Quantifying Classroom Instructor Dynamics with Computer Vision

Nigel Bosch[1], Caitlin Mills[2], Jeffrey D. Wammes[3], and Daniel Smilek[4]

[1] University of Illinois at Urbana-Champaign, Urbana, IL 61820, USA
[2] University of British Columbia, Vancouver, BC V6T 1Z, Canada
[3] Yale University, New Haven, CT 06511, USA
[4] University of Waterloo, Waterloo, ON N2L 3G1, Canada
pnb@illinois.edu

Abstract. Classroom teachers utilize many nonverbal activities, such as gesturing and walking, to maintain student attention. Quantifying instructor behaviors in a live classroom environment has traditionally been done through manual coding, a prohibitively time-consuming process which precludes providing timely, fine-grained feedback to instructors. Here we propose an automated method for assessing teachers' non-verbal behaviors using video-based motion estimation tailored for classroom applications. Motion was estimated by subtracting background pixels that varied little from their mean values, and then noise was reduced using filters designed specifically with the movements and speeds of teachers in mind. Camera pan and zoom events were also detected, using a method based on tracking the correlations between moving points in the video. Results indicated the motion estimation method was effective for predicting instructors' non-verbal behaviors, including gestures (kappa = .298), walking (kappa = .338), and camera pan (an indicator of instructor movement; kappa = .468), all of which are plausibly related to student attention. We also found evidence of predictive validity, as these automated predictions of instructor behaviors were correlated with students' mean self-reported level of attention (e.g., r = .346 for walking), indicating that the proposed method captures the association between instructors' non-verbal behaviors and student attention. We discuss the potential for providing timely, fine-grained, automated feedback to teachers, as well as opportunities for future classroom studies using this method.

Keywords: Instructor non-verbal behaviors, attention, motion estimation.

1 Introduction

Classroom lecturing can be a daunting task. Presenting the learning material in a meaningful way is only half the battle, as maintaining students' attention and engagement is perhaps equally challenging. One way in which instructors might manage students' attention is through titrating their own behaviors during the lecture (e.g. moving around, altering their volume) in response to their perception of student attentiveness. Indeed, research has shown that an instructor's behavior (e.g., head nodding) is related not only to students' learning [1–4], but also to their characterizations of instructors (i.e. competence and enthusiasm). To date, however, very little work has focused on quantifying the moment-to-moment dynamics of instructors' behavior in the classroom. Similarly, there is a paucity of work developing tools that can provide live feedback to classroom instructors about their behaviors. The ultimate goal of this work is therefore to fulfill this need and provide automated feedback to instructors regarding their behavioral dynamics in the classroom. A critical first step toward this goal, and the focus of the current paper is to build an automatic quantification system which employs a video-based method in the wild.

Previous research evaluating teachers' non-verbal behaviors has primarily focused on either simulations of instructor behavior by professional actors [5], manual evaluations of behavior [6, 7], human-like avatars of teachers in e-learning [8], or laboratory environments that may not adequately approximate actual classrooms [9, 10]. Although these methods have led to some valuable empirical insights about how instructor behaviors influence learning, they cannot be easily parlayed into feedback systems for instructors. We addressed this gap by creating an automated approach to estimating instructor behaviors (e.g., walking, gesturing, interacting with a presentation) with techniques from artificial intelligence and computer vision. The method we present in this paper also has the advantage that it does not require any specialized sensors, such as depth sensors [10]. Instead, it requires only a video camera. Furthermore, real classroom videos were recorded from a vantage point at the back of the room, thus alleviating privacy concerns related to images of student faces being recorded.

2 Related Work

2.1 Impact of instruction behaviors

Instructor behavior is critical to assess given its consistent relationship with various aspects of learning. Witt et al. [2] conducted a meta-analysis of 81 studies including over 24,000 students, and discovered a significant correlation between teachers' non-verbal behaviors and students' self-reported perceptions of how much they were learning ($r = .510$). Furthermore, teachers' non-verbal behaviors were correlated almost as strongly with students' affective learning ($r = .490$), which is a strong indicator of one's enjoyment of the course, and likelihood of enrolling in a future related course.

