# Visualizing Classroom Dynamics with COILS: the Classroom Observation Interactive Learning System

Category: Research

Paper Type: application/design study

**Abstract**— K12 teachers and administrators routinely review video recordings of classroom interactions as part of training, professional development, and evaluation to help ensure positive classroom dynamics. However, reviewing videos for key moments of teacher-student and student-student interaction takes substantial time and effort, due to the large amount of raw footage and inherent complexities of social interactions. In this paper, we explore the potential of machine learning and visualization to bridge this gap through the design and evaluation of COILS, the Classroom Observation Interactive Learning System. COILS uses a pipeline developed with input from teachers and domain experts to automatically annotate classroom videos with detected eye-gaze interactions and estimated facial emotions. We describe how teachers contributed to the design of COILS, influencing the resulting tasks, goals, and visual encodings. In a controlled mixed-methods experiment with 11 in-service teachers, we explore how COILS shapes teachers' video review process and strategies. Experiment results paint a complex picture of using ML-driven interfaces in the classroom. Teachers remarked on the real-life applicability of COILS not only for teacher development, but also for identifying how student interactions evolve in the classroom and how they can better support individual students (*e.g.* English language learning (ELL) students). At the same time, we occasionally observed misalignments of teacher perceptions and ML predictions, suggesting a need to further explore model improvements, design strategies, and training efforts. We discuss broader implications for similar visualization and ML efforts, and emerging possibilities for future work in supporting teachers in video review contexts.

**Index Terms**—Visualization, Education, Machine Learning, Computer Vision

## 1 Introduction

add update Classrooms with consistent, quality teacher-student interactions and positive emotional climates are a critical component for successful student learning [28]. Many teacher education programs now include extensive training in assessing and shaping classroom climate. In these programs, teachers learn to better perceive subtle interactions and interpersonal dynamics in students, and to reflect on their own behaviors and strategies in the classroom [18, 36]. Attention equity is another emerging component in teacher training, as studies have shown that teachers sometimes give more attention to particular students over others, which can have a negative impact on those who are left out [54]. In the education community, positive classroom climate and associated trainings comprise a substantial portion of curricula and research initiatives (*e.g.* [30, 60]).

K12 teachers and administrators primarily use video recording and systematic review protocols to support positive classroom climate training and goal-setting efforts. Common video review scenarios include teachers in training, *e.g.* as part of a graduate course, in-service teachers setting personal goals for classroom climate and equity, or administrators assessing teacher performance and climate. Protocols used in these reviewing scenarios, such as the widely used Classroom Assessment Scoring System (CLASS), guide reviewers in assessing particular states, actions, and interactions in the classroom [50]. However, this manual review process can be costly in terms of time and effort, potentially requiring hours for a single classroom video, further compounding the substantial challenges K12 teachers already face.

Advances in computer vision and machine learning present a promising, yet challenging, means for supporting teachers in video review activities. Deep learning has produced more capable computer vision techniques for object detection, activity recognition, and more targeted social features like gaze detection [3, 53]. Affective computing models aim to estimate people's emotional state through facial features [14, 23, 35, 43]. However, vision technologies are increasingly scrutinized due to documented biases (*e.g.* [22, 55]) and potential ethical and privacy concerns [5, 32, 33]. A further challenge is that it is not clear which video-extracted features, if any, can help teachers in classroom review activities, nor is it clear what transformations, visual encodings, and interactions would constitute an effective interface.

In this paper, we explore challenges in classroom video review support for teachers through the design and evaluation of COILS, the Classroom Observation Interactive Learning System. We describe a video processing pipeline which focuses on eye-gaze detection (*i.e.* eye contact between people) and facial emotion estimation, in an IRB-approved study using classroom videos created specifically for research and training. We report a design process in which teachers and other education professionals shape the tasks, goals, and prototype designs of COILS through brainstorming sessions and focus groups[1]. We evaluate COILS alongside control conditions in a mixed-methods experiment with $n = 11$ in-service teachers of varying backgrounds, using a qualitative coding methodology to develop themes from the resulting transcripts, and quantitative interaction log analysis to explore how teachers might differ in how they use COILS for classroom video review. From these activities we contribute:

- Design findings from several activities with teachers and educational professionals, comprising goals, tasks, and guidance for visualization-enabled classroom video review interfaces.
- COILS, the Classroom Observation Interactive Learning System, an interface that combines traditional video review with visualizations showing both eye-gaze interaction data and estimated emotion data in classroom videos.
- Results from a controlled mixed-methods study with $n = 11$ in-service K12 teachers of varying backgrounds, with quantitative evidence suggesting teachers employ different video review strategies with COILS, and qualitative evidence highlighting both opportunities and challenges for machine learning-driven visualization for supporting teachers in classroom video review.

We conclude with a discussion following the results of the study, including potential reasons why teachers consistently preferred the eye-gaze visualizations, opportunities for new visual metaphors for video analysis and visualization that align with teachers' needs, and other possibilities for future work.

## 2 Related Work

In designing and developing COILS, we draw on three areas of prior research, including classroom dynamics analysis, machine learning for

[1] Education researchers are also part of the research team and authors, but this information is anonymized for review.

video analysis, and visual analytics for video data. We also situate our work in the context of critiques of facial and emotion recognition, and explore related implications in our results.

### 2.1 Classroom Dynamics Analysis

Classroom dynamics analysis focuses on the assessment of classroom climate and the behavior of teachers and students. Research in classroom dynamics typically measures the quantity and quality of teacher-student interactions, student-student interactions, and students' attitudes and emotions [41]. Several studies suggest that the quality of students-teacher interactions in the classroom impacts student learning outcomes [28, 36, 51], and positive interactions with peers aid student learning and development [21].

Multiple correlational studies have demonstrated links between emotional and instructional support in the classroom and children's cognitive, social, and emotional skills [28, 36, 48, 51]. Large-scale causal initiatives, such as the Bill and Melinda Gates Foundation's *Measures of Effective Teaching Project*, have yielded extensive evidence that positive climate positively impacts students' academic performance [26, 44]. Other studies have shown that student emotion is an essential part of study motivation, intertwined with both teachers' instructional responses and students' beliefs and actions [39]. These factors constitute an integral part of the interpersonal processes that create positive and effective classroom climates.

Systematic classroom observation is one of the most common assessment methodologies used by researchers and for training initiatives [17, 37]. In terms of methodology, some efforts involve analyzing interactions based text transcriptions of classroom activity [49] and audio recordings [47]. In contrast, evidence from Mehrabian *et al.* indicates that when still images or videos of participants' faces is combined with audio, analyzers weighed it more when deciding whether an interaction was positive or not [38]. Video recordings are pervasive in educational research, providing means for both shared observations and for individual observers to revisit key interaction moments [31]. Video analysis is used in for training and assessment scenarios, such as identifying children in need of emotional or behavioral support, and for assessing classroom climate as part of academic interventions [16]. In-service teachers often use recording for assessing their own classroom climate as part of goal-setting and providing student support [29].

### 2.2 Machine Learning for Video Processing

Researchers from computer science, cognitive science, and psychology have explored how to use machine learning to detect interactions between students and teachers in classrooms for over 20 years [27]. Several efforts have used machine learning-based perceptual systems to analyze entire classrooms in school [1, 7, 12, 13, 46, 61]. D'Mello *et al.* [12, 13] explored how to segment and recognize students' and teachers' speech in unconstrained classrooms based on different microphone configurations. Wang *et al.* [61] segmented teachers' speech by deploying small wearable recording devices in math classrooms. Ahuja *et al.* [1] developed a combined hardware and software toolkit called *EduSense* that detects students' body and facial movements automatically. Their system uses *OpenPose* [7], as well as multiple classifiers (random forests, support vector machines, multi-layer perceptrons) trained on top of its outputs, to track each student in each video frame as well as their body posture, hand gestures, and facial expressions. OpenPose also analyzes audio features recorded from different microphones to determine whether speech was produced by students versus the instructors. They further extend the system that could capture 3D classroom environment and provide more Degrees of Freedom (DOF) of gaze. [2] The machine learning architecture in [46] is based on an ensemble of decision trees that analyze the volume and standard deviation of classroom sound in 15 second intervals, where the goal is to classify different classroom activities. These works inform our video processing pipeline.

### 2.3 Visual Analytics for Video Analysis

Video visualization techniques have been developed to provide compact overviews of video content by summarizing and encoding key features or events [11]. Romero *et al.* studied how to abstract and visualize activities in video. [56]. Many research explored the application of video analysis in different fields. Haotian *et al.* built the visual analytics system assist Proctoring of Online Exams. Aoyu and Qu developed the visual analytics system that empowers users to explore presentation techniques in TED Presentations. Tam *et al.* [59] used interactive parallel coordinates to analyze facial dynamics data, with the goal of developing better analysis algorithms. Zadeh *et al.* [62] develop histogram techniques to visualize relationships of sentiment intensity between visual gestures and spoken words in videos.

Most relevant to our work are efforts from Zeng *et al.*, who developed a interactive visual analytic systems and techniques for emotion analysis in presentation videos [64] and classroom videos [63]. The EmotionCues system, for example, uses multi-linked views to show student emotions over time in a classroom. Their system also includes a novel sunburst-like diagram to show the quality of video frames. Major differences in the present work include the use of eye-gaze interactions as well as estimated emotion data, additional design activities with teachers and educational professionals, and a controlled user study with in-service teachers to explore the how such visualizations might shape the classroom video review experience.

