Computer Vision and Human Behaviour, Emotion and Cognition Detection: A Use Case on Student Engagement

Pieter Vanneste, José Oramas, Thomas Verelst, Tinne Tuytelaars, Annelies Raes, Fien Depaepe, Wim van den Noortgate

mathematics

MDPI

Article

Computer Vision and Human Behaviour, Emotion and Cognition Detection: A Use Case on Student Engagement

Pieter Vanneste [1,2,*], José Oramas [3], Thomas Verelst [4], Tinne Tuytelaars [4], Annelies Raes [1,2,5], Fien Depaepe [1,2] and Wim Van den Noortgate [1,2]

1 KU Leuven, Faculty of Psychology and Educational Sciences, 3000 Leuven, Belgium; annelies.raes@kuleuven.be (A.R.); fien.depaepe@kuleuven.be (F.D.); wim.vandennoortgate@kuleuven.be (W.V.d.N.)
2 KU Leuven, imec research group itec, 8500 Kortrijk, Belgium
3 University of Antwerp, Department of Computer Science, Internet Data Lab (IDLab), 2000 Antwerpen, Belgium; Jose.Oramas@uantwerpen.be
4 KU Leuven, Department of Electrical Engineering, research group on Processing Speech and Images (PSI), 3000 Leuven, Belgium; thomas.verelst@kuleuven.be (T.V.); tinne.tuytelaars@kuleuven.be (T.T.)
5 CIREL—Centre Interuniversitaire de Recherche en Education de Lille (ULR 4354), 59650 Villeneuve-d'Ascq, France
* Correspondence: pieter.vanneste@kuleuven.be

Abstract: Computer vision has shown great accomplishments in a wide variety of classification, segmentation and object recognition tasks, but tends to encounter more difficulties when tasks require more contextual assessment. Measuring the engagement of students is an example of such a complex task, as it requires a strong interpretative component. This research describes a methodology to measure students' engagement, taking both an individual (student-level) and a collective (classroom) approach. Results show that students' individual behaviour, such as note-taking or hand-raising, is challenging to recognise, and does not correlate with students' self-reported engagement. Interestingly, students' collective behaviour can be quantified in a more generic way using measures for students' symmetry, reaction times and eye-gaze intersections. Nonetheless, the evidence for a connection between these collective measures and engagement is rather weak. Although this study does not succeed in providing a proxy of students' self-reported engagement, our approach sheds light on the needs for future research. More concretely, we suggest that not only the behavioural, but also the emotional and cognitive component of engagement should be captured.

Keywords: student engagement; synchronous hybrid learning; computer vision

Citation: Vanneste, P.; Oramas, J.; Verelst, T.; Tuytelaars, T.; Raes, A.; Depaepe, F.; Van den Noortgate, W. Computer Vision and Human Behaviour, Emotion and Cognition Detection: A Use Case on Student Engagement. *Mathematics* **2021**, *9*, 287. https://doi.org/10.3390/math9030287

Academic Editors: Heui Seok Lim, Danial Hooshyar, Kyu Han Koh and Michael Voskoglou
Received: 23 December 2020
Accepted: 27 January 2021
Published: 1 February 2021

1. Introduction

1.1. Background and Rationale behind the Study

Computer vision techniques have been shown to be successful for tasks such as classification, segmentation or object recognition. There are plenty of examples across multiple domains on how computer vision could impact our daily lives [1,2]. In mobility, for example, computer vision and radar technologies are key components of the future self-driving car. In industry, feedback loops with computer vision can contribute to controlling the quality of products. Additionally, in security applications, facial recognition is increasingly being used for identity control.

For tasks that require a stronger interpretative component and more contextual assessment, computer vision tends to become less accurate. An example of a subtle task that may be hard for machines and even for humans concerns recognising human behaviour, cognition and emotions. In addition, data for subtle constructs are more scarce, which makes collecting data more labour intensive. Nonetheless, grasping subtle behaviour, cognition and emotions will also be necessary in order for computers to attain a level of intelligence that approaches that of humans. Imagine, for instance, computers that can be

operated with small gestures, or that account for a user's emotional or cognitive state [3]. Research on using computer vision to identify such more subtly articulated behavioural, cognitive or emotional states can progressively contribute to machine intelligence.

This study addresses an example of a multifaceted subtle construct as a case study: student engagement. The aim of this study is to elaborate a methodology that applies computer vision techniques to video recordings in an attempt to measure students' engagement. Interestingly, engagement is usually conceptualised as a three-dimensional construct consisting of a behavioural, a cognitive and an emotional component [4], which makes it a suitable concept to uncover how well computer vision can grasp these three components. In doing so, we take both an individual (student-level) and a collective (classroom-level) approach.

The video recordings concern lectures in higher education that take place in a hybrid virtual classroom [5]. In this relatively new learning setting (see Figure 1), on-site and remote learners are synchronously connected, which makes it even more difficult for the teacher to grasp the engagement of the student group (see [5] for a more detailed description of the learning space). If engagement could be measured automatically, this could be used as immediate feedback to the teacher and support him/her in educational decision making.

Figure 1. The hybrid virtual classroom, a relatively new learning space to synchronously connect remote and face-to-face learners.

1.2. Possibilities of Computer Vision Techniques to Detect Human Behaviour, Cognition or Emotion: State of the Art

Regarding recognising human behaviour, computer vision has been shown to be very successful in pose estimation (thanks to applications like OpenPose, see [6]) and in action recognition.

Regarding assessing human emotion, a large body of research has mainly addressed the seven universal emotions: fear, anger, sadness, disgust, contempt, surprise and enjoyment. These emotions can be related to movements of the facial muscles described in the Facial Action Coding System [7], which, in turn, can be recognised by computer vision. State-of-the-art results [8] have shown accuracies that typically exceed 90% for subject-dependent analyses (when applying different classifiers to different individuals), and exceed 70% for subject-independent approaches (when applying the same classifier across all individuals). However, these accuracies depend on the method used, the size of the training set, the emotion that is addressed and the extent to which that emotion is pronounced. More subtle emotions tend to be much more difficult to identify, as they are less linked to the muscles of the human face, and as training data are less abundant.

As far as human cognition is concerned, research employing computer vision is scarce. Mental states are more difficult to infer from facial expression than emotional states, for two main reasons. A first reason is that whereas emotional states can be identified from a few frames or even a single frame, mental states show more temporal dependency, requiring multiple frames to be identified. A second reason is that, whereas emotional states can be analysed solely based on information from facial muscles, analysing mental states also requires information from head gestures and eye-gaze measures (pupil dilation, micro-saccades, etc.) [9]. The spatial resolution that is required to assess eye-gaze measures

is usually not (yet) achieved in classroom settings, and taking pupillometry from a controlled lab environment to the real world involves quite some challenges [10]. Because of the additional complexities that come with recognising mental states, cognition is rarely measured by computer vision, but rather by self-reporting or by neuroscientific methods such as electroencephalography. Still, Kaliouby and Robinson [9], who aimed to assess (dis)agreement, concentration, uncertainty, thinking and interest from facial expression and head movement, reported an average correct classification rate of 89.5%.

1.3. Engagement and Its Relevance for Learning

Engagement in learning is often considered as a multifaceted construct [4,11,12], particularly consisting of three dimensions: a behavioural, a cognitive and an emotional dimension. The behavioural dimension represents all actions related to students' participation and involvement during their learning. The cognitive dimension covers the mental resources that students invest during their learning processes. Finally, the emotional dimension concerns positive and negative affective states and reactions to teachers, fellow students and the school.