Computerized learning environments have allowed researchers to precisely manipulate the non-verbal behaviors that (virtual) teachers exhibit and test their influence on students' perceptions. Alseid & Rigas [8] studied the effects of facial expressions (e.g., happy, interested), hand gestures (e.g., pointing, chin stroking), and walking in the context of a computerized learning environment with a virtual teaching agent. Students rated their perceptions of these instructor activities before and after the study. Several of the teacher activities were rated significantly more positively after the study than before, including the two most well-liked activities, pointing (100% positive rating post-study) and walking (98% positive rating post-study). While it is unclear what the impact of these positive ratings would be on learning, it is clear that students develop a preference for particular instructor behaviors, which may in turn foster improved attention or learning.

2.2 Automatic evaluation of teaching behaviors

Multiple efforts have been made to develop automated methods for evaluating teaching and presentation style, including evaluation of non-verbal behaviors. TeachLivE is one example of a software platform designed to capture the behaviors of teachers as they interact with 3D virtual students, and to automatically provide real-time feedback about those behaviors [10]. Participants were 34 teachers in training, half of whom received automated real-time feedback on non-verbal behaviors (specifically, open body postures such as arms hanging down versus arm-crossing and other closed postures) in their first of two sessions with TeachLivE, and half who received feedback in their second session. The participants who received feedback in the first session displayed more open postures in the second session than the other participants, despite receiving no further feedback. This study demonstrated that it is possible to perform real-time assessment of teaching behaviors, and that instructors can modify their behavior based on this feedback. However, TeachLivE requires a close, unobstructed view of the instructor, to allow their behaviors to be tracked with a depth-sensing camera. This is a limitation which would prevent TeachLivE from being broadly applicable in many common lecture hall classroom environments.

Presentation Trainer is another training platform designed to assess non-verbal behaviors and give corresponding feedback to presenters [9]. While it is not specifically intended for teacher training, it does include relevant feedback about the behaviors quantified by TeachLivE (open versus closed posture), as well as stance (shifting side to side, which conveys being uncomfortable). In one empirical investigation of Presentation Trainer, university professionals who received this feedback about behavior self-reported significantly more learning than a matched control group (24.5% more, $p < .05$), indicating that they found the automated feedback helpful for improving their presentations. In a follow-up study, nine students gave presentations to their peers before and after completing a training session with Presentation Trainer [11]. The group of peers rated the quality and frequency of gesture use for both presentations, and ratings were significantly higher *after* training (27.9% improved, $p < .01$). Together, these studies demonstrate that automated evaluation methods exhibit strong potential for the improvement of non-verbal communication skills. However, similar to TeachLivE, Presentation Trainer suffers the shortcoming that it requires a depth-sensing camera and a clear, close and unobscured video of the speaker, both of which are unlikely to be available in typical classroom environments.

2.3 Feasibility of camera-based motion tracking

Tracking motion in video is a well-studied problem in the field of computer vision. Applications include tracking the movement of people [12], tracking the movement of key visual points (e.g., corners, edges [13]), and background subtraction to find visual changes over time [14]. Camera-based human motion tracking research has typically focused on skeletal tracking and detection of individual body parts [15–17]. However, unobstructed high-quality video of instructors is difficult to acquire, so alternative methods are needed. Kory-Westlund et al. [18] made strides toward this goal with a human motion tracking method that does not require detection of people or individual body parts. Their motion measure correlated quite well with other markers of movement, including posture changes measured by a pressure-sensitive chair (mean r = .708), and hand gestures measured by a wrist-mounted accelerometer (mean r = .720). This method was optimized for measuring the motions of a person seated in front of a computer with a stationary camera, however, and is unlikely to be capable of effectively tracking more dynamic classroom behaviors.

