# EduSense: Practical Classroom Sensing at Scale

KARAN AHUJA†, Carnegie Mellon University
DOHYUN KIM†, Carnegie Mellon University
FRANCESKA XHAKAJ, Carnegie Mellon University
VIRAG VARGA, ETH Zurich
ANNE XIE, Carnegie Mellon University
STANLEY ZHANG, Carnegie Mellon University
JAY ERIC TOWNSEND, Carnegie Mellon University
CHRIS HARRISON, Carnegie Mellon University
AMY OGAN, Carnegie Mellon University
YUVRAJ AGARWAL, Carnegie Mellon University

Providing university teachers with high-quality opportunities for professional development cannot happen without data about the classroom environment. Currently, the most effective mechanism is for an expert to observe one or more lectures and provide personalized formative feedback to the instructor. Of course, this is expensive and unscalable, and perhaps most critically, precludes a continuous learning feedback loop for the instructor. In this paper, we present the culmination of two years of research and development on EduSense, a comprehensive sensing system that produces a plethora of theoretically-motivated visual and audio features correlated with effective instruction, which could feed professional development tools in much the same way as a Fitbit sensor reports step count to an end user app. Although previous systems have demonstrated some of our features in isolation, EduSense is the first to unify them into a cohesive, real-time, in-the-wild evaluated, and practically-deployable system. Our two studies quantify where contemporary machine learning techniques are robust, and where they fall short, illuminating where future work remains to bring the vision of automated classroom analytics to reality.

CCS Concepts: • **Human-centered computing** → **Human computer interaction** → Interactive systems and tools.

Additional Key Words and Phrases: Classroom, Sensing, Teacher, Instructor, Pedagogy, Computer Vision, Audio, Speech Detection, Machine Learning.

**ACM Reference Format:**

Karan Ahuja, Dohyun Kim, Franceska Xhakaj, Virag Varga, Anne Xie, Stanley Zhang, Jay Eric Townsend, Chris Harrison, Amy Ogan, Yuvraj Agarwal. 2019. "EduSense: Practical Classroom Sensing at Scale". *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, Vol. 3, No. 3, Article 71 (September 2019). 26 pages. https://doi.org/10.1145/3351229

---

†**Authors with equal contribution**. Author's addresses: Carnegie Mellon University, Pittsburgh, PA, USA, kahuja@cs.cmu.edu, dohyunk@cs.cmu.edu, francesx@cs.cmu.edu, szz@andrew.cmu.edu, ax@andrew.cmu.edu, jtownsen@andrew.cmu.edu, chris.harrison@cs.cmu.edu, aeo@cs.cmu.edu, yuvraj@cs.cmu.edu. ETH Zurich, Zurich, Switzerland, virag.varga@inf.ethz.ch

---

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from Permissions@acm.org.

Copyright © ACM 2019 2474-9567/2019/9-71 $15.00
https://doi.org/10.1145/3351229

## 1 INTRODUCTION

Quality university education has significant personal and macroeconomic implications. For this reason, there is an extensive literature on ways to improve university instruction [29][35][56][69][72]. For instance, increasing student engagement and participation in class has been shown repeatedly to significantly improve learning outcomes [24][69]. However, most university classes still rely on less-effective approaches where students passively receive information, even in small classes where greater interaction could occur [59]. This situation is difficult to change. Professors are typically hired and promoted for their domain expertise, and they typically view themselves as domain experts and not teaching experts [3]. Unlike K-12 teachers, who almost always receive professional education on how to teach in addition to gaining expertise in a content domain, university faculty typically receive no training on how to teach and have no official time allocated for doing so. Instead, they are expected to learn how to teach when they work as a teaching assistant while pursuing an advanced degree [33][34], often continuing models practiced by their own professors [10].

For instructors who want to improve their practice, there remain significant challenges that have led to this systemic problem with the current state of learning. The most critical of these is the lack of sufficient feedback opportunities on pedagogical skill, due to unscalable resources. Regular feedback on ones' current practice is an essential component of improving any skill [9]. Teachers need to see how their practices (mis)align with effective pedagogy in order to change [35]. However, unlike learning algebra, acquiring regular, accurate data on teaching practice is currently not scalable. When it is requested, acquiring this data currently relies on professional human observers to provide individualized formative feedback [22][28][29][39][67]. Unfortunately, the high cost in situ experts precludes any continuous instructional feedback loop, as typically just one or two lectures are observed for an instructor in a year.

To investigate how automated approaches could support professional development for university teachers at scale, we developed a holistic, classroom sensing system called EduSense. This system captures a wide variety of classroom facets shown to be actionable in the learning science literature, at a scale and temporal fidelity many orders of magnitude beyond what a traditional human observer in a classroom can achieve. Automated class analytics can provide a continuous feed of data on which instructional support systems can be built – in much the same way a Fitbit sensor tracks step count that feeds into end user applications that e.g., encourage healthy behavior. More specifically, EduSense captures both audio and video streams using low-cost commodity hardware that views both the instructor and students. We build upon state-of-the-art computer vision and audio classifiers, adapting them to the classroom domain, on top of which we developed custom classifiers that detect theoretically-motivated features associated with effective instruction from prior work (Table 1). These include detection of hand raises, body pose, body accelerometry, and speech acts.

While some of these features have been demonstrated individually in prior work, EduSense is the first to unify them into an extensible and cohesive system that has been deployed in the wild. We also describe how our system scales across many simultaneous class sessions, helps preserve student privacy, and maintains a coherent datastore across different temporal resolutions. Taken together, we believe EduSense constitutes a significant advance beyond prior efforts to make pedagogically-relevant classroom data available to instructors about their own practice on a regular basis, while also providing an extensible platform for others to deploy and build upon (to be open-sourced at publication). We systematically evaluate the efficacy of each EduSense feature, in both a controlled classroom study as well as in real-world classrooms. Our results show where state-of-the-art machine learning succeeds and fails, underscoring important avenues of future work.

## 2 RELATED SYSTEMS

Feedback on teaching is most effective when it contains accurate data and evidence, and when it is focused on specific topics, such as classroom discussion or question wait time [9]. There is an extensive learning science literature on methods to improve instruction through training and feedback that informed our system's features (broken down in Table 1). Much prior work has investigated how such data can be shared with teachers through feedback systems (e.g., "dashboards"), and studies have shown that teachers make use of the data and that it results in positive effects in changing teacher behavior (see e.g., [32][37][38][77][78]). Several successful

feedback systems were powered through "Wizard of Oz" means (see e.g., Classroom Discourse Analyzer [15] and Gerritsen et al. [26][27]), which could be powered by systems such as EduSense in the future. Most directly relevant to EduSense are automated sensing systems, which we now review in greater detail.

## 2.1 Instrumented Classrooms

In order to achieve robust detection of events in a practical manner, most prior sensing systems that attempt to collect data about the classroom have elected to directly instrument the physical fabric of the classroom, typically the furniture, such as student desks and chairs. For example, several projects have used chairs instrumented with pressure sensors [2][58]. The latter systems demonstrated recognition of various student poses that characterize varying levels of interest and engagement, such as slumped back vs. sitting upright.

There have also been innumerable efforts to make student desks interactive, usually by adding computing to the tabletop (e.g., buttons, touchscreens, etc.), or with response systems like "clickers" [1][12][20][21][68]. Such approaches are inflexible and can be expensive to deploy and maintain. To reduce cost and avoid direct instrumentation, some systems replace physical, electronic clickers with low-cost printed responses using color markers [25], QR Codes [17] or ARTags [57], combined with a computer vision system for detection.

Another approach is to locate the sensing apparatus on students or instructors themselves (i.e., "wearables"), which can offer robust detection of fine-grained signals. Notable among these systems is Affectiva's wrist-worn Q sensor [62], which senses the wearer's skin conductance, temperature and motion (via accelerometers); such data has been used to infer engagement level [2]. EngageMeter [32] used electroencephalography headsets to detect shifts in student engagement, alertness, and workload. As a practical tradeoff, there have also been efforts that instrument just the teacher, with e.g., microphones [19]. Of course, instrumenting users with non-personal, accessory hardware carries a social, aesthetic and practical cost.

## 2.2 Non-Invasive Class Sensing

Non-invasive user sensing avoids the social and practical costs of instrumenting teachers and/or students with hardware. For this reason, our goal from the outset was to be minimally invasive with respect to hardware, so as to be maximally practical. While there are many classes of non-invasive sensors, two in particular have been brought to bear for classroom sensing: acoustic and visual.

Speech is a rich signal source that can inform superior instruction (e.g., turn-taking [14][40], question asking [74], and pauses [49]). For instance, [19] used an omnidirectional room microphone and head-mounted teacher microphone to automatically segment teacher and student speech events, as well as intervals of silence (such as after teacher questions). Oral presentation practice systems such as AwareMe [11], Presentation Sensei [46] and RoboCOP [75] compute speech quality metrics, including pitch variety, pauses and fillers, and speaking rate.

Equally versatile are systems that employ cameras and computer vision in the classroom. Early systems, such as [23], targeted coarse tracking of people in the classroom, in this case using background subtraction and color histograms. Movement of students has also been tracked with optical flow algorithms, as was demonstrated in [54][63], though neither of these systems attempted automatic segmentation of individuals, and instead tracked audience-scale movement or used human-labeled bounding boxes. Computer vision has also been applied to automatic detection of hand raises, including classic methods such as skin tone and edge detection [41], as well as newer deep learning techniques [51].

Robust face detection has been of great interest for classroom sensing; not only can it be used to find and count students, but also estimate their head orientation, coarsely signaling their area of focus [63][73][80]. Facial landmarks can offer a wealth of information about students' affective state, such as engagement [76] and frustration [6][31][43], as well as detection of off-task behavior [7] (see e.g., [18] for a review of affect-sensitive instructional strategies). The Computer Expression Recognition Toolbox (CERT) [52] is most widely used in these educational technology applications, though it is limited to videos of single students (i.e., not classroom scale). Finally, cameras have also been used to detect activity and objects on student work surfaces [5][79].

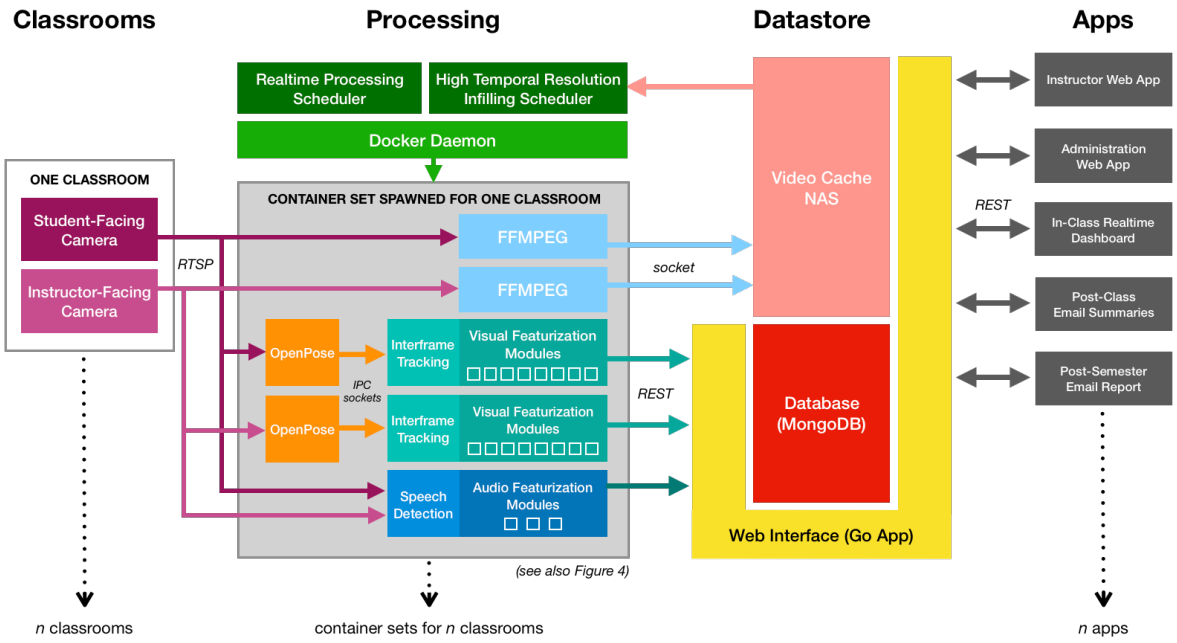



Fig. 1. High-level architecture of EduSense, illustrating the key components of our full-stack system.