Given the relevance of engagement in terms of students' learning outcomes [13,14] and retention in education [15], many researchers have attempted to keep track of students' engagement. It is generally acknowledged that retention and engagement are fundamental weaknesses of distance education compared to conventional education [16]. Fostering engagement in an attempt to decrease drop-out is therefore a key challenge in distance learning. It is important that teachers can keep track of the engagement of their students if they want to be able to act upon it.

Unfortunately, teachers' options to grasp students' engagement that exist in face-to-face settings partly disappear in distance education. Although it is not straightforward to understand why this happens, we put forward two main reasons for this phenomenon. A first reason is that teachers' and students' non-verbal immediacy behaviours (e.g., eye-contact, movement, facial expression, vocal variety) that generate the perception of closeness between students and the teacher are less prominent in distance settings [17]. These non-verbal immediacy behaviours are not only known to foster learning [18], but presumably also help teachers to sense the engagement of their classroom (see also the systematic review on synchronous hybrid learning by Raes [19]). A second reason is that in distance settings, the behaviour of students is reflected through various channels, to which the teacher cannot possibly pay attention at the same time. In the hybrid virtual classroom setting, as examined in the current study, for example, students are simultaneously present in the classroom and on LCD screens, participate in interactive quizzes and in a chatroom, etc. Because there are fewer non-verbal immediacy behaviours and because students' behaviour is scattered across different channels, monitoring students' engagement becomes more difficult, and teachers need to rely on other methods to do so. The next paragraphs discuss the most common traditional methods to measure engagement, and a novel method to do so, namely computer vision.

1.4. Measuring Engagement

One of the traditional approaches to measure engagement is via self-reporting [20]. However, self-reports interrupt the learning process, and may as such also influence learning itself. In addition, the frequency at which self-reports can be collected is limited, as subjects cannot be interrogated on a permanent basis. Collecting self-reports after the learning process requires a post hoc reconstruction of one's own engagement, which may be more biased. Moreover, the information on the classroom engagement is then not available during the classroom activity itself, so teachers cannot act upon it.

Observations are another traditional approach to measure engagement. They consist of having teachers or external observers complete a questionnaire on students' engagement [21]. An advantage of observations compared to self-reports is that they do not

interrupt the learning process. Nonetheless, the temporal resolution of observations is limited, and collecting them is labour intensive.

Another traditional way to measure engagement is by keeping track of time-on-task. As an example, Spanjers et al. [22] established a small partial correlation (r = .30) between students' self-reported engagement and the time spent on a task. Time-on-task in learning exercises has also been shown to strongly influence academic achievement [23]. In a context of computer-based learning, it is possible to automatically keep track of time-on-task. However, this is not possible when students do not merely learn through an electronic device, but rather in a class setting in which all students work in a common time period. Because this study looks into the automatic measurement of engagement during synchronous learning, time-on-task is not further investigated.

In addition to the aforementioned "practical" limitations, self-reports are also prone to several more fundamental limitations. As an example, it is not straightforward to assess one's own mental or emotional state. Furthermore, subjects cannot capture their unconscious emotions [24]. Brown et al. [25] acknowledge that self-reports are prone to many complex pitfalls, most of which cannot be easily resolved. Interestingly, the authors also mention pitfalls that researchers can actually avoid to some extent, such as issues related to reliability, grading, social response bias, response style, influences of peers and the environment in general. As an example, researchers are encouraged to create a climate that allows students to make honest, insightful and evidence-based judgements.

Despite these practical and fundamental limitations, self-reporting remains a valuable and often-used instrument in many research areas. A particular strength of self-reports is their relatively high construct validity, as self-reports at least directly inquire the construct of interest. The subjective nature of self-reports is not necessarily an obstacle, as it is learners' subjective perception that actually determines their behaviour during learning.

It is primarily the aforementioned "practical" limitations of self-reports in combination with their rather high construct validity that have inspired some of today's innovative research projects. This made researchers wonder if self-reports could somehow be replaced by a model consisting of directly measurable engagement proxies.

Unlike machines, humans can use their intelligence to give meaning to a complex interrelatedness of behaviour, cognition, emotion and context. Nonetheless, driven by the drawbacks that are inherent to self-reporting and observations, several researchers have attempted to measure engagement using machine intelligence, through computer vision. In doing so, two main approaches can be distinguished. In a first and most often addressed approach, individual learners are analysed. In a second approach, the collective behaviour of a group of learners is analysed.

When analysing individual learners' engagement through computer vision, researchers either focus on their behaviour (the actions they exhibit), or on their emotions. As previously argued, cognition is usually not addressed, as this comes with additional complexities. The behavioural dimension of engagement may include effort, attention, participation and persistence [26], concepts which are manifested by actions such as students raising their hand, taking notes, wrinkling, yawning, interacting with each other or with the teacher, exhibiting a certain body pose, etc. Several studies have attempted to recognise such actions using computer vision techniques and evaluate the obtained accuracy.

An important nuance to the majority of the studies mentioned below is that although these studies in general reached relatively high action recognition precision, they did not address the strength of the connection between these indicators and engagement (e.g., by comparing computer vision measures to self-reports or observations).

An often-addressed action is hand-raising. Wang et al. [27] elaborated a two-stage method consisting of body pose estimation and hand-raising detection and achieved up to 95% precision and 90% recall. Lin, Jiang and Shen [28] analysed a large-scale dataset consisting of 40,000 examples of hand-raising gestures and achieved 85% overall detection accuracy. Böheim, Knogler, Kosel and Seidel [29] did not employ computer vision, but human observers, with an aim to study the association between hand-raising

and motivation (which covaries with engagement to an important extent [30]). Their results indicated that students' hand-raising explains—depending on the lecture topic—between 11 and 15% of the variance in their motivation. In sum, it seems that hand-raising can be detected relatively well by means of computer vision, and there is some evidence for a weak but significant association with students' engagement.

Students' eye-gaze directions have also been used as an indicator of engagement. Barbadekar et al. [31] analysed whether students' heads were directly facing the teacher, and if so, students were assumed to be engaged. Canedo, Trifan and Neves [32] built a prototype to track students' faces and their eye gaze direction, while assuming that the more students' eye-gazes differ from the camera direction (the location of the teacher, the slides or the blackboard), the lower their level of attention is.

Researchers have also attempted to detect actions that indicate disengagement, such as yawning and sleeping. Li, Jiang and Shen [33] investigated sleep detection and applied it to a dataset containing 3000 images of students in a real classroom setting, including 5000 sleep gesture targets (some pictures included multiple targets). Their method obtained an average precision of 75%. Wang, Jiang and Shen [27] addressed yawning with a dataset of 12,000 samples and reached up to 90% detection accuracy.

Regarding the emotional aspect of engagement, the most commonly used method that does not interrupt learners is facial expression analysis. As the majority of research focuses on the facial expression of the seven universal emotions [7], there is less knowledge on other, more subtle and multifaceted emotional or cognitive states, such as engagement, which are typically less pronounced in the face. Nonetheless, some researchers have attempted to recognise engaged faces through computer vision.

Nezami et al. [34] annotated facial expressions of over 4600 samples of students' faces in terms of emotional and behavioural engagement. Students were annotated as behaviourally engaged when they were looking at the screen or down to the keyboard. Accuracies of around 72% were achieved to distinguish engaged from disengaged samples (note that the chance level equals 50%). Machardy, Syharath and Dewan [35] classified images across 20 individuals as bored or engaged. Labels were annotated by the users themselves. The overall accuracy reached 72%. Bosch et al. [36] seeked out to detect several affective states of 137 students playing educational games. Classification accuracies for engagement reached around 68% compared to trained observers' ratings. Note that these studies distinguished engaged from disengaged samples, and did not consider more fine-grained engagement scales.

