Toward Automated Classroom Observation: Predicting Positive and Negative Climate

Anand Ramakrishnan[1], Erin Ottmar[1], Jennifer LoCasale-Crouch[2], and Jacob Whitehill[1]
[1] Worcester Polytechnic Institute, Worcester, MA, USA
[2] University of Virginia, Charlottesville, VA, USA

Abstract— **We devised and evaluated a multi-modal machine learning-based system to analyze videos of school classrooms for "positive climate" and "negative climate", which are two dimensions of the Classroom Assessment Scoring System (CLASS) [1]. School classrooms are highly cluttered audiovisual scenes containing many overlapping faces and voices. Due to the difficulty of labeling them (reliable coding requires weeks of training) and their sensitive nature (students and teachers may be in stressful or potentially embarrassing situations), CLASS-labeled classroom video datasets are scarce, and their labels are sparse (just a few labels per 15-minute video clip). Thus, the overarching challenge was how to harness modern deep perceptual architectures despite the paucity of labeled data. Through training low-level CNN-based facial attribute detectors (facial expression & adult/child) as well as a direct audio-to-climate regressor, and by integrating low-level information over time using a Bi-LSTM, we constructed automated detectors of positive and negative classroom climate with accuracy (10-fold cross-validation Pearson correlation on 241 CLASS-labeled videos) of 0.40 and 0.51, respectively. These numbers are superior to what we obtained using shallower architectures. This work represents the first automated system designed to detect specific dimensions of the CLASS.**

I. Introduction

One of the most important variables of school classrooms that predict students' learning is the nature and quality of *interactions* between teachers and their students. Numerous correlational [2], [3], [4], [5], [6]) and some large-scale causal [7], [8] studies have demonstrated the link between emotional and instructional support in the classroom and children's cognitive, social, and emotional skills.

In order to characterize classroom interactions precisely, educational researchers have developed a variety of classroom observation protocols. One of the most widely used protocols is the Classroom Assessment Scoring System [1] (CLASS). The CLASS measures the quality of teacher-student interactions within ten different dimensions, including (1) **positive climate**, which measures the "warmth, respect, and enjoyment communicated by verbal and nonverbal interactions" between students and teachers; and (2) **negative climate**, which measures the "overall level of expressed negativity in the classroom" [1].

A typical CLASS coding session requires human coders – who could be teachers, educational researchers, or school administrators – to examine specific characteristics of the states, actions, and interactions among the classroom participants (children and teachers) during the class session. Classroom interactions can be scored either during live observation or by watching recorded videos. CLASS coders assign a *single* score (1-7 scale) for each dimension to each 15-minute chunk of video; hence, CLASS codes are very sparse.

The official CLASS manual provides guidelines for how to summarize the discrete observations into an aggregate score (see Table I). Each judgment is based on the relative presence or absence of *behavioral markers* that belong to a specific *indicator* of a particular CLASS dimension; in this sense, CLASS is organized hierarchically. The behavioral markers can span auditory, visual, linguistic, and pedagogical dimensions. For example, when assessing positive climate, CLASS coders are instructed to consider how frequently smiles are exhibited by classroom participants; whether the teacher calls his/her children by name and looks them in the eye; whether the emotions between teachers and students are congruent; etc. Negative climate can be signified when a teacher raises his/her voice in anger at a student; makes threats to punish them if they do not behave; etc. Table I shows a small subset of the behavioral markers to which CLASS coders should attend for positive and negative climate.

Classroom observation protocols such as the CLASS are invaluable for providing feedback to teachers on what went well in their teaching and what could be improved. However, manual classroom observation has important limitations (1) *Cost*: Careful coding of videos is very laborious and can take many person-hours and cost hundreds of dollars per day. (2) *Reliability*: Scores often vary significantly across coders, and multiple codes per video must be collected to obtain a reliable estimate. Coders are also prone to form early judgments based on just a few minutes of video (primacy effects [9]) and may be reluctant to change their minds [10]. (3) *Temporal resolution*: classroom observation videos are typically scored in relatively long (15-20 minute) chunks, partially due to the high cost of coding, which may not be ideal for giving teachers specific feedback on how to improve their teaching. These limitations, along with dramatic advances in machine learning and deep learning during the past 5 years, raise the question: Could machine perception be harnessed to enable more precise, efficient, reliable, and fine-grained feedback to teachers? The **goal of this paper** is to take some steps toward answering this question by developing a machine learning architecture to recognize CLASS positive climate and negative climate.

This material is based upon work supported by the National Science Foundation under Grant Nos. #1551594 and #1822768, and by Spencer Small Research Grant No. #201800131

978-1-7281-0089-0/19/$31.00 ©2019 IEEE

Structural challenges: Two inter-related challenges posed by this task are *data privacy* and *data scarcity*: Classroom observation videos are sensitive – they show adults and children in sometimes emotionally distressed situations. For example, especially in preschool classrooms (as we analyze in this study), a child may start crying or screaming, or a teacher may occasionally become visibly frustrated in front of his/her classroom. Moreover, careful coding of classroom observation videos – whether with the CLASS or another protocol (e.g., Framework for Teaching [11], UTeach [12]) – is highly laborious and requires coders to undergo many weeks of training or more. Both these factors mean that the available video datasets for training and testing are typically modest in size (a few hundred videos), and that researchers are very hesitant to share them with others. One of the overarching research questions we tackled in this study was how to harness modern deep learning perceptual architectures, via a combination of transfer learning from public datasets and fine-tuning on a modest-sized classroom video dataset, to predict CLASS positive and negative climate accurately.

