiments also show how user-personalized training unlocks the effectiveness of our affect representation. We also study early

prediction as well as present ablation studies for the dimensionality reduction that is accomplished by the proposed fully-connected layer. In this work we limit ourselves to the webcam video stream of the student.

A. Model Training

a) ***Training the Affect Representation Network****:* For source domain affect training, we selected a ResNet-50 network. We pre-trained the affect network on a subset of 50,000 randomly sampled images from the AffectNet dataset and validated the network on 5,000 randomly selected images. We limited ourselves to a subset since the dataset contains more than one million examples. Note that our training and validation data subsets are not the same as used by [35]. On our subset, our network achieves a mean accuracy of 47.3%, which is close to the accuracy reported by [35] on their skew-normalized validation set of 54%, and much higher than the random baseline of 9.0%. The relatively low accuracy scores can be contributed to a data that is unbalanced, noisy, and overall challenging.

We extracted the target domain affect features from our videos by performing inference of the affect network on every frame. We chose a granularity of three frames per second, down from 30 frames per second in our videos, in order to save on processing time and storage space. We found that this granularity was a good compromise between performance and cost. The affect network uses each frame as an input and the last-layer features are extracted as a vector of dimension $8,192$.

We trained the affect network with the Adam optimizer with a learning rate of 3×10^{-4}, β_1 of 0.9, and β_2 of 0.999. The standard batch normalization layers of the ResNet-50 were used and fixed throughout training.

b) ***Training ATL-BP to Predict Student Exercise Outcome****:* For each frame used, the feature vector computed is $\phi(\boldsymbol{X}_t) = \psi(\boldsymbol{X}_t) \oplus c_a(\rho(\boldsymbol{X}_t)) \oplus c_v(\xi(\boldsymbol{X}_t))$. We observed that the dimensionality reduction due to the compression layer stabilizes training and improves performance. The feature vector ϕ is used to train the LSTM with two stacked layers. Specifically, at each instant t, features $\phi(\boldsymbol{X}_t)$ are fed to the LSTM. The LSTM is trained on all the video segments. It outputs a class probability for each problem outcome. The LSTM is trained using the cross-entropy loss function. The Adam optimizer is used for training. We use a learning rate of 3×10^{-5} for 30 epochs, and a batch size of 1.

B. Experimental Setup for Testing

a) ***Model Variations for Testing****:* In addition to our main proposed ATL-BP, shown in Figure 1 and which we call "ATL-BP with affect embedding" for clarity, we implemented and test two variants of ATL-BP. The first variant is *ATL-BP without transfer learning.* In this model, the LSTM directly interprets the output ψ of the facial analysis network and does not use the embedding scheme we propose in this work. The second variant is *ATL-BP with VGG-Face embedding.* In this model, the LSTM interprets the output $c_v(\xi(\boldsymbol{X}_t))$ concatenated with the output ψ of the facial analysis network.

Furthermore, for comparison baselines, we reproduced the method described by Joshi et al. [22] and show results for a majority vote classifier. The majority vote classifier simply selects the most prevalent class in our dataset, "Solved on First Try," for every video.

b) ***Random Dataset Split****:* Following the experimental setup in [22], we performed five-fold cross validation on our dataset by randomly shuffling video segments and constructing five different train and test splits. The train splits contain 80% of the data while the test splits contain the rest.

Experiments conducted using this random splitting experimental setup cannot reliably measure generalization to new users since videos of problems from the same student can be present in both the training and test set. This means that the network does not have to learn how to generalize to a completely new identity. We propose an improved experimental setup next.

c) ***User Generalization Split****:* In order to test generalization to new users we propose a leave-users-out experimental setup where users are exclusively split into either the training or test set. In other words, we enforce the rule that no video clips of the same user can be in both the test and training sets. In this manner we can measure how the system performs when applied to an unseen user. This is a substantially more challenging task since the network has to generalize to new identities and features. We suggest that all future research on this dataset use this type of setup. We created five leave-users-out splits for five-fold cross-validation and train different model variations for each split.