### 2.4 Privacy, Bias, and Ethics Considerations

The expanding use of technology such as facial and emotion recognition has led to research outlining cases where these systems exhibit bias and other ethical concerns (*e.g.* a review from Crawford [10]). For example, an emotion recognition audit found that some models estimate black faces as showing anger more than white faces in the same dataset [55]. An extensive review from Barrett *et al.* highlights many instances in which outward facial emotion does not necessarily reflect what people are actually feeling [4]. We adopted Barrett's recommendations for future research in the use of affective technology as part of our work. In particular, we incorporate "Support larger scale studies that bridge the lab and the world, that study individual people across many contexts, and that measure emotional episodes in high dimensional detail, including physical, psychological, and social features". In our case, we involve teachers and educational professionals in the design of COILS, and explicitly aim to highlight the social context behind estimated interactions in the video. Furthermore, we provide opportunities in the evaluation for teachers to share what aspects of the data/interface did *not* work well or as expected.

## 3 Data Characterization and Transformation

In developing COILS, we used a data transformation pipeline that processes classroom videos for climate-related features. Videos were pre-processed, then transformed into time series containing estimated emotion and eye-gaze events. We describe this pipeline, including prior machine-learning approaches that inform our approach, in the following sections.

### 3.1 Data Sources and Pre-processing

The videos used in COILS were created expressly for educational research and training, including parental permission for publication. Consent was obtained from both teachers and parents of students that appear in the videos, each of whom were informed the videos would be used for educational purposes and materials. In each video, a single camera recorded a classroom throughout the class session. For pre-processing and to reduce computation cost while keeping sufficient information, we adopt the method introduced by Zeng *et al.* [63], where we re-sample the video to remove redundant frames. Given a video with 30 frames per second, we extracted three frames per second that are passed to the machine learning pipeline, making the sampling ratio $1/10$.

### 3.2 Machine Learning Pipeline

The machine learning pipeline processes the re-sampled frames individually in two major steps. First, the pipeline extracts the faces in a frame using a Faster R-CNN face detector [24]. Next, the detected faces are clustered semi-automatically into separate tracks (one for each student

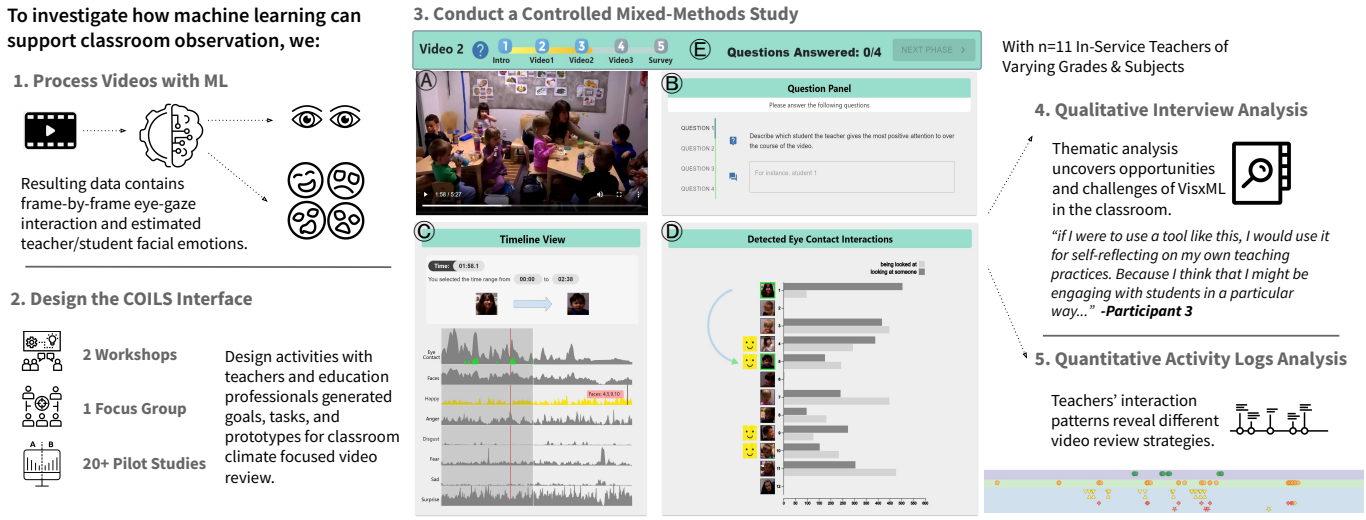


Figure 1. We investigate the role of machine learning-supported classroom video review through design activities, the development of COILS, and a mixed-methods user study. The COILS interface includes a video player, a questions panel showing prompts designed by teacher classroom climate review experts, a timeline view with area charts for estimated emotion data and other video features, and an eye-gaze visualization showing a hybrid bar and arc chart of teacher/student and student/student interactions. Results of the mixed-methods study show both tangible benefits and future challenges for using machine-learning and visualization to support teachers in classroom video review.

and teacher). Finally, the faces detected by the face detector are fed into an emotion detector as well as an eye-gaze detector [3].

*Face clustering*: For semi-automatic clustering of which face belongs to which person in the classroom, we used a pre-trained face embedding neural network (FaceNet [57]), which is designed to map faces into a metric space so that faces from the same person are "pulled together" and faces from different people are "pushed apart". *FaceNet* was mainly trained on adults from images available on the Web while, in contrast, the people in the classroom videos from our study were mostly of children whose faces were often non-frontal relative to the camera. Hence, we found that, while the FaceNet embeddings provided a useful starting point for clustering the faces by identity, they needed substantial manual correction to be accurate. After performing this machine learning-aided labeling process, we arrived at a set of face tracks (one for each student and teacher) of where each person is at each moment in time.

*Emotion detection*: The emotion detector consists of a VGG16 [58] backbone pretrained on ImageNet. The network was then fine-tuned on a customized training dataset consisting of the *AffectNet* dataset [40] and a custom YouTube-based dataset of classroom images that were labeled for two sets of classes – smile/non-smile and adult/child – using Mechanical Turk [52]. The emotion classifier predicted the probabilities that each image belonged to one of six mutually exclusive emotion categories: Happy, Anger, Disgust, Fear, Sad, and Surprise.

*Eye-gaze detection*: The eye-gaze detector fuses two different VGG16-based pathways – one based on just the detected face image, as well as the whole image frame – to predict the eye-gaze fixation [3]. The eye-gaze fixation is formulated as the $(x, y)$ location in each frame of where a particular person is looking; hence, it is a 2-D prediction task. Based on the gaze target, and based on the locations (from the face detector) of where each person is in each frame, we can estimate who is looking at whom and when.

The data generated from the machine learning pipeline in the form of face detection, emotion prediction, and eye-gaze estimations, are then utilized by the COILS interface in the format specified below.

### 3.3 Resulting Data and Format

As a result, we have a time-varying dataset of gaze events and estimated emotion data for a class video. The estimated emotion data is processed on a person-by-person basis. When someone is present in the frame, there is a vector of emotion data consisting of estimates following the circumplex model of emotions, including happiness, surprise, fear, disgust, sadness, and anger. Gaze events contain three primary fields: timestamp, source (the person looking), and target (the person looked at). Gaze events can also be mutual, when the source and target are looking at each other. With this dataset, our task is to determine how it maps to possible goals and tasks teachers have when reviewing classroom videos.

## 4 Designing COILS

In designing COILS, we adopted a multi-stage, user-centered iterative design methodology. We work closely with domain experts (X professional teacher trainers and X senior K12 teachers with average X years of teaching experience) to determine system scope, identify tasks and goals, and perform user studies.

**Brainstorming Stage:** In this preliminary stage, we ran multiple on-site workshops which include formative and evaluative activities with teachers and teacher trainers. The goal of the workshops was to gauge use cases for the system, distill concrete goals and tasks of reviewing classroom videos, and craft the intial design process in accordance with this feedback. At these workshops, we first collect and summarize teacher's demand, then build multiple low-fidelity prototypes of COILS, and present these prototypes for brainstorming and discussion. Participants provided feedback on specific visualization ideas as well as reinforce potential applications of COILS. At one of these workshops, we conducted several pilot studies with teachers who gave feedback on visual encoding details and interaction schemes.

From the discussion and brainstorming at workshops, we record real-life use cases and end-user desired features on classroom video reviewing. Based on these outcomes, we summarize them into a structured goals and tasks collection

**Iterative Design Stage:** After the brainstorming stage, we performed multiple designing-implantation-evaluation cycles iteratively. we held a multi-week focus-group with four in-service teachers. In these hour-long sessions, we provided teachers with walkthroughs of the COILS interface at that time, example usage scenarios, and more information about the machine learning pipeline. Teachers had open-ended and comprehensive discussions about the efficacy and limitations of the COILS interface and the AI system, come up with new interface features and evaluate the existing ones, and gauge potential real-life use cases for the system.

This round of sessions was very iterative, where the feedback from

a session was used to guide the development process and the updated version of the interface was evaluated in the following session. The goal of this stage was to fine-tune the COILS interface to fit the feedback received during these sessions and the discussion was guided to hit major points of interest. One such point of interest was pertaining to the AI system, specifically its advantages and drawbacks, how its decisions will be perceived by the teachers, how it will be used in real life, and most importantly, how the limitation of such an AI system in its real-world ambiguous decision-making should be communicated to the teachers. The sessions also served as a source of discourse about real-life use cases for the system including the need for such a system to support teacher agency in the context of school administration initiatives

Drawing on these activities, we conducted a goal and task analysis following recommendations from Hindalong *et al.* [20]. Goal and task analysis allows us to externalize what we learn about teachers' needs, and how to connect them to particular aspects of the data, before finalizing visual encodings. With the involvement of domain experts and the feedback received from in-service teachers during the researcher-teacher brainstorm sessions and user studies, we summarize the generic goals (Table 1) and low-level tasks (Table 2). In this section, we will elaborate how we finalize the goals and tasks for class video reviewing.