II. Related Work

There has been substantial prior work [13], [14], [15], [16], [17], [18], [19], [20], [21] on using machine learning to analyze learners' affective states, from either video of students' faces (e.g., [13], [15], [14], [16]) or log files of students' interactions (e.g., [19], [20], [21]). Much of this work has focused on intelligent tutoring systems (ITS), in which each student mostly interacts with the computer alone, without much interaction with others.

More recently, researchers in multi-modal machine learning and educational data mining have investigated how to characterize the dynamics of an entire classroom. D'Mello, Donnelly, and colleagues [22], [23], for example, have recently explored how to segment and recognize students' and teachers' speech in unconstrained classrooms based on different configurations of Kinect cameras. Wang and colleagues [24] obtained high accuracy in segmenting teachers' speech by deploying small wearable recording devices in math classrooms. For the specific application of automated classroom observation scoring, we are only aware of one prior work: Qiao and Beling [25] developed a computer vision system, optimized within a multiple-instance learning framework, to estimate which 3-minute snippets of classroom videos were most relevant for CLASS coders to code manually. However, in contrast to our study, however, their algorithm did not actually predict the CLASS scores themselves.

III. High-Level Approach

In our study we focus on automatically estimating the positive climate and negative climate dimensions of the CLASS (see Table I). Unlike simpler labeling tasks such as smile, anger, crying, etc., the CLASS dimensions are high-level semantic states that are evaluated holistically by watching an entire 15-minute video segment and making an overall judgment. In our exploratory work – in fact, to the best of our knowledge ours is the first study to attempt to measure CLASS dimensions automatically using computer vision and computer audition – we used a hierarchical, multimodal machine learning approach consisting of two broad classes of features (see Figure 1 for an overview):

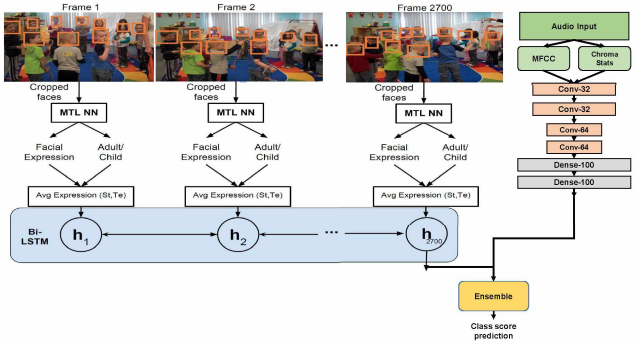

Fig. 1: Ensemble of visual and audio-based models to better predict CLASS climate. The left side of the ensemble shows the architecture which predicts CLASS climate scores using an LSTM network on top of facial expression scores (smile, sadness, anger). The right side of the ensemble shows the architecture that predicts CLASS climate scores using a convolutional neural network on top of MFCC and Chroma features extracted from audio. In the left part of the figure, St=student, Te=teacher.

Positive Climate	
Indicators	**Behavioral Markers**
Relationships	Physical proximity, matched affect
Positive Affect	Smiling, laughter
Respect	Eye contact, warm voice
Negative Climate	
Indicators	**Behavioral Markers**
Negative Affect	Irritability, harsh voice, anger
Punitive Control	Yelling, threats
Teacher Negativity	Sarcastic voice, humiliation
Child Negativity	Victimization, bullying

TABLE I: Positive and negative climate dimensions of the Classroom Assessment Scoring System (CLASS) [1]. Each dimension is sub-defined in terms of indicators, each of which has multiple behavioral markers.

Visual features: There are a variety of visual behavioral markers that suggest positive climate. For instance, positive affect is signaled (in part) by smiling and laughter, and positive relationships are associated with *congruent* facial expression between the teacher and her/his students, i.e., the teacher shows positive emotion when the students show positive emotion. Similarly, overt displays of anger, frustration, or sarcasm indicate negative climate. Although automatic facial expression recognition is a mature field, the vast majority of available systems (e.g., Amazon Rekognition, or OpenFace [26]) are trained mostly on adult faces. In contrast, our population consists of very young (2-3 years old) children. Both due to the demographics of the target population and to data privacy constraints (which precludes the use of a cloud-based facial expression recognition service trained on millions of faces), we decided to train custom CNN-based

Fig. 2: Three representative images for the "unclear" label for smile/non-smile annotation task.

binary detectors of facial expression (smile, sadness, and anger), as well as child/adult (useful for determining who is the teacher in a classroom based just on the face). We then integrate the facial expression estimates – in particular, the average (within each video frame) smile, sadness, and anger scores of all the adults, as well as the average expression scores of all the children – over time using (Bi-)LSTM neural networks.

Auditory features: Low-level audio features can capture paralinguistic and prosodic features such as sarcasm, laughter, yelling, screaming, crying, etc.; see Table I. In this work we focus on how low-level auditory features can directly predict CLASS positive and negative climate scores of the classroom videos.