C. Results and Discussion

a) ***ATL-BP Results for Random Splits****:* Using the experimental protocol of a random dataset split, our ATL-BP for problem outcome prediction on MathSpringSP+ achieves an accuracy of 60.2% (Table II). Compared to the previous state-of-the-art method [22], this is an increase of 14 percent points (pp) in accuracy. ATL-BP also achieves a 44% relative increase in mean F-score improving from 0.238 to 0.330. The mean F-score is computed by first computing the individual F-score for all classes and averaging over all classes. By comparing the results for ATL-BP without transfer learning and those by Joshi et al. [22], we can see that by integrating an LSTM architecture that allows for modeling the temporal component in the videos we can achieve a marked increase in performance (5.6 pp). We achieve a further increase in performance by using deep embeddings (8.6 pp for using the VGG-Face embedding ξ), and especially our proposed affect embedding ψ (as mentioned, 14 pp).

b) ***MathSpringSP Results****:* We conducted experiments on the original MathSpringSP dataset in order to verify that our proposed ATL-BP model with affect embeddings achieves improved results in the same testing environment presented by Joshi et al. [22]. Our results show a consistent improvement in mean F-score and accuracy of our method (Table III).

TABLE II

RESULTS FOR PROBLEM OUTCOME PREDICTION ON THE MATHSPRINGSP+ DATASET USING FIVE-FOLD CROSS-VALIDATION AND RANDOM DATA SPLITS

Method	Mean F-Score	Accuracy
Majority Vote Classifier	0.103	56.1%
Joshi et al. [22]	0.228	46.2%
ATL-BP w/o transfer learning	0.295	51.8%
ATL-BP w/ VGG-Face embedding	0.304	54.8%
ATL-BP w/ affect embedding	**0.330**	**60.2%**

TABLE III

RESULTS FOR PROBLEM OUTCOME PREDICTION ON THE ORIGINAL MATHSPRINGSP FOR ATL-BP FOLLOWING THE DATA SETUP FROM JOSHI ET AL. [22]

Method	Mean F-Score	Accuracy
Joshi et al. [22]	0.270	54.0%
ATL-BP w/ affect embedding	**0.362**	**61.0%**

*c) **Early Prediction of Problem Outcome**:* We experimented with obtaining prediction using only the five first seconds of each video clip (Table IV). Early outcome prediction is important since the ITS should have time to react and deliver the intervention should it be decided to do so. It turns out that is straightforward to do early prediction using an LSTM since it outputs a prediction at every time step, as opposed to the method proposed by Joshi et al. [22], where each video has to be summarized into a fixed-sized vector before being fed into a multilayer perceptron. We observe that ATL-BP achieves a large increase (6.7 pp) in performance over [22]. ATL-BP without transfer learning obtains the best F-score (0.295) in this experimental setup.

*d) **Deep Embedding Dimensionality Reduction**:* We performed an ablation study on the fully-connected layer that is used for reducing the dimensionality of the deep embeddings that are used as inputs for our LSTM architecture (Table V). While the mean F-score does not change on both the VGG-Face and proposed affect embedding ATL-BP variants, dimensionality reduction does improve the accuracy of the models by 3.5 pp and 1.5 pp, respectively.

*e) **ATL-BP Results for User Generalization**:* For the user generalization split of the training and testing data, we report the mean F-score and mean accuracy in Table VI for the "Majority Vote Classifier" benchmark, Joshi et al. [22] and our proposed model with different combinations of embeddings. We observe that the temporal modeling improves results from Joshi et al. [22] substantially (12.1 pp in accuracy). We observe that ATL-BP without transfer learning outperforms the ATL-BP version with our proposed affect embedding with regards to the F1 score. We hypothesize that leveraging affect embeddings is more difficult in this setup since the model does not have access to baseline levels of expression for each user.

TABLE IV

RESULTS FOR EARLY PREDICTION OF PROBLEM OUTCOME USING ONLY THE FIRST FIVE SECONDS OF VIDEO FOOTAGE ON THE MATHSPRINGSP+ DATASET (FIVE-FOLD CROSS-VALIDATION, RANDOM DATA SPLITS).