Besides analysing individual learners, some researchers have also taken a collective approach, analysing groups of learners as a whole. In a study from Raca, Tormey and Dillenbourg [37], the level of attention was measured based on the synchronisation of students' actions and reaction times (referred to as sleepers' lag by the authors). More specifically, Raca et al. [37] verified whether there was a correlation between the average reported level of attention and the reaction speed, obtained by analysing movement intensity graphs. Although the obtained Kendall correlation had the expected trend ($\tau = -0.259$), it was not significant ($p = 0.06$) for the observed sample ($n = 29$). The idea of Raca et al. [37] to use the variance in reaction time of the classroom as a whole, in an attempt to obtain a measure for the classroom, is equally used in this study. However, this study does not examine reaction times based on movement intensity graphs, but based on transitions between latent students' states. These latent states are obtained via unsupervised clustering of students' joints and body parts (see further).

2. Research Gaps and Aim of This Study

There is relatively little research on how to measure subtle human behaviour, cognition or emotion through computer vision. That is why this study elaborates a methodology to measure engagement, as an example of one such subtle states.

The field still wonders if there is a single "gold standard", pinpointing multidimensional concepts like engagement or cognitive load [38]. Being aware of the many limitations

of self-reporting, it is still one of the best available criteria to measure engagement. That is why we decided to use self-reports as a gold standard for engagement. By simultaneously collecting engagement self-reports and monitoring different kinds of computer vision measures, this study aims to investigate if computer vision measures could serve as proxies for engagement. If effect sizes are strong enough, a combination of computer vision proxies could potentially be used to measure engagement, and self-reports could be omitted. This would then enable one to keep track of engagement continuously and in an automatic way.

The study may be interesting for computer vision as well as educational researchers and practitioners, by providing insights into the strengths and shortcomings of computer vision in view of measuring the multifaceted construct of engagement.

More precisely, this study sets the following objectives:

(1) At the individual (student) level:
   - (1a) Describing a methodology to apply computer vision to detect student-level indicators for engagement (hand-raising, note-taking, etc.) and evaluating how well computer vision can detect these, in terms of precision and recall;
   - (1b) evaluating how well these indicators can measure self-reported engagement.

(2) At the collective (classroom) level:
   - (2a) Describing a methodology to apply computer vision to detect classroom-level indicators for engagement (a measure for synchronicity, students' reaction times and shared eye-gaze intersections);
   - (2b) evaluating how well these indicators can measure self-reported engagement.

3. Methodology

3.1. Participants, Setting, Procedure, Self-Reporting and Annotations

This study's empirical data originate from two different groups of participants. The first group of participants includes 14 students of grade 12 from a Belgian secondary education school. These students follow a general education programme that prepares them for higher education, and includes Latin as a major topic. There are 4 female and 10 male students (average age = 17.5 years). These students are followed during six lectures on economics and marketing, an elective subject they voluntarily applied for. The lectures last around 70 min each and take place in a hybrid virtual classroom. In this relatively new educational setting, both face-to-face and virtual students can simultaneously attend a lecture. In the first two lectures, all students follow face-to-face. The third and fourth lectures are mixed, in that half of the students attend the lecture face-to-face, the other half virtually. In the last two lectures, the teacher is alone in the classroom, and all students attend the lecture virtually (the way in which these different learning settings affect students' relatedness, intrinsic motivation and learning achievement has been investigated by Raes [5]). Five relatively large (55 inch) screens at the back of the classroom can each simultaneously display four virtual students. The classroom can give space to about 25 face-to-face students.

During these lectures, student engagement is inquired via digital pop-ups that appear at random moments on their computers, with interval times that differ from student to student and range from 5 to 12 min. Students were explained that they should position the slider fully to the left if they felt totally disengaged (a score of 0), and fully to the right if they felt totally engaged (a score of 2). When students were in a neutrally engaged state, we explained to them to position their slider in the middle. As the scale was quasi-continuous, all other positions could be used to represent intermediate engagement states. These self-reports are used as a gold standard for engagement. All six lectures resulted in a total amount of 580 digital self-reports for students' engagement (on average about 40 measurements per student).

The second group of participants includes 51 first year university students (38 female, 13 male, aged around 18.5 years). Most participants are enrolled in a programme on educa-

tional sciences, with a smaller number of participants studying mathematics, engineering or chemistry. These students attended six traditional face-to-face lectures, each lasting around 1 h 20 min. The topic of the lecture is educational technology and mainly deals with educational theories that underpin how technology can foster learning. During these lectures, students' engagement was not repeatedly inquired. The classroom is fairly traditional and can fit about 50 students within six different rows. The only reason to include this group of participants was to test the collective approach to measuring engagement (see further), and to increase the amount of annotations for students' individual actions. In total, 1031 annotations were made for students' individual actions, for the entire set of recordings.

All students signed an informed consent form indicating their voluntary participation in the study. The study was approved by the ethical commission of the university (reference G- 2018 06 1264).

Before the start of the study, students were informed about the learning settings that they would experience. Students did not know that the aim of this research in particular was to study whether the monitored manifest variables could measure their engagement. Nonetheless, students realised that they were monitored, which most probably had some effect on their behaviour in the classroom. However, as students were followed during multiple lectures, we believe that their behaviour was quite natural, although we assume that explicit off-task behaviour was greatly reduced.

3.2. High-Level Overview of Data-Processing Pipeline

A camera mounted in front of the classroom provides a video stream of all students. In order to extract meaningful higher-level representations from this video stream, the first processing stage applies a body pose estimator on the image, extracting students' body keypoint locations (joints and other parts of the body). Interesting advantages of body pose estimators are that they can be trained on a large variety of data and are robust against camera viewpoint changes.

The second stage uses the body keypoint locations to estimate students' engagement levels. Training a neural network that directly estimates engagement levels from these keypoint locations was considered. However, this would require annotations of students' engagement levels by independent observers that watched the videos of the lectures, a task which is, even for human observers, not straightforward. Nonetheless, these manual annotations could then serve as a ground truth to train a computer vision algorithm for detecting engagement, based on the position of students' body parts. A major concern of this approach is that it would require a huge amount of training data, for two reasons. First, the high dimensionality of the space: there are $2N$ dimensions representing the 2D positions of N different parts of the body. Second, we do not expect a strong association between the position of a single part of the body and engagement, but rather a complex association between the entirety of body parts and engagement. Because covering the entire n-dimensional space with training data (manual annotations) was far from possible, the dimensionality of the latent space has to be reduced. To that aim, our approach extracts meaningful and unidimensional intermediate measures that could be derived from the N positions of the parts of the body.

The indicators for student engagement can be split into two groups: measures of individual behaviour and measures of collective behaviour. The predicted poses are taken as input by an action recognition component, which is in charge of providing cues for behaviour at the individual level. These cues represent action units of which the literature has shown their possible relevance for student (dis)engagement (actions such as raising hands, taking notes, etc.). In parallel, the predicted poses are also given as input to a collective analysis component which will aggregate the body poses of all students and which will provide cues related to the collective/group behaviour of students in the classroom. Such measures are related to synchrony or reaction times [37]. This high-level approach is illustrated in Figure 2.