Ensemble: To estimate CLASS climate scores, we compute the unweighted average of the visual and auditory predictors of each dimension. (In pilot experimentation we found that weighting resulted in almost no improvement.) Note that, while multi-task learning (MTL) could be used to predict positive and negative climate jointly, we purposely avoided this approach because, according to the CLASS definition, absence of positive climate is not evidence of negative climate – they are independent dimensions.

IV. Datasets

To train our system, we used (1) a private dataset of CLASS-labeled classroom videos; (2) YouTube videos of school classrooms; and (3) the AffectNet dataset [27].

A. CLASS-labeled classroom videos

The first data source is from the University of Virginia (UVA) and consists of videos of toddler classrooms (children under 3 years old), each of which is $45-60$ minutes long. Each video is split into 15-minute chunks, and each chunk is labeled for the 10 dimensions of the CLASS-Toddler coding protocol. In total we have 241 labeled 15-minute videos distributed across the 7 classes as shown in table II.

	Score						
Dimension	**1**	**2**	**3**	**4**	**5**	**6**	**7**
Positive Climate	0	6	16	67	68	78	6
Negative Climate	196	36	7	2	0	0	0

TABLE II: # labeled video segments for each CLASS score.

Preprocessing: All videos were split into frames by sampling at a rate of 3 frames per second. To enable automated face analysis, the frames were then processed by the state-of-the-art CNN-based face detector (Faster RCNN, [28]), which is highly robust to non-frontal faces. The image frames and associated face coordinates constitute the **CLASS-Images** dataset.

Fig. 3: Classroom video from our YouTube dataset (Sec. IV); `https://youtu.be/cjNv2dQCFEk`.

In addition, from each of the UVA videos, we also extracted two sets of audio-based features: (1) the Mel-Frequency Cepstral Coefficients (MFCC), which represents the coefficients of the highest energy frequencies present in the audio; and (2) Chroma features, which convert the audio frequency into the 12 musical octaves bins. These features constitute our second dataset, which we call **CLASS-Audio**, and are used to train the audio-based CNNs that directly predicts CLASS positive and negative climate scores.

B. YouTube classroom videos

Modern deep perceptual architectures typically need a large amount of training data. For this reason, we collected an additional dataset to train the smile/non-smile and child/adult detectors, both of which were trained specifically to perform well in cluttered classroom environments with highly non-frontal faces. In particular, we harvested 70 publicly available classroom observation videos of young children from YouTube. The videos were split into frames at 3 frames per second, and each frame was processed with the face detector to obtain the location of all the faces. In total, 1.5 million cropped faces images were collected from these 70 videos with around 600 unique people in total. From these 1.5 million face images, 15000 images were sub-sampled randomly while maintaining an equal number of images from each video (to ensure a high number of unique faces).

The YouTube face images and associated labels (see next section) constitute the **YouTube** dataset. We note that, the YouTube dataset, in contrast to the CLASS-Images and CLASS-Audio, is publicly available, which enabled us to use Mechanical Turk for efficient annotation.

C. AffectNet [27]

While the YouTube dataset we collected was useful for collecting smiles, it contained relatively few negative expressions. Hence, we also used the emotion-labeled AffectNet dataset to train detectors of anger and sadness.

V. Data Annotation

Positive and negative climate depend (in part) on the emotional state of the students and teacher(s), as well as on the *interaction* between them. We thus trained binary face classifiers of smile, sadness, anger, and child/adult (to distinguish students from teachers) that were specifically tailored

to highly cluttered classroom environments with toddlers (< 3 years old) and highly non-frontal faces (sometimes even exceeding 90°).

Smile: We designed a crowdsourcing task to label the face data of the YouTube dataset and designed an annotation protocol with which we annotated the 15000 cropped face images from the YouTube videos on Amazon Mechanical Turk. Each face image was labeled by 3 different annotators. We asked annotators to assign each face image to one of four categories using the following instructions:

- **Smile** if the face subjectively looks joyful with evidence of lip-corner pull (i.e., AU 12 of FACS [29]).
- **Non-Smile** if there is no evidence of joy. The mouth might still be open.
- **Unclear** if it is hard to tell as smile or non-Smile.
- **Invalid** if there is no face in the image; if there are multiple faces in the same image; or if the face is not a human face (e.g., cartoon).

The final label for each face image was decided by majority vote across the 3 labelers. In total we had 1644 faces labeled as smile, 5578 as non-smile, 3704 images as unclear, and the rest as invalid.

Child/Adult: We asked annotators to label each face image of the YouTube dataset as "child" or "adult". Similar to the smile labeling task above, labelers could also respond with "unclear" or "invalid". After performing majority vote, we had 8769 images labeled as child 2329 as adult, 3522 as unclear, and the rest as invalid.

VI. Training the face classifiers

A. Smile and child/adult detectors

For training, validation, and testing, we first removed all images from the YouTube dataset that were labeled as "invalid". We explored two particular approaches: (1) multi-task learning (MTL) to train a classifier for both tasks simultaneously, and (2) supervised pre-training on ImageNet.