### 4.1 Goals Analysis

| Generic Goals | | |
|---|---|---|
| **G1** | **Analyze teacher/student interactions** | |
| | a | Analyze interactions between students and teacher |
| | b | Analyze interactions between students |
| | c | Identify if some students have more attention from teacher than others |
| **G2** | **Analyze classroom emotional climate** | |
| | a | Summarize the trend of overall emotions during class |
| | b | Identify moments of interest in the class |
| **G3** | **Evaluate class quality and teacher performance** | |
| | a | Evaluate student emotion during class |
| | b | Summarize the teacher interaction with students |
| **G4** | **Discover nuances in the class session** | |

Table 1. Through multiple design activities with teachers and education professionals, we develop a list of goals for climate-focused classroom video review.

The COILS goals analysis developed an end-user endorsed list of generic classroom climate-oriented goals to serve as the basis for interface design and interaction schemes. These goals also informed our evaluation strategy.

With the help of domain experts, we developed an initial draft of goals for video-focused classroom climate analysis. We further refined this list using feedback received during researcher-teacher brainstorm sessions and the subsequent pilot user studies. The feedback specifically targeted how often and to what extent the in-service teachers would tend to align with these goals while analyzing a classroom recording. Additionally, with the involvement of the in-service teachers, we expanded the preliminary high-level list of generic goals into a more detailed list with lower-level sub-goals. These low-level sub-goals were refined to match the user strategies and use cases highlighted by the in-service teachers. Table 1 lists both the high-level goals and the more granular low-level sub-goals.

*G1: Analyze teacher/student interactions.* From previous focus group study, we learnt interactions are the essential goals across all participating teachers. When reviewing classroom recordings, teachers need to grasp and summarize interactions between students and between students and the teacher. During their visual analysis on video recordings, teachers are interested in when, where and how long do they interact with a certain student. They also want to capture these information of interactions that happened between students, which could be very difficult to spot and remember when they are teaching. We generalize these goals as "analyze" interactions in video. Further, teachers mentioned they will specifically looking for any students receive much less attention than others, which is a key aspect in evaluating teachers' performance.

*G2: Analyze classroom emotional climate.* Emotion is another major information teachers will try to extract from most classroom recordings. Teachers are more interested in the overall classroom emotion climate, rather than individual emotion fluctuation. And they indicated interest in comprehending in-classroom dynamics and prefer to adopt a teaching strategy that fosters positive classroom climate. Summarize the trend of emotions during the entire class and identify the peak of certain emotion are the two common sub-goals in emotion analysis.

*G3:Evaluate class quality and teacher performance.* This goal is proposed by teacher trainers, might looks similar to G1 and G2, is more application-focused and directly tackles the evaluation of the teaching performance and class quality using the trends in teachers' attention and student emotions as metrics. Teachers, teacher-evaluators and teacher-trainees are interested in the distribution of teacher's attention towards individual students and potential corresponding bias, and the variation in student emotions during a classroom session is often observed as a means for teacher evaluation and training.

### 4.2 Task Analysis

| Tasks | |
|---|---|
| **G1** | **Analyze teacher/student interactions** |
| T1 | Identify which student has the most/least interactions with others |
| T2 | Identify which student received the most/least attention from the teacher |
| T3 | Locate what time a specific teacher and students interacted |
| T4 | Compare two students, Identify which pairs of students had more interaction during the class. |
| **G2** | **Analyze classroom emotional climate** |
| T5 | Locate emotion peaks for moments of interest |
| T6 | Summarize how estimated emotion changes throughout the class video |
| **G3** | **Evaluate class quality and teacher performance** |
| T7 | Evaluate the overall class climate. |
| T8 | Summarize the quality of interactions between teachers and students |
| T9 | Discover key events during the class |

Table 2. After developing goals, we identify lower-level tasks aligning with each. These tasks serve as a basis for considering alternative visual encodings and interactions. Goal 4, discovering nuance, is supported through the overall exploratory features of the COILS design.

Task analysis targets more granular and specific delineation of generic goals, serving as a means for exploring and justifying visual encodings and interactions in the design of COILS. In terms of refinement, the task list was updated throughout design activities based on reported strategies and needs as specified by participants. We use Brehmer *et al.*'s multi-level task abstraction topology [6] to shape the language of the tasks, defining major tasks that the users will typically undertake to realize the previously specified goals. Task analysis is intended to operationalize a specified goal into actions that include specific elements of the available data. For example, consider G1 which deals with analyzing classroom interactions. T2 aims to identify which student received the most attention from the teacher, a specific instance of G1, and one that connects to available data, the source/target pairs in the eye-gaze interaction data.

The task analysis started at brainstorming stage and continued towards end of iterative design stage. Teachers and teacher trainers came up with an initial list of tasks based on their past experience, then updating and refining the list during workshops and user studies to keep only non-trivial tasks. Table 2 represents the final version of the task list.

## 5 COILS Visual Encodings and Interaction Design

COILS was designed to assist teachers in the review and analysis of classroom recordings by integrating interactive data visualizations of detected eye-gaze interactions and estimated emotion data. The main design challenge was to determine how to visualize these data streams, while aligning with teachers' experiences, comfort with visualization

tools, and providing sufficient functionality to explore classroom videos in new ways.

COILS consists of three major components: The video component (Figure 1A) shows classroom recordings in a video player, with standard seek, play/pause functionality. We use this same video player for the controlled comparative study. Visualization components (Figure 1C and 1D) contain multiple linked views that encode the estimated emotion and eye-gaze data, with associated interaction schemes. Finally, as part of the user study, we add a question panel (Figure 1B), which contains four task questions that the participants were tasked with answering during the mixed-methods user study.

### 5.1 Eye-Gaze Interaction View

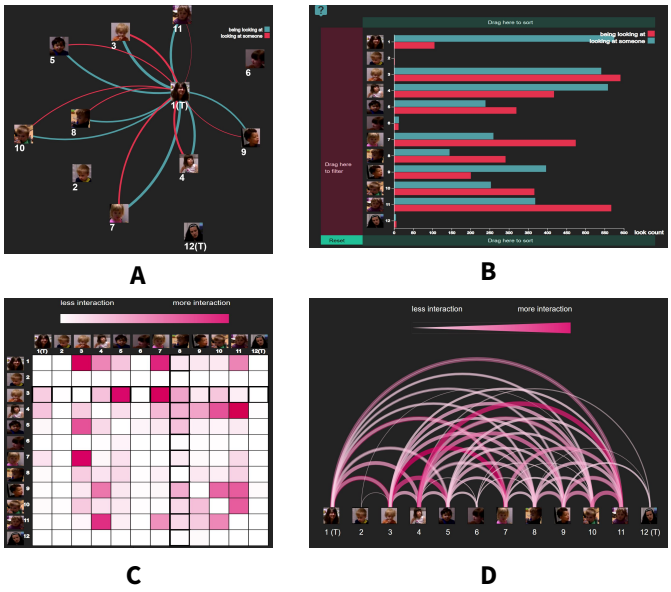


Figure 2. We explored four alternative designs for the gaze interaction display. Pilot studies and workshops with teachers led us to combine the bar chart (B) and arc chart (D) views, to align with the teachers' goals and tasks.

To assist users in discovering and examining moments of interaction between students and teachers (Task T1), we designed a hybrid arc diagram and bar-chart display (Figure 1D). This view includes a grouped bar chart to visualize the frequency of incoming gaze interactions (others to target) and outgoing (source to others). The bars are colored using a gray-scale color scheme to preserve color for use in interaction events, with dark gray showing incoming gaze and light gray the outgoing gaze. Additionally, bars are ordered by students or teachers as they appear in the video. For purposes of the user study, names and demographic information is hidden, however, for individual classroom use this information could easily be made available. Each student and teacher are given an avatar to support reasoning across the abstract visualizations and the video.

To help users accomplish Task T2, identifying which student received the most or least attention from the teacher, we designed interaction schemes to support focused gaze interaction querying. This is driven by the arc chart, which visualizes how much others are detected as looking at the student/teacher of interest. In the arc chart, we use multiple encodings including arc thickness and color saturation for gaze amount, and arrows for gaze direction. When a COILS user selects a person in this view (by click the avatar in bar chart), the encoding change to only display the gaze interactions involving the selected person (Figure 5 A, Mode 1). This mode also helps reviewer complete Task 4 by checking arc chart for different individuals. Clicking on the second person will refine the query even further. While only one arc between the selected people will remain, triggering an update on the timeline view, highlighting timestamps of all detected events between the two people(Figure 5, Mode 2). The Mode 2 also inspired by the Task T3, reviewers need a functionality to quick locate where are those interactions happened, and use it as an index to re-check certain part of the video. Evaluators/Teacher trainers could use feature to locate when teacher interact with students, then evaluate quality of interaction from the video (Task T8). With this information, a COILS user can navigate to potential moments of interest between two people in the classroom.