2.4 Research Questions

We extracted novel estimations of instructor motion, camera pan (rotation back and forth), and camera zoom from classroom videos taken with a standard digital camera. These estimations were then used as features in machine-learned models that detected teacher activities including walking, gesturing, and presentation usage (slide changes). As a proof of concept, we aimed to answer the following research questions: 1) How well can we automatically detect instructors' non-verbal behaviors using only amateur videos taken from the back of a classroom?; and 2) Do the instructor activities detected with this method correspond to student attention in the classroom?

3 Method

We took a multi-step approach to answering the foregoing research questions. We collected classroom videos and manually annotated them for instructors' non-verbal behaviors, then applied methods to estimate motion, camera pan, and camera zoom in the videos.[1] Finally, we applied supervised machine learning methods test whether these estimated features could reliably predict the instructors' non-verbal behaviors. These steps are described in detail below.

3.1 Classroom videos and students' self-reported attention

Nine classroom lectures were recorded over the span of six days at a Canadian university. There were three videos each of three different instructors, all of whom taught undergraduate psychology classes. The lectures were completely naturalistic and were not manipulated in any way for the study. Lectures included the common elements one might expect: speaking, question answering, referencing presentation slides, and occasionally watching videos on a projector screen. Recordings occurred in two different classrooms, each of which was equipped with a similar setup: a lectern/podium and a stage-like space for the instructor to walk. A researcher started the video camera, which

[1] We have made the code for motion, camera pan, and zoom estimation available online at https://github.com/pnb/classroom-motion

was placed at the back of the classroom, and actively panned and zoomed the camera to keep the instructor and presentation slides in frame.

We also obtained self-reports of attention from the students during the lectures. Students who agreed to participate (N = 76) downloaded a thought-probe application to their laptop computers. This application displayed a notification in one corner of the computer screen up to five times per class. The notification prompted students to report their level of attention using a continuous scale ranging from *Completely mind wandering* to *Completely on task*. Students were instructed to introspect about their mental state just before the thought-probe appeared, with *mind wandering* defined to the students as "thinking about unrelated concerns" and *on task* defined as "thinking about the lecture".

3.2 Video annotation

To determine whether our motion estimation method was associated with the non-verbal behaviors of the recorded instructors, ground truth labels of these behaviors and related events were required. To this end, classroom videos were retrospectively coded to describe instructors' non-verbal behaviors (i.e. gesturing or movement), environmental changes (e.g. lighting or camera pan), and student interaction (e.g. asking or fielding questions). A total of 5,415 annotations were made, of which 24.9% were camera pan, 0.3% camera zoom, 0.7% room lighting change, 0.1% instructor playing video, 3.8% student asking question, and 70.2% other (which were then explained in greater detail in the coder's comments). Because there were relatively few instances of camera zoom, room lighting changes, video playing, and question asking, these annotations were not considered further. We extracted additional annotations from the comments documented by the coder, including 13.1% instructor walking, 33.3% gesturing, and 6.2% presentation slide change. Finally, we expected that camera pan events would be closely related to walking, so we created an additional set of annotations pooling across cases where either the camera panned or the instructor was coded as walking (37.3% of annotations).

Although video annotations were made at the precise moment that a relevant event was observed, the videos were pooled into 30-second segments (n = 1,431) to reduce noise for automatic classification[2]. Each 30 s segment was assigned a binary label for each annotation category. For instance, if the first and only camera pan occurred 43 s into a particular lecture, the first 30 s epoch (0-30 s) would be coded as 0, the second (30-60 s) as 1 and the third as 0 (60-90 s). In this manner, we derived a time series of binary ratings at regular intervals, for each possible annotation class. Preprocessing the data this way allowed for automatic classification via supervised machine learning with features estimated from the videos.

[2] We experimented with a range of segment lengths from 10 to 60 s, finding that 30-60 s segments provided equivalent results and were consistently better than 10 or 20 s. We thus segmented at 30 s intervals to provide the finest granularity from the 30-60 s range.

3.3 Motion estimation

We estimated motion in two steps: 1) raw motion detection and 2) human-oriented filtering. Raw motion was detected by applying a background subtraction method which learns a Gaussian mixture model describing the video's pixel values [14]. This method learns descriptive statistics about pixels in the video, and tags pixels with low variance over time as *background*, and those with high variance as *motion* pixels. Consider the circumstance where the instructor moves across a particular region of the video frame. Intuitively, the pixels the instructor crosses over will change dramatically over time and be correctly labelled as motion. However, other artifacts may be erroneously labelled as motion, so the raw motion data must be filtered.