Architecture: All networks we investigated take cropped RGB face images of size $100 \times 100 \times 3$ as inputs. We explored around 30 different network designs; in the end we settled on VGG16 [30]. We compared training the network from scratch to pre-training on ImageNet [31]. For the latter approach, we replaced the final fully connected ("dense") layers of the pre-trained VGG16 with 2 dense layers trained from scratch on YouTube. In the end, the network outputs the probability of smile/non-smile and of child/adult.

"Unclear" images: Some of the face images (Figure 2) that were cropped from the YouTube dataset were labeled as "unclear". We compared two approaches: (1) removing the unclear images from the dataset, and (2) setting the target label (smile/non-smile, or child/adult) to 0.5.

Data partition: The 70 videos from the YouTube dataset were partitioned using an 80-20 split on a video level between training and testing. This ensured that no video used for training or validation was used for testing. For hyperparameter selection, we further split the training data using an 80-20 split on an image level for a validation dataset. See Table III for details.

Face Type	Train	Valid	Test
Smile, Child	803	201	125
Smile, Adult	270	68	52
Non-smile, Child	2746	687	629
Non-smile, Adult	583	146	172
Unclear	725	140	238

TABLE III: The number of face images of each class within each subset (train, validation, test) of the YouTube dataset.

Hyperparameters: The model was trained using Adam, with default parameters suggested in [32], for 100 epochs. The learning rate is annealed by a factor of 0.1 with a patience of 3 on the validation loss.

Comparison to Amazon Rekognition: Rather than collect emotion-labeled face images and train a neural network, one might consider using off-the-shelf software such as OpenFace [26] or even a cloud-based facial expression recognition platform such as Amazon Rekognition. Rekognition has presumably been trained on a very large dataset of images, but not specifically for classroom environments. In contrast, our MTL pre-trained network was optimized specifically for young children (and their teachers) in cluttered classrooms. In pilot experimentation, we found OpenFace to be less accurate on the YouTube face images than Rekognition and thus abandoned the approach.

In addition to recognizing smile/non-smile, Rekognition also estimates the age (an integer value) of each face. We compared the accuracy of our custom-trained neural networks with Rekognition. Our accuracy metric is the Area Under the ROC Curve (AUC). Since Rekognition outputs an integer value for the estimated age of the face, and since AUC is invariant to monotonic transformations of the predictions, we can use the age values directly as the likelihood of "adult". Also, since Rekognition uses a different face detector, we first matched the set of detected faces from Rekognition with those with our face detector [28] using the Intersection over Union (IOU). For any faces not detected by Rekognition, we took the mean prediction (over all faces in the test set) as the estimate. We then compared the accuracy on the test set of smile/non-smile and child/adult.

Approach	Child/ Adult	Smile/ Non-smile
MTL (pretrained on ImageNet, with "unclear" images)	0.942	0.879
MTL (pretrained on ImageNet, without "unclear" images)	0.957	0.889
MTL (no pretraining, without "unclear" images)	0.931	0.822
Individual networks (no pretraining, without "unclear" images)	0.863	0.755
Amazon Rekognition	0.935	0.837

TABLE IV: Accuracy (AUC) of the different neural networks we compared for child/adult and for smile/non-smile classification. MTL is a single network trained using multi-task learning for both tasks jointly.

Results on YouTube images: Accuracy results (measured

as AUC) are shown in Table IV. There are several interesting trends: (1) The magnitude of the smile AUC scores – about 89% for the MTL-based NN trained without unclear images – is lower compared to the accuracy levels (typically mid/high 90%) found on more conventional frontal datasets of adults (e.g., [33]). This underlines the difficulty of the recognition task, which is likely due to the highly cluttered and non-frontal nature of classroom faces. (2) Corroborating many other works [34], pre-training on ImageNet significantly boosted recognition accuracy of the neural networks. (3) Training custom face classifiers, tuned specifically to classroom environments on modest-sized datasets, can result in higher accuracy than a commercial-grade system (such as Amazon Rekognition) trained on millions of images. (4) Multi-task learning significantly enhanced the accuracy of the classifiers for both tasks. (5) Training with "unclear" images (whose labels were set to 0.5) lowered the test accuracy of the child/adult and smile/non-smile CNNs slightly. However, we discovered that it actually increased (compared to using a CNN trained without unclear images) the accuracy of the overall CLASS estimates from the ensemble network (see next paragraph).

"Unclear" images: Without adding "unclear" faces to the training data, our MTL-trained model tended to assign a low probability of "smile" too images that were labeled by humans as "unclear": the mean estimated smile probability was 0.096 (s.d. 0.180). But this may be a mischaracterization of "unclear" faces, whose facial expression is, by definition, uncertain and thus might be more aptly represented by a probability of 0.5. After adding "unclear" images (with target probability of 0.5) to the training set, the new detector gives a mean probability of 0.476 (s.d. 0.112) for smile/non-smile, and of 0.519 (s.d. 0.113) for child/adult, to the "unclear" images in the test set. This could be important for environments – such as ours (see Table III) – in which "unclear" face images commonly occur.

B. Sadness and anger detectors

Using the AffectNet dataset and the same neural network architecture as for smile detection, we trained binary detectors of sadness and anger (jointly using MTL); their accuracies (AUC) on the test set were 0.872 and 0.884, respectively.