### 2.3 System Contribution

The above advances in classroom sensing have been published individually, often developed and tested in isolation, with few systems drawing together more than a handful of featurization facets. This limits their practical real-world utility in classrooms, precludes holistic evaluation, and obscures their true potential. Moreover, all of the aforementioned systems are single-classroom scale (e.g., requiring a server per classroom), and do not present a scalable system architecture that could approach campus-scale deployments – a key goal of our system. Simultaneously, there have been recent, tremendous strides in computer vision and deep learning, but these literatures rarely touch on educational uses, and thus offer little insight into how such technologies work in complex classroom settings. Thus, we believe EduSense is unique in putting together disparate advances from several fields into a comprehensive and scalable system, paired with a holistic evaluation combining both controlled studies and months-long, real-world deployments.

## 3 EDUSENSE SYSTEM

EduSense is a full-stack system comprised of four key layers, illustrated in Figure 1. The physical sensors that power our system are the lowest *classrooms layer*. This is followed by a *processing layer*, in which the audio-visual scene is parsed to generate initial data, after which a series of specialized featurization modules convert and classify into educationally-relevant features. This digested data is then saved for long term storage and information retrieval in the *datastore layer*. The final *Apps layer* is comprised of end-user applications, which is our focus for the next year of research and development. We now review these key layers in greater detail, focusing on distinguishing features.

### 3.1 Sensing

Our decision to use vision and acoustic sensing was driven by a desire for low-cost, wide-angle, long-range sensing that was minimally intrusive to the physical fabric of a classroom. We tested scores of cameras

Fig. 2. Left: Early version of EduSense, using a Microsoft Kinect One depth camera and Intel NUC. Center: Current system, which uses networked 4K cameras that are accessed from a central server *(partially anonymized for review)*. Right: example classroom from our deployment with instructor-facing camera circled in blue.

(including depth cameras – early system shown in Figure 2, left) before selecting Lorex LNE8950AB cameras, which offer a 112° field of view and feature an integrated microphone, costing around $150 in single unit retail prices. These are connected to our campus network using ethernet, which also provides power (i.e., POE). We configured these networked cameras to capture 3840x2160 video (i.e., "4K") at 15 FPS with 16 kHz mono audio. They are sufficiently compact and lightweight to enable mounting directly onto electrical boxes, providing a clean and inexpensive install (Figure 2, center and right). In each classroom (Figure 1, purple boxes), we deploy one camera at the front (looking towards students) and another at the rear (looking towards the front of the class where the instructor typically stands).

## 3.2 Compute

Early versions of EduSense used small Intel NUCs to provide compute power in classrooms (Figure 2, left). However, this hardware approach was expensive to scale, deploy and maintain. Even with only five augmented classrooms early in our development, there was a significant time investment just to keep the machine operational and reliably connected to the network. On the software front, this first iteration was built as a monolithic C++ application, in which we encountered practical software engineering problems such as dependency conflicts and development overhead to integrate new modules. Remote software deployment was also frustrating, given these complex dependencies. It also made it more time consuming and fragile to incorporate modules relying on other languages such as Python, which is popular in the computer vision community. Additionally, errors and inefficiencies in any of the modules caused the entire system to break due to the lack of isolation.

Although successful as a proof-of-concept, these frustrations ultimately led to an architectural redesign, which offered isolation and modularity. We moved to networked cameras (akin to a "thin-client" model), as described in the previous section. We found these cameras to be highly reliable, with all having run for nearly a year without any maintenance. These IP cameras are accessed from a centralized, dedicated on-campus server, which pulls audio and video streams on demand using the Real-Time Streaming Protocol (RTSP). Our custom GPU-equipped EduSense server has 28 physical cores (56 cores with SMT), 196GB of RAM and nine NVIDIA 1080Ti GPUs.

We designed the different layers, and the interfaces between them (Figure 1), with modularity and isolation in mind. We leverage container-based virtualization extensively to isolate our different processes, and indeed entire classroom processing instances (which launch as a set of containers; Figure 1, large grey box). We use Docker, with each module packaged as a Docker image that includes the module code and its specific dependencies. With this design, developers of individual components can develop using whatever language or third-party libraries they desire, without worrying about dependencies of other modules or what is available on

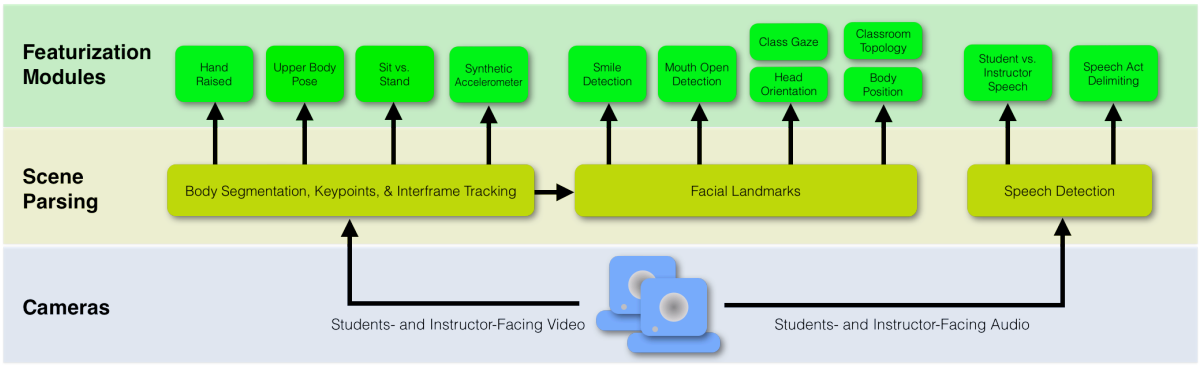



Fig. 3. Processing pipeline. Video and audio from classroom cameras first flows into a scene parsing layer, before being featurized by a series of specialized modules. See also Figure 1 and Table 1.

the host machine. These containerized processes communicate with each other via efficient interprocess communication (IPC) sockets.

Each active classroom has an associated classroom processing instance (Figure 1, large grey box). This executes at best effort, load balanced across whatever physical CPU and GPU resources are available on the machine. The available resources are elastic depending on how many simultaneous classes are being processed. Since several of our featurization and scene parsing modules use the NVIDIA machine learning framework, we used the NVIDIA Visual Profiler [60] extensively to improve their performance.

### 3.3 Scene Parsing

With visual and audio data streaming in, the next step is to parse the scene into data primitives that facilitate subsequent feature extraction (Figures 1 and 3). The first stage of our visual scene parsing pipeline is a multi-person body keypoint (i.e., joint) labeler build on top of OpenPose 1.4 [13]. These are launched in their own containers (Figure 1, orange), allowing them to be independently restarted in the case of failure, and also easily swapped as new versions of OpenPose (or other software) become available.

We extensively tested and tuned OpenPose parameters to achieve the best performance in our classroom context (e.g., heavy occlusion, distant people) and with our high, wall-mounted (i.e., non-frontal) and slightly fish-eyed view of the class (examples offered in Figure 4, top row). We also added additional logic to reduce false positive bodies (e.g., bodies too large or small), as well as interframe persistent person IDs with hysteresis

Fig. 4. Top row: Example classroom scenes processed by EduSense (image data is not archived; shown here for reference and with permission). Bottom row: Featurized data, including body and face keypoints, with icons for hand raise, upper body pose, smile, mouth open, and sit/stand classification.

(to mitigate e.g., momentary occlusion or loss of tracking) using a combination of Euclidean distance and body inter-keypoint distance matching. Please see our opensource repository for the above implementation details. The result is an array of body data structures, populated with IDs and keypoint metadata, for each frame of video. Head keypoints are used to search for faces more intensively, using OpenPose's face landmarking engine (OpenFace [4] as well as dlib 68-point face landmarks [44] were integrated as alternative landmarking engines, selectable as a command line argument or in the debug interface). Any landmarks found are added to a body's metadata. To facilitate movement of data, both to later featurization modules and also to persistent storage, we serialize our body array as a JSON object.

Running in parallel (in separate container) is our acoustic scene parsing pipeline (Figure 1, blue), which uses a deep learning model paired-down from a 30-class, acoustic model developed by Laput *et al.* [48]. Instead of predicting all 30 classes, we use the network to predict only silence and speech. The process described in [48] also makes it relatively straightforward to train new sound classes of interest in the future. An adaptive background noise filter is applied to classroom audio to remove persistent background noises (e.g., HVAC, projector fans). In addition to outputting a predicted class, a confidence score is also generated.

### 3.4 Featurization Modules

Our scene parsing stage provides the raw material for a series of subscriber featurization modules, responsible for a particular classroom facet of interest (Figure 1, cyan and dark blue; Figure 3). These are launched as containers and receive data over an IPC socket using a standard API which affords modularity. This architecture also allowed us to swap in new implementations as they become available, as well as toggle individual featurization modules on and off (as can be done with checkboxes in our debug GUI interface). Figure 4, bottom row, illustrates the output of various featurization modules. Table 1 provides a listing of all features of interest we implemented, along with citations that theoretically motivate or experimentally demonstrate the value of such sensed dimensions. For specific implementation details, please refer to our open source code repository (http://www.EduSense.io). We note that our feature set, while diverse, is not exhaustive. There are many other valuable dimensions of data that could be gleaned through video and audio processing; our present implementation is one set of features that we believed were a natural starting point and potent proof-of-concept.

*Sit vs. Stand Detection*: This featurization module uses body keypoints to predict if a person is sitting or standing. It requires seven keypoints to make an accurate prediction: neck (1), hips (2), knees (2), and feet (2). The relative geometry of these points is encoded by computing direction unit vectors between all pairs of these keypoints. To this feature vector, we also add the ratio of distances between the chest and foot, and chest and knee for both legs. This combined feature vector is passed to an MLP classifier (sklearn, default parameters) we trained with pilot data. In cases where the lower body is occluded, and the above keypoints are not available, we bypass our classifier and predict that the user is seated.

*Hand Raise Detection:* In addition to occlusion, hand raise detection is even more challenging due to the variation in the way students participate. We experimented with a number of techniques, including training a deep neural net on a cropped region above and to the sides of detected faces, but we found that other student's hands entering the frame led to many errors. Similar to sit/stand, we found the best result by relying on body keypoints, which avoids a lot of visual noise. Specifically, this module ingests eight body keypoints per body: neck (1), chest (1), shoulder (2), elbow (2), and wrist (2). We compute direction unit vectors between all pairs of these points. We also compute the distance between all pairs of points, normalized by the distance between the nose and neck points. Note these features are essentially scale invariant to both body distance from camera and physical body size. These values are used as input to an MLP classifier (sklearn, default parameters), which predicts either hand raised or not. Importantly, during training (and later, in our evaluations), we exposed our classifier to a wide range of partially and fully raised hands (see Figure 5, right four images).

*Upper Body Pose:* As shown in [2], which used physical chair sensors, body pose can be indicative of student affective and attentional state. To explore the feasibility of sensing similar attributes, but in a non-invasive computer-vision driven manner, we trained an upper body pose classifier. Due to heavy visual occlusion, we did not attempt lower body pose. For this module, we utilize the same eight upper body keypoints we found to be successful in our hand raise detection module. As before, we compute direction unit vectors between all pairs of points, and distance between all pairs of points, normalized by the distance between the nose and neck points. These values are used as input to a multiclass MLP model (sklearn, default hyper parameters), which was trained during development to predict three proof-of-concept classes: arms at rest, arms closed (e.g., crossed), and hands on face (Figure 5, left three images).