We observed several different sources of such visual noise in classroom videos (illustrated in Fig. 1), each of which caused background pixels to be incorrectly labelled as motion pixels: 1) electronic noise, causing random pixel variation especially under low-light conditions, 2) camera vibration, caused by brief movements in the camera mount, and 3) intentional camera pan and zoom actions performed to follow the teacher more closely. To combat the first two flaws in particular, we developed a novel filtering approach intended specifically for capturing human motion.

Human motion in the classroom can be distinguished from these sources of noise using the speed and duration of the motion. Morphological filters were developed to distinguish motion at different speeds [19]. Specifically, every pixel where motion was detected was dilated, such that a surrounding circular area with a 5-pixel radius was also labelled as motion. This was conducted with every frame of the video, and the motion label was only upheld if the dilated area overlapped in temporally contiguous frames. By manipulating the length of these time frames (the filter history length), motion with different speeds could be captured. Fast motions were captured with a short filter history length, because even rapid human movements are likely to fall within a five-pixel distance of their position in the immediately prior video frames. Conversely, faster-moving motions were filtered out (leaving only slow motions) by a longer filter history, which would only register a motion occurring in the same area of the video.

We distinguished instructor motions of different durations by also varying the history length in the background removal process. The background removal history length determines how much time the process considers when calculating the variance of pixels. With a short history length, the background removal process quickly 'forgets' what is moving and what is part of the background. Brief movements will register more strongly with a longer history length, as the background removal process remembers these brief changes for a longer amount of time. If the teacher briefly gestures, the motion pixels from the gesture will only be briefly counted as motion and soon be reabsorbed back into the background (non-motion) pixels. Conversely, a sustained action like walking will continue to register as motion. In practice, we applied filter history lengths of 100ms and 200ms, and background removal history lengths of 500ms, 1000ms, 2000ms, and 4000ms, for a total of [2 filters × 4 background removals] = 8 smoothed motion estimates. We estimated eight values of motion to account for the fact that there are several different types of motion in the classroom, ranging from long and slow (e.g., walking) to short and fast (e.g., a hand wave). Finally, we included unfiltered motion estimates, for a total of 12 estimates.

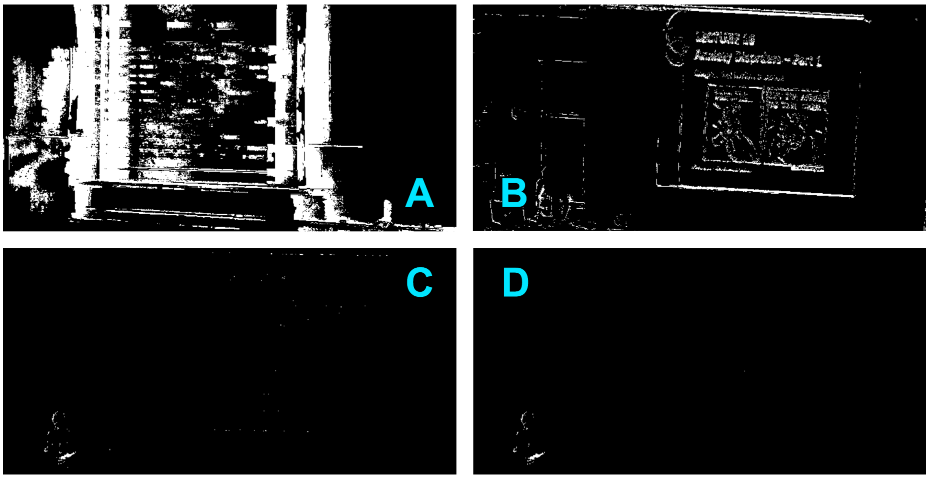

Fig. 1. Examples of three kinds of visual noise in videos: A) camera pan, B) camera vibration, and C) electronic low-light noise. Image D contains an example of smoothed motion from C with lower-left instructor motion preserved and noise visible in the upper-right of C correctly removed (zoom in for detail).