VII. Predicting CLASS scores from visual cues

Given the trained face classifiers of smile, sadness, anger, and child/adult, we created predictors of the CLASS positive and negative climate scores by applying the classifiers to every detected face in each frame of the CLASS-Image dataset, and integrating the results over each 15-minute CLASS-labeled video. Since the sadness and anger detectors were trained later in the course of the project, most of the architectures we compare use only the smile detector. In Section IX we show how the two additional detectors boost the overall accuracy.

We considered two temporal integration approaches: simple average, and a recurrent neural network. For both approaches, we applied 10-fold cross-validation to the CLASS-Images dataset subject to the following stratification constraints: (1) Whenever possible, all climate levels (1-7) must be represented in each fold; and (2) No two folds contain a video clip from the same classroom observation video.

1) Simple average: Since the frequency of facial expression in classroom videos is one of the behavioral indicators associated with positive climate [1], it seemed plausible that the *average* expression, across all participants and all frames of the video, might be predictive. To explore this, we trained a decision tree-based regressor (using the CART algorithm [35]) that took as input the average smile and predicted the CLASS score. In contrast to simple linear regression, the decision tree can capture non-linear relationships.

2) Recurrent neural networks: Based on the intuition that the *sequence* of facial expression events could be important for estimating CLASS climate scores, we explored several temporal integration models. The general approach was to compute the average smile scores *within* each video frame, and then pass these scores to an LSTM recurrent neural network with one hidden layer containing 100 units. The number of recurrent steps was 2700 (900sec for a 15-min video chunk at 3 frames/second). At the end of the time series, a single output is predicted which is the CLASS score. See the architecture in the left half of Figure 1.

We explored several input representations: (1) The average smile (within each frame) of *all* participants (both teachers and students); (2) the average smile of just the students (i.e., we use the smile scores of only those faces that are considered "child" by the child/adult detector); (3) the average smile of just the teachers; (4) the average smile of students and teachers *separately* (i.e., as two different input features). Moreover, since there is no constraint to code CLASS climate in a causal manner, we also tried using a bi-directional LSTM. A Bi-LSTM utilizes the knowledge of the future events to understand the context of current events.

Results: The average (over 10 folds) Pearson correlations (r) for the different approaches for predicting positive and negative climate, along with their p-values (for the null-hypothesis that the true correlation is 0), are shown in Table V. The accuracy of the simple average approach (shown in the first row) was very low, and the results were not statistically significant.

The LSTM-based predictor of climate delivered higher accuracy than the decision tree-based predictor (which simply analyzed the average smile of the *whole* video). Moreover, analyzing students' and teachers' smiles *separately* using the LSTM was even more (and statistically signficantly so) accurate. Analyzing the video from both directions using a Bi-LSTM gave yet another small accuracy boost. This suggests that the facial expression *dynamics* are more important than their mean values for predicting CLASS scores.

Hyperparameters of the Bi-LSTM: Optimizer=Adam, lr=0.001 (annealed by factor of 0.1 with a patience of 15 by monitoring the validation loss), epochs=500, with early stopping patience of 25 epochs by monitoring training loss.

Class imbalance: To address the class imbalance, we tried

oversampling the stratified folds using SMOTE (Synthetic Minority Oversampling Technique) [36] and then retrained our models. But by doing this we saw a drop in performance; see Table V.

VIII. Training the audio-to-climate classifiers

We trained a CNN that analyzes the audio of each classroom observation to predict positive and negative climate. In particular, we applied 10-fold cross-validation to the CLASS-Audio dataset and extracted both MFCC and Chroma features using the Librosa package [37]. From the two feature sets we took the top 100 most significant coefficients to form two feature vectors of 100 features in each. By treating the spectrograms features obtained as images, we can use a CNN to look for patterns that are predictive of CLASS scores.

Results are shown in the fourth-to-last row of Table V: the correlation of the predicted scores with the ground-truth positive and negative climate were 0.308 and 0.29, respectively, and were statistically significant. These provide evidence that low-level audio features, even without downstream speech recognition or NLP, can be useful for CLASS prediction.

We also tried two alternative approaches for using audio-based features: (1) Similar to how we trained face classifiers by pre-training a VGG16 network on ImageNet, we can pre-train an audio-to-climate network on a large audio dataset such as Audioset [38]. In practice, we were unsuccessful with this approach, and the predictive accuracy was barely above chance. A plausible explanation is that the expected input to the Vggish network [39] a 5-second audio clip sampled randomly from a given video. It might be conceivable that 5-seconds of classroom audio would be very hard to classify for a particular climate and was the result for the low accuracy we had got. (2) We trained auditory emotion classifiers from datasets such as RAVDESS [40], SAVEE [41], and UMSSED [42] datasets. We then used the resulting detector as mid-level features to predict CLASS scores. However, this approach also resulted in low accuracy – possibly due to the quite different environments in which the audio was collected compared to school classrooms – and we abandoned it.

Hyperparameters: Optimizer was Adam, lr=0.0001, epochs=500, with early stopping patience of 25 epochs by monitoring the training loss.

Class imbalance: Similar to the previous section, we also tried using SMOTE to handle class imbalance but found it negatively impacted accuracy.