Table 1. Features of interest we selected for implementation, along with motivating literature.

| **Feature of Interest** | **Motivation and Sources** |
|---|---|
| Sit vs. Stand | In addition to demarking the start and end of class, this feature could provide the basis for noting changes in class activity (e.g., breaking into groups). Further, studies of classroom proxemics [36] revealed that teachers who sat behind/beside/on desks were rated by students as low in both affection and inclusion. In contrast, teachers who stood in front of desks or walked among students were more likely to be perceived as warm, friendly, and effective. |
| Hand Raised | Frequency and quantity of hands raises is a good indicator of student participation [69] and effective lecture design [14][59], both of which correlate with positive instructional outcomes. |
| Upper Body Pose | Instructors who have an open body position communicate to their students that they are receptive and immediate, whereas teachers who fold in or keep a closed body position are perceived as nonimmediate and unreceptive [65]. Student body pose can be indicative of student affective and attentional state, as demonstrated in [2][58], which instrumented classroom chairs with sensors. |
| Smile | Facial landmarks can offer a wealth of information about students' affective state, such as engagement [76] and frustration [6][31][43], as well as detection of off-task behaviors [7]. To evaluate the feasibility of facial affect analysis at classroom scales, we selected smile as a proof of concept. Instructors who smile and have positive-valence facial affect are perceived as more "immediate" and "likeable" than those who do not, engendering affiliation [65]. Adapting instruction to affect has been shown to improve educational outcomes [18]. |
| Head Orientation; Attention; Class Gaze | At a high level, attention is a pre-requisite for learning [30]. Head orientation has been shown to be a proxy for gaze attention [63][64][73][80], e.g., toward the instructor, educational materials (e.g., on the table), and other classroom foci (e.g., whiteboard, projection screen). In addition, low mutual gaze with students suggests that a teacher is not interested and not approachable [65]. Conversely, teachers who look at their students (i.e., rather than at the board or down) are perceived as more animated, more interested and have more rapport [65]. |
| Body Position; Classroom Topology | Student location has been shown to impact participation [61], and such data can be used to detect if actions are occurring more frequently in one area of the classroom (e.g., hand raises). Studies have also shown that instructors who move equally between the right and left sides of a classroom are more effective [36]. This spatial data is also useful in visualizing the distribution of other features as a "birds eye view" of the classroom. |
| Accelerometry | Studies have found different kinesthetic patterns between effective and "average" teachers, with effective teachers moving more [55][70]. Additionally, student attitudes were positively correlated with increased movement by instructors. For students, prior work (using worn accelerometers) demonstrated that movement could be used to estimate affect [2][62]. |
| Student vs. Instructor Speech | Effective instruction incorporates student discussion and questions [14][40][69], and thus it is useful to know the patterns and ratio of student vs. instructor speech. With such data, it may be possible to prompt the instructor to elicit more student participation, which has been shown to produce deeper learning than simply requiring them to listen and take notes [16][59]. |
| Speech Act Delimiting | Frequency and duration of speech have been shown to be an indicator of participation [19], which strongly corelate to educational outcomes [69]. Studies have investigated effective intervals of silence during instruction [14][49], for example after a question is asked [74] and to facilitate effective turn taking [40]. |

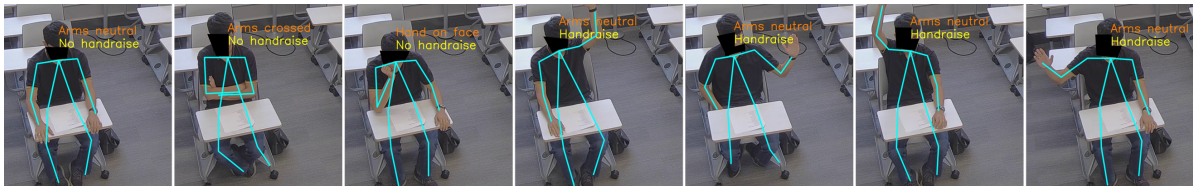



Fig. 5. Example participant from our controlled study. EduSense recognizes three upper body poses (left three image) and various hand raises (right four images). Live classification from our upper body pose (orange text) and hand classifiers (yellow text) are shown.

*Smile Detection:* As a proof-of-concept of class-scale facial affect analysis, we built a module that detects smiles (which has been shown to correlate to useful facets in previous work such as [18]). For this, we use ten mouth landmarks on the outer lip and ten landmarks on the inner lip. We compute direction unit vectors from the left lip corner to all other points and use a SVM (sklearn, poly kernel with degree 3, default parameters) for binary classification.

*Mouth Open Detection:* With our two audio channels, it is not possible to localize a speaker in the classroom. As a potential, future way to identify speakers, we developed a module that estimates if a mouth is open, the confidence of which could be tracked over many frames (per person) to produce a talking confidence. We use a binary SVM (sklearn, linear kernel, default parameters) and two highly descriptive features adapted from [71] (which predicted eyes open vs. closed): the height of the mouth to the left and right of center, divided by the width of the mouth.

*Head Orientation & Class Gaze:* Head orientation can be used as a coarse proxy for gaze attention; e.g., toward the instructor and other classroom foci. Using a perspective-n-point algorithm [50] in combination with anthropometric face data [53], EduSense produces a coarse 3D orientation of the head for each body. This process requires an accurate measurement of camera intrinsics before use, which we performed using OpenCV's calib3d module [8]. While our estimation code can run with as little as two facial landmarks, in practice, ≥75% are needed for any degree of accuracy. Once found, head orientations for individuals can be aggregated into a classroom histogram of foci, or even a combined mean gaze vector.

*Body Position & Classroom Topology:* Our facial landmark, perspective-n-point computation not only produces a 3D orientation for each head, but also an estimated 3D position in real world coordinates. We save this metadata for each body found in a scene, allowing us to visualize data not only from the camera's perspective, but also, e.g., from a synthetic top-down view. This can be used to reveal the classroom topologies (i.e., student layout), and in the future, help illuminate spatial patterns in the class (e.g., fewer hand raises in the back of class or where the instructor spends the most time).

*Synthetic Accelerometer:* Worn accelerometers have been used previously to infer student engagement and affect [62]. To achieve a similar result, but without the need for worn sensors, we simply track the motion of bodies across frames. Similar to the previous module, we use the 3D head position produced during scene parsing, and calculate a delta X/Y/Z normalized by the elapsed time since the previous frame. This affords us a 3D motion and acceleration vector in real world units (e.g., m/s). As one might expect, in-camera-plane motion (i.e., X/Y/) is more robust then Z-axis estimations.

*Student vs. Instructor Speech:* This module builds on top of the sound and speech detector running as part of the scene parsing pipeline. When speech is detected, this module computes three features: 1) the RMS of the student-

facing camera's microphone (closest to the instructor), 2) the RMS of the instructor-facing camera's microphone (closest to the students), and the ratio between the latter two values. A random forest classifier (sklearn, default parameters) is then used predict whether the current speech is coming from the instructor or student(s).

*Speech Act Delimiting:* This module ingests per-frame speech detection results, and computes speech act delimiters (e.g., 1:05:34 to 1:05:42 is one continuous speech act). We apply basic hysteresis to mitigate single frame classification errors.

### 3.5 Training Data Capture

Most of the featurization modules described in the previous section required extensive labeled data in order to train classifiers. There was the immediate and obvious challenge of needing to recruit a multitude of participants to help capture the wide variety of manners in which e.g., humans raise their hands. However, a secondary challenge was capturing data at a wide variety of viewpoints, as our cameras were being deployed in classrooms at many heights and horizontal positions (though we recommended central locations whenever possible). This is contrast to many other computer vision systems, that assume head-on orientations.

To overcome this challenge, we built a custom, multi-viewpoint capture rig, consisting of three heavy duty stage tripods and 12 Lorex LNE8950AB cameras (Figure 6). The outer two tripods carried three cameras each, while the center tripod carried six cameras. Cameras were placed between 150 and 350 cm above the ground, alternating sides of the tripod, in a distributed fashion in order to provide a variety of viewpoints. All cameras were connected to gigabit ethernet switch, to which a laptop was also connected. Custom software opened all 12 camera streams simultaneously at 4K resolution at 3 FPS, and saved video to disk for later processing and machine learning use.

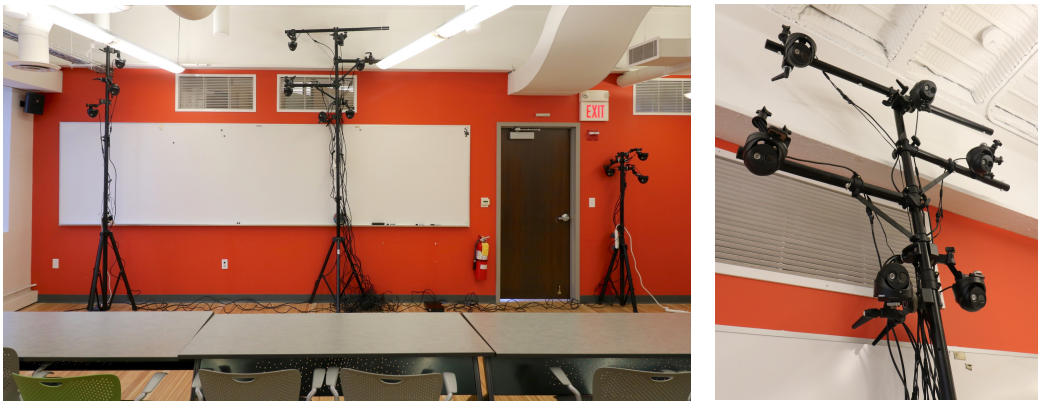



Fig. 6. Left: Training data capture rig in an example classroom. Right: Closeup of center mast, with six cameras.

### 3.6 Datastore

There are several types of data generated by EduSense. The first is non-image classroom data, which is the output of our various scene parsing and featurization modules. Importantly, this featurized data does not contain identifying information. A typical class session, lasting around 80 minutes with 25 students, generates roughly 250MB of data in uncompressed ASCII JSON format when processing in realtime (around 0.3-2.0 FPS; see also Figure 15). Infilled data (15 FPS, see Section 3.8) is roughly sixty times larger (i.e., ~16GB per typical class session, at 15 FPS for both front and back cameras). It is also important to note that EduSense may be processing upwards of a dozen classes (live or infill) simultaneously. Managing this volume of data required us to build a custom backend to efficiently store, organize and retrieve EduSense data. For this, we wrote a custom backend server in the Go language (Figure 1, yellow), which communicates with a dedicated MongoDB instance (Figure 1, red).

Both class processing instances and user-facing applications use a well-defined, narrow REST API on our GO backend, using Transport Layer Security (TLS) for encrypted communication.

The other data generated is raw classroom video/audio, which is saved long enough for temporal infilling to occur (or in the case of our studies, for human coders to label). We do not save these frames long-term to mitigate obvious privacy concerns. It is also possible to discard video/image frames as soon as they are processed live, with no data infilling occurring at a later time, if lower temporal resolution is acceptable. For these recordings, we use a secure Network Attached Storage (NAS; Figure 1, pink) box on our local network. We use flat files, organized in a directory structure by course number, classroom, time of class and date of capture.

### 3.7 Automated Scheduling & Classroom Processing Instances

We built a scheduler (Figure 1, dark green) using the open source version of SOS JobScheduler [42], which includes many useful features, such as a web interface that facilitates monitoring and tracking job status. EduSense maintains a list of enrolled classes, which includes the time, day(s) and room where each meet, as well as metadata such as instructor name, department, course number and name. The room UID (BuildingAbbreviation_Number, e.g., CRG_172) is used to lookup the network addresses of its two IP cameras. Using this data, EduSense automatically launches real-time processing instances at the start of each class, which consists of a set of containers for our various scene parsing and featurization modules (Figure 1, large grey box; see also Sections 3.3 & 3.4). Additionally, two FFMPEG instances (Figure 1, light blue) are launched to record the front and back camera streams, which are used for subsequent infilling, described next.

### 3.8 High Temporal Resolution Infilling

As noted previously, EduSense's current real-time performance is around 0.5 FPS (while processing up to 9 simultaneous classes). As resources become available, an infilling scheduler (Figure 1, dark green) opportunistically launches EduSense processing instances (i.e., same container set as Section 3.7) to re-process classes using recorded footage. This infills a class' data at maximum temporal resolution in the database – 15 FPS in our current implementation – which takes longer than real-time to process. Note that our EduSense architecture is designed to scale horizontally; by adding more EduSense servers and load balancing which classes are handled by each server, campus-scale sensing should be achievable.

This infilling mechanism is a natural complement to real-time processing. The latter gives us real-time class data at reduced temporal resolution to power immediate in-class and after-class feedback. On the other hand, infilling provides superior temporal resolution for non-real-time uses, such as end-of-day reports or even semester-long analytics. Whenever a report or statistic is generated, it uses all available data, even if the full framerate infilling is only partially completed.

### 3.9 Privacy Preservation

There are natural privacy concerns when capturing audio and video data. Deploying on our campus required buy-in from administration, registrar, facilities management, computing services, instructors and students. These conversations were important in underscoring the sensitivity of the data, as well as identifying stakeholders. As such, privacy preservation was a first-class design constraint in our system.