Method	Mean F-Score	Accuracy
Majority Vote Classifier	0.103	56.1%
Joshi et al. [22]	0.173	46.7%
ATL-BP w/o transfer learning	**0.295**	51.8%
ATL-BP w/ VGG-Face embedding	0.239	47.0%
ATL-BP w/ affect embedding	0.270	**53.4%**

TABLE V

EMBEDDING DIMENSIONALITY REDUCTION ABLATION STUDY. WE SHOW RESULTS FOR PROBLEM OUTCOME PREDICTION ON THE MATHSPRINGSP+ DATASET USING FIVE-FOLD CROSS-VALIDATION AND RANDOM DATA SPLITS

Method	Mean F-Score	Accuracy
ATL-BP w/ VGG-Face	0.304	51.3%
ATL-BP w/ VGG-Face & dim. reduction	0.304	**54.8%**
ATL-BP w/ affect	0.330	58.7%
ATL-BP w/ affect & dim. reduction	0.330	**60.2%**

TABLE VI

RESULTS FOR PROBLEM OUTCOME PREDICTION ON THE MATHSPRINGSP+ DATASET USING FIVE-FOLD CROSS-VALIDATION AND THE MORE CHALLENGING LEAVE-USERS-OUT SPLITS

Method	Mean F-Score	Accuracy
Majority Vote Classifier	0.102	55.9%
Joshi et al. [22]	0.182	41.9%
ATL-BP w/o transfer learning	**0.270**	50.3%
ATL-BP w/ VGG-Face embedding	0.246	51.8%
ATL-BP w/ affect embedding	0.251	**54.0%**

*f) **Personalization of Prediction**:* An effective real-time tutoring system would benefit from personalizing its prediction using initial data captured from a specific user stream. People have different emotional and expression baselines that can be learned using data collected in a trial run of the system. Specifically, we want the model to act on the variations of our affect embedding compared to the mean affect embedding, since each person will have a different baseline expression and thus a different baseline affect embedding. The model does not have any way to integrate this information without it being personalized for each user.

We propose a personalization scheme in which our system can be tailored to individual users and can fully utilize our proposed affect embedding. In this scheme, the network is fine-tuned on the initial problems corresponding to 20% of the session for users in the test set for 30 epochs. Our experiments show that user personalization unlocks the potential of the affect features (Table VII). ATL-BP with affect embedding achieves the highest F-score of 0.308 and the highest accuracy of 55.1% compared to the other methods. Our full method achieves a relative increase of 50% in mean F-score as well as an absolute increase in accuracy of more than 11 pp compared to the previous state of the art [22]. Our full method also outperforms variants of ATL-BP, which do not use our proposed affect representation.

*g) **Outcome Prediction for Children**:* As a final experiment we tested our method on a new dataset of children

TABLE VII

RESULTS FOR PROBLEM OUTCOME PREDICTION (7-CLASSES) ON THE MATHSPRINGSP+ DATASET AFTER USER PERSONALIZATION (FIVE-FOLD CROSS-VALIDATION AND LEAVE-USER-OUT SPLITS)

Method	Mean F-Score	Accuracy
Majority Vote Classifier	0.090	45.3%
Joshi et al. [22]	0.206	43.8%
ATL-BP w/o transfer learning	0.278	48.4%
ATL-BP w/ VGG-Face embedding	0.262	48.7%
ATL-BP w/ affect embedding	**0.308**	**55.1%**

TABLE VIII

RESULTS FOR PROBLEM OUTCOME PREDICTION (7-CLASSES) ON THE CHILDREN DATASET (FIVE-FOLD CROSS-VALIDATION, RANDOM DATA SPLITS).

Method	Mean F-Score	Accuracy
Majority Vote Classifier	0.070	32.3%
Joshi et al. [22]	0.202	32.0%
ATL-BP w/o transfer learning	0.238	33.4%
ATL-BP w/ affect embedding	0.260	39.6%
ATL-BP w/ LIRIS children affect embedding	0.272	45.2%
ATL-BP w/ CAFE children affect embedding	0.273	44.4%
ATL-BP w/ LIRIS+CAFE affect embedding	**0.278**	**45.2%**

working on math problems. Following the same data collection protocol as MathSpringSP+, we collected 968 recorded problem-solving interaction samples of fifty-one K12 students who used MathSpring. We show some extracted frames from the dataset in Figure 4. Results on this Children dataset show that our model consistently outperforms the baseline and previous state-of-the-art method (Table VIII).