3.4 Pan and zoom detection

For this technique to be effective, we also needed it to detect pan and zoom in addition to motion because the foregoing techniques would label both of these as substantial motion. However, these events are clearly different from, and potentially less informative than, other types of motion (e.g., gesturing). We therefore sought to automatically measure these events so that machine-learned models (section 3.5) could distinguish this type motion from instructor behaviors. For example, a model might learn that brief, high-intensity motion indicates a gesture if and only if a zoom event is not concurrent. Furthermore, camera pan may be an indicator of instructor movement even when walking cannot be detected, so this is an important additional element to track.

To detect these events, we automatically selected 50 salient points to track in the video with a common corner detection algorithm [20]. We then tracked the points (Fig. 2), and computed the correlation between the velocities of salient points in the video (typically known as “optical flow” [13]). Over time some of these points would be lost—for example, when the camera panned so far to one side that a tracked point was no longer in view. When this occurred, we chose a replacement point to track such that it was at least 10 pixels away from another tracked point. The number of tracking points replaced in each video frame was included in the output, since tracking points being lost may be an additional indicator of events such as pan and zoom.

Pan events were captured by measuring the standard deviation (SD) of all tracked points in the preceding one second of video, selecting the middle 50% (25 tracked points) to eliminate outliers, and then measuring the mean SD of those points and the mean pairwise correlation (Pearson's *r*) between those points. In theory, pan events should have two measurable characteristics: 1) high standard deviations, because the points moved across the video, and 2) large *r,* because the points moved together as the camera panned.

Zoom events were rare in our data, as noted in section 3.2, but could still be important contextual information, and importantly, will be necessary for applying these techniques to other video sets. To detect zoom events, we measured mean point SD in the same way as pan events, but point correlation was measured differently. In a zoom event, all tracked points appear to move radially, either toward the edge of the frame (zoom in) or toward the center of the frame (zoom out). We thus converted tracked point coordinates from Cartesian to polar coordinates and measured the mean correlation between the point radii. Analogous to pan events, zoom events should have two measurable characteristics: 1) large standard deviation, and 2) radial correlation of points. However, we included the mean SD and mean *r* values as features for both pan and zoom detection, so that our machine learning methods would not be restricted by certain cutoff values for what qualifies as pan or zoom.

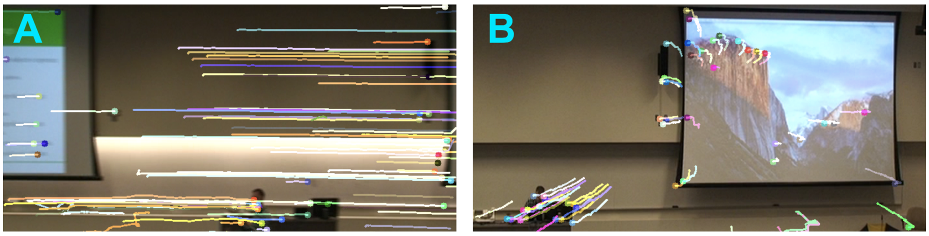

Fig. 2. Example (A) pan and (B) zoom events. In the pan event, most tracked points are moving together to the right as the camera pans left, while in the zoom event points are moving outward radially as the camera zooms in.

3.5 Machine learning for prediction of video annotations

We created features at the same timescale as the annotations made manually by coders (every 30 seconds) by calculating the mean and standard deviation of frame-level estimates in each 30-second window, yielding 38 features in total.

We tested the capabilities of our motion estimation methods by training logistic regression models to predict the video annotations. Separate models were trained to detect each annotation type in a one-vs-other scheme: camera pan, instructor walking, gesturing, slide change, and walking + pan combined. Models were cross-validated by training on data from five of the six days of video and testing on the remaining day. This process was repeated six times so that each day's data served as the test-set once. We also applied a forward feature selection process with nested cross-validation to select predictive features [21]. Feature selection took place in training data only, so that features would not be selected based on performance in the testing data.