Foremost, in actual (i.e., not study) deployments, EduSense does not archive classroom video. Instead, incoming audio and images are immediately parsed and featurized into a format that is privacy preserving (e.g., body keypoints, but no image). If post-class infilling is enabled (entirely optional), the video will persist in a temporary cache until it is processed, after which it is automatically deleted. If post-class infilling is disabled, no image or audio is ever saved to the system, temporary or otherwise. Only featurized data is saved long-term in our database, accessible through our REST-based Web API, which includes user authentication and access control (i.e., only the authenticated instructor for a class is able to access their class's data). Figure 4 offers several example classroom scenes – the bottom row illustrates the type of anonymized data saved by EduSense

(note the images are not saved; included here only for reference and with permission). In the future, EduSense can be extended to reduce privacy concerns even further. For example, from our current set of features, we could expose only higher-level "class aggregates" (e.g., mean class gaze location, total number of hands raised), rather than featurized data for each individual.

In order to track metrics of individuals across frames, EduSense does assign a "person ID". This is not based on any visual or audio data, but rather body keypoints, which are tracked over time via a distance-based and keypoint similarity matching algorithm. This ID tracks with a body as long as it stays in view, and thus there are no IDs that persist across more than one class session. Note that an instructor (or someone else in the classroom) may still be able to de-anonymize the ID and associate it with a real identity based on the coordinates stored by EduSense and where they remember someone sitting. However, this risk is no worse than someone being in class and being able to observe the behavior of various students and instructor directly.

Finally, we note that for purposes of development, we temporarily stored classroom video data in order to manually annotate ground truth in a subset of frames. We used this data in many ways, including unit tests, training machine learning models, and studying the efficacy of our featurization modules. In accordance with our IRB, video data was deleted once testing or analysis was complete.

## 3.10 Debug and Development Interface

EduSense is primarily designed to be launched as a headless process by an automated scheduler (described in the next section). However, EduSense can also be launched with a minimal user interface, designed for debugging and demonstration (Figure 7). This was built in Qt5 and runs in an independent thread, so as not to block any live processing. Users can connect to any RTSP stream or browse their local filesystem to select a video or folder of images. There are widgets to configure most options in the system, for example, toggling featurization modules on or off. A live featurized view is provided in the center of the application, which can be panned and zoomed as needed to limit the total information shown and more closely integrate a scene or feature. Finally, a detail inspector can be summoned by clicking on the skeleton of any person in the scene.

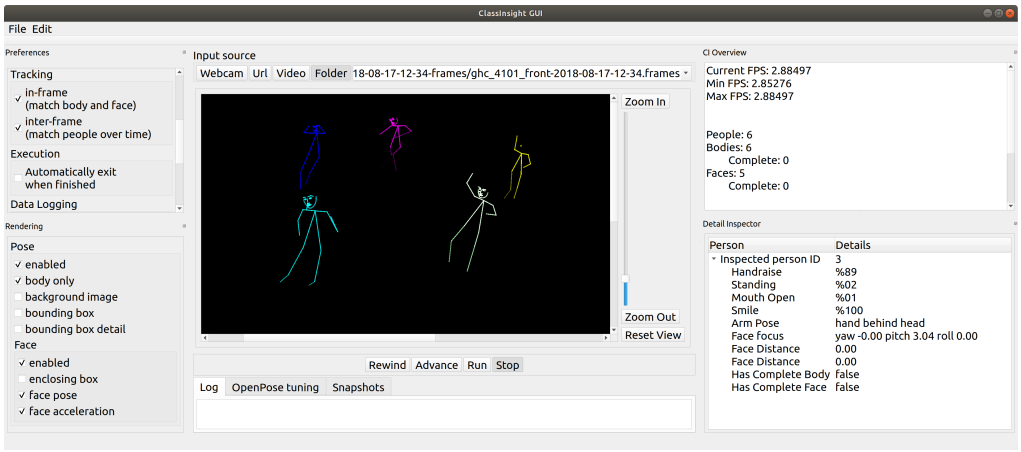



Fig. 7. Although EduSense is mostly launched as a headless process, we built a utilitarian graphical user interface for debugging and demonstration.

## 3.11 Open Source and Community Involvement

EduSense is open sourced with several goals in mind. Foremost, we hope that others will deploy the system and gain value not only in the data generated, but also as an opportunity to engage with topics surrounding smart classroom sensing (e.g., responsive pedagogy, professional development, privacy, automation, sensing fidelity).

Second, EduSense can serve as a comprehensive springboard for testing and continually refining the underlying computer vision algorithms in the classroom (and similar settings). Finally, we designed EduSense with modularity in mind, hoping to cultivate a community that can help improve and contribute new features and modules. EduSense can be downloaded from: http://www.EduSense.io.

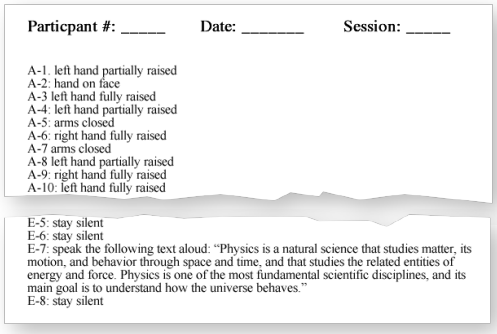
## 4 CONTROLLED STUDY

Systems research, such as this work, is challenging to assess holistically due to the highly faceted nature of the research and the lack of immediate baselines. In response, our evaluation strategy was to assess the technical efficacy of our individual featurization modules described above. Of course, testing these in a live classroom setting is challenging, as the range of behaviors and frequency of some events can be low. Likewise, establishing a ground truth can be subjective and requires human observers to label data. Thus, as a first evaluation, we used a controlled experimental procedure that allowed us to systematically quantify the performance of each featurization module. In our second study, we move to real classrooms with uncontrolled instructors and students, offering greater ecologically valid (albeit with other limitations). However, taken together, these two studies provide a holistic view of EduSense's performance.

### 4.1 Overall Procedure

We deployed EduSense in five exemplary classrooms with a variety of sizes, configuration, and lighting at our institution. For each classroom, we recruited one participant to play the role of the instructor, and between 3 and 6 participants to play the role of students. In total, we had 5 instructors and 25 student participants. Each instructor stood at the front of their class, while the students were randomly seated in the classroom. Participants were given a brief orientation to familiarize them with the actions they would perform and the language the experimenter would be using to proceed through the experiment.

Now situated, an experimenter used a surveyor's rope to measure the approximate distance of each participant from the camera (front-facing camera for instructors; rear-facing for students). Participants were then given a printed script, which provided a numbered list of actions to perform, which was randomized per participant (an illustration of this ordering can be seen in Figure 8). The experimenter would verbally announce the current stage (e.g., "We're now on B-17"), and once participants had complied with the instruction, the EduSense debug interface was used to flag data at that instant in time (e.g., B-17.json) for later analysis. Non-flagged frames (i.e., no ground truth) were discarded. In total, there were five phases of data collection, listed in Table 2 and described in detail subsequently.


Particpant #: _____ Date: _____ Session: _____

A-1: left hand partially raised
A-2: hand on face
A-3 left hand fully raised
A-4: left hand partially raised
A-5: arms closed
A-6: right hand fully raised
A-7 arms closed
A-8 left hand partially raised
A-9: right hand fully raised
A-10: left hand fully raised

E-5: stay silent
E-6: stay silent
E-7: speak the following text aloud: "Physics is a natural science that studies matter, its motion, and behavior through space and time, and that studies the related entities of energy and force. Physics is one of the most fundamental scientific disciplines, and its main goal is to understand how the universe behaves."
E-8: stay silent

Fig. 8. Illustration of paper "script" given to participants, leading them through a per-participant randomized (known) action sequence.

Table 2. Overview of experimental phases. In total, our 30 participants (5 instructor and 25 student roles) provided 1545 body instances and 60 speech/silence audio instances with labeled ground truths for analysis.

| Study Phase | Experimental Focus | Example Action | Trials per "student" participant | Trials per "instructor" participant |
|---|---|---|---|---|
| **A** | Upper body pose | Hand on face | 21 | 6 |
| **B** | Mouth state | Smile with teeth showing | 12 | 12 |
| **C** | Sit vs. stand | Stand | 6 | – |
| **D** | Head orientation | 15° pitch and -30° yaw | 16 | 16 |
| **E** | Silence *vs.* speech | Silence | 2 | 2 |

Using this corpus of ground-truthed, flagged frames, we were able to benchmark the performance of all major features of EduSense. Student and instructor results are combined where there was no significant performance difference. After discussing general body keypointing performance, we describe the experimental procedure of each phase, followed by results. Figure 9 provides an overview of results.

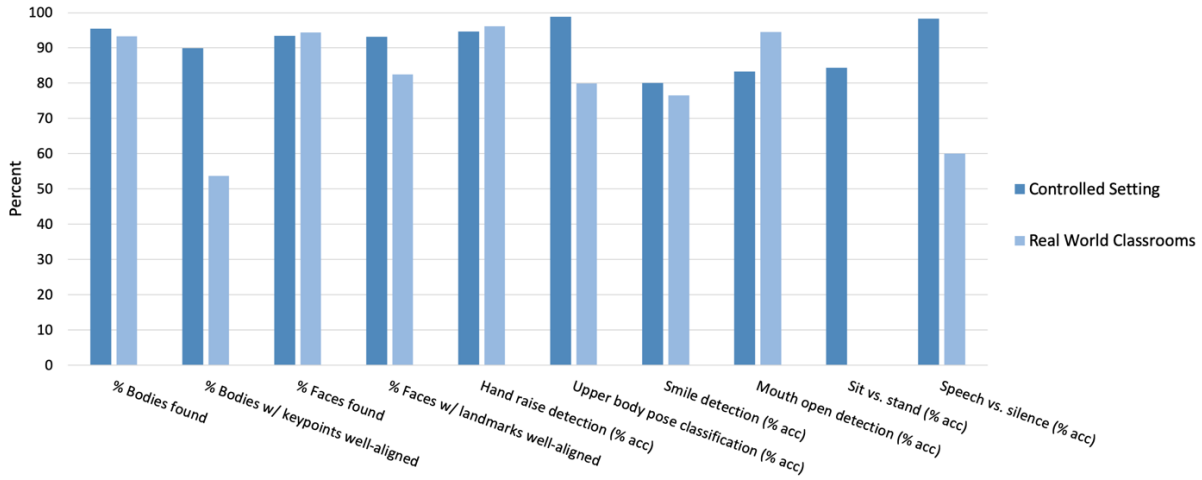



Fig. 9. Overview of experimental results for our controlled study (this section) and our subsequently described real-world classrooms study. Note sit vs. stand was not evaluated in the latter study. Please see text for procedural details.

## 4.2 Body Keypointing

Although OpenPose reports a rigorous evaluation of its body keypoint tracking performance [13], it offers limited insight into how well the model would perform in a classroom setting, with cameras operating high on walls, and where students are seated and often occluded by other people and furniture. Second, we made a number of small improvements to OpenPose during development, tweaking parameters and adding extra pose logic, which we found to improve stability and accuracy in pilot testing.

To best convey performance in the context of a classroom, we break out results by error type. We found that 4.5% of bodies were missed entirely, representing a low-level failure to recognize a person. However, 95.5% of bodies were correctly found. Of these, 10.1% were found to have at least some misalignment (defined as at least one body keypoint exceeding 10% of true body joint position). The remaining 89.9% of body-instances were judged to be accurately keypointed. Finally, EduSense found eleven bodies where there was no human, for a false positive rate of less than 1%.

In a classroom setting, visual occlusion is unavoidable. To quantify the effect of occlusion on body keypoint visibility, we provide the success rate of finding keypoints in Figure 10. Unsurprisingly, the upper body is most readily captured, with the feet and legs posing a (likely impossible) challenge. Fortunately, most classroom interactions are upper-body driven.

## 4.3 Phase A: Hand Raises & Upper Body Pose

"Student" participants were requested to perform one of seven possible upper body poses: arms resting (e.g., on table, on arm rests, or by side), left hand raised, left hand raised partial, right hand raised, right hand raised partial, arms closed, and hands on face. These seven upper body poses were requested three times each, for a total 21 instances per student participant. Instructor participants only performed arms resting and arms closed, requested three times each in different standing locations at the front of the classroom (left front, center front, and right front), for a total of six trials per participant.