Since the AffectNet dataset mainly captures facial expressions of adults, we further finetuned the affect representation network using two datasets of children facial expressions, LIRIS [26] and CAFE [31], in order to tailor the model specifically for children. LIRIS contains 208 video clips of 6-to-12-year-old children showing six basic spontaneous facial expressions, while CAFE dataset contains 1,192 images of 2-to-8-year-old children posing for seven facial expressions. We trained three variants of models using LIRIS only (frames), CAFE only, and a combination of both datasets. The best model among the three achieves the highest accuracy (45.2%) and mean F-score (0.278), improving on the previous state-of-the-art [22] (13.2 pp absolute increase in accuracy and 38% relative increase in mean F-score) on the challenging task of predicting future outcome using only student face movements and gestures. The prediction task has 7 classes which contributes to the difficulty.

VI. CONCLUSION

We introduced a large labeled dataset of student interactions with an intelligent online math tutor that consists of 68 sessions, where 54 individual students solved 2,749 math problems. Using this dataset we design a transfer learning model ATL-BP that improves problem outcome predictions for students interacting with the ITS and answering math problems. By modeling the temporal structure of the videos with ATL-BP, we achieved a substantial increase in classification F-score and accuracy compared to previous state of the art in this task. Additionally, using a novel affect representation along with user personalization, we achieved a further increase in performance. More generally, these promising results suggest that leveraging affect representations might be valuable in behavior analysis applications more generally. Our final method achieves a 50% relative increase in mean F-score as well as an absolute 11 percentage point increase in accuracy compared to previous work. Finally, we collect a dataset of children student interactions and present results on this dataset. We show that finetuning of the Affect network with age-appropriate images and video further improves performance in this scenario. These results pave the way for future improvements in solutions for this task. Future tutor systems may use our proposed outcome prediction model in order to deliver real-time interventions to improve the learning of students.

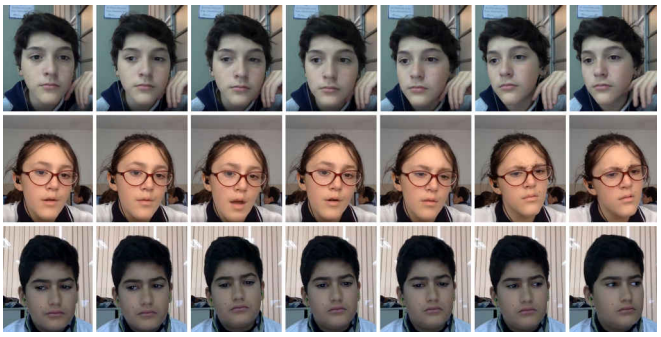

Fig. 4. Example face-cropped images from the Children dataset showing the evolution of student expressions.

ACKNOWLEDGEMENTS

We thank the participants of our experimental study and acknowledge partial funding for this work by the National Science Foundation, grant 1551572.

REFERENCES

[1] T. Almaev, B. Martinez, and M. Valstar. Learning to transfer: transferring latent task structures and its application to person-specific facial action unit detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3774–3782, 2015.

[2] S. Amershi, C. Conati, and H. Maclaren. Using feature selection and unsupervised clustering to identify affective expressions in educational games. In *Workshop on Motivational and Affective Issues in ITS. 8th International Conference on ITS*, pages 1–8, 2016.

[3] I. Arroyo, D. G. Cooper, W. Burleson, B. P. Woolf, K. Muldner, and R. Christopherson. Emotion sensors go to school. In *Proceedings of the 2009 conference on Artificial Intelligence in Education: Building Learning Systems that Care: From Knowledge Representation to Affective Modelling*, pages 17–24. IOS Press, 2009.

[4] I. Arroyo, B. P. Woolf, W. Burelson, K. Muldner, D. Rai, and M. 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.

[5] R. S. Baker, S. K. D'Mello, M. M. T. Rodrigo, and 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(4):223–241, 2010.

[6] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L. Morency. OpenFace 2.0: Facial behavior analysis toolkit. In *13th IEEE International Conference on Automatic Face & Gesture Recognition, FG 2018, Xi'an, China, May 15-19, 2018*, pages 59–66, 2018.

[7] J. Chen, X. Liu, P. Tu, and A. Aragones. Person-specific expression recognition with transfer learning. In *The 19th IEEE International Conference on Image Processing*, pages 2621–2624. IEEE, 2012.

[8] J. Chen, X. Liu, P. Tu, and A. Aragones. Learning person-specific models for facial expression and action unit recognition. *Pattern Recognition Letters*, 34(15):1964–1970, 2013.