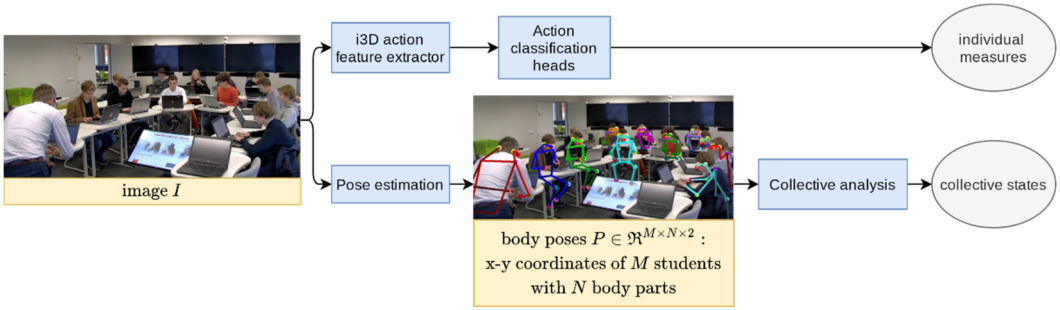


Figure 2. Data-processing pipeline for engagement estimation.

This study does not address facial expression, for two main reasons. A first reason is that facial expression analysis requires a relatively high spatial resolution, which may be hampered in real-world classroom settings that include rather cheap cameras, large groups of students that may be positioned relatively far away and bandwidth limitations. A second reason is that more subtle emotional or cognitive states, such as engagement, tend to be less pronounced in the face.

The body pose estimation is described in more detail in Section 3.3, the action detector in Section 3.4 and the collective analysis in Section 3.5.

3.3. Estimating Body Poses of Students in the Classroom

Body pose estimation algorithms estimate the locations of N human keypoint locations, based on an image. These are joints and other parts of the body, including the nose, ears, chin, shoulders, elbows, wrists, hips and legs. Contemporary computer vision algorithms are based on deep learning methods. We use the popular OpenPose library [6].

In order to identify internal states that can represent every student appearing in the videos, and potentially new students, we compute the pose representation $p_{m,t}$ for every student e_m at time t in every training video. We represent the location of keypoint n at time t as a 2D image $(x_{n,t}; y_{n,t})$. A pose, being a set of keypoints, for a student e_m in the classroom at time t is then given by $p_{m,t} = [(x_{1,t},y_{1,t}); (x_{2,t},y_{2,t}); \ldots ;(x_{N,t},y_{N,t})]$. This produces, at each time t, a data matrix $P = [p_{1,t}; p_{2,t}; \ldots ; p_{N,t}.]$ in $R^{M \times N \times 2}$ where M represents the total number of students in the videos and N corresponds to the dimensionality of the pose representation p.

Given that in the classroom setting, keypoints related to legs are not visible, we will limit further analysis to keypoints from the upper body (torso, head and arms). This reduces the total set of keypoints per student to 13, thus producing a 26-dimensional pose representation p_m per student e_m.

In order to compensate for the perspective effects introduced by the position of the camera, we conduct a two-step normalisation process on the pose representation $p_{m,t}$ of each student e_m at time t. First, we centre every predicted pose ($p_{m,t}' = p_{m,t} - c_{m,t}$) w.r.t. its neck keypoint $c_{m,t}$. Then, we divide the centred coordinates of each keypoint from $p_{m,t}'$ by the width and height of the rectangle defined by the shoulder and hips keypoints from $p_{m,t}'$.

3.4. Recognising Individual Behaviour from Students' Body Poses

Manual annotations are made for 8 actions that are relevant for student engagement: hand-raising, note-taking, hand on face, working with laptop, looking back, fiddling with hair, playing with cellphone and crossing arms. In total, we annotated 1031 sample clips. For each action, 10% of the clips are used as a test set. The other clips serve as training data for computer vision techniques in an attempt to learn to recognise these actions automatically.

We trained a deep learning-based action classifier following a transfer learning approach: a general action classification network, trained on a large variety of data, is retrained to detect actions in this specific classroom setting. Such an approach reduces the

amount of labeled training data, since the pretrained backbone network already extracts meaningful features. The general action classifier is based on the state-of-the-art i3D model [39], which takes as input a clip of about 2 seconds and generates a rich representation with 1024 dimensions. We train a separate classification head for each sample class. The classification head is a multilayer fully connected neural network with three layers of 512, 256 and 2 neurons, respectively.

As well as the accuracy in detecting students' individual and collective behaviour, we also attempt to investigate if and how these behaviours relate to the engagement scores. For this purpose, multilevel regression is used, to account for the nesting of the engagement scores (level 1) within students (level 2). The following equation applies:

$$y_{ti} = \beta_0 + \beta_1 \text{time}_{ti} + \beta_2 \text{ computer\_vision\_measure}_{ti} + r_{0i} + e_{ti} \quad (1)$$

y_{ti} refers to the engagement score of student i at time t. The first two terms ($\beta_0 + \beta_1 \text{time}_{ti}$) in the equation model the average trajectory of engagement over time. The third term (β_2 computer_vision_measure$_{ti}$) models the average effect of students' behaviour (detected by computer vision) on their engagement.

The random intercept, r_{0i}, captures individual differences in engagement levels. In this way, the model can also account for the possible individual nature of students' self-reporting, in that some students may systematically report higher or lower engagement scores (this phenomenon is known as grade inflation).

Finally, the residual e_{ti} represents the deviation in y at time t around a subjects' individual trajectory, due to noise, measurement error or the influence of other confounding variables.

3.5. Quantifying Collective Behaviour from Students' Body Poses

3.5.1. Estimating Intermediate Individual State Representations

After obtaining the normalised pose representation of each student, we attempt to cluster keypoint combinations, dividing the data contained in the pose representation matrix P into groups that represent similar poses students can take. In our experiments, this grouping process is conducted via k-means clustering, minimising the squared Euclidean distance between points and centroids. As a result of the grouping (clustering) process, each pose representation $p_{m,t'}$ in P is assigned a corresponding cluster id s_i which indicates its state at a given frame (time t). We decided to start our experiments with a relatively smaller number of $k = 12$ states. This number was large enough to distinguish sufficient states and, at the same time, low enough to ease the generation of discernible visualisations that will be used in later steps of the analysis. No other number for k was investigated.

In order to verify this observation, we run the clustering approach, focusing on three different subsets of keypoints: upper body (14 keypoints), head (6 keypoints) and arms (6 keypoints). This allows us to allocate students to different clusters based on the position of their joints and body parts.

3.5.2. Measuring Collective States

Consequently, the proportion of students over the possible states s_i within a temporal window Δt is computed. This distribution will provide insight on whether students, or subsets of them, are behaving in synchrony or in a random fashion.

To this end, we compute the collective state distribution $\Psi_{s_i,t}$, representing the proportion of students in state s_i at time (frame) t over a temporal window Δt composed of T frames. As illustrated in Figure 3, $\Psi_{s_i,t}$ is computed by assigning to each student the state label s_i that he or she adopted with the highest frequency within Δt after time t. Using a temporal window Δt to calculate the collective state distribution is advantageous over using real-time calculations because temporal windows average out possible noisy predictions introduced by the pose estimator.