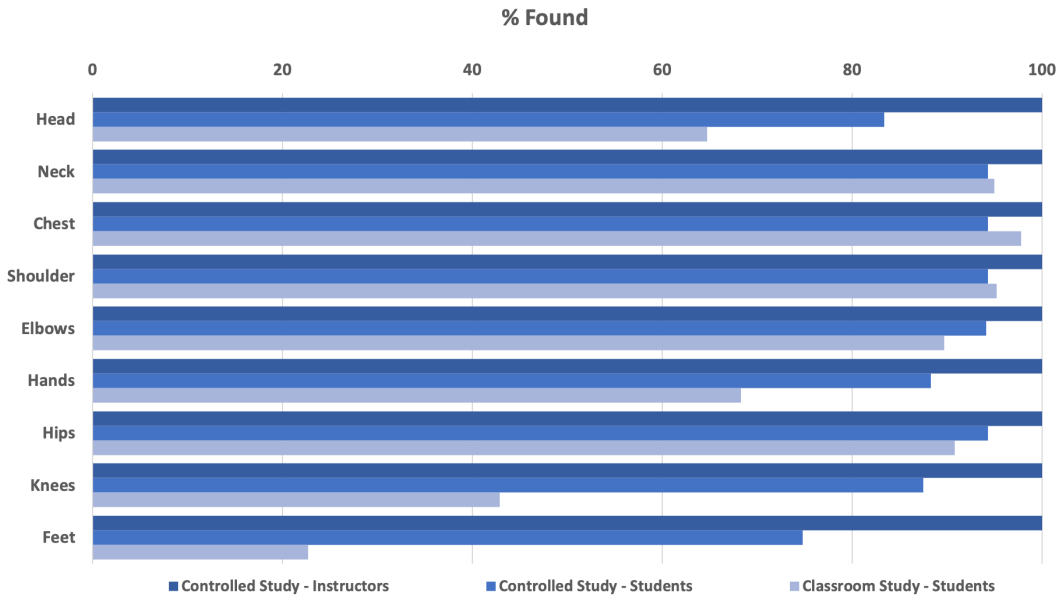



Fig. 10. Histogram showing the percent of different body keypoints found in three of our experimental contexts.

In Section 4.2, we quantified keypoint tracking quality. To understand the performance of EduSense's hand raise detection, we only studied frames where participants' upper bodies were captured (consisting of head, chest, shoulder, elbow, and wrist keypoints – without these eight keypoints, our hand raise classifier returns null). This prevents compounding several sources of error and offers a more direct evaluation of our hand raise classifier. To assess hand raise detection accuracy, we combine full and partial hand raises along with left and right hand use, whilst all other upper body poses we captured in Phase A are used for the negative class. On this data, hand raises were detected with 94.6% accuracy. There were no false positive instances (i.e., any other pose being incorrectly detected as a hand raise). In regard to our three upper body poses (arms resting, arms closed, and hands on face), again looking only at instances with correctly identified keypoints, mean accuracy was 98.6% for students and 100% for instructors.

## 4.4 Phase B: Mouth State

Similar to the above procedure, all participants were instructed to perform one of four possible mouth poses: *neutral* (mouth closed), *mouth open* (teeth apart, as if talking), *closed smile* (no teeth showing), and *teeth smile*

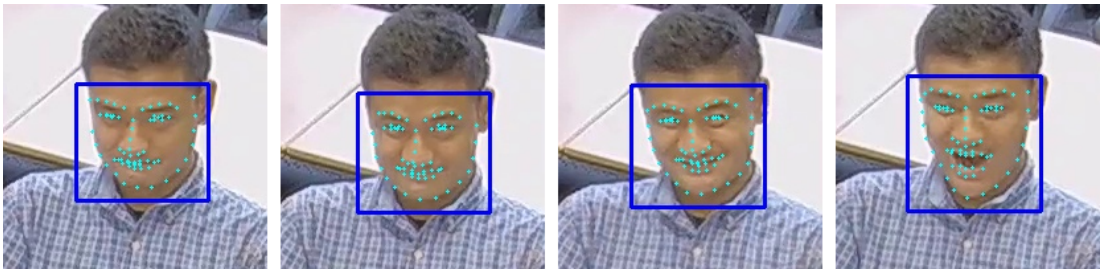


Fig. 11. The mouth states captured in our controlled study: *mouth closed*, *closed smile*, *teeth smile*, and *mouth open*.

(with teeth showing). For student participants, the 4 mouth poses were requested three times each, in a random order, for a total 12 trials per participant. Instructors also performed the four mouth poses, repeating the process three times at different standing positions at the front of the classroom, for a total of 12 trials per participant. Figure 11 illustrates the facial landmarks used to determine mouth state.

In frames where mouth landmarks were found (discussed in greater depth in Section 4.8), overall smile classification accuracy was 78.6% for students and 87.2% for instructors (positive classes: *closed smile* and *teeth smile*; negative classes: *neutral* and *mouth open*). A separate classifier predicts if the mouth is opened or closed, which had a mean accuracy of 83.6% for students and 82.1% for instructors (positive class: *mouth open*; negative classes: *neutral*, *closed smile* and *teeth smile*). We suspect the main limitation on performance is lack of camera resolution – faces in the rear of a classroom (e.g., 6m away) might only be 40 pixels wide, which is insufficient for accurate facial landmarking, which causes later modules like these to fail (mouths were as small as 17×5). Fortunately, this issue should be mitigated as higher resolution cameras become available.

### 4.5 Phase C: Sit vs. Stand

Participants were asked to either sit or stand, 3 times each, in a random order (i.e., 6 instances per participant). "Instructor" participants skipped this phase and remained standing throughout data collection. On these frames, EduSense entirely missed one body-instance (0.7% of our data) and had serious keypoint alignment errors on 4.6% of bodies, often due to lower body occlusion. On the remaining 94.7% of correctly keypointed instances, EduSense was 84.4% accurate at predicting sitting versus standing. We found the chief source of confusion stemmed from our angled (i.e., "3/4ths") view of the classroom, where the difference between a front-on reclined and standing pose was often indiscernible when using only 2D skeletal data. In the future, depth data could be used to help disambiguate these poses.

### 4.6 Phase D: Head Orientation

In this phase of data collection, the experimenter verbally requested eight head orientations: three possible pitches ("down" -15°, "straight" 0°, "up" +15°) × three possible yaws ("left" -20°, "straight" 0°, "right" +20°), omitting directly straight ahead (i.e., 0°/0°). To perform this with some level of accuracy, a sheet of paper was affixed to student desks with printed lines denoting the three yaws. For pitch, participants were given a smartphone running a custom application that displayed a large, real-time pitch measurement (smartphone held at the chin roughly perpendicular to the plane of the face; Figure 12). For a single trial (e.g., pitch -15°, yaw "right" +20°), participants first oriented their heads to look directly into the camera. They then tapped the smartphone screen, which zeroed the pitch value, with all subsequent values being shown relative to this vector. Then, using both the paper yaw guide and live smartphone pitch value, they aligned their head to the requested orientation. Each yaw-pitch combination was requested twice, for a total of 16 head orientation trials.

For analysis, we compared EduSense's predicted head pitch and yaw against the orientations requested in our user study. Unfortunately, in many frames we collected, ~20% of landmarks were occluded by the smartphones we gave participants – an experimental design error in hindsight. If we limit our analysis to body instances where at least 90% of landmarks were found (i.e., ≥ 63 landmarks out of 70), roughly a quarter of our collected data remains. On this subset, mean angular error is 12.6° (SD=8.2) for yaw and 14.5° (SD=11.7) for pitch,

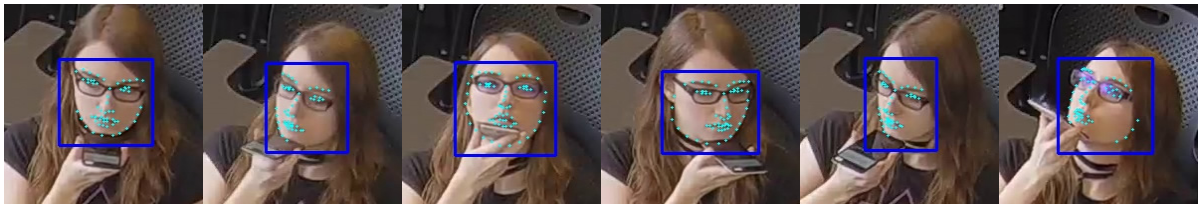


Fig. 12. Example head orientations requested in our study, with detected face landmarks shown.

which should be sufficient for coarse estimation of attention. Nonetheless, we ran the same analysis on all head orientation study data (Phase D), in which faces had an average of 40.9 (SD=25.6) landmarks available to EduSense's head pose estimator (out of a full set of 70). Using this less than ideal data, EduSense achieved a mean angular error of 23.6° (SD=18.8) for yaw and 44.9% (SD=39.4) for pitch. These results suggest that EduSense should only attempt head orientation estimations with sufficient facial landmarks.

### 4.7 Phase E: Speech Procedure

Finally, the script instructed one participant at a time to read aloud a paragraph of provided text (both student and instructor participants). While each participant spoke, one frame of data was flagged as "speech" by the experimenter. In between speakers, one frame of data was flagged as "no speech". After the study session concluded, the flagged frames were used to extract five-second audio clips from recordings. Thus, our 30 participants yielded 30 speaking and 30 non-speaking instances on which to evaluate our audio features. All *no speech* trials were correctly classified, and all but one *speech* trial was correct, for a mean accuracy of 98.3%.

### 4.8 Face Landmarks Results

Like body keypoint tracking, facial landmark performance has been well studied in previous work [4][13][44]. Here we provide results to quantify performance specifically in our classroom context. We dropped Phase D data from this corpus as there was significant occlusion of the face from participants holding a smartphone to their face as part of the experiment. In response, we used data from Phases A through C for this analysis. EduSense found 93.5% of student faces. Of those that were found, 61.8% had a majority of landmarks correctly registered, 37.5% had correct alignment on less than half of the landmarks, and 0.7% had poor alignment. For our instructor participants, all faces were found, but one instance was misaligned; 79.0% had a majority of landmarks correctly registered, with the remaining faces correctly aligned, but with less than half the landmarks found. In general, poor registration of landmarks was due to limited resolution (the farthest faces in our study had resolutions around 340×340). Lastly, in the combined instructor and student corpus, eight false positive faces were found.

### 4.9 Classroom Position & Sensing Accuracy vs. Distance

EduSense uses the 3D position of students to generate and maintain a real-time classroom topology, which is intended to power higher-level analysis (question rate vs. distance from instructor) and visualization modules (e.g., top down heatmaps). X/Y position (i.e., movement parallel to the image plane) is relatively straightforward to estimate if distance is known (i.e., with trigonometry). However, distance from the camera is more challenging to estimate from 2D image data alone. To evaluate the quality of EduSense's estimation, we manually recorded the distance of all participants from the camera using a surveyors' rope, as noted in our

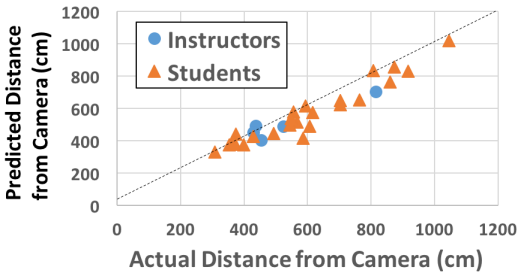



Fig. 13. Predicted distance vs. actual participant distance.

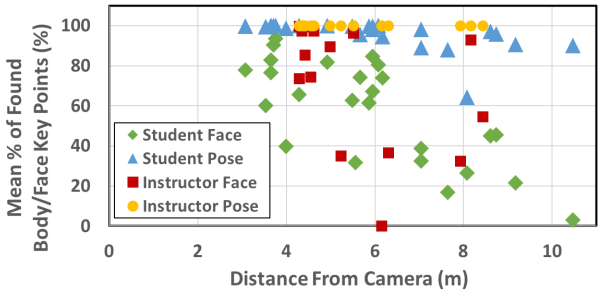



Fig. 14. Percentage of body keypoints and face landmarks found plotted against participant distance.

procedure above. Using this ground truth, we found a mean Z-distance error of 50.3 cm (SD=39.8) across all participants (see also per participant plot in Figure 13). This spatial accuracy should be sufficient to allow for students to be clustered into e.g., rows or tables.