[9] A. T. Corbett and J. R. Anderson. Student modeling and mastery learning in a computer-based programming tutor. In *International conference on intelligent tutoring systems*, pages 413–420. Springer, 1992.

[10] S. Craig, A. Graesser, J. Sullins, and B. Gholson. Affect and learning: an exploratory look into the role of affect in learning with autotutor. *Journal of educational media*, 29(3):241–250, 2004.

[11] S. D. Craig, A. C. Graesser, and R. S. Perez. Advances from the Office of Naval Research STEM grand challenge: Expanding the boundaries of intelligent tutoring systems. *International Journal of STEM Education*, 5(1):11, 2018.

[12] K. Delgado, J. M. Origgi, T. Hasanpoor, H. Yu, D. A. Allessio, I. Arroyo, W. Lee, M. Betke, B. P. Woolf, and S. A. Bargal. Student engagement dataset. In *ICCV 2021: 2nd Workshop and Competition on Affective Behavior Analysis in-the-wild (ABAW)*, pages 1–9, Oct. 2021.

[13] S. D'Mello, R. W. Picard, and A. Graesser. Toward an affect-sensitive autotutor. *IEEE Intelligent Systems*, 22(4):53–61, 2007.

[14] S. K. D'Mello, A. Olney, C. Williams, and P. Hays. Gaze Tutor: A Gaze-reactive Intelligent Tutoring system. *Int. J. Hum.-Comput. Stud.*, 5:377–398, 2012.

[15] I. Dua, A. U. Nambi, C. Jawahar, and V. Padmanabhan. Autorate: How attentive is the driver? In *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, pages 1–8. IEEE, 2019.

[16] S. D'Mello, B. Lehman, J. Sullins, R. Daigle, R. Combs, K. Vogt, L. Perkins, and A. Graesser. A time for emoting: When affect-sensitivity is and isn't effective at promoting deep learning. In *International conference on intelligent tutoring systems*, pages 245–254. Springer, 2010.

[17] G. Gordon, S. Spaulding, J. K. Westlund, J. J. Lee, L. Plummer, M. Martinez, M. Das, and C. Breazeal. Affective personalization of a social robot tutor for children's second language skills. In *Thirtieth AAAI Conference on Artificial Intelligence*, pages 3951–3957, 2016.

[18] A. C. Graesser, K. VanLehn, C. P. Rosé, P. W. Jordan, and D. Harter. Intelligent tutoring systems with conversational dialogue. *AI magazine*, 22(4):39–39, 2001.

[19] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[20] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[21] N. Jaques, C. Conati, J. M. Harley, and R. Azevedo. Predicting affect from gaze data during interaction with an intelligent tutoring system. In *International conference on intelligent tutoring systems*, pages 29–38. Springer, 2014.

[22] A. Joshi, D. Allessio, J. Magee, J. Whitehill, I. Arroyo, B. Woolf, S. Sclaroff, and M. Betke. Affect-driven learning outcomes prediction in intelligent tutoring systems. In *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, pages 1–5. IEEE, 2019.

[23] A. Kapoor, W. Burleson, and R. W. Picard. Automatic prediction of frustration. *International journal of human-computer studies*, 65(8):724–736, 2007.

[24] S. Karumbaiah, R. Lizarralde, D. Allessio, B. P. Woolf, I. Arroyo, and N. Wixon. Addressing student behavior and affect with empathy and growth mindset. In *Proceedings of the 10th International Conference on Educational Data Mining (EDM)*, pages 96–103, 2017.

[25] H. Kaya, F. Gürpınar, and A. A. Salah. Video-based emotion recognition in the wild using deep transfer learning and score fusion. *Image and Vision Computing*, 65:66–75, 2017.

[26] R. A. Khan, A. Crenn, A. Meyer, and S. Bouakaz. A novel database of children's spontaneous facial expressions (liris-cse). *Image and Vision Computing*, 83:61–69, 2019.

[27] 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*, pages 9–16, 2005.

[28] J. Kulik and J. D. Fletcher. Effectiveness of intelligent tutoring systems: A meta-analytic review. *Review of Educational Research*, 86, 04 2015.

[29] J. A. Kulik and J. Fletcher. Effectiveness of intelligent tutoring systems: a meta-analytic review. *Review of Educational Research*, 86(1):42–78, 2016.