Model accuracies were measured with Cohen's kappa and area under the receiver operating characteristic curve (AUC), both of which are commonly employed to evaluate predictive models in educational contexts [22–24]. Kappa measures the agreement between two sets of labels; in this case, those labels are the manual video annotations and the automatically predicted annotations. Kappa = 0 represents random chance-level accuracy, while kappa = 1 represents perfect agreement between ground truth and predictions. AUC, on the other hand, measures the tradeoff in accuracy between true positive classifications and false positives across all possible false positive rates, rather than one specific rate. AUC = .5 represents chance-level accuracy, while AUC = 1 represents perfect accuracy.

4 Results

We set out to answer two research questions using these techniques. Below we unpack the results with respect to these two questions.

How well does the proposed method detect teachers' non-verbal behaviors? Table 1 displays the results of supervised classification models trained to predict teachers' non-verbal behaviors and associated classroom activities, from the motion-related features. Overall, models predicted instructor behaviors at levels well above chance, with the exception of slide changes (kappa = .048). Furthermore, predicted rates were similar to the actual base rates present in the video annotations. Thus, we have confidence that these models provide reliable video annotations, and may be applied in authentic classroom settings to examine the role of instructor behaviors (as we have done below).

The models predicted longer-lasting behaviors such as walking with high accuracy, whereas briefer activities were not predicted as effectively (gesture kappa = .298, slide change kappa = .048). Video data and annotations were processed at a 30-second granularity, which may have caused the motion of these briefer events to be lost in the larger period of non-motion. It is possible that with narrower time windows, these finer-grained movements would be detected as accurately. However, we note that our gesture detection models are on par with or superior to previous video-based modeling in classrooms [22].

Table 1. Results of models for automatically predicting video annotations.

Annotation	Kappa	AUC	Base Rate	Predicted Rate
Camera pan	.468	.768	.419	.460
Gesture	.298	.705	.393	.550
Slide change	.048	.595	.173	.011
Walk	.338	.748	.294	.217
Walk + pan	.397	.683	.516	.549

Do the instructor activities detected with this method correspond to student attention in the classroom? It is also important to establish if predicted instructor behaviors relate to students' self-reported attention. We examined the correlation between the predictions made by the models in Table 1 and students' self-reported levels of attention (see section 3.1). We divided every class lecture into 72 consecutive segments, each of which was 500 s long[3], and calculated the mean self-reported attention level across all students within each segment. We then similarly calculated the mean prediction of each annotation as well as standard deviation (to capture variation in non-verbal behaviors) within each 500 s segment. Finally, we did the same with the ground truth manually-coded annotations to provide a comparison to the automatic method. We thus measured 20 correlations (5 activities × automatic and manual annotation × mean and SD of each) to determine which measures of classroom activity were reliably correlated with student attention. Given the large number of correlations, we applied a post-hoc Benjamini-Hochberg procedure to control for multiple tests [25].

Table 2 contains the correlations that were significant at $p < .05$ after controlling for multiple tests. Two compelling patterns emerge from these results. First, every significant correlation was positive, indicating that increased non-verbal activity from the instructor was generally related to better (not worse) student attention. Second, the automatic annotations were more consistent predictors of attention than even the ground truth manual video annotations.

Table 2. Significant correlations between automatic/manual annotations of instructors' non-verbal behaviors and students' self-reported attention.

Annotation	Method	Pearson's r	N
Camera pan (mean)	Manual	.299	72
Camera pan (SD)	Automatic	.311	72
Gesture (SD)	Automatic	.346	72
Walk (SD)	Automatic	.331	72
Walk OR pan (SD)	Automatic	.303	72
Walk OR pan (mean)	Automatic	.293	72

5 General Discussion

We were interested in automatic evaluation of instructors' non-verbal behaviors as an initial step towards providing useful feedback to instructors. We developed a video-based motion estimation method tailored for classroom videos, and evaluated its effectiveness compared to manual annotations of video, and the ability of these features to predict student attention. In this section we discuss the implications of our findings, as well as limitations and opportunities for future work in this area.