All of our aforementioned computer-vision-driven modules are sensitive to image resolution, and thus vary in accuracy as a function of distance from the camera. This performance behavior can be seen in Figure 14, where the percent of body and face landmarks found (per participant mean) are plotted against participant distance. There is an unsurprising negative trend in accuracy as distance increases, which is most pronounced for face landmarks, which need high resolution image data. We also see that instructor data is more robust than students, chiefly because instructors are less occluded at the front of class.

### 4.10 Framerate and Latency

Figure 15 provides a runtime breakdown of various scene parsing and featurization modules. For this experiment, EduSense processed recorded videos, which ran one at a time, so there was no CPU/GPU contention (see Section 5.8 for a discussion of real-time performance). We can see that body keypointing runtime is mostly constant. However, facial landmarking increases linearly with respect to the number of bodies. For the five controlled study classes with less than ten students in the scene, the two scene parsing functions consume 73.1% of our total compute time on average. All of our featurization modules combined consume roughly 9.7%. Formatting and transmitting the output to the storage backend over the network takes roughly 16.7% of the compute time. Later, in Section 5.8, we discuss real-time performance.

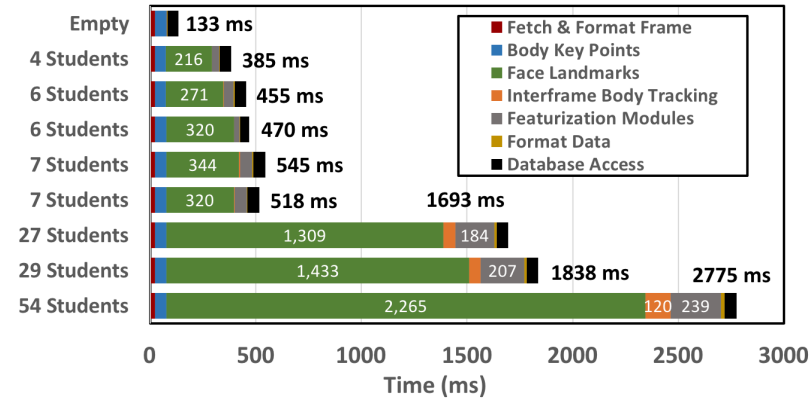



Fig. 15. Runtime performance of EduSense's various processing stages at different loads (i.e., number of students).

## 5 REAL-WORLD CLASSROOMS STUDY

As already discussed, our first study was a controlled experiment, with well-behaved participants that moved in lockstep through a series of defined ground truth states, allowing us to accurately benchmark features such as the estimation of head orientation (in concert with an accessory smartphone app) and distance from camera (verified with a surveyor's rope). Of course, in a live classroom setting with dozens of individuals, such heavyweight methods are impossible. The only practical solution is to post hoc label frames using human coders and compare them to EduSense output. Although this approach has limitations, it also offers the best evaluation of the real-world feasibility. Figure 9 provides an overview of the results.

### 5.1 Deployment and Procedure

We deployed EduSense in 13 classrooms at our institution and recruited 22 courses for an "in-the-wild" evaluation (with a total student enrollment of 687). Over the course of two semesters, EduSense captured and

processed 360.8 hours of classroom data. In total, 438,331 student-facing frames and 733,517 instructor-facing frames were processed live, with a further 18.3M frames infilled after class to bring the entire corpus up to a 15 FPS temporal resolution. The difference in live frame count is due to EduSense operating at a slower framerate with more people in the scene (i.e., student view).

After verifying all students were over 18 years of age, and with assent from both instructor and all students, a class was enrolled in the EduSense scheduler for automatic processing. After this point, there was no further contact with the class. At the end of the study period, we randomly pulled 100 student-view frames (containing 1797 student body instances) and 300 instructor-view frames (containing 291 instructor body instances; i.e., nine frames did not contain instructors) from our corpus. Though only a small subset, this random cross-section of classroom scenes is still sufficiently large and diverse so as to be representative of the full population of frames. To provide the ground truth labels, we hired two human coders, who were not involved in the project. It was not possible to accurately label head orientation and classroom position (see previous controlled study instead).

### 5.2 Body Keypointing Results

EduSense found 92.2% of student bodies and 99.6% of instructor bodies. This is similar to the 95.5% detection accuracy found in our controlled study. Though body detection was accurate, keypoint alignment was worse than our controlled study: 59.0% of student and 21.0% of instructor body instances were found to have at least one visible keypoint misalignment. We believe this is due to partial (but not full) occlusion of most student lower bodies (instructors to a lesser extent, standing behind podiums), which produces noisier (or erroneous) results. Similar to our first evaluation, we computed the success rate of finding various body keypoints, with results plotted in Figure 10 (student data only). Finally, looking at both student and instructor datasets, coders recorded 15 false positive bodies (vs. 11 in our smaller controlled study corpus).

We were surprised that our real-world results were comparable to our controlled study, despite operating in seemingly much more challenging scenes. However, upon closer inspection, there appeared to be a mitigating factor – in real world classrooms, students generally look straight ahead, with the arms resting in front of the body. This is in contrast to our controlled study, which had much greater variability in arm pose and head direction (by design). So although our real-world environments were more chaotic, with greater body occlusion, the poses were also generally easier.

### 5.3 Face Landmarking Results

In student frames, 94.3% of faces were correctly detected, with 4.3% having keypoints partially misaligned and 10.8% badly misaligned. Instructor faces were correctly found 94.6% of the time, with 10.8% of keypoints partially misaligned and 21.2% badly misaligned. The reduced keypointing accuracy for instructors is likely due to the distance (i.e., low image resolution) they are from the instructor-facing camera (generally attached to the rear wall of classroom), which is equivalent to the back row of seats for students. However, although keypointing was noisy for both students and instructors, face detection was strong, despite the fact classroom scenes were more complex and busier than in our controlled study. We also found that faces were rarely occluded, even partially, in real world scenes, which aided real-world accuracy.

### 5.4 Hand Raise Detection & Upper Body Pose Classification

Overall student hand raise detection accuracy was 96.1%. Unfortunately (and another good motivator for automated classroom analytics at our institution), hand raises in our real-world dataset were exceedingly rare. We suspected this would be the case, and as such, please refer to the controlled study results for a better estimate of performance. Our of our 1797 student body instances, we only found 6 body instances with hand raised (representing. less 0.3% of total body instances). Of those six hand raised instances, EduSense correctly labeled three, incorrectly labeled three, and missed zero, for an overall true positive accuracy of 50.0%. There was also 58 false positive hand raised instances (3.8% of total body instances). Detecting of our three poses (arms at rest,

arms closed, hands on face) was 78.4% accurate for students. Instructors fare better, due to reduced occlusion, with an accuracy of 88.8%. Note that in our corpus, arms at rest accounted for 88.3% of ground truth poses, with the other poses being far less common (hands on face 4.3%; arms closed 3.0%).

### 5.5 Mouth Smile and Open Detection

Only 17.1% of student body instances had the requisite mouth landmarks present for EduSense's smile detector to execute. On these instances, mean smile vs. no smile classification accuracy was 77.1% when compared to human labels. The results for instructors are comparable, with only 21.0% of body instances having the required facial landmarks. On these instances, mean smile vs. no smile classification accuracy was 72.6% when compared to human labels. For mouth open/closed detection, accuracy was stronger – 96.5% – though we note the data is heavily skewed to mouths being closed (94.8% of our coded data). Instructor accuracy was slightly worse, 82.3% accuracy on the instances with the requisite facial landmarks. We suspect the superior accuracy versus smile detection is chiefly because opening one's mouth is much less subtle than a smile, in which only the corners of the lips might move, which at a resolution of 10×3 constitutes a sub-pixel change.

We are confident the main obstacle to higher accuracy for mouth state detection is limited camera resolution; bodies situated close to a camera had mouths that were around 25×10 pixels in size, while our farthest bodies had mouths as small as 10×3 pixels. Such limited resolution does not permit robust facial landmarking, nor subsequent mouth classification. Indeed, we note this task was a significant (and subjective) challenge for our human coders as well, and thus should be regarded as a preliminary result.

### 5.6 Sit vs. Stand Classification

We found that a vast majority of student lower bodies were occluded, which did not permit our classifier to produce a sit/stand classification, and thus we omit these results. Instructors were more visible, though in 33.7% of frames their lower bodies were occluded (by a podium or table), precluding prediction. If we only consider frames (66.3% of our dataset) where the instructor's body was fully visible, sit and stand detection was 90.5% and 95.2% accurate, respectively.

### 5.7 Speech/Silence & Student/Instructor Detection

We used a slightly different procedure to evaluate our speech detector. Our coders randomly pulled 50 five-second clips of "speech" and 50 five-second clips of "no speech" (i.e., HVAC hum, papers rustling, distant chatter) from recordings of class. These labeled clips were passed, one at a time, into the detector, which outputted a prediction. On this audio corpus, accuracy was 82.0%. To evaluate student vs. instructor detection, our coders randomly pulled 25 ten-second clips of instructors talking and 25 ten-second clips of one or more students talking. These labeled clips were passed, one at a time, into our student vs. instructor speech classifier, which was 60.0% accurate at determining the speaker class. This poor result appears to stem from the fact our classifier was trained on data from a subset of deployment rooms. Unfortunately, it seems each classroom has a unique student-instructor amplitude ratio threshold based on the geometry of the room and placement of the camera/microphone. A per-room threshold or model would likely be needed to improve accuracy if only two microphones are available (without resorting to more sophisticated methods like speaker identification).

### 5.8 Framerate & Latency

Similar to our controlled study results, we analyzed the runtime performance of EduSense's main features using video recordings of three large classes with 27, 29 and 54 students in the scene. We add these results to Figure 15, which offers a greater variety in the number of students. Note that EduSense spends up to 84.5 % of its compute time parsing the scene (body keypointing and face landmarking) in a large class with 54 students in the scene. Overall, we achieve a mean student view processing framerate of between 0.3 and 2.0 FPS. Not shown

in Figure 15 is instructor view performance, which is typically 2-3 times faster due to far fewer face landmarks to process in the scene (i.e., most often only backs of student heads are visible).

When running in real-time, processing live frames from an active classroom, EduSense has an end-to-end latency (i.e., real-world state to value saved in database) of roughly 3-5 seconds. The largest contributor of latency – approximately 2.5 seconds – comes from our IP cameras, which are propriety and thus largely a black box. However, we suspect that buffering, compression and encoding (for both audio and video data) is intensive for the low-cost, embedded processor. Next is transmission over a wired network, which is negligible. Then, once data has been received by the EduSense backend, it takes between 0.3-2 seconds to process moderate-sized classes, as noted previously. Finally, storing the processed data takes another 70-100ms. This does mean that our current, proof-of-concept system is less suitable for rapid interventions, such as showing a heatmap of student participation to the instructor when hands are raised. However, such latency would be acceptable for keeping a running clock of how long the instructor has been speaking without a student question or discussion. Fortunately, newer and higher-end IP cameras offer dramatically reduced latency (~0.5 seconds), which should bring down EduSense's end-to-end latency to around 1-2 seconds, which should be sufficient for a wide range of realtime feedback and interventions.

## 6 END-USER APPLICATIONS

Our future goal with EduSense is to power a suite of end-user, data-driven applications. Over the next 18 months, we plan to shift of our focus from the backend system described in this paper, to such frontend applications. These, of course, have their own special considerations and development challenges, and will also require their own in-depth deployment studies.