[30] S. Lallé, C. Conati, and R. Azevedo. Prediction of student achievement goals and emotion valence during interaction with pedagogical agents. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, pages 1222–1231, 2018.

[31] V. LoBue and C. Thrasher. The child affective facial expression (cafe) set: Validity and reliability from untrained adults. *Frontiers in psychology*, 5:1532, 2015.

[32] MathSpring math tutor. https://www.mathspring.org Last accessed on October 15, 2018.

[33] M. Mayo and A. Mitrovic. *International Conference on Artificial Intelligence in Education AIED 2020: Artificial Intelligence in Education*. International Artificial Intelligence Education Society, 2001.

[34] A. Mitrović and J. Holland. Effect of non-mandatory use of an intelligent tutoring system on students' learning. In *International Conference on Artificial Intelligence in Education*, pages 386–397. Springer, 2020.

[35] A. Mollahosseini, B. Hassani, and M. H. Mahoor. AffectNet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, abs/1708.03985, 2017.

[36] A. L. Mondragon, R. Nkambou, and P. Poirier. Evaluating the effectiveness of an affective tutoring agent in specialized education. In *European conference on technology enhanced learning*, pages 446–452. Springer, 2016.

[37] R. Nkambou. A framework for affective intelligent tutoring systems. In *2006 7th International Conference on Information Technology Based Higher Education and Training*, pages nil2–nil8. IEEE, 2006.

[38] B. D. Nye, S. Karumbaiah, S. T. Tokel, M. G. Core, G. Stratou, D. Auerbach, and K. Georgila. Engaging with the scenario: Affect and facial patterns from a scenario-based intelligent tutoring system. In *International Conference on Artificial Intelligence in Education*, pages 352–366. Springer, 2018.

[39] S. J. Pan and Q. Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2009.

[40] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In X. Xie, M. W. Jones, and G. K. L. Tam, editors, *Proceedings of the British Machine Vision Conference (BMVC)*, pages 41.1–41.12. BMVA Press, September 2015.

[41] R. W. Picard. Affective computing: From laughter to IEEE. *IEEE Transactions on Affective Computing*, 1(1):11–17, 2010.

[42] R. W. Picard, E. Vyzas, and J. Healey. Toward machine emotional intelligence: analysis of affective physiological state. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(10):1175–1191, 2001.

[43] E. Sangineto, G. Zen, E. Ricci, and N. Sebe. We are not all equal: Personalizing models for facial expression analysis with transductive parameter transfer. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 357–366. ACM, 2014.

[44] A. C. Strain and S. K. D'Mello. Emotion regulation during learning. In *International Conference on Artificial Intelligence in Education*, pages 566–568. Springer, 2011.

[45] K. VanLehn. The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational Psychologist*, 46(4):197–221, 2011.

[46] B. Woolf, W. Burleson, I. Arroyo, T. Dragon, D. Cooper, and R. Picard. Affect-aware tutors: recognising and responding to student affect. *International Journal of Learning Technology*, 4(3-4):129–164, 2009.

[47] B. P. Woolf, I. Arroyo, K. Muldner, W. Burleson, D. G. Cooper, R. Dolan, and R. M. Christopherson. The effect of motivational learning companions on low achieving students and students with disabilities. In *International Conference on Intelligent Tutoring Systems*, pages 327–337. Springer, 2010.

[48] M. Xu, W. Cheng, Q. Zhao, L. Ma, and F. Xu. Facial expression recognition based on transfer learning from deep convolutional networks. In *2015 11th International Conference on Natural Computation (ICNC)*, pages 702–708. IEEE, 2015.

[49] H. Yu, A. Gupta, W. Lee, I. Arroyo, M. Betke, D. A. Allesio, T. Murray, J. J. Magee, and B. P. Woolf. Measuring and integrating facial expressions and head pose as indicators of engagement and affect in tutoring systems. In *Adaptive Instructional Systems. Adaptation Strategies and Methods, Third International Conference, AIS 2021, Held as Part of the 23rd HCI International Conference, HCII 2021, Virtual Event, July 24–29, 2021, Proceedings, Part II*, pages 219–233, 2021.

[50] G. Zen, L. Porzi, E. Sangineto, E. Ricci, and N. Sebe. Learning personalized models for facial expression analysis and gesture recognition. *IEEE Transactions on Multimedia*, 18(4):775–788, 2016.