

Webcam-Based Observation of Student Engagement Patterns in Face-to-Face Classroom Learning

Research Overview

This study examines student engagement patterns in face-to-face classroom learning through webcam-based observation. Long classroom videos were segmented into 10-second video clips, and each clip was annotated using four engagement levels. The study focuses on the label distribution, and temporal transitions of engagement.

Purpose of the Study

Student engagement is an important indicator of how students pay attention, participate, and remain involved during classroom learning. In face-to-face classrooms, engagement is dynamic and may change depending on teacher explanations, peer presentations, classroom interactions, fatigue, and individual learning behavior. However, engagement is not always easy to observe because students may express attention in different ways.

The purpose of this study is to empirically examine student engagement transitions in real face-to-face classroom learning using structured observation of webcam-based classroom video recordings. The focus is understanding how engagement remains stable or changes over time. Long classroom videos were divided into consecutive 10-second clips, and each clip was treated as one observation unit. This design allowed engagement to be analyzed from one clip to the next rather than from isolated images. Therefore, the main contribution of this study is its empirical finding on student engagement stability and transition patterns across consecutive 10-second clips in real classroom learning.

Method

The study used webcam video recordings collected during real face-to-face classroom learning sessions at Shibaura Institute of Technology. A single RGB webcam was placed in the classroom to record students' visible learning behavior without wearable sensors. The videos were collected from two course contexts: *Advanced Information Systems and Engineering (AISE)* and *History of Japan (HOJ)*. The AISE sessions included lectures, discussions, quizzes, group presentations, and student presentations, whereas the HOJ session focused on student presentations.

The original recordings were obtained as long continuous videos. For annotation, each long video was segmented into consecutive 10-second clips, as shown in Figure 1. Each clip became one unit of observation and received one engagement label. The analyzed dataset was

collected from 10 classroom sessions: one session from the *HOJ* class and nine sessions from the *AISE* class. Each session was recorded for approximately 100 minutes. After segmentation and preparation for annotation, the dataset consisted of 5,591 annotated 10-second video clips.

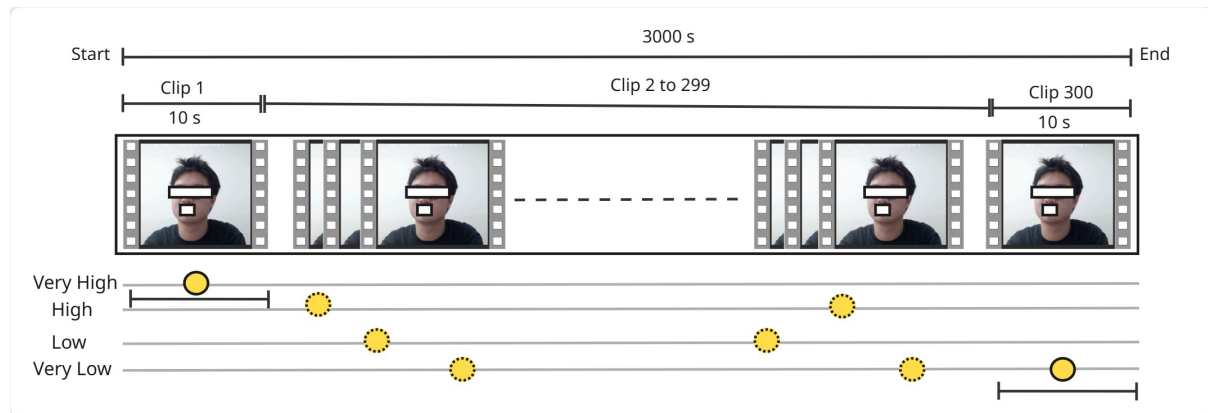


Figure 1: *Temporal segmentation and engagement annotation process.* A video sequence with a total duration of 3000 seconds is divided into 300 clips, each with a duration of 10 seconds. Engagement labels are assigned to each clip and classified into four discrete levels: Very High, High, Low, and Very Low. The yellow markers represent annotation points.

Student engagement was assessed using a four-level ordinal scale: Very Low, Low, High, and Very High. Very Low indicates clear disengagement, such as closed eyes, a bowed head, lack of attention, or visible boredom. Low indicates limited engagement, where the student is physically present but shows weak attention. High indicates normal engagement, such as paying attention to the teacher. Very High indicates deep and focused engagement, such as a concentrated gaze and strong attention.

Three university lecturers served as annotators. Before annotation, they received training on the engagement scale and observable indicators. A custom annotation interface supported synchronized video playback and category selection. Each annotator independently labeled the same clips. Inter-annotator agreement was evaluated using ICC(2,1), and the final label was determined using majority voting.

To examine engagement patterns over time, the final labels were arranged according to their original order in the long video. Transitions between consecutive clips were calculated.

Results

The annotation procedure produced 5,591 labeled clips. The final distribution consisted of 1,570 clips labeled Very Low, 1,054 labeled Low, 1,438 labeled High, and 1,529 labeled Very High. This distribution shows that classroom engagement appeared across all four levels. The Very Low, High, and Very High categories had relatively similar proportions, whereas the Low category had the smallest number of clips.

The inter-annotator agreement reached $ICC(2,1) = 0.83$, indicating strong agreement

among the three annotators. This result suggests that the four-level engagement scale can be applied consistently by trained observers and supports the reliability of the proposed observation procedure.

A transition analysis was conducted across four ordinal engagement levels. Stability, defined as no label change between consecutive 10-second clips, was the most dominant pattern, with an average of 53.59%. Among changing labels, one-level transitions were the most common pattern. The average proportion of one-level decreases was 20.91%, whereas one-level increases reached 19.15%. Two-level transitions were less frequent, with average decreases of 2.60% and increases of 3.00%. Extreme three-level transitions were very rare, with decreases of only 0.23% and increases of 0.52%.

Discussion

These findings suggest that engagement rarely shifts instantly from Very High to Very Low or from Very Low to Very High. Instead, engagement tends to move through adjacent levels in a gradual manner. Overall, the webcam-based observation procedure produced engagement labels with a logical temporal pattern. Engagement appeared dynamic but generally gradual, and sudden transitions can be used as candidate segments for annotation verification.

Keywords

student engagement; webcam observation; face-to-face learning; human annotation; engagement patterns