The motion estimation method we proposed was effective for detecting non-verbal activities and related events that had been manually annotated by humans (see Table

[3] Changing the segment length does not have a dramatic effect on results. Longer segment lengths (e.g., 700 s) produce slightly stronger correlations, but we report our original segment length in this paper (500 s) to avoid overfitting the analysis to desirable results.

1). It shows great promise for teachers who wish to get feedback on their non-verbal behaviors without requiring manual coding, advanced technical tools or professional videographers. It can provide them feedback as to how they might improve their lecturing to create a more engaging experience for students. Furthermore, we showed that the automatic assessments of instructors' non-verbal behaviors were significantly correlated with students' self-reported attention in the expected direction (i.e. more activity = more attention). In fact, the automatic annotations of non-verbal behaviors were more strongly related to attention than the manual annotations (which showed only one significant correlation). One possible explanation for this is that the motion estimation method works on every frame of classroom video, whereas humans produced much sparser annotations that may not have been sufficiently fine-grained estimates of non-verbal behaviors in every 500 s segment. The proposed method, being automatic, is also easier to apply frequently and in many classrooms compared to manual annotation – making it suitable for the eventual goal of providing feedback to teachers.

This paper is the first to attempt automatically estimating teacher behaviors from classroom videos taken in the wild. This is a particularly challenging problem because the videos analyzed in this study are from ordinary cameras placed in the back of the classroom, which could pan and zoom at the discretion of the operator. Furthermore, videos were recorded in multiple classrooms with varying designs. Nevertheless, instructor behavior was accurately detected with our methods.

There are two key limitations that should be noted. First, few (i.e. three) teachers were examined in this study. It may be that increased non-verbal activity was related to student attention for these instructors (Table 2), but that replication is warranted since these instructors may not be representative of others' teaching styles. Future work evaluating many more teachers with this method will be able to address this issue by searching for consistent differences in style, and the relationships of these styles with student attention. Furthermore, we will collect data in different classrooms and topics, to measure the robustness of the method across these dimensions and others (e.g., lighting conditions) that might be encountered in classroom applications. Second, certain activities such as slide changes and question answering were not detected as well as motion. Future work with different visual features may be able to capture slide changes, but it is likely that multimodal analysis involving audio will be needed to detect questions effectively [26].

Improving the quality of classroom lectures is a difficult process that can require years of teaching experience to overcome a lack of feedback about one's teaching style. As a step toward ameliorating this difficulty, we developed video-based methods for detecting teachers' non-verbal behaviors and showed that detected behaviors related to student attention. In the future, these methods will enable wide-scale research into assessment for teachers, thus improving the technique of teachers and the learning experiences of students.