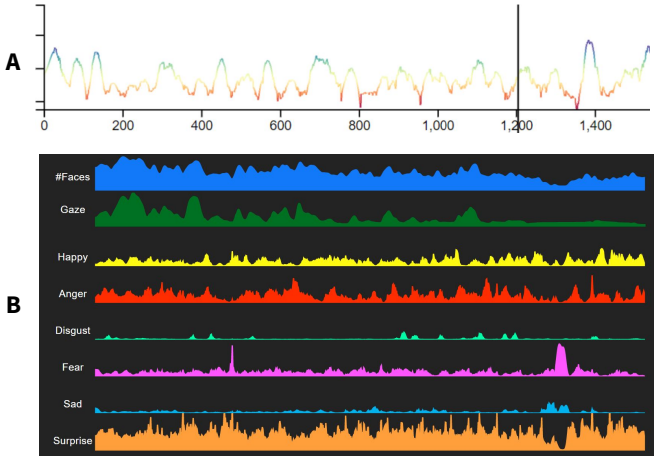


Figure 3. Two alternative design for the Climate Timeline View. A: A line chart showing "detailed" emotion values over time (teachers found area charts more comfortable, one indicated that this view seemed too "scientific"). B: An area chart where each emotion has an associated color (found to interfere with interaction state indicators).

We explored several alternatives for the gaze interaction display, shown in Figure 2. These include a force-directed node link chart of detected interactions (Figure 2A) and a matrix view (Figure 2C). While teachers in workshops and pilot studies specified some desirable facets of these views such as the person-by-person focus in the matrix view, the majority of positive comments were directed at the bar (Figure 2B) and arc (Figure 2D) diagrams. We hypothesize that this may be due to their balance of familiarity and perceived utility. We also experimented with direct manipulation features such as sorting (*e.g.* [8, 9, 45]), however, because pilot studies showed that they were rarely used and possibly confusing, these were removed to help focus the user study.

### 5.2 Climate Timeline View

The climate timeline view focuses on the time dimensions of the eye-gaze and estimated emotion data. Given the number of dimension we need to visualize and their temporal attributes, we adopt juxtaposition techniques to place small multiples area chart beneath the video and align with the video timeline [42]. Each area chart serves as a video controller, where participants can check the video progress and click to seek to a moment of interest in the video. Displaying both aggregate detected gaze interaction data and emotion estimates over time, the climate timeline view shows several peaks and valleys in the video that may guide users to more moments of interest. Similar to the gaze view, the climate timeline uses a gray-scale color scheme and adds color upon interaction events. In addition to gaze interaction events between people, participants can hover over particular emotion moments in the timeline and see "emotion" icons indicating which classroom participants were estimated to be showing that particular emotion. Through pilot testing, we also found that showing the count of detected faces for a particular timeframe in the video can be useful for detecting situations such as scene changes (*e.g.* where students leave the frame). To support this type of query, we added another area chart indicating the number of detected faces over time.

These linked interactions aim to aid users in addressing several tasks, for example T5 "Locate emotion peaks for moments of interest", and T6 "Summarize how estimated emotion changes throughout the class

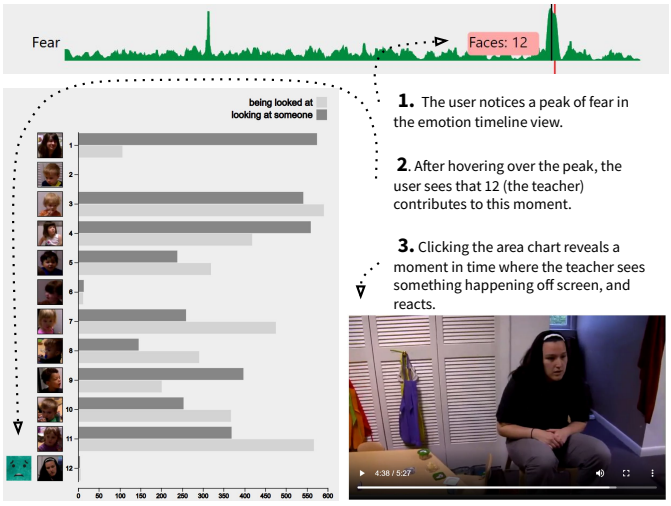


Figure 4. COILS users can select peaks of estimated emotion to explore possible moments of interest. Here, a spike in estimated fear shows an instance where a teacher responds to an off-screen incident.

video". An example usage pattern for Task T5 is shown in Figure 4. In this instance, a user may notice the spike in detected fear, hovering to quickly inspect who might be involved. In this case, a single person – the teacher – is shown, suggesting high confidence. Upon clicking the area chart, the video scrubs to this moment in time displaying an instance where something happens off screen and the teacher's reaction to it. With similar operations, evaluators/teacher trainers could either evaluate class climate from the emotion area charts directly or use these charts to discover key events during the class for further checking in the video. (Task T7&T9)

In designing the climate timeline view, we explored other time-oriented prototypes (see Figure 3). Our first design was a line chart (Figure 3A), chosen because it showed the change in estimated emotion and other values over time with detail. However, particularly when multiple lines were shown at once, participants expressed confusion about what the detailed lines represented and how to interpret them. We then opted for an area chart with color encoding (Figure 3B). While participants indicated more comfort with area charts, we discovered that the color encodings interfered with intended interaction schemes, like clicking on the arc chart to update the timeline view, which could be missed. We therefore adopted a similar grayscale design, using color (*e.g.* green line marks) to show the location of detected events on the area chart timeline.

Other interface considerations include the colorscheme, which evolved over the course of design from a dark to light background to better align with "classroom" color schemes. A small text and visual dashboard appears above the area charts, showing video playback information in addition to other aspects of visualization state, like when two people are selected and interaction events are shown on the timeline. Finally, through pilot testing we developed colloquial language for some of the features, such as "eye contact", "looking at someone" and "being looked at" for gaze interaction instead of the original "incoming" and "outgoing" labels, which were found to be overly formal language. While the question panel is part of the interface, we discuss it's motivation and the design of the questions as part of evaluation.

## 6 Evaluation

To evaluate how these data and interface designs might shape how actual teachers review classroom videos, we constructed a controlled, mixed-methods experiment.

### 6.1 Study Design

Throughout the study, we tested three separate classroom videos. Two of these videos comprised the control condition, meant to simulate reviewing a standard classroom video. The third video is associated with the COILS condition. In each condition, teachers are given the same four question prompts, encouraging in-depth exploration of the videos. The order of the conditions was semi-randomized: a control condition always came first allowing participants to get familiar with the tasks, followed by a random branch of the COILS or the second control condition. Quantitative measures include activity logs throughout the experiment, such as when questions were answered and edited, video interaction events, and various events associated with each of the visualizations. Qualitative measures include a semi-structured interview protocol which immediately followed the experiment, which we will discuss in the subsequent sections.

### 6.2 Classroom Climate Questions

For each of the conditions, we include four classroom climate questions which were designed by researchers in Education and Learning Sciences. These served as a major source of the qualitative user experience data. Pilot studies showed that teachers answered the questions with various levels of detail, and in some cases, preferring to verbally discuss about what they observed in the classroom. Therefore, instead of grading these questions directly, which may be of limited utility and high variance, we use the resulting observations teachers made as part of a larger qualitative analysis between the control and COILS conditions. The questions include:

1. Describe which student the teacher gives the most positive attention to over the course of the video.
2. Describe which student the teacher speaks to the most over the course of the video.
3. Overall, describe which students show the most positive emotion over the course of the video.
4. Overall, describe which pairs of students interact with each other the most over the course of the video.

### 6.3 Study Procedure

Prior to the user studies[2], potential participants in several school districts were sent invitation emails with instructions about the study format and a link to sign up for study slots. User studies were conducted over Zoom, in part to support geographically distributed interest, and due to the novel Coronavirus-19 pandemic restrictions in place at the time of the study. Sessions were recorded to facilitate qualitative analysis. The overall study was designed to be driven by the participant, with little researcher intervention. However, a researcher was always present for any questions or technical support needs.

Participants began with a consent form and an introduction video describing the study, followed by the study conditions. At the beginning of each condition, a walk-through tutorial appeared that guided participants through each of the main interface features. Condition order included one of the control videos (always presented first), then a random branch to either the second control video and COILS condition. At the beginning of the COILS condition, a brief video is shown as a supplementary walk-through specific to the COILS interface. This video shows COILS "in action" to provide participants with a more concrete idea of what the interface looks like as it is being used. In each condition, participants are asked to watch and explore the video in the manner they prefer, before answering the classroom climate questions questions. Participants were invited to think-aloud during the study. After the three conditions, a post-study survey was conducted including general questions about participant demographics and prior visualization tool usage experience. Finally, the researcher invited each participant to discuss their experience and thoughts about classroom observation and COILS. These transcripts form the basis of much of the qualitative data collected.

### 6.4 Quantitative Measures

Throughout the experiment, key interface and study events were logged. These include question answering and editing events, video playback

[2]This study was reviewed and approved by the institutional review board of a university in the Northeastern United States.