IX. Ensemble Model: Expression + Audio

Given the Bi-LSTM that examined the smile dynamics of children and adults separately, as well as the audio-to-climate regressor that examined low-level audio features, we created two ensemble networks (one each for positive and negative climate; see Figure 1) that simply outputs the average of its two inputs as the CLASS score. Combining the audio and visual channels yielded an improve accuracy over either one. See Table V. Finally, we also tried training a visual detector of negative climate using not just smile, but also sadness and anger as facial expression inputs (see Section VI-B). We found that including these new features boosted the accuracy of the negative climate significantly.

Results: The best ensemble model predicted the ground-truth CLASS positive and negative climate scores with a Pearson correlation of 0.40 and 0.51, respectively, both of which were statistically significantly better than chance. Table VI shows the confusion matrices obtained over the 10-fold cross validation for positive and negative climate respectively on the UVA CLASS-labeled video dataset. Finally, Figure 4 shows the trajectory of smiles, separately for teachers and students, on one of the UVA toddler classroom videos. Each dot represents one smile estimate for one face detected at a particular video frame. The solid line represents the smoothed smile trajectory over time.

Model	Positive Climate		Negative Climate	
	r	p	r	p
DT (Avg Smile)	0.08	0.178	0.02	0.360
RandomForest (Audio)	0.28	0.004	0.22	0.007
HMM (States:13; Smile:St, Te)	0.15	0.092	0.14	0.101
LSTM (Smile: All)	0.13	0.102	0.15	0.091
LSTM (Smile: St)	0.12	0.108	0.13	0.045
LSTM (Smile: Te)	0.11	0.137	0.13	0.032
LSTM (Smile: St, Te)	0.15	0.097	0.20	0.009
Bi-LSTM No Unclear (Smile: St, Te)	0.15	0.092	0.22	0.007
Bi-LSTM (Smile: St, Te)	0.17	0.052	0.24	0.006
SMOTE-Bi-LSTM (Smile: St, Te)	0.14	0.098	0.21	0.007
CNN (Audio)	0.31	0.002	0.29	0.003
SMOTE-CNN (Audio)	0.29	0.004	0.26	0.006
Ensemble: Bi-LSTM + CNN (Audio)	0.38	0.002	0.45	0.001
Ensemble: Bi-LSTM (Smile,Anger,Sadness) + CNN (Audio)	**0.40**	0.001	**0.51**	0.001

TABLE V: CLASS climate prediction: average 10-fold Pearson (r) correlations and associated two-tailed p-values. DT=decision tree-based regressor, St=student, Te=teacher.

Positive Climate

	1	2	3	4	5	6	7
1	0	0	0	0	0	0	0
2	0	0	3	1	2	0	0
3	0	2	3	4	6	1	0
4	0	0	23	19	20	5	0
5	0	0	1	14	29	23	1
6	0	4	4	18	23	28	1
7	0	0	0	1	3	1	0

Negative Climate

	1	2	3	4	5	6	7
1	139	54	2	0	0	0	0
2	12	22	1	1	0	0	0
3	4	2	1	0	0	0	0
4	0	2	0	0	0	0	0
5	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0

TABLE VI: Confusion matrices of CLASS predictions. Rows are ground-truth; columns are the (rounded) predictions.

X. Comparison to shallow architectures

To validate the need for a deep learning approach we performed experiments with 2 commonly used *shallow* architectures: (1) Random Forest to analyze audio features. (2) Hidden Markov Model to analyze the facial expressions. For the auditory features we applied a Random Forest model on the MFCC+Chroma features. We optimized the number trees

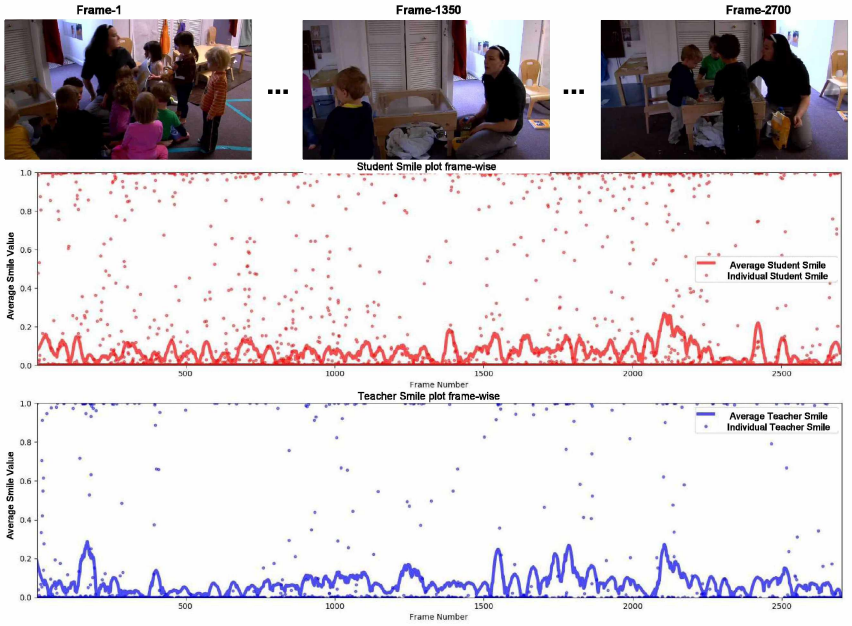

Fig. 4: The smoothed average student and the teacher smile vectors (red and blue lines) along with the individual smile values (dots) over the 2700 frames for one representative video. This vector is given as input to our Bi-LSTM model to predict for Positive and Negative Climate. For this particular video we obtained a Positive Climate score of 4.59 and Negative Climate score of 1.78. The ground-truth scores, as obtained from human CLASS experts, were 5 and 2, respectively.