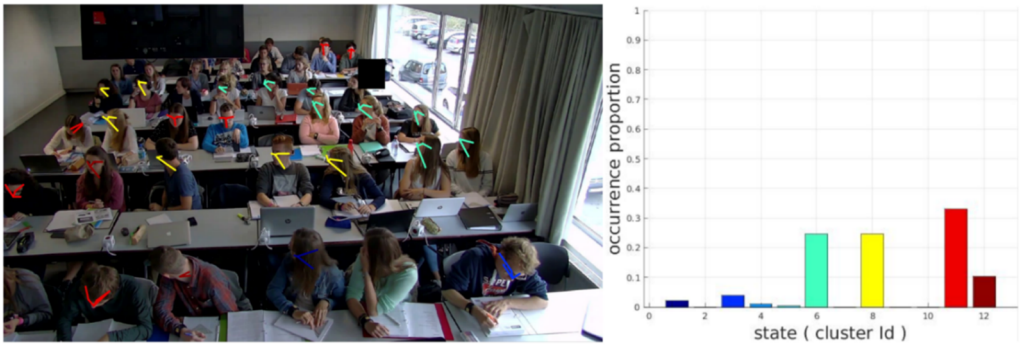


Figure 3. Example of the clustering approach applied to the keypoints of students' heads. Students are allocated to different clusters depending on their eye-gaze direction (here in a face-to-face setting). Students allocated to the red cluster tend to look to the front. The yellow, light green and dark blue clusters group students that are looking to the side or backwards. The feature spaces covered by each of these clusters only differ in a subtle and not necessarily meaningful way. As an example, the students allocated to the yellow cluster seem to have a larger nose-to-chest distance. The blue cluster includes a student that is looking down. Results are less consistent for the dark red cluster that includes two front-row students that were wrongly detected. The graph in the top right corner shows $\Psi_{s_i,t}$, the collective state distribution.

We use $\Psi_{s_i,t}$ to compute two collective measures that may be relevant in view of the classroom engagement: a measure for collective behaviour, and a measure for the reaction time to educational events in the classroom. This procedure is visualised in Figure 4.

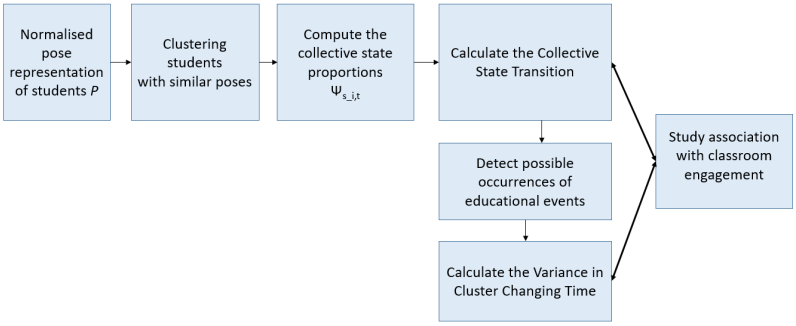


Figure 4. Method to calculate a measure for collective behaviour and a measure for the reaction time to classroom events.

3.5.3. Collective State Transition: A Measure to Detect Synchrony and Educational Events

An educational event is defined as any specific event that may simultaneously impact the engagement of the students in the classroom (e.g., when a student or the teacher asks a question, when students start interacting with each other, when an interactive quiz, a video or a new part of the lecture being started, etc.). The hypothesis put forward is that when educational events occur, the way in which students are distributed over the k different clusters changes in a relatively short period of time. Indeed, one could assume that an educational event may cause students to react and eventually change their body pose $p_{m,t'}$, which may in turn result in allocation to a different cluster s_i. Collective state transition (CST) is defined as the sum, over all k clusters, of the squared differences between the proportion $\Psi_{s_i,t}$ of the students that are allocated to cluster s_i at time t and the students that are allocated to cluster s_i at time $t + \Delta t$, divided by the interval time Δt. As such, CST represents the extent to which students simultaneously switch to similar postures (in indication for collective behaviour, or symmetry). In a formula, this becomes:

$$CST_t = \sum_{i=1}^{i=k} \frac{\left(\Psi_{s_i,t} - \Psi_{s_i,t+\Delta t}\right)^2}{\Delta t} \quad (2)$$

where:

- k is the total number of clusters, obtained by applying unsupervised clustering to students' joints and body parts;
- Δt is the time interval between two consecutive calculations (e.g., 0.5 s for a stride of 2 frames per second);
- $\Psi_{s_i,t}$ is the proportion of students that are assigned to cluster s_i at time t;
- $\Psi_{s_i,t+\Delta t}$ is the proportion of students that are assigned to cluster s_i at time $t + \Delta t$.

The hypothesis put forward is that an educational event occurs in the classroom when students simultaneously switch to similar postures, which we can detect by CST exceeding a certain threshold value τ. In addition, we investigate if higher values of CST will correspond to higher engagement levels.

3.5.4. From Collective Behaviour to Individual Behaviour: The Variance in Students' Reaction Time to Classroom Events

When CST exceeds τ and an educational event may be happening, it is checked which students changed cluster between the moment CST was exceeded and the moment CST dropped again below another threshold value τ'.

Then, it is calculated how much time it took for each student to change cluster. Let us indicate the time for student e_m to change cluster upon educational event e by $Time_{m,e}$.

Finally, the variance in cluster changing time (VCCT) is calculated. This is the variance in $Time_{m,e}$ required for all M students to change cluster s_i after the occurrence of an educational event e. The following formula applies:

$$VCCT_e = \sum_{m=1}^{m=M} \frac{\left(Time_{m,e} - \underline{Time_e}\right)^2}{M-1} \quad (3)$$

where:

- M is the total number of students.
- $Time_{m,e}$ is the time for student m to change cluster upon educational event e.
- $\underline{Time_e}$ is the average time across students to change cluster upon educational event e.

Inspired by the research of Raca et al. [37], the hypothesis put forward is that VCCT is an indicator of students' engagement. More specifically, we assume that the classroom is more engaged at time t_a than at time t_b, when at t_a the variance in reaction time is lower and the proportion of students that react to the event is higher. Figure 5 shows a visualisation of this idea.

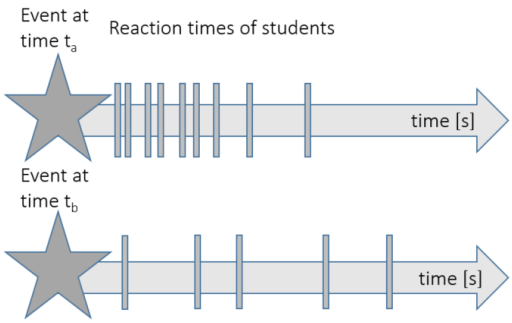


Figure 5. Illustration of a hypothesis of this study: when students are more engaged, they are more likely to react, and in a faster way.

3.6. Assessing the Number of Eye-Gaze Intersections That Students Share

This study also addresses eye-gaze interactions, by detecting the number of eye gazes that each pair of students share within a certain time window. Both the eye-gaze intersections at the individual (student) level and at the collective (classroom) level are quantified. The aim is to study the relatedness between the eye-gaze intersection frequency and the engagement self-reports. As such, the time window is centred around the occurrence of the self-reports, ranging from one minute before to one minute after the self-report.

4. Results

4.1. Recognising Individual Behaviour

Following our experiments, we observed that the action classifier does not perform equally well for all kinds of actions. We distinguish four reasons that have caused these low precision and recall values.

First, for all of the four actions that were mentioned, there was only a small number of annotations (between 37 and 185) available for training the computer vision algorithm. This reduces its capability to generalise and its robustness. To alleviate the effect of few annotations, the model follows a transfer learning approach by means of an i3D feature extractor pretrained on the Kinetics dataset. Despite this approach, the number of annotated samples still remains rather small.

Second, most of these actions are quite subtle, in that the position of students' body keypoints (used for the analysis) does not change much compared to a neutral position. This is, for example, the case for the action "playing with cellphone". When a certain action is so subtle, the discriminative visual clue to detect is too small, and the computer vision algorithm sometimes fails to do so.