There are numerous in-class instructional aids that we envision, for example, tracking the elapsed time of continuous speech, to help instructors inject lectures with pauses, as well as opportunities for student questions and discussion. Similarly, pauses in speech by the instructor, for instance after posing a question to a class, should follow the recommended wait time of three seconds, which has been shown to significantly raise student participation [49][69]. These timers could pop-up on a carefully designed, low-visual-complexity instructor tablet. Other simple cues that could be automatically generated include suggestions to increase movement at the front of the class (increasing student attention [55][70]), and modify the ratio of facing the board vs. facing

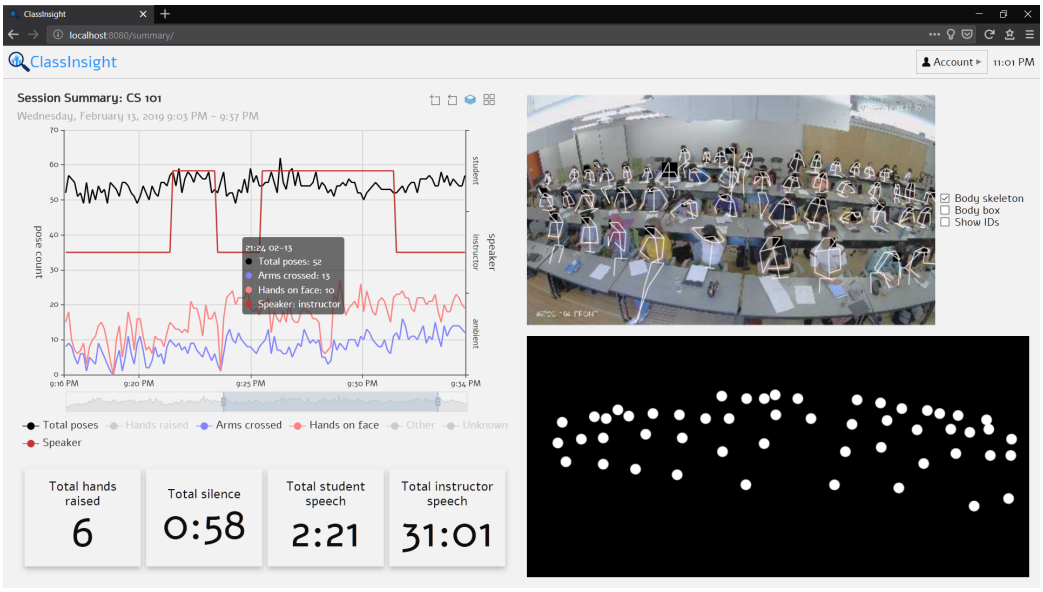



Fig. 16. Preliminary classroom data visualization app.

the class. More complex visualizations are possible too, for example, a cumulative heatmap off all student hand raised thus far in the lecture, which could facilitate selecting a student who has yet to contribute. Similarly, a histogram of the instructor's gaze could highlight areas of the classroom receiving less visual attention, which has been shown to decrease learning [65].

In additional to in-class, real-time feedback, we also see great value in after-class and end-of-semester reports, that could be generated as PDFs and sent in email. This could provide opportunities to reflect on teaching practices [15][27], or perhaps even the efficacy of interventions throughout the semester, such as increasing wait time after posing a question. Prior work, using similar, but manually coded data, has already shown such reports change instructor behavior and improve teaching efficacy [27][46].

Our open source system also serves as a springboard for more advanced features, for example, models of student engagement and attention. For example, prior work has shown that nonverbal communication on the teacher side (such as gaze direction [65], gesticulation though hand movement [81], smiling [65], and moving around the classroom [55][70] – all features EduSense tracks) can boost student attention and engagement, and has been shown to improve learning outcomes and improve likability of the instructor [66].

We have started work on one tool, a web-based data visualizer (Figure 16), which serves as a proof of concept. This was built using Node.js for communicating with our EduSense backend and handling data, and ECharts and React for front end components. Although early in development, it demonstrates the potential of our full stack system. We chose to develop this tool first, as the development team desired a way to easily load a class' data and explore the output of the featurization pipeline during iterative development. Instructors can login to review their own classes data, which is managed through EduSense's backend access control mechanisms.

## 7 DISCUSSION

Taken together, our controlled and real classroom studies offer the first comprehensive evaluation of a holistic audio- and computer-vision-driven classroom sensing system, offering new insights into the feasibility of automated class analytics. A key contribution of this paper is a detailed characterization of all layers of our EduSense stack in real classroom use. While many of our features were able to achieve reasonable accuracy (e.g., student segmentation ≥95%, upper body pose ~80%, gaze estimation ≤15°, speech detection ≥80%), other modules need further improvement to be practical. For some modules, it was a function of limited camera resolution (e.g., mouth state detection), while for others innate occlusion challenges will be hard to overcome (e.g., sit/stand detection for students). For this reason, we recommend that EduSense only be deployed in classroom with maximum front-back length of 8m, and with a sufficiently high mounting point to afford a good view of the classroom.

We also saw cascading effects, where each layer of our stack introduced some degree of error, which accumulated as data moved up the component food chain. For instance, our cameras had limited resolution at longer distances, which led to poor student segmentation, which lead to poor facial landmarking, which lead to poor smile detection. Even if each layer operates at 90% equivalent accuracy, the end result is at best 65% accurate due to compounding errors. As much as possible, we separated sources of error in our analyses to elucidate where individual features succeeded and failed.

Overall, based on our evaluations, it is clear that all layers of the stack – from sensors to high-level meta features – need continued attention from the research community. As accuracy is a continuum, it is always challenging to state definitively what level of performance (e.g., 90%, 99%, 99.9% accuracy?) is needed before value can be extracted, which is itself not binary. Additionally, each feature of the system almost certainly has a different threshold of utility, perhaps ±15° for gaze, but 99.9% for hand raises. In the latter case, false positive hand raises likely carry a high penalty with instructors than true negatives. The only way to understand where these accuracy-utility thresholds lie is with extensive, longitudinal deployments, which we have planned over the next few years with multiple institutions, both universities and high schools, and with a variety of end user tools, described in the previous section.

Nonetheless, we believe that our system has immediate utility for human observers, augmenting their notes with data they cannot easily capture themselves, or provide data at greater frequency. We also envision

EduSense as a stepping stone towards the furthering of a university culture that values professional development for teaching. It may reduce current barriers such as time and effort needed to collect, process, and view fine-grained data that leads to quality feedback on teaching.

## 8 CONCLUSION

We have presented our work on EduSense, a comprehensive classroom sensing system that produces a wide variety of theoretically-motivated features, using a distributed array of commodity cameras. We deployed and tested our system in a controlled study, as well as real classrooms, quantifying the accuracy of key system features in both settings. We believe EduSense is an important step towards the vision of automated classroom analytics, which hold the promise of offering a fidelity, scale and temporal resolution, which are impractical with the current practice of in-class observers. To further our goal of an extensible platform for classroom sensing that others can also build on, EduSense is open sourced and available to the community.

## ACKNOWLEDGMENTS

Our sincere thanks to the McDonnell Foundation and National Science Foundation for supporting this work (IIS-1822813, IIS-1747997 and CSR-1526237) for their generous support of fostering new research in education technologies. We would also like to thank the staff of the Eberly Center at Carnegie Mellon University, particularly Marsha Lovett and David Gerritsen, who have been ardent supporters of this project since its inception. We also appreciate all the help that the staff of the Media Services provided us with early installation of cameras, particularly Brian Fitzgerald and Dan Noulett. We also would like to acknowledge Avi Romanoff and Cyrus Tabrizi, who worked on an early versions of the system. Lastly, we would like to sincerely thank all the students and instructors who participated in this study, allowing us to collect data in their classrooms.

## REFERENCES

[1] A. L. Abrahamson (1998). “An Overview of Teaching and Learning Research with Classroom Communication Systems,” *International Conference on Teaching of Mathematics,* (1998), Village of Pythagorion, Samos, Greece.

[2] I. Arroyo, D.G. Cooper, W. Burleson, B.P. Woolf, K. Muldner and R. Christopherson (2009). Emotion Sensors Go To School. In *Proceedings of the 2009 conference on Artificial Intelligence in Education*. IOS Press, Amsterdam, The Netherlands, The Netherlands, 17-24.

[3] A.E. Austin (2002). Preparing the Next Generation of Faculty: Graduate School as Socialization to the Academic Career. *The Journal of Higher Education*, *73*(1), 94–122.

[4] T. Baltrušaitis, A. Zadeh, Y.C. Lim, and L.P. Morency (2018) OpenFace 2.0: Facial Behavior Analysis Toolkit. *IEEE International Conference on Automatic Face and Gesture Recognition*.

[5] R. Beckwith, G. Theocharous, D. Avrahami, M. and Philipose. Tabletop (2010). ESP: Everyday Sensing and Perception in the Classroom. *Intel® Technology Journal*, Volume 14, Issue 1, pages 18-33.

[6] N. Bosch, Y. Chen and S. D'Mello (2014). It's written on your face: detecting affective states from facial expressions while learning computer programming. In *Proceedings of the 12th International Conference on Intelligent Tutoring Systems (ITS 2014)* Switzerland: Springer International Publishing, pp. 39-44.

[7] N. Bosch, S.K. D'Mello, J. Ocumpaugh, R.S. Baker V. and Shute (2016). Using Video to Automatically Detect Learner Affect in Computer-Enabled Classrooms. *In Proc. of ACM Trans. Interact. Intell. Syst*. 6, 2, Article 17 (July 2016), 26 pages. DOI: http://dx.doi.org/10.1145/2946837

[8] G. Bradski (2000). The OpenCV Library. Dr. Dobb's Journal of Software Tools. Article 2236121.

[9] K.T. Brinko (1993). The Practice of Giving Feedback to Improve Teaching: What Is Effective? *The Journal of Higher Education*, *64*(5), 574.

[10] S.E. Brownell and K.D. Tanner (2012). Barriers to faculty pedagogical change: Lack of training, time, incentives, and...tensions with professional identity? *CBE—Life Sciences Education*, *11*(Winter), 339–346.

[11] M. Bubel, R. Jiang, C.H. Lee, W. Shi and A. Tse. 2016. AwareMe: addressing fear of public speech through awareness. In Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems. ACM, 68-73

[12] J.E. Caldwell (2007) Clickers in the large classroom: current research and best-practice tips. CBE— Life Sciences Education 6, 1, 9-20.

[13] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. (2017). Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[14] C.D. Cazden (2001). Variations in lesson structure. In Classroom discourse: The language of teaching and learning (pp. 53–79).