and depth of the trees using 10 fold double cross validation to give it the best chance of success. For the facial expression features we trained a separate HMM for each CLASS score (1, 2, ..., 7) and each dimension (Positive, Negative). For prediction, we passed each sequence to all the HMMs and then attributed the sequence to the model that gave the highest probability. The trend observed is that the shallow architectures did not perform as well as the deeper models, particularly the Bi-LSTM (Smile: St, Te) (see Table V).

XI. Conclusions

We devised a multi-modal neural network-based system to detect automatically the levels (1-7) of positive and negative climate of the CLASS protocol for classroom observation. Such a system could, with further development, be used to provide teachers with more specific and frequent feedback about the quality of their classroom interactions. From a machine learning perspective, the chief challenge was how to harness modern deep perceptual architectures, which are very data-hungry, for an application domain in which the data are sensitive and scarce; difficult to label (requiring significant training); and sparsely labeled (only a few numbers per 15-minute video clip).

Lessons learned: (1) Automatic estimates of individual participants' facial expression, as well as low-level audio features, provided enough signal to predict CLASS positive and negative climate well above chance – even though these constitute just a few of the behavioral markers associated with classroom climate. Moreover, the dynamics of facial expression were important – the average "smile" value alone had very little predictive power. (2) Corroborating many prior works on image recognition, we found that transfer learning via pre-training a VGG16 classifier on ImageNet was highly effective at creating a domain-specific face classifier of smile/non-smile and of child/adult. The resulting system outperformed a state-of-the-art cloud-based facial expression recognition service (Amazon Rekognition). This is likely due to the characteristics of classrooms which involve less frontal faces, faces that are mostly of children whose face features and emotion response is different compared to adults, as well as high clutter in the background. For audio-based CLASS prediction, however, we did not observe any advantage in pre-training on standard audio datsets. (3) Despite the modest size of our training datasets, we found that – by using a combination of multi-task learning, supervised pre-training, and augmentation with supplementary data – modern deep perceptual architectures of both visual and auditory information delivered higher accuracy than commonly used shallow models.

Future work: It may be useful to track the expression trajectories of individual people in the classroom over time, rather than just treating each frame as a "bag" of expressions.

References

[1] Robert C Pianta, Karen M La Paro, and Bridget K Hamre. *Classroom Assessment Scoring System™: Manual K-3.* Paul H Brookes Publishing, 2008.

[2] Susan Kontos and Amanda Wilcox-Herzog. Teachers' interactions with children: Why are they so important? research in review. *Young Children*, 52(2):4–12, 1997.

[3] Andrew J Mashburn, Robert C Pianta, Bridget K Hamre, Jason T Downer, Oscar A Barbarin, Donna Bryant, Margaret Burchinal, Diane M Early, and Carollee Howes. Measures of classroom quality in prekindergarten and children's development of academic, language, and social skills. *Child development*, 79(3):732–749, 2008.

[4] Claire Cameron Ponitz, Megan M McClelland, JS Matthews, and Frederick J Morrison. A structured observation of behavioral self-regulation and its contribution to kindergarten outcomes. *Developmental psychology*, 45(3):605, 2009.

[5] Deborah Lowe Vandell, Jay Belsky, Margaret Burchinal, Laurence Steinberg, and Nathan Vandergrift. Do effects of early child care extend to age 15 years? results from the nichd study of early child care and youth development. *Child development*, 81(3):737–756, 2010.

[6] Ellen S Peisner-Feinberg, Margaret R Burchinal, Richard M Clifford, Mary L Culkin, Carollee Howes, Sharon Lynn Kagan, and Noreen Yazejian. The relation of preschool child-care quality to children's cognitive and social developmental trajectories through second grade. *Child development*, 72(5):1534–1553, 2001.

[7] Barbara Nye, Spyros Konstantopoulos, and Larry V Hedges. How large are teacher effects? *Educational evaluation and policy analysis*, 26(3):237–257, 2004.

[8] Thomas J Kane, Daniel F McCaffrey, Trey Miller, and Douglas O Staiger. Have we identified effective teachers? validating measures of effective teaching using random assignment. In *Research Paper. MET Project. Bill & Melinda Gates Foundation*. Citeseer, 2013.

[9] Hermann Ebbinghaus. Memory (ha ruger & ce bussenius, trans.). *New York: Teachers College.(Original work published 1885)*, 39, 1913.

[10] Andrew D Ho and Thomas J Kane. The reliability of classroom observations by school personnel. research paper. met project. *Bill & Melinda Gates Foundation*, 2013.

[11] Charlotte Danielson. *Enhancing professional practice: A framework for teaching*. ASCD, 2011.

[12] Candice Walkington, Prerna Arora, Shasta Ihorn, Jessica Gordon, Mary Walker, Larry Abraham, and Mary Marder. Development of the uteach observation protocol: A classroom observation instrument to evaluate mathematics and science teachers from the uteach preparation program. *preprint*, 2012.