Third, some of these actions are quite similar to others, in particular from certain viewpoints. As an example, fiddling with hair can be easily confused with hand on face or with raising a hand. The similarity of these actions results in some actions being erroneously recognised as other actions.

A final reason is that, in our classroom setting, some of the actions are prone to occlusion. This is, for example, the case for "working with laptop", as students' hands and cellphones are often occluded behind their laptops. Moreover, students sitting at the back of the classroom are often occluded by students that are sitting more to the front of the classroom.

The number of annotated samples and the obtained precision and recall are mentioned in Table 1. Given that the performance was found to be unsatisfactory for six of the eight actions, we decided only to withhold the two most accurate ones in our further analysis: hand-raising and note-taking.

Table 1. Binary action classification on top of i3D features for each class.

	Annotated Samples	Precision	Recall
Raising hand	85	0.67	0.59
Taking notes	315	0.69	0.63
Hand on face	185	0.35	0.32
Working with laptop	76	0.26	0.31
Looking back	84	0.50	0.53
Fiddling with hair	45	0.17	0.11
Playing with cellphone	37	0.22	0.38
Crossing arms	48	0.27	0.50

4.2. Measuring Engagement through Students' Individual Behaviour

As well as accuracies in detecting this behaviour through computer vision, we are also interested in the correlation between this behaviour and students' engagement. Results from the multilevel analysis indicate that both hand-raising ($t(284) = -0.20$, $p = 0.99$) and note-taking ($t(284) = 0.00$, $p = 0.83$) are not related to students' individual self-reported engagement scores.

4.3. Recognising Collective Behaviour

4.3.1. Unsupervised Clustering

As regards unsupervised clustering, a detailed analysis demonstrated that clusters related to students' heads were more consistent and more informative than clusters related to students' upper bodies and arms. Clusters related to students' upper body and arms

are not considered for further analysis because these were highly subject to "random" movements that were not meaningful in terms of engagement. In addition to that, joints and body parts belonging to these clusters were more often invisible because other students that were sitting more to the front of the classroom occluded them. Therefore, only the clusters related to students' heads were kept for further analysis.

4.3.2. Collective State Transition: A Measure to Detect Synchrony and Educational Events

First, the optimal temporal window to calculate CST (collective behaviour) was investigated. Human observations of several samples across videos of the face-to-face lectures (a total length of 103 min was analysed) showed that a Δt of 4 s is optimal to detect classroom events. If Δt is too short, e.g., 1 s, some classroom events are overlooked, because only a small proportion of students react within that short timeframe, resulting in the threshold value τ of the CST not being exceeded. If Δt is too long, e.g., 15 s, the classroom event cannot be detected either because the length of the window is much longer than the reaction time of most students, resulting in the flattening of peaks in the value of CST.

In addition, an optimal value for the threshold value τ was addressed. Further human observations of recordings showed that a good threshold value for τ in order to be a reliable indicator for classroom events is 0.2×10^{-3} s^{-1}. If the threshold value is lower, it sometimes gets exceeded, while no educational event happened (false positives). If the threshold value is higher, some educational events are missed (false negatives). The analysed sample of 103 min indicated a recall of 63% for classroom events being visible, as peaks in CST exceeded τ, and a precision of 45% for peaks in CST corresponding to classroom events.

From that analysis, it is also concluded that when CST drops again below a threshold value of 0.1×10^{-3} s^{-1}, this is a good indication for the end of that classroom event.

4.4. Measuring Engagement through Students' Collective Behaviour

4.4.1. CST

The multilevel analysis that applies to the individual engagement scores for each student did not establish a significant relationship between engagement and CST (($t(284) = 1.24$, $p = 0.22$). Next to individual engagement, the average engagement score across the entire classroom was also studied, and its association with CST. Figure 6 indicates that high degrees of synchrony (CST > τ) correspond to high levels of classroom engagement. A multilevel analysis confirmed this observation, in that a significant effect of a dichotomous indicator based on CST (equal to 1 when CST exceeds τ, and 0 otherwise) on classroom engagement ($t(387) = 2.54$, $p = 0.01$) was established.

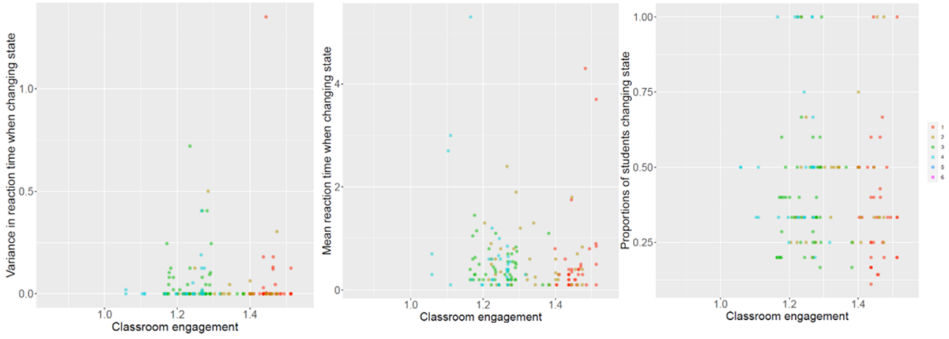
Figure 6. Synchronous movements in the classroom (quantified by collective state transition (CST)) as a function of the average classroom engagement. The colours refer to the different lectures.

Despite this significant finding, this relationship is not very strong and we found that it is also possible that the classroom is engaged while students do not show any collective

behaviour. In addition, the events characterised by high values in CST all correspond to the start of the first lecture (where students were seemingly both synchronous and engaged) and not to multiple moments spread out in time, which makes the evidence of the significance of the association between engagement and collective behaviour much weaker.

4.4.2. The Variance in Students' Reaction Time to Educational Events

The results show that the variance in students' reaction time to educational events does not correlate with the average engagement of the classroom ($r = 0.03$, $p = 0.69$). The same observation applies to the mean of students' reaction times ($r = -0.07$, $p = 0.42$) and to the proportion of students that react to an educational event by changing cluster ($r = -0.08$, $p = 0.33$). The scatterplots in Figure 7 illustrate these observations.



(**a**) (**b**) (**c**)

Figure 7. Reaction times as a function of the classroom engagement: (**a**) the variance in reaction time; (**b**) the mean reaction time and (**c**) the proportion of students changing state. The colours refer to the different lectures.

4.4.3. Shared Eye-Gaze Intersections

Separate plots and analyses are foreseen for the pure face-to-face setting (lectures 1 and 2) and for the hybrid setting (lectures 3 and 4). The pure virtual lectures 5 and 6 are not analysed, as virtual students cannot share eye gazes.

The results suggest that eye gazes in the classroom do not covary with engagement for lectures 1 and 2 ($r = -0.13$, $p = 0.21$), but do covary for lectures 3 and 4 ($r = 0.59$; $p < 0.001$). When analysing the graphs (see Figure 8) more closely, it seems that there is no correlation within a lecture, but there is some correlation between lectures, in that students are more engaged and at the same time share more eye gazes during lecture 3 than during lecture 4. It seems that lecture 3 triggered the students more than lecture 4, both in terms of engagement and the eye gazes students share.