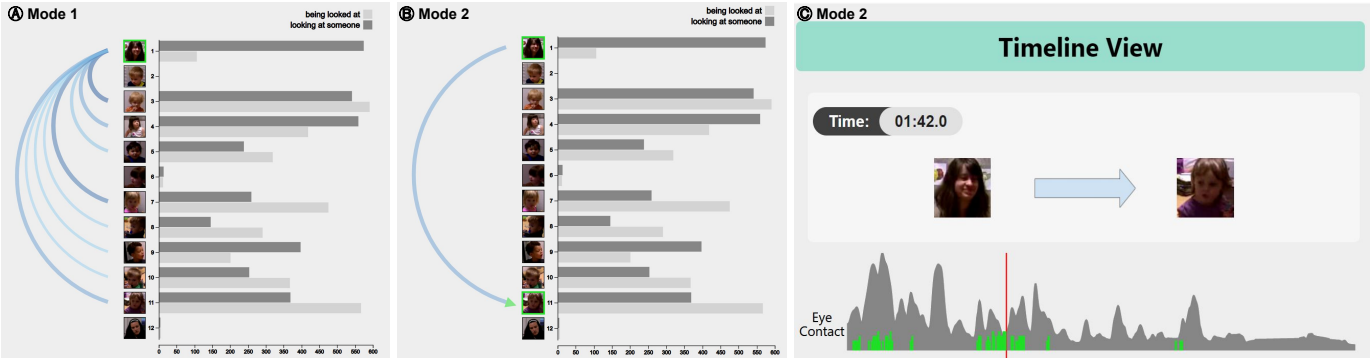


Figure 5. The interaction and linking of the gaze summary view and the timeline view. A) No-one is selected, so the arc chart is hidden. B) One avatar is selected, and the arcs appear to show all outgoing eye-gaze. C) With a second person selected, now only mutual gaze is shown. Teachers can now see these events visualized with green bars in the eye contact area chart, and click to navigate the video to those moments.

changes, and several visualization interface events. Visualization events were distinguished by the focus interface, *e.g.* checking gaze interactions for one or two people, seek events on various charts, *etcetera*. Our goal with this data is to determine which visualization is used most, and whether participants display any specific interaction strategies in classroom video analysis.

### 6.5 Qualitative Measures

Qualitative analysis targets our research questions about evaluating COILS in particular, and the role of AI-driven data and interfaces in the classroom more generally. Each session with a participant produced an (approximately) hour long recording, which we transcribed as part of a qualitative coding analysis, following similar efforts in other visualization studies [15, 25]. To transcribe the studies, we follow a two-step process. First, automatic software was used to create an initial transcript. Then, several researchers were recruited to fix ambiguities in the transcript, making final versions available for the qualitative coding analysis.

We constructed an initial codebook by recruiting several teachers for pilot studies, in a procedure similar to the main study [34]. These recordings and analyses also served to familiarize the research team with the final coding process. Minor changes were made to the codebook throughout the qualitative analysis process as coders worked through each of the interviews and met frequently to review the process. The resulting codebook is shown in Table 4. Themes were developed around participant comments about the performance benefits and limitations of the underlying machine learning pipeline, the COILS interface itself, the use of these technologies in the classroom, and various other considerations.

Each transcript was coded independently by three separate members of the research team. Individual codes were then compared and any coding conflicts, as determined by Inter-Rater Reliability metrics, were resolved through deliberations between the coders. Specifically, any transcript excerpts that were coded but with different codes were resolved to follow the same code. Any disparity in the size or location of the excerpt was resolved to encompass the same portion of text. Finally, codes were analyzed for common themes and consequently divided into sub-categories (see Table 4). In addition to the codes themselves, researchers also recorded direct participant quotes.

## 7 Results

In total, $n = 11$ teachers participated in the study, each contributing a session approximately one hour long [19]. Participant demographics are summarized in Table 3. Ten participants identified as female and one as male, with a mean age of 39.3 years old (SD: 9.16). Participants reported their experience using data visualization tools on a scale of 1 to 7, 1 meaning "*I have never used data visualization*" and 7 meaning "*I use data visualization in my daily work*". The mean of their experience with data visualization tools was 2.45 (SD: 1.72). Finally, the participants also reported their teaching experience in terms of years as a teacher and what grades they have taught, with a mean teaching experience of 13.59 years (SD: 6.09). To preserve anonymity, we refer to the participants as "Participant #".

| Participant ID | Reported Gender | Age | Years Teaching | Data Visualization Experience | Grades Taught |
|---|---|---|---|---|---|
| Participant 1 | F | 31 | 3.5 | 5 | Pre-K |
| Participant 2 | F | 38 | 16 | 1 | Pre-K to 3rd |
| Participant 3 | F | 32 | 14 | 1 | 1st |
| Participant 4 | F | 36 | 11 | 3 | 8th-12th |
| Participant 5 | F | 49 | 19 | 4 | Pre-K to 4th; College |
| Participant 6 | F | 49 | 22 | 1 | 8th |
| Participant 7 | F | 48 | 20 | 2 | 6th to 12th |
| Participant 8 | F | 40 | 18 | 6 | Pre-K to 12th |
| Participant 9 | M | 55 | 15 | 1 | 3rd to 8th |
| Participant 10 | F | 28 | 5 | 1 | 3rd and 6th |
| Participant 11 | F | 27 | 6 | 2 | 9th to 12th; College |
| Stats | | M: 39.3 SD: 9.16 | M: 13.59 SD: 6.09 | M: 2.45 SD: 1.72 | |

Table 3. Study Participant Demographics

## 8 Qualitative Results

Several themes were developed following the transcription of the $n = 11$ teachers that participated in the study. Findings common to multiple participants centers on support and opportunities to improve the COILS interface itself, thoughts on the role of similar machine learning tools in classroom review support, and perceived use cases for COILS in teachers' classroom video review process. In the subsequent sections, we display quotes from the user studies conducted with COILS. We note that the quotes were wordsmithed by the research team to improve readability by removing filler words and adding references to interface elements when possible. For un-edited quotes, please see our supplemental materials.

### 8.1 How Teachers use COILS

While the introduction videos and on-boarding tutorials show COILS functionality, this still leaves participants with a wide range of choices on how to use the system. We observed some users who immediately begin interacting with COILS features to navigate the video, or to explore context while the video is playing.

Some teachers answer questions related to teacher-student interactions by directly using the eye-gaze interaction view: both as a summary tool and to query detected interactions in the video to view specific events. This view received the most positive comments of any other part of the COILS interface, for example:

*I think [ COILS] really puts into perspective rather than just*

| Codebook | |
|---|---|
| Code Name | Details |
| Machine Learning Pipeline | |
| ML Wish list | Desired future features that ML might support |
| Perceived Benefits of ML | How ML helped / could help in classroom analysis |
| Perceived Limitation of ML | Limitations of ML that impeded analysis |
| COILS Interface | |
| Participants' Strategy | How participants use COILS to answer questions |
| Interface Usability | Usability issues in COILS |
| Interface Wish List | Desired future functions for COILS |
| Technologies in the Classroom | |
| Real-life Use Cases | How and where the COILS might be used by in-service teachers |
| Research Observation | Observations and thoughts on the research study |
| Other | |
| Demographics | Participant comments on their own background |
| Experiment Usability | Usability issues for the study interface (*i.e.* not COILS) |

Table 4. We develop a qualitative codebook covering themes observed in the COILS user study sessions. Themes teachers commented on include the capabilities of the machine learning pipeline, the COILS interface and feature set, how such technology could shape their classroom review practice, and miscellaneous other issues.

> *reading... you know, a whole bunch of statistics and data that say, you know, this did was engaged or they weren't engaged. So just being able to really pinpoint where it is, I think is huge.*
>
> *... it was hard for me to tell which kids were talking to each other.* ***So I felt like the eye detection on the right was very helpful because I couldn't [tell] with all the noise, you can't necessarily tell who's talking to who****. -Participant 3*

> *... how many interactions are occurring between each that's really fascinating because it does give you that.* ***[COILS] gives you a more concrete representation of how many times you're actually interacting with the students, or even when the students are interacting with each other****. -Participant 4*

Teachers in our study repeatedly suggested that a primary challenge for teachers is that they can only pay attention to one student at a time. Perhaps more importantly, some also suggested that this was the same for classroom video review, for example:

> *Yeah, I think it's helpful... because like putting myself in the teacher's perspective in that moment, you can't ... look at every face at the same time. And even watching the video, you can't look at every face at the same time...* ***so it was helpful to kind of see, you know, [the] big picture*** *-Participant 10*

> *[COILS] is really remarkable. This would be so helpful especially in a setting where as a teacher, we go back and question, if we've connected with the students enough. To actually have [COILS] and be able to* ***track how many times you've actually interacted with the students compared to what we think we do*** *-Participant 4*

The classroom climate timeline displays the emotion data. We observed teachers using these charts as a live video progress bar and as a controller to explore particular moments of interest. Since one of the prompts was related to happiness, we observed strategies in which participants would watch the video to form an initial opinion about the classroom climate, then use the emotion timeline to validate their thoughts and explore other options:

> *Yeah, so* ***I went back and like kept dragging back and forth of the happy one until I could... and I thought I noticed a couple of kids that were there a lot****. So then I kept staring at them so I picked the [student] I thought that might be the most. -Participant 7*

> *That was helpful. I did find it kind of hard... the emotions were changing so quickly. Like I went in and was trying to kind of toy around with like, you know, positive,* ***confirming my sense of when positive interactions were taking place between people.*** *-Participant 11*

### 8.2 Teacher Perceptions of Machine Learning for Classroom Video Review Support

The data displayed in COILS is the result of a computer-vision focused machine learning pipeline that estimate eye-gaze interactions and potential emotional state from video frames. Teachers were comfortable talking about this pipeline in terms of "AI", a term which the research team adopted for simplicity.

A perceived benefit was related to a more efficient use of time. In control conditions, teachers had to watch a video all the way through, and several noted that this is not always feasible with busy schedules. With COILS, however, several teachers remarked on potential time-saving benefits:

> *So that* ***reflective piece*** *might be helpful, and it might be good to get a quick snapshot. Like you said, I can't always look at everything, but if I looked at the data real fast and then you can go back and watch the video and see, oh yeah, that's what I was doing there, and that's what was happening there...*
>
> *But the fact that* ***[COILS] ability to scan the entire room*** *versus, you know, my attention was basically on the teacher... -Participant 2*

One participant remarked on COILS as acting as an always-available, neutral observer:

> *... this gets developed, it's less personal... you're arguing with a computer versus like somebody who came in and observed you, because then it can be argued, it's subjective versus a computer says to you, well, you talked to this student 15 times and this student two times, and that's just black and white and it was a computer measuring,* ***it's a little harder to argue with, and it maybe makes it a little easier to hear that*** *feedback or to understand it and take it. -Participant 2*

This aligns with observations made during the focus-groups and workshops, where several participants noted that teachers often set concrete goals for themselves related to equity of attention and positive climate in the classroom.