[15] G. Chen, S.N. Clarke & L.B. Resnick (2015). Classroom Discourse Analyzer (CDA): A Discourse Analytic Tool for Teachers. *Technology, Instruction, Cognition & Learning*, *10*(2).
[16] M.T.H. Chi and R. Wylie (2014). The ICAP Framework: Linking Cognitive Engagement to Active Learning Outcomes. Educational Psychologist, 49(4), 219–243.
[17] A. Cross, E. Cutrell and W. Thies (2012). Low-cost audience polling using computer vision. *In Proceedings of the 25th annual ACM symposium on User interface software and technology (UIST '12).* ACM, New York, NY, USA, 45-54. DOI: https://doi.org/10.1145/2380116.2380124
[18] S. D'Mello, N. Blanchard, R. Baker, J. Ocumpaugh & K. Brawner (2014). I feel your pain: A selective review of affect-sensitive instructional strategies. In *Design Recommendations for Intelligent Tutoring Systems*, Volume 2: Instructional Management, pp. 35-48.
[19] S. D'Mello, A.M Olney, N. Blanchard, B. Samei, X. Sun, B. Ward, and S. Kelly (2015). Multimodal Capture of Teacher-Student Interactions for Automated Dialogic Analysis in Live Classrooms. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction* (ICMI '15). ACM, New York, NY, USA, 557-566. DOI: http://dx.doi.org/10.1145/2818346.2830602
[20] S. Draper, J. Cargill, and Q. Cutts (2002). "Electronically Enhanced Classroom Interaction," *Australian Journal of Educational Technology*, pp. 13–23.
[21] C. Fies and N. Jill (2006). "Classroom Responses Systems: A Review of the Literature," *Journal of Science Education and Technology*, 101–09.
[22] C.J. Finelli, M. Ott, A.C. Gottfried, C. Hershock, C. O'neal, & M. Kaplan (2008). Utilizing instructional consultations to enhance the teaching performance of engineering faculty. *Journal of Engineering Education*, *97*(4), 397–411.
[23] J. Flachsbart, D. Franklin, and K. Hammond (2000). Improving human computer interaction in a classroom environment using computer vision. *In Proceedings of the 5th international conference on Intelligent user interfaces (IUI '00)*. ACM, New York, NY, USA, 86-93. DOI=http://dx.doi.org/10.1145/325737.325790
[24] S. Freeman, S.L. Eddy, M. McDonough, M.K. Smith, N. Okoroafor, H. Jordt and M.P. Wenderoth (2014). Active learning increases student performance in science, engineering, and mathematics. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(23), 8410–8415.
[25] J.E. Gain (2013) Using poll sheets and computer vision as an inexpensive alternative to clickers. *In Proceedings of the South African Institute for Computer Scientists and Information Technologists Conference (SAICSIT '13)*. ACM, New York, NY, USA, 60-63. DOI=http://dx.doi.org/10.1145/2513456.2513468
[26] D. Gerritsen, J, Zimmerman, & A. Ogan (2018). Towards a Framework for Smart Classrooms that Teach Instructors to Teach. In *International Conference of the Learning Sciences* (Vol. 3).
[27] D. Gerritsen (2018) "A socio-technical approach to feedback and instructional development for teaching assistants" PhD diss., Carnegie Mellon University, 2018,
[28] G. Gibbs, & M. Coffey (2004). The impact of training of university teachers on their teaching skills, their approach to teaching and the approach to learning of their students. *Active learning in higher education*, 5(1), 87-100.
[29] C. Gormally, M. Evans, & P. Brickman (2014). Feedback about Teaching in Higher Ed: Neglected Opportunities to Promote Change. *Cell Biology Education*, *13*(2), 187–199.
[30] D. Gottlieb & H. Moreira (2012). Should Educational Policies Be Regressive?. *Journal of Public Economic Theory*, *14*(4), 601-623.
[31] J.F. Grafsgaard, J.B. Wiggins, K.E. Boyer, E.N. Wiebe, and J.C. Lester (2013). Automatically recognizing facial expression: Predicting engagement and frustration. In *Proceedings of the 6th International Conference on Educational Data Mining*
[32] M. Hassib, S. Schneegass, P. Eiglsperger, N. Henze, A. Schmidt, and F. Alt. EngageMeter: A system for implicit audience engagement sensing using electroencephalography. In Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, pp. 5114-5119. ACM, 2017.
[33] P.L. Hardré (2005). Instructional design as a professional development tool-of-choice for graduate teaching assistants. *Innovative Higher Education*, *30*(3), 163–175.
[34] P.L. Hardré & A.O Burris (2012). What contributes to teaching assistant development: Differential responses to key design features. *Instructional Science*, *40*(1), 93–118.
[35] C. Henderson, A. Beach, & N. Finkelstein (2011). Facilitating change in undergraduate STEM instructional practices: An analytic review of the literature. *Journal of Research in Science Teaching*, *48*(8), 952–984.
[36] M.W. Hesler (1972). An Investigation of Instructor Use of Space. Dissertation Abstracts International, 1972, 33, 3055A
[37] K. Holstein, G. Hong, M. Tegene, B. McLaren, and V. Aleven, 2018. The classroom as a dashboard: co-designing wearable cognitive augmentation for K-12 teachers. In Proceedings of the 8th International Conference on Learning Analytics and Knowledge. ACM, 79–88.
[38] K, Holstein, B. McLaren, and V. Aleven (2018). Student learning benefits of a mixed-reality teacher awareness tool in AI-enhanced classrooms. In International Conference on Artificial Intelligence in Education. Springer, 154–168.
[39] D.R. Ilgen, C.D. Fisher, & M.S. Taylor (1979). Consequences of individual feedback on behavior in organizations. *Journal of Applied Psychology*, *64*(4), 349–371.
[40] J. Ingram, & V. Elliott (2014). Turn taking and "wait time" in classroom interactions. *Journal of Pragmatics*, 62, 1–12.
[41] J. Jesna, A.S. Narayanan, & K. Bijlani (2016) Automatic Hand Raise Detection by Analyzing the Edge Structures. In Emerging Research in Computing, Information, Communication and Applications (ERCICA '16). Springer. pp 171-180.
[42] JobScheduler. Software- und Organisations-Service (SOS). Last retrieved September 12, 2018. Thttp://www.sos-berlin.com/jobscheduler

[43] A. Kapoor, W. Burleson, R.W. Picard (2007). Automatic prediction of frustration, *International Journal of Human-Computer Studies*, v.65 n.8, p.724-736, August.

[44] V. Kazemi, and J. Sullivan (2014) One Millisecond Face Alignment with an Ensemble of Regression Trees. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition* (CVPR).

[45] D.E. King (2009). Dlib-ml: A Machine Learning Toolkit. *J. Mach. Learn. Res*. 10 (December 2009), 1755-1758.

[46] P. Koshland-Crane (2008). The effect of professional development of nonverbal communication behaviors of participants' regognition and understanding of these behaviors. (University of San Francisco)

[47] K. Kurihara, M. Goto, J. Ogata, Y. Matsusaka, and T. Igarashi (2007). Presentation Sensei: a presentation training system using speech and image processing. In Proceedings of the 9th international conference on Multimodal interfaces. ACM, 358-365.

[48] G. Laput, K. Ahuja, M. Goel, and C. Harrison. (2018). Ubicoustics: Plug-and-Play Acoustic Activity Recognition. *In Proc. of the annual ACM symposium on User interface software and technology (UIST '18).* ACM, New York, NY, USA, 45-54.

[49] L.R. Larson & M.D. Lovelace (2013). Evaluating the Efficacy of Questioning Strategies in Lecture-Based Classroom Environments: Are We Asking the Right Questions? *Journal of Excellence in College Teaching*, 24, 1–18.

[50] V. Lepetit, F. Moreno-Noguer, & P. Fua (2009). EPnP: Efficient perspective-n-point camera pose estimation. International Journal of Computer Vision, 81(2), 155-166.

[51] J. Lin, F. Jiang & R. Shen. (2018). Hand-Raising Gesture Detection in Real Classroom. 6453-6457. 10.1109/ICASSP.2018.8461733.

[52] G. Littlewort, J. Whitehill, T. Wu, I. Fasel, M. Frank, J. Movellan, and M. Bartlett (2011). The computer expression recognition toolbox (CERT). *IEEE International Conference on Automatic Face Gesture Recognition and Workshops (FG 2011)*. pp. 298–305.

[53] P. Martins. "Anthropometric Model". Last retrieved May 15, 2018. aifi.isr.uc.pt/Downloads/OpenGL/glAnthropometric3DModel.cpp

[54] D. Maynes-Aminzade, R. Pausch, and S. Seitz (2002). Techniques for Interactive Audience Participation. *In Proceedings of the 4th IEEE International Conference on Multimodal Interfaces (ICMI '02)*. IEEE Computer Society, Washington, DC, USA, 15-. DOI: http://dx.doi.org/10.1109/ICMI.2002.1166962

[55] A. Mehrabian (1971). *Silent messages* (Vol. 8). Belmont, CA: Wadsworth.

[56] J. McShannon, P. Hynes, N. Nirmalakhandan, G. Venkataramana, C. Ricketts, A. Ulery, & R. Steiner (2006). Gaining retention and achievement for students program: A faculty development program. *Journal of Professional Issues in Engineering Education and Practice*, *132*(3), 204–208.

[57] M. Miura and T. Nakada (2012). Device-Free Personal Response System Based on Fiducial Markers. *In Proceedings of the 2012 IEEE Seventh International Conference on Wireless, Mobile and Ubiquitous Technology in Education (WMUTE '12)*. IEEE Computer Society, Washington, DC, USA, 87-91. DOI=http://dx.doi.org/10.1109/WMUTE.2012.22

[58] S. Mota and R.W. Picard (2003). Automated Posture Analysis for Detecting Learner's Interest Level. *In Proc. of Conference on Computer Vision and Pattern Recognition Workshop*, Madison, Wisconsin, USA, 2003, pp. 49-49. doi: 10.1109/CVPRW.2003.10047

[59] C.E. Nunn (1996). Discussion in the College Classroom: Triangulating Observational and Survey Results. *The Journal of Higher Education*, *67*(3), 243–266.

[60] NVIDIA Visual Profiler. Last Retrieved, September 13, 2018. https://developer.nvidia.com/nvidia-visual-profiler.

[61] T. Parker, O. Hoopes, & D. Eggett (2011). The effect of seat location and movement or permanence on student-initiated participation. *College teaching*, *59*(2), 79-84.

[62] R.W. Picard (2011). Measuring affect in the wild. *In Proc. of International Conference on Affective Computing and Intelligent Interaction* (pp. 3-3). Springer, Berlin, Heidelberg.

[63] M. Raca and P. Dillenbourg (2013). System for assessing classroom attention. *In Proceedings of the Third International Conference on Learning Analytics and Knowledge* (LAK '13), ACM, New York, NY, USA, 265-269. DOI=http://dx.doi.org/10.1145/2460296.2460351

[64] M. Raca and P. Dillenbourg (2015). "Camera-based estimation of student's attention in class." EPFL Thesis 6745. urn: urn:nbn:ch:bel-epfl-thesis6745-4

[65] V.P. Richmond (2002). Teacher nonverbal immediacy. Communication for teachers, 65, 82.

[66] V.P. Richmond, J.S. Gorham, & J.C. McCroskey (1987). The relationship between selected immediacy behaviors and cognitive learning. Annals of the International Communication Association, 10(1), 574-590.

[67] K.D. Roach (1991). Graduate teaching assistants' use of behavior alteration techniques in the university classroom. *Communication Quarterly*, 39(2), 178–188. http://doi.org/10.1080/01463379109369795

[68] S. Robbins (2011). Beyond clickers: using ClassQue for multidimensional electronic classroom interaction. *In Proceedings of the 42nd ACM technical symposium on Computer science education (SIGCSE '11)*. ACM, New York, NY, USA, 661-666. DOI=http://dx.doi.org/10.1145/1953163.1953347

[69] K.A. Rocca (2010). Student Participation in the College Classroom: An Extended Multidisciplinary Literature Review. *Communication Education*, *59*(2), 185–213.

[70] J.M. Seals & P.A. Kaufman (1975). Effects of nonverbal behavior on student attitudes in the college classroom. *Humanist Educator, 14*(2), 51-55.

[71] T. Soukupová, & J. Cech. (2016). Real-time eye blink detection using facial landmarks. *In Proc. of 21st Computer Vision Winter Workshop*. Rimske Toplice, Slovenia, February 3–5, 2016.

[72] A. Stes, M. Min-Leliveld, D. Gijbels, & P. Van Petegem (2010). The impact of instructional development in higher education: The state-of-the-art of the research. *Educational Research Review*, *5*(1), 25–49.

[73] R. Stiefelhagen (2002). Tracking Focus of Attention in Meetings. *In Proceedings of the 4th IEEE International Conference on Multimodal Interfaces (ICMI '02).* IEEE Computer Society, Washington, DC, USA, 273-. DOI: http://dx.doi.org/10.1109/ICMI.2002.1167006
[74] A. Stolcke, N. Coccaro, R. Bates, P. Taylor, C. Van Ess-Dykema, K. Ries, E. Shriberg, D. Jurafsky R. Martin and M. Meteer (2000). Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational Linguistics*, 26(3): 339–373.
[75] H. Trinh, R. Asadi, D. Edge and T.W. Bickmore (2017). RoboCOP: A Robotic Coach for Oral Presentations. PACM Interact. Mob. Wearable Ubiquitous Technol. 1, 2, Article 1 (June 2017), 22 pages. DOI: 10.1145/1234
[76] J. Whitehill, Z. Serpell, Y.-C. Lin, A. Foster & J.R. Movellan (2014). The faces of engagement: Automatic recognition of student engagement from facial expressions. *IEEE Transactions on Affective Computing*, 5(1), 86-98.
[77] F. Xhakaj, V. Aleven and B. McLaren (2016). How Teachers Use Data to Help Students Learn: Contextual Inquiry for the Design of a Dashboard. In Adaptive and Adaptable Learning. Springer International Publishing, 340–354.
[78] F. Xhakaj, V. Aleven and B. McLaren (2017). Effects of a Teacher Dashboard for an Intelligent Tutoring System on Teacher Knowledge, Lesson Planning, Lessons and Student Learning. In Data Driven Approaches in Digital Education. Springer International Publishing, 315–329.
[79] N. Yannier, K. Koedinger, and S. Hudson (2013), "Tangible Collaborative Learning with a Mixed-Reality Game: EarthShake", Artificial Intelligence in Education, 131-140.
[80] J. Zaletelj and A. Košir (2017). Predicting students' attention in the classroom from Kinect facial and body features. EURASIP Journal on Image and Video Processing. Volume 1, pages 80.
[81] C.P. Zeki (2009). The importance of non-verbal communication in classroom management. Procedia-Social and Behavioral Sciences, 1(1), 1443-1449.