References

1. Ambady, N., Rosenthal, R.: Half a minute: Predicting teacher evaluations from thin slices of nonverbal behavior and physical attractiveness. Journal of Personality and Social Psychology. 64, 431–441 (1993).
2. Witt, P.L., Wheeless, L.R., Allen, M.: A meta-analytical review of the relationship between teacher immediacy and student learning. Communication Monographs. 71, 184–207 (2004).
3. Babad, E., Avni-Babad, D., Rosenthal, R.: Teachers' brief nonverbal behaviors in defined instructional situations can predict students' evaluations. Journal of Educational Psychology. 95, 553–562 (2003).
4. Pogue, L.L., Ahyun, K.: The effect of teacher nonverbal immediacy and credibility on student motivation and affective learning. Communication Education. 55, 331–344 (2006).
5. Andersen, J.F., Withrow, J.G.: The impact of lecturer nonverbal expressiveness on improving mediated instruction. Communication Education. 30, 342–353 (1981).
6. Allen, J.L., Shaw, D.H.: Teachers' communication behaviors and supervisors' evaluation of instruction in elementary and secondary classrooms. Communication Education. 39, 308–322 (1990).
7. Menzel, K.E., Carrell, L.J.: The impact of gender and immediacy on willingness to talk and perceived learning. Communication Education. 48, 31–40 (1999).
8. Alseid, M., Rigas, D.: Empirical results for the use of facial expressions and body gestures in e-learning tools. International Journal of Computers and Communications. 2, 87–94 (2008).
9. Schneider, J., Börner, D., van Rosmalen, P., Specht, M.: Presentation Trainer, your public speaking multimodal coach. In: Proceedings of the 2015 ACM on International Conference on Multimodal Interaction. pp. 539–546. New York, NY: ACM (2015).
10. Barmaki, R., Hughes, C.E.: Providing real-time feedback for student teachers in a virtual rehearsal environment. In: Proceedings of the 2015 ACM on International Conference on Multimodal Interaction. pp. 531–537. ACM, New York, NY, USA (2015).
11. Schneider, J., Börner, D., Rosmalen, P. van, Specht, M.: Enhancing public speaking skills - An evaluation of the Presentation Trainer in the wild. In: Adaptive and Adaptable Learning. pp. 263–276. Springer, Cham (2016).
12. Ramanan, D., Forsyth, D.A.: Finding and tracking people from the bottom up. In: Proceedings of the 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. pp. 467–474. IEEE (2003).
13. Bouguet, J.-Y.: Pyramidal implementation of the Lucas Kanade feature tracker. Intel Corporation, Microprocessor Research Labs (1999).
14. Zivkovic, Z., van der Heijden, F.: Efficient adaptive density estimation per image pixel for the task of background subtraction. Pattern Recognition Letters. 27, 773–780 (2006).
15. Yun, X., Bachmann, E.R.: Design, implementation, and experimental results of a quaternion-based Kalman filter for human body motion tracking. IEEE Transactions on Robotics. 22, 1216–1227 (2006).
16. Stoll, C., Hasler, N., Gall, J., Seidel, H.P., Theobalt, C.: Fast articulated motion tracking using a sums of Gaussians body model. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV 2011). pp. 951–958 (2011).
17. Rautaray, S.S., Agrawal, A.: Vision based hand gesture recognition for human computer interaction: A survey. Artif Intell Rev. 43, 1–54 (2015).

18. Westlund, J.K., D'Mello, S.K., Olney, A.M.: Motion Tracker: Camera-based monitoring of bodily movements using motion silhouettes. PLoS ONE. 10, (2015).
19. Maragos, P.: Tutorial on advances in morphological image processing and analysis. Optical Engineering. 26, 267623 (1987).
20. Shi, J., Tomasi, C.: Good features to track. In: Proceedings of 1994 IEEE Conference on Computer Vision and Pattern Recognition. pp. 593–600 (1994).
21. Guyon, I., Elisseeff, A.: An introduction to variable and feature selection. Journal of Machine Learning Research. 3, 1157–1182 (2003).
22. Bosch, N., D'Mello, S.K., Ocumpaugh, J., Baker, R.S., Shute, V.: Using video to automatically detect learner affect in computer-enabled classrooms. ACM Transactions on Interactive Intelligent Systems (TiiS). 6, (2016).
23. Paquette, L., de Carvahlo, A., Baker, R., Ocumpaugh, J.: Reengineering the feature distillation process: A case study in detection of gaming the system. In: Proceedings of the 7th International Conference on Educational Data Mining (EDM 2014). pp. 284–287. Educational Data Mining Society (2014).
24. Jeni, L.A., Cohn, J.F., De la Torre, F.: Facing imbalanced data–Recommendations for the use of performance metrics. In: Proceedings of the 5th International Conference on Affective Computing and Intelligent Interaction. pp. 245–251 (2013).
25. Benjamini, Y., Hochberg, Y.: Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society. Series B (Methodological). 57, 289–300 (1995).
26. Blanchard, N., Donnelly, P.J., Olney, A.M., Samei, B., Ward, B., Sun, X., Kelly, S., Nystrand, M., D'Mello, S.K.: Identifying teacher questions using automatic speech recognition in classrooms. In: Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL). pp. 191–201. Association for Computational Linguistics (2016).