### 8.3 Teacher Perceptions of Machine Learning Limitations

Teachers in our study also commented on cases where they believed the machine learning pipeline could be improved. Notably, some teachers also described *adaptive strategies* they employed to mitigate these shortcomings.

> *It might be a starting off point, and like you said,it's a computer software.* ***[COIL's] gonna miss things. It's not a human, it's not going to catch all the nuances****, but it's a good starting off for a teacher to look and go, ooh, it's telling me I didn't, I'm going to go rewatch and see, and maybe I'm not. -Participant 2*

Another concern is the alignment between what machine learning determines to be an interaction compared to the teachers' personal perceptions of what an interaction in a classroom is.

> *Yeah, I don't know enough about the technology that really speak to this, but it kind of seemed like the technology, if it would breeze back and forth and catch eye contact, it would count it as one, but that's not necessarily a personal interaction... so,* ***while the data may say two people made eye contact****, especially at this level of, you know, students,* ***they're not necessarily really engaging in communication with one another****. -Participant 3*

Other comments were directed at the detection of emotional states:

> *I was interested in the surprised thing. It had a lot of surprise [emotion], and* ***I don't know what the computer is basing surprised on versus sadness****... I'd be interested what factors [ COILS's] looking at in terms of emotion and what it's determining. Is it based on face? Is it based on what they're saying, tone of voice, cadence, all of that kind of stuff. -Participant 2*

### 8.4 COILS in Real Classrooms: Reported Use Cases

Teachers also reported on how they might use a COILS-like interface in their own day-to-day classroom review activities. We categorize these observations into: (1) Teacher Reflection, (2) Administrative Support, and (3) Improving Student Support.

#### 8.4.1 Support for Ongoing Teacher Reflection

Teachers recognized the possible applications of the COILS system in supporting ongoing training and reflection, especially if they can review their teaching more frequently. While there is no replacement for colleagues and educational professionals, such reviews are costly and time-consuming. Participants noted how COILS could help them reflect on the quality of interactions during their lessons, both between teachers and students and students to students:

> *... like an email told me at the end of the day, which students I interacted the most with that day, and then* ***I could reflect on that for the next day****. -Participant 10*

> *I want to reflect on my own teaching* ***based on this instructional technique that I'm implementing****. I think I would have been really happy to do that on like a weekly basis -Participant 11*

Additionally, the participants detailed the impact COILS could have on improving teaching of second language learners.

> *... their name or their student number or whatever it is, being able to see that on every angle, especially for like the special ed and the ELL students that, you know, you can see that they looked sad the whole time or* ***they looked angry the whole time and really kind of pinpointing exactly like what was going on.****. -Participant 8*

> *Like that's an insecurity of mine about my teaching and I'm like, oh my gosh, how do I scaffold for them?* ***How do I provide support for them in these like really heavy literary discussions when they're second language?*** *-Participant 11*

Teachers appear to want to use tools like COILS to explore their interactions with particular students, who they know need extra support.

#### 8.4.2 Administrative Support and Training

In the educational environment, just like with students, teachers are observed and evaluated on how they are running their classroom. Teachers throughout the development of COILS mentioned that these types of reviews could be started at the individual level (by request), the school level, the school district level, or as part of statewide initiatives. Multiple participants discussed the impact COILS could have on feedback by highlighting how administrators could provide more targeted feedback and support for teachers:

> *[ COILS] would actually give really good* ***immediate data and immediate feedback*** *as to what the interactions our in person observer is saying, hey, you didn't interact with 70% of the kids. You only interacted with 20%. Well, now* ***I have data that shows like I actually did interact with all these kids****. -Participant 4*

> *I was a counselor for 10 years in elementary schools and I used to come in sometimes and observe teachers who were having problems with a kid or kids. I mean,* ***this could help counselors too****. -Participant 9*

#### 8.4.3 Improved Student Support

Teachers discuss the implications of using the system to better understand how their students are interacting in class and how these interactions can support their teaching style. Primarily, the teachers see this as a tool to recognize which students may be struggling in class:

> *...picking up on student interactions, kind of just as* ***like a screening tool*** *to see if there are kids who are routinely not making eye contact with anybody else, cause sometimes those kids can kind of fall under the radar too. -Participant 1*

> *Not only at the people right here, but everyone in the classroom see that a lot.* ***If you're not in their eyeshot, they're kind of getting ignored.*** *-Participant 5*

Further building upon the element of being able to recognize who is struggling, participants detailed how the interface can help to better understand how a child interacts over the course of a school day and what may influence their participation levels.

> *So maybe* ***they're asleep first period****, but by the end of the day they've had lunch and* ***recess or something, and they're more active****, you know? So I think it's really cool. -Participant 8*

## 9 Quantitative Activity Log Analysis

For eight users, we have complete logs of their interaction activity throughout the study[3]. This data includes eye-gaze view operations like focusing on interactions between two people, and clicking on the estimated emotion charts to go to a moment in the video. To examine potential differences in COILS visualization usage strategies, we craft an event log plot shown in Figure 6. Figure 6 suggests that seven participants actively used COILS data visualizations at some point in the video review process. One person, P8, does not interact and instead talks about the visualizations in their interview.

From the remaining event plots, we can observe different COILS usage strategies. The gap along the left side in most charts indicates that many participants watched most of the video before interacting. These participants then begin actively exploring the video as a sort of post-hoc analysis. In this analysis, participants appeared to form opinions about the classroom interactions before using the COILS visualizations, where they would then navigate the visualization to confirm their suspicions and explore other possibilities. Participant 6 is a notable exception, who began interacting with the visualizations from the beginning, actively switching between video watching and

[3]Three users were missing due to a data collection error

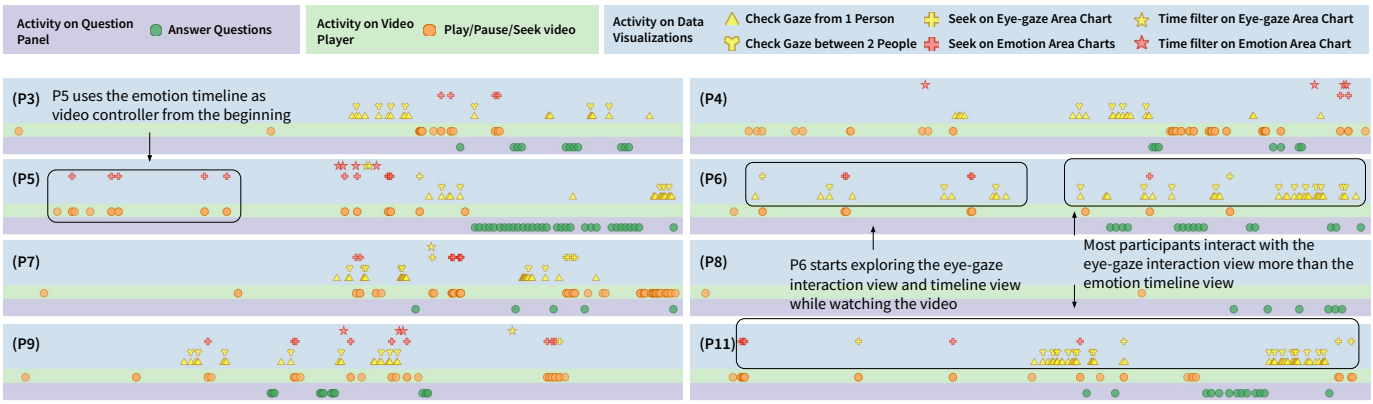


Figure 6. In addition to the qualitative analyses, we log participants activity throughout the experiment. Activity logs of eight participants show that teachers use different strategies when reviewing classroom videos with COILS. Best viewed in color.

examining eye-gaze interactions, and using the area charts as a video controller to navigate between moments of interest.

Looking across participants, examining estimated gaze interactions for one person (incoming/outgoing) and pairs (gaze between two people) were the two most commonly used functions. This usage is also reflected in teacher comments, who specified that being able to identify moments in time where either teachers/students or students/students interacted was a benefit, particularly given their limited capacity to investigate interactions throughout while simply watching the video. Seven participants also used the emotion climate timeline to investigate potential moments of interest in the video. However, the eye-gaze features were used more frequently.

During the qualitative evaluation, participants were instructed to explore the COILS interface freely. Participants generally used the interface to compare or confirm their perception and opinion on classroom dynamics. For instance, they used the eye-gaze interaction to verify whether an interaction actually took place as they observed, between whom, and for how long. While a limitation of the underlying machine learning to capture all the nuances in classroom interactions were raised, teachers are confident that the interactivity, linked views, and the provided controls are helpful to provide an increased perspective about what is going on in the classroom, and reflect on important moments they can miss while being recorded in a real-life setting. Comparing the qualitative observation with the quantitative metrics shows that the views that participants visited to form their answers to our evaluation questions align with a reasonable pattern that was expected while designing the tasks for this study. This suggests that the COILS interface is intuitive and has the potential in accompanying teachers in their routine video review activities.