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

[14] Sidney D'Mello, Rosalind W Picard, and Arthur Graesser. Toward an affect-sensitive autotutor. *IEEE Intelligent Systems*, 22(4), 2007.

[15] Ivon Arroyo, David G Cooper, Winslow Burleson, Beverly Park Woolf, Kasia Muldner, and Robert Christopherson. Emotion sensors go to school. In *AIED*, volume 200, pages 17–24, 2009.

[16] Jacob Whitehill, Zewelanji Serpell, Yi-Ching Lin, Aysha Foster, and Javier R Movellan. The faces of engagement: Automatic recognition of student engagementfrom facial expressions. *IEEE Transactions on Affective Computing*, 5(1):86–98, 2014.

[17] Joseph Grafsgaard, Joseph B Wiggins, Kristy Elizabeth Boyer, Eric N Wiebe, and James Lester. Automatically recognizing facial expression: Predicting engagement and frustration. In *Educational Data Mining*, 2013.

[18] Nigel Bosch, Sidney D'Mello, Ryan Baker, Jaclyn Ocumpaugh, Valerie Shute, Matthew Ventura, Lubin Wang, and Weinan Zhao. Automatic detection of learning-centered affective states in the wild. In *Intl. conference on intelligent user interfaces*, 2015.

[19] Anthony F Botelho, Ryan S Baker, and Neil T Heffernan. Improving sensor-free affect detection using deep learning. In *International Conference on Artificial Intelligence in Education*, 2017.

[20] Ryan SJ d Baker, Sujith M Gowda, Michael Wixon, Jessica Kalka, Angela Z Wagner, Aatish Salvi, Vincent Aleven, Gail W Kusbit, Jaclyn Ocumpaugh, and Lisa Rossi. Towards sensor-free affect detection in cognitive tutor algebra. *Educational Data Mining*, 2012.

[21] Zachary A Pardos, Ryan SJD Baker, Maria OCZ San Pedro, Sujith M Gowda, and Supreeth M Gowda. Affective states and state tests: Investigating how affect throughout the school year predicts end of year learning outcomes. In *International Conference on Learning Analytics and Knowledge*, pages 117–124. ACM, 2013.

[22] Sidney K D'Mello, Andrew M Olney, Nathan Blanchard, Borhan Samei, Xiaoyi Sun, Brooke Ward, and Sean Kelly. Multimodal capture of teacher-student interactions for automated dialogic analysis in live classrooms. In *Intl. conference on multimodal interaction*, 2015.

[23] Patrick J Donnelly, Nathaniel Blanchard, Borhan Samei, Andrew M Olney, Xiaoyi Sun, Brooke Ward, Sean Kelly, Martin Nystrand, and Sidney K D'Mello. Multi-sensor modeling of teacher instructional segments in live classrooms. In *ACM international conference on multimodal interaction*, pages 177–184. ACM, 2016.

[24] Zuowei Wang, Xingyu Pan, Kevin F Miller, and Kai S Cortina. Automatic classification of activities in classroom discourse. *Computers & Education*, 78:115–123, 2014.

[25] Qifeng Qiao and Peter A Beling. Classroom video assessment and retrieval via multiple instance learning. In *International Conference on Artificial Intelligence in Education*, pages 272–279. Springer, 2011.

[26] Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency. Openface: an open source facial behavior analysis toolkit. In *Applications of Computer Vision (WACV)*, 2016.

[27] Ali Mollahosseini, Behzad Hasani, and Mohammad H Mahoor. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *arXiv preprint arXiv:1708.03985*, 2017.

[28] Huaizu Jiang and Erik Learned-Miller. Face detection with the faster r-cnn. In *IEEE Automatic Face & Gesture Recognition*, 2017.

[29] Paul Ekman and Wallace V Friesen. *Facial action coding system: Investigator's guide*. Consulting Psychologists Press, 1978.

[30] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[31] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

[32] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[33] Youngkyoon Jang, Hatice Gunes, and Ioannis Patras. Smilenet: Registration-free smiling face detection in the wild. 2018.

[34] Dhruv Mahajan, Ross Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, and Laurens van der Maaten. Exploring the limits of weakly supervised pretraining. *arXiv preprint arXiv:1805.00932*, 2018.

[35] Leo Breiman. *Classification and regression trees*. Routledge, 2017.

[36] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

[37] Brian McFee, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. librosa: Audio and music signal analysis in python. In *python in science conference*, 2015.

[38] Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *ICASSP*, 2017.

[39] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. Cnn architectures for large-scale audio classification. In *ICASSP*, 2017.

[40] Steven R Livingstone, Katlyn Peck, and Frank A Russo. Acoustic differences in the speaking and singing voice. In *Proceedings of Meetings on Acoustics ICA2013*, volume 19, page 035080. ASA, 2013.

[41] P Jackson and S Haq. Surrey audio-visual expressed emotion(savee) database. *University of Surrey: Guildford, UK*, 2014.

[42] Biqiao Zhang, Emily Mower Provost, Robert Swedberg, and Georg Essl. Predicting emotion perception across domains: A study of singing and speaking. In *AAAI*, pages 1328–1335, 2015.