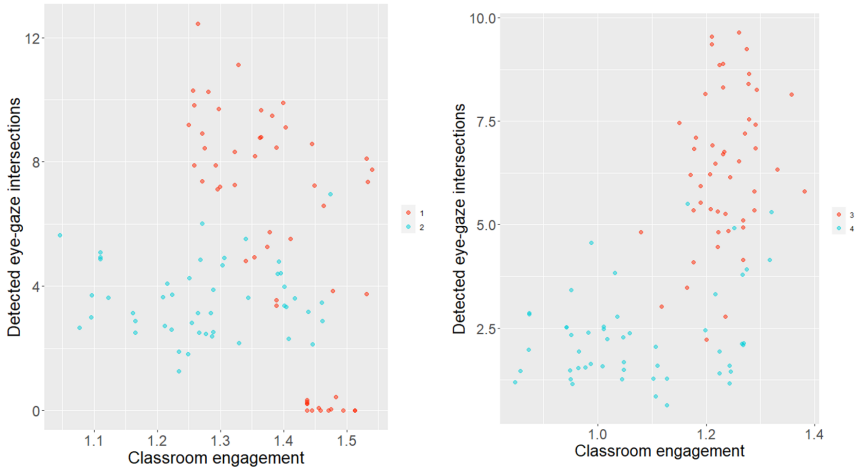


Figure 8. Detected eye-gaze intersections as a function of the classroom engagement.

5. Discussion

5.1. Discussion of This Study's Main Findings on Individual Behaviour

Regarding the recognition of students' individual behaviour, the results show that it is not yet straightforward to recognise students' actions in the classroom through computer vision, and some actions are more difficult to recognise than others. Similarly, as with Wang et al. [40], we found that recognising actions typically becomes more challenging as the degree of variation with which the action can be exhibited (both within and between students) increases, the number of joints that are involved increases, there is more interference with other similar actions and there are fewer annotations available. The hand-raising study by Wang et al. [27], for example, uses a much larger dataset (>10,000 hand-raising samples versus 85 in our study), which resulted in a higher accuracy of 95%.

Regarding the measurement of engagement through students' individual behaviour, the results are somewhat disappointing. Although there is a theoretical ground for a connection between several of the investigated actions (note-taking and hand-raising) and engagement, no significant correlations were established in our sample.

5.2. Discussion of This Study's Main Findings on Collective Behaviour

This study quantified students' collective behaviour by applying unsupervised clustering to students' body keypoints. In this way, measures could be obtained for students' synchrony, eye-gazes and reaction times.

Regarding the measurement of engagement through these measures for collective behaviour, this study found a significant association between the engagement of the classroom and the extent to which students simultaneously change body poses, a measure for their synchrony. Despite this interesting finding, this correlation was only found between the first lecture and the other lectures, which makes the evidence of this association rather weak. Students' reaction times did not relate to their self-reported engagement either. As for students' eye-gaze intersections, a significant correlation was established with the classroom engagement, but this was merely a correlation between lectures rather than within a lecture. A possible explanation for this is that lecture 3 was the first time students experienced the hybrid setting. It is plausible that this novelty effect both resulted in higher engagement levels and more interaction, manifested by more eye-gaze intersections.

5.3. Limitations of the Study

An important limitation is that the way in which this study measures individual and collective behaviour is tightly connected to the behavioural component of engagement, but does not cover the emotional and cognitive components, although these components are represented in students' self-reported score. As a result, students who are mainly cognitively and emotionally engaged, but not exhibiting explicit engaged behaviour, may not be recognised as actually being engaged.

Capturing the different components of engagement is important as they only covary to a limited extent [26]. Interrogating the different engagement components may be interesting, as this would enable us to reveal how different indicators relate to the behavioural, emotional and cognitive component separately.

Another limitation is the limited sample size, both in view of the number of participants and the number of lectures that were recorded, which reduces the study's statistical power.

The way in which engagement is inquired also has its limitations, as the criterion validity of self-reporting is not perfect. Self-reporting scales may cause some subjects to consistently report lower or higher scores than others. This phenomenon creates some additional variance in the data, which makes it more difficult to detect significant effects, if these exist. Furthermore, this additional variance puts a certain "upper limit" on the variance in the self-reported engagement that the manifest variables can eventually explain.

Although this phenomenon cannot be entirely mitigated, multilevel regression models are well suited to deal with this, as they include person-specific intercepts to represent

differences between subjects, which can, for example, account for students consistently reporting higher or lower engagement levels. This functionality improves the robustness of the parameter estimation, even if some subjects tend to indicate higher/lower scores than others.

We expect the social response bias in reporting engagement to be small, as participants did not benefit from reporting high scores. Participants were also sufficiently old to understand abstract concepts such as engagement. Apart from this, there may have been an influence of peers, as students had the possibility to exchange experiences about the study in between different lectures.

Another limitation concerns the conceptualisation of engagement: as a trait, or as a state [41]. This conceptualisation affects the extent to which engagement can fluctuate. When engagement is operationalised as a trait (e.g., at the institute level), it refers to how engaged someone is with a certain educational programme in general [42]. This conceptualisation implies that engagement is relatively constant over time, as it is influenced by students' characteristics (academic self-concept, sense of belonging, interest in the academic programme, etc.) that do not change rapidly. Another way to operationalise engagement is as a state, i.e., the extent to which students engage during a lecture. If students participate more actively (behaviour), invest more mental resources (cognition) and show more positive reactions to the topic of the lecture, the teacher and the fellow students (emotion), we assume students are more engaged. The second conceptualisation is the one that is used in our research. It implies that engagement is malleable and can fluctuate over the course of a lecture [43]. Nonetheless, we agree that the variation of engagement within a lecture may be somewhat limited. From a research perspective, this is a pity, as the somewhat limited variations in engagement during a lecture make the proxies for engagement less visible, thereby reducing statistical power.

A final limitation arises from the cultural lens through which engagement is measured. As engagement is socially constructed, it is perceived differently across cultures [44]. Perceptions of emotion are not universal, but depend on cultural and conceptual contexts [44]. Parents also have an effect on their children, by transferring a "cultural toolkit" to them, that impacts their attitude and preferences [45]. As an example, Western countries may tend to associate engagement more with its behavioural component (e.g., assertiveness, hand-raising and discourse), whereas other cultures may value the cognitive component more (thoughtful attentiveness). As a result, self-reporting scores as well as associations between engagement and its proxies may differ across cultures. In a similar line of thought, engagement may also depend on students' personality and the context of the classroom as a whole (fellow students, teaching style, etc.).

5.4. Assets of the Study

Irrespective of these results, the methodology to quantify collective behaviour may be a promising approach for future research. Clustering the most meaningful keypoints and detecting changes in the distribution across clusters over time has several advantages. A first advantage is its generalisability, which enables us to apply this method to variables other than engagement or to settings other than the hybrid virtual classroom. A second advantage is that the collective approach does not require the identification of individuals, which makes it less privacy intrusive. A third advantage is that the uni-dimensional (or low-dimensional) collective measures require fewer data to investigate the problem at hand than the initial high-dimensional latent space. A final advantage is that a measure that is inaccurate in a non-systematic way, and thus not suitable on an individual level, may still provide an accurate estimation for groups, as the aggregation tends to cancel out individual differences.

This work may also inspire researchers in the teaching and learning field regarding methods that are used to conduct manual video analysis [46], in two particular ways. First, computer vision may automate the recognition of several actions, which could make the counting process more efficient. Second, computer vision enables us to quantify certain

complex measures in a more precise way than is possible with manual video analysis, expanding the number of events that can be kept track of (think about measures related to synchrony, eye gaze or reaction times).

5.5. Implications for Practice and Future Research

The nonexistent or weak associations that this study obtained do not yet allow us to measure students' engagement in an accurate way. As a consequence, and even irrespective of ethical concerns, we recommend being very careful when monitoring students in terms of their engagement through computer vision. Based on this study's results, we advise not to display any of the addressed measures (e.g., synchrony, eye-gaze intersections, hand-raising, note-taking, etc.) on teacher dashboards, as they may give teachers the unjustified impression that they represent an accurate measure of the engagement of their classroom.