## 10 Discussion

Through the design, development, and evaluation of COILS, we learn much about teacher needs and goals in classroom video analysis, and how visualization can support them. More broadly, we also reflect on the role of visualization and machine learning in the classroom and ongoing challenges.

**Relative Effectiveness of COILS Views** Quantitative and qualitative results of the COILS user study suggests that interactive visualizations can shape teachers' video review process in beneficial ways. Teachers made several positive remarks on the prospect of using such tools as a means for on-demand feedback for their teaching, and in service of examining their interactions with students for equity, climate, *etcetera*.

In exploring classroom dynamics, teachers most frequently turned to the eye-gaze interaction view. This view gave machine-estimated summaries of the count and direction of interactions, but perhaps more importantly, it allowed teachers to navigate to where machine learning "thought" these interactions had occurred to investigate for themselves. Teachers' frequent mention of these features and elaboration on the strategies they used suggest that it may be a valuable area for future design efforts. For example, future efforts might transform the eye-gaze interaction data into summaries for pairs of students, especially teachers and students, to provide teachers with a more rapid way to examine how they interact with their students.

The estimated emotion data and views, while generally comprehensible, were used less frequently overall. Part of this may be the design and interaction set of the emotion climate view. We have a comparison point in the *EmotionCues* work from Zeng *et al.* [63], which uses a more complex storyline-based timeline view for estimating emotions over time. Our design activities showed that teachers were somewhat uncomfortable with unfamiliar visualizations (*e.g.* node-link, matrix diagrams), leading us to develop area charts, which teachers rated more highly by comparison. At the same time, there is likely room for more complex views in future iterations in this space as multiple teachers indicated that they would expect training workshops to be provided with a tool like COILS. This raises the prospect in the future of introducing more feature-rich visualizations and data, perhaps with uncertainty measures.

**Limitations of Machine Learning in the Classroom** Another possibility for the observed differences in teacher usage patterns and preferences is that the eye-gaze interaction data is simply more useful in its current state than the estimated emotion data. Estimating emotion data remains a challenging problem in machine learning, particularly for videos where people may be moving around in the frame or conducting different activities. In our case, several teachers noticed in the emotion visualizations a constant high level of estimated *surprise*, which would normally indicate faces detected with mouth-open and eyebrows raised. Investigating this, we found that talking and eating could also trigger the surprise detector leading to false positives. While this is a single example, it did appear to impact some teachers' perception of the utility of the emotion data and views given their mentions of it in the interview session. Given ongoing concerns and challenges with the use of emotion data in automated systems (*e.g.* [4]), design and training efforts should focus on adequately providing teachers with information about the limitations of such detectors, even as the state-of-the-art continues to improve.

In the interview sessions, teachers mentioned another limitation of machine learning pipeline: its reliance on visual information. Other contextual cues, like speech and background noise, can impact classroom dynamics. Such features could readily be integrated into a climate-focused video review tool. For instance, visually encoding sound information might help teachers distinguish between a student who is eating and a student expressing surprise, reducing false positives. Furthermore, as teachers reflect on their own behavior and interactions with students, speech could be combined with eye-gaze detection to determine when people are speaking to each other, narrowing the moments of interest a teacher would inspect and improving efficiency.

Perceived Uses of COILS in the Classroom Teachers offered multiple ideas about how COILS and similar systems could be used to benefit their day to day teaching practice. Teacher training was one of the most commonly suggested use cases. Participants indicated that video recording and review was more frequent for trainee teachers, requiring extensive effort on the part of the trainee and reviewers (usually more senior teachers). Teachers envisioned COILS playing a dual role: as a means for trainees to get rapid feedback on their teaching, and as a tool for reviewers to guide the trainees through a video reflection.

Teachers were enthusiastic about the prospect of using COILS for their own personal reflection and goal setting. Teachers mentioned school- and district-wide initiatives targeting particular aspects of classroom climate, which would include video recording and review as part of assessing progress. In these scenarios, teachers envisioned being able to periodically review their interactions with specific students, to evaluate themselves in relation to initiative goals. Teachers also thought of COILS as a way for them to make better use of sporadic in-person reviews from school administration. Teachers noted that a reviewer might feel that the classroom climate is different on the day of the visit than it usually is. (For example, a visiting principal would likely impact students' behavior.) However, with COILS, teachers described the possibility of having conversations with the reviewer using the interface to explore past interactions with students, providing deeper context.

In either case, supporting these use cases would likely require additional design and evaluation activities. For one, alongside these use cases teachers also mentioned additional feature needs to realize them, like annotation and various customization options for their classrooms. For the administration-review scenario, it is also likely that teachers would need to be able to bookmark particular interactions in preparation for meetings with people reviewing their classroom.

## 11 Conclusion and Future Work

In this paper, we describe the design and evaluation of COILS, a visualization system targeting classroom-climate recording review. Across multiple design activities with teachers and education professionals, we crafted a set of tasks, goals, and design constraints for visualizations supporting classroom video review. In a controlled mixed-methods study, we evaluated how visualizations of eye-gaze interactions and estimated emotion data shape teachers' video review process. Results of these studies suggest that COILS brings several tangible benefits to teachers' video review process, while also highlighting future opportunities for design and directions for the underlying machine learning models.

Future work in this area could include similar studies with teachers, but using videos from their own classrooms instead of training videos. This addition could elicit more in-depth use cases and feature requests given teachers' familiarity with students in their own classrooms. Another possibility is to move beyond directly visualizing eye-gaze and estimated emotion data, instead creating models that segment the data into models of "moments of interest" based on what teachers look for in classroom videos. Such an approach would expand the design space of possible visualizations and interaction schemes, potentially enabling teachers to more readily hone in on meaningful interactions in their classrooms.

## References


[1] K. Ahuja, D. Kim, F. Xhakaj, V. Varga, A. Xie, S. Zhang, J. E. Townsend, C. Harrison, A. Ogan, and Y. Agarwal. EduSense: Practical Classroom Sensing at Scale. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 3(3):1–26, 2019.

[2] K. Ahuja, D. Shah, S. Pareddy, F. Xhakaj, A. Ogan, Y. Agarwal, and C. Harrison. Classroom digital twins with instrumentation-free gaze tracking. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pp. 1–9, 2021.

[3] A. M. Aung, A. Ramakrishnan, and J. R. Whitehill. Who are they looking at? automatic eye gaze following for classroom observation video analysis. *International Educational Data Mining Society*, 2018.

[4] L. F. Barrett, R. Adolphs, S. Marsella, A. M. Martinez, and S. D. Pollak. Emotional Expressions Reconsidered: Challenges to Inferring Emotion from Human Facial Movements. *Psychological science in the public interest*, 20(1):1–68, 2019.

[5] N. Bostrom and E. Yudkowsky. The Ethics of Artificial Intelligence. In *Artificial intelligence safety and security*, pp. 57–69. Chapman and Hall/CRC, 2018.

[6] M. Brehmer and T. Munzner. A Multi-Level Typology of Abstract Visualization Tasks. *IEEE transactions on visualization and computer graphics*, 19(12):2376–2385, 2013.

[7] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh. Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7291–7299, 2017.

[8] M. Cavallo and Ç. Demiralp. A Visual Interaction Framework for Dimensionality Reduction Based Data Exploration. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pp. 1–13, 2018.

[9] D. Coffey, C.-L. Lin, A. G. Erdman, and D. F. Keefe. Design by Dragging: An Interface for Creative Forward and Inverse Design with Simulation Ensembles. *IEEE transactions on visualization and computer graphics*, 19(12):2783–2791, 2013.

[10] K. Crawford et al. Time to Regulate AI that Interprets Human Emotions. *Nature*, 592(7853):167–167, 2021.

[11] G. Daniel and M. Chen. *Video Visualization*. IEEE, 2003.

[12] S. K. D'Mello, A. M. Olney, N. Blanchard, B. Samei, X. Sun, B. Ward, and S. Kelly. Multimodal Capture of Teacher-Student Interactions for Automated Dialogic Analysis in Live Classrooms. In *Intl. conference on multimodal interaction*, 2015.

[13] P. J. Donnelly, N. Blanchard, B. Samei, A. M. Olney, X. Sun, B. Ward, S. Kelly, M. Nystrand, and S. K. D'Mello. Multi-Sensor Modeling of Teacher Instructional Segments in Live Classrooms. In *ACM international conference on multimodal interaction*, pp. 177–184. ACM, 2016.

[14] N. Esau, E. Wetzel, L. Kleinjohann, and B. Kleinjohann. Real-Time Facial Expression Recognition using a Fuzzy Emotion Model. In *2007 IEEE international fuzzy systems conference*, pp. 1–6. IEEE, 2007.

[15] M. Fan, K. Wu, J. Zhao, Y. Li, W. Wei, and K. N. Truong. VisTA: Integrating Machine Intelligence with Visualization to Support the Investigation of Think-Aloud Sessions. *IEEE Transactions on Visualization and Computer Graphics*, 26(1):343–352, 2019.

[16] R. Flewitt. Using Video to Investigate Preschool Classroom Interaction: Education Research Assumptions and Methodological Practices. *Visual communication*, 5(1):25–50, 2006.

[17] Y. M. Goodman. Observing Children in the Classroom. *Making sense of learners making sense of written language: The selected works of Kenneth S. Goodman and Yetta M. Goodman*, pp. 197–214, 2014.

[18] B. K. Hamre, R. C. Pianta, J. T. Downer, J. DeCoster, A. J. Mashburn, S. M. Jones, J. L. Brown, E. Cappella, M. Atkins, S. E. Rivers, et al. Teaching Through Interaction: Testing a Developmental Framework of Teacher Effectiveness in Over 4,000 Classrooms. *The Elementary School Journal*, 113(4):461–487, 2013.

[19] M. Hennink and B. N. Kaiser. Sample Sizes for Saturation in Qualitative Research: A Systematic Review of Empirical Tests. *Social Science & Medicine*, p. 114523, 2021.

[20] E. Hindalong, J. Johnson, G. Carenini, and T. Munzner. Towards Rigorously Designed Preference Visualizations for Group Decision Making. In *2020 IEEE Pacific Visualization Symposium (PacificVis)*, pp. 181–190. IEEE, 2020.

[21] C. Howe and N. Mercer. Children's social development, peer interaction and classroom learning. *The Primary Review*, Research Survey 2/1b, 2007.

[22] R. E. Jack, O. G. Garrod, H. Yu, R. Caldara, and P. G. Schyns. Facial Expressions of Emotion are not Culturally Universal. *Proceedings of the National Academy of Sciences*, 109(19):7241–7244, 2012.

[23] D. K. Jain, P. Shamsolmoali, and P. Sehdev. Extended Deep Neural Network for Facial Emotion Recognition. *Pattern Recognition Letters*, 120:69–74, 2019.

[24] H. Jiang and E. Learned-Miller. Face Detection with the Faster R-CNN. In *IEEE Automatic Face & Gesture Recognition*, 2017.

[25] S. Kandel, A. Paepcke, J. M. Hellerstein, and J. Heer. Enterprise Data Analysis and Visualization: an Interview Study. *IEEE transactions on visualization and computer graphics*, 18(12):2917–2926, 2012.

[26] M. T. Kane. Validating the Interpretations and Uses of Test Scores. *Journal of Educational Measurement*, 50(1):1–73, 2013.

[27] A. Kapoor, S. Mota, and R. W. Picard. Towards a Learning Companion that Recognizes Affect. In *AAAI Fall symposium*, vol. 543, pp. 2–4, 2001.

[28] S. Kontos and A. Wilcox-Herzog. Teachers' Interactions with Children: Why Are They So Important? Research in Review. *Young children*, 52(2):4–12, 1997.

[29] P. C. Kumar, J. Vitak, M. Chetty, and T. L. Clegg. The Platformization of the Classroom: Teachers as Surveillant Consumers. *Surveillance & Society*, 17(1/2):145–152, 2019.

[30] L. Kyriakides, B. P. Creemers, and P. Antoniou. Teacher Behaviour and Student Outcomes: Suggestions for Research on Teacher Training and Professional Development. *Teaching and teacher education*, 25(1):12–23, 2009.

[31] D. Lee, I. T. Arthur, and A. S. Morrone. Using Video Surveillance Footage to Support Validity of Self-Reported Classroom Data. *International Journal of Research & Method in Education*, 40(2):154–180, 2017.

[32] D. Leslie, L. Holmes, C. Hitrova, and E. Ott. Ethics Review of Machine Learning in Children's Social Care. *Leslie, D., Holmes, L., Hitrova, C. & Ott, E.(2020). Ethics review of machine learning in children's social care. [Report] London, UK: What Works for Children's Social Care*, 2020.

[33] X. Li and T. Zhang. An Exploration on Artificial Intelligence Application: From Security, Privacy and Ethic Perspective. In *2017 IEEE 2nd International Conference on Cloud Computing and Big Data Analysis (ICCCBDA)*, pp. 416–420. IEEE, 2017.

[34] K. M. MacQueen, E. McLellan, K. Kay, and B. Milstein. Codebook Development for Team-Based Qualitative Analysis. *Cam Journal*, 10(2):31–36, 1998.

[35] A. Martinez and S. Du. A Model of the Perception of Facial Expressions of Emotion by Humans: Research Overview and Perspectives. *Journal of Machine Learning Research*, 13(5), 2012.

[36] A. J. Mashburn, R. C. Pianta, B. K. Hamre, J. T. Downer, O. A. Barbarin, D. Bryant, M. Burchinal, D. M. Early, and C. Howes. Measures of Classroom Quality in Prekindergarten and Children's Development of Academic, Language, and Social Skills. *Child development*, 79(3):732–749, 2008.

[37] D. M. Medley and H. E. Mitzel. A Technique for Measuring Classroom Behavior. *Journal of Educational Psychology*, 49(2):86, 1958.

[38] A. Mehrabian and S. R. Ferris. Inference of Attitudes from Nonverbal Communication in Two Channels. *Journal of consulting psychology*, 31(3):248, 1967.

[39] D. K. Meyer and J. C. Turner. Discovering Emotion in Classroom Motivation Research. *Educational psychologist*, 37(2):107–114, 2002.

[40] A. Mollahosseini, B. Hasani, and M. H. Mahoor. Affectnet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild. *arXiv preprint arXiv:1708.03985*, 2017.

[41] M. Montague and C. Rinaldi. Classroom Dynamics and Children at Risk: A Followup. *Learning Disability Quarterly*, 24(2):75–83, 2001.

[42] D. Moritz and D. Fisher. Visualizing a million time series with the density line chart. *arXiv preprint arXiv:1808.06019*, 2018.

[43] S. C. Neoh, L. Zhang, K. Mistry, M. A. Hossain, C. P. Lim, N. Aslam, and P. Kinghorn. Intelligent Facial Emotion Recognition Using a Layered Encoding Cascade Optimization Model. *Applied Soft Computing*, 34:72–93, 2015.

[44] B. Nye, S. Konstantopoulos, and L. V. Hedges. How Large are Teacher Effects? *Educational evaluation and policy analysis*, 26(3):237–257, 2004.

[45] D. Orban, D. F. Keefe, A. Biswas, J. Ahrens, and D. Rogers. Drag and Track: A Direct Manipulation Interface for Contextualizing Data Instances within a Continuous Parameter Space. *IEEE transactions on visualization and computer graphics*, 25(1):256–266, 2018.

[46] M. T. Owens, S. B. Seidel, M. Wong, T. E. Bejines, S. Lietz, J. R. Perez, S. Sit, Z.-S. Subedar, G. N. Acker, S. F. Akana, et al. Classroom Sound can be Used to Classify Teaching Practices in College Science Courses. *Proceedings of the National Academy of Sciences*, 114(12):3085–3090, 2017.

[47] N. Peachey. Using Audio or Video to Record Classroom. *British Council Action Research Articles*, 2008.

[48] E. S. Peisner-Feinberg, M. R. Burchinal, R. M. Clifford, M. L. Culkin, C. Howes, S. L. Kagan, and N. 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.

[49] J. W. Pennebaker, R. L. Boyd, K. Jordan, and K. Blackburn. The Development and Psychometric Properties of LIWC2015. 2015.

[50] R. C. Pianta, K. M. La Paro, and B. K. Hamre. *Classroom Assessment Scoring System™: Manual K-3*. Paul H Brookes Publishing, 2008.

[51] C. C. Ponitz, M. M. McClelland, J. Matthews, and F. J. Morrison. A Structured Observation of Behavioral Self-Regulation and its Contribution to Kindergarten Outcomes. *Developmental psychology*, 45(3):605, 2009.

[52] A. Ramakrishnan, E. Ottmar, J. LoCasale-Crouch, and J. Whitehill. Toward Automated Classroom Observation: Predicting Positive and Negative Climate. In *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, pp. 1–8. IEEE, 2019.

[53] A. Recasens, C. Vondrick, A. Khosla, and A. Torralba. Following Gaze in Video. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1435–1443, 2017.

[54] R. J. Reed and N. Oppong. Looking Critically at Teachers' Attention to Equity in their Classrooms. *The Mathematics Educator*, 2005.

[55] L. Rhue. Racial Influence on Automated Perceptions of Emotions. *Available at SSRN 3281765*, 2018.

[56] M. Romero, J. Summet, J. Stasko, and G. Abowd. Viz-a-vis: Toward visualizing video through computer vision. *IEEE Transactions on Visualization and Computer Graphics*, 14(6):1261–1268, 2008.

[57] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 815–823, 2015.

[58] K. Simonyan and A. Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[59] G. K. Tam, H. Fang, A. J. Aubrey, P. W. Grant, P. L. Rosin, D. Marshall, and M. Chen. Visualization of Time-Series Data in Parameter Space for Understanding Facial Dynamics. In *Computer Graphics Forum*, vol. 30, pp. 901–910. Wiley Online Library, 2011.

[60] P. C. van der Sijde and W. Tomic. The Influence of a Teacher Training Program on Student Perception of Classroom Climate. *Journal of Education for Teaching*, 18(3):287–295, 1992.

[61] Z. Wang, X. Pan, K. F. Miller, and K. S. Cortina. Automatic Classification of Activities in Classroom Discourse. *Computers & Education*, 78:115–123, 2014.

[62] A. Zadeh, R. Zellers, E. Pincus, and L.-P. Morency. Multimodal Sentiment Intensity Analysis in Videos: Facial Gestures and Verbal Messages. *IEEE Intelligent Systems*, 31(6):82–88, 2016.

[63] H. Zeng, X. Shu, Y. Wang, Y. Wang, L. Zhang, T.-C. Pong, and H. Qu. Emotioncues: Emotion-Oriented Visual Summarization of Classroom Videos. *IEEE transactions on visualization and computer graphics*, 27(7):3168–3181, 2020.

[64] H. Zeng, X. Wang, A. Wu, Y. Wang, Q. Li, A. Endert, and H. Qu. EmoCo: Visual Analysis of Emotion Coherence in Presentation Videos. *IEEE transactions on visualization and computer graphics*, 26(1):927–937, 2019.