Extending the current approach to better cover the behavioural component of engagement, as well as to cover parts of the emotional and cognitive components of engagement, could result in better estimations. Therefore, additional data sources could be monitored, taking a more multimodal approach. Regarding behavioural engagement, more actions could be monitored, such as students' participation in a chat room and their participation in interactive quizzes. To cover the emotional component, facial recognition could still be considered, as far as the spatial resolution and the classroom setting (occlusion) allow it. Covering the cognitive component remains challenging, as neuroscientific methods are typically intrusive and expensive, and their accuracy is still limited. Conducting discourse analysis to capture students' thoughts and therefore part of their emotional and cognitive engagement is another option to further extend the multimodal approach. To operationalise this idea in an automatic way, students' voices could be captured by means of natural language processing. Alternatively, one may also think of an application to analyse students' notes in real time.

Even with a multimodal approach, estimating the engagement of a specific student will most likely still be difficult. Estimating the average engagement of a group may be more feasible as this may cancel out non-systematic inaccuracies of individual measurements, but even this approach is uncertain. If the engagement measurement could be improved, it could be given as feedback to teachers, who could combine this information with an interpretation of the classroom context. In this way, computer vision could provide objective information to a human agent, who could employ his or her ability to interpret this information together with a complex interplay of context, emotions and past events. In this way, in a similar vision as that of Luckin [47], AI does not replace teachers, but supports them, to solve educational challenges together.

It is important to be aware of aspects related to ethics and privacy when monitoring students and analysing their data, and accounting for them before deploying a monitoring system of any kind. More specific guidelines, however, are not the focus of this study, and are therefore not further expanded upon.

6. Conclusions

This study has elaborated a methodology to detect individual behaviour (hand-raising, note-taking) as well as collective behaviour (symmetry, reaction times, eye-gaze intersections) of students in a hybrid virtual classroom. Computer vision techniques allow us to recognise individual behaviour with reasonable accuracy and precision, and they can also quantify collective behaviour, via unsupervised clustering. However, none of the investigated measures for individual behaviour is found to correlate significantly with students' self-reported engagement. Nonetheless, at the collective level, a weak but significant connection is established between the classroom engagement on the one hand and students' collective behaviour and students' eye-gaze intersections on the other hand.

Based on this and some other studies' findings, we recommend not to use computer vision techniques to estimate students' engagement. Only if it can be clearly demonstrated that engagement can be measured sufficiently precisely, which includes also grasping

the cognitive and emotional component, does it make sense to provide teachers with that information. Even then, we suggest that teachers combine this estimate with their perception of the entire classroom context, emotion and past events, in an aim to obtain a reliable engagement estimate of their classroom.

Author Contributions: Conceptualisation, J.O., T.T., A.R. and P.V.; methodology, J.O., T.V. and P.V.; software, J.O., T.V. and T.T.; investigation, P.V., J.O., A.R. and T.V.; writing—original draft preparation, P.V., J.O. and T.V.; writing—review and editing, T.T., A.R., F.D. and W.V.d.N. All authors have read and agreed to the published version of the manuscript.

Funding: The project was funded by the Flemish Government (through imec, the research center on nanoelectronics and digital technologies and through VLAIO, the Flemish agency for innovation and entrepreneurship).

Institutional Review Board Statement: The study was approved by the ethical commission of KU Leuven (reference G- 2018 06 1264).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: Anonymised datasets can be made available upon request.

Acknowledgments: This work was executed within LECTURE+, a two-year research project executed by several partners from industry (Barco, Televic and Limecraft) and academia (KU Leuven research groups Distrinet, ESAT-PSI and Itec). In this research, Itec's expertise related to educational technology and PSI's expertise related to computer vision techniques are brought together.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Zhang, H.; Zhang, Y.-X.; Zhong, B.; Lei, Q.; Yang, L.; Du, J.-X.; Chen, D.-S. A Comprehensive Survey of Vision-Based Human Action Recognition Methods. *Sensors* **2019**, *19*, 1005. [CrossRef] [PubMed]
2. Muthalagu, R.; Bolimera, A.; Kalaichelvi, V. Lane detection technique based on perspective transformation and histogram analysis for self-driving cars. *Comput. Electr. Eng.* **2020**, *85*, 106653. [CrossRef]
3. Peter, C.; Beale, R. *Affect and Emotion in Human-Computer Interaction: From Theory to Applications*, 1st ed.; Springer: Berlin/Heidelberg, Germany, 2008.
4. Fredricks, A.J.; Blumenfeld, P.C.; Paris, A.H. School Engagement: Potential of the Concept, State of the Evidence. *Rev. Educ. Res.* **2004**, *74*, 59–109. [CrossRef]
5. Raes, A.; Vanneste, P.; Pieters, M.; Windey, I.; Noortgate, W.V.D.; Depaepe, F. Learning and instruction in the hybrid virtual classroom: An investigation of students' engagement and the effect of quizzes. *Comput. Educ.* **2020**, *143*, 103682. [CrossRef]
6. Cao, Z.; Hidalgo Martinez, G.; Simon, T.; Wei, S.-E.; Sheikh, Y.A. OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *43*, 172–186. [CrossRef] [PubMed]
7. Ekman, R. *What the Face Reveals: Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System (FACS)*, 2nd ed.; Series in affective science; Oxford University Press: New York, NY, USA, 1997.
8. Tarnowski, P.; Kołodziej, M.; Majkowski, A.; Rak, R.J. Emotion recognition using facial expressions. *Procedia Comput. Sci.* **2017**, *108*, 1175–1184. [CrossRef]
9. El Kaliouby, R.; Robinson, P. Real-Time Inference of Complex Mental States from Facial Expressions and Head Gestures. In Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2004, Washington, DC, USA, 27 June–2 July 2004; Volume 2. [CrossRef]
10. Van Acker, B.B.; Bombeke, K.; Durnez, W.; Parmentier, D.D.; Mateus, J.C.; Biondi, A.; Saldien, J.; Vlerick, P. Mobile pupillometry in manual assembly: A pilot study exploring the wearability and external validity of a renowned mental workload lab measure. *Int. J. Ind. Ergon.* **2020**, *75*, 102891. [CrossRef]
11. Dobbins, C.; Denton, P. MyWallMate: An Investigation into the use of Mobile Technology in Enhancing Student Engagement. *TechTrends* **2017**, *11*, 142–549. [CrossRef]
12. Gobert, J.D.; Baker, R.S.; Wixon, M.B. Operationalizing and Detecting Disengagement within Online Science Microworlds. *Educ. Psychol.* **2015**, *50*, 43–57. [CrossRef]
13. Connell, J.P.; Spencer, M.B.; Aber, J.L. Educational Risk and Resilience in African-American Youth: Context, Self, Action, and Outcomes in School. *Child Dev.* **1994**, *65*, 493. [CrossRef]
14. Marks, H.M. Student Engagement in Instructional Activity: Patterns in the Elementary, Middle, and High School Years. *Am. Educ. Res. J.* **2000**, *37*, 153–184. [CrossRef]

15. Furlong, M.J.; Christenson, S.L. Engaging students at school and with learning: A relevant construct forall students. *Psychol. Sch.* **2008**, *45*, 365–368. [CrossRef]
16. Boyle, F.; Kwon, J.; Ross, C.; Simpson, O. Student–student mentoring for retention and engagement in distance education. *Open Learn. J. Open Distance e-Learn.* **2010**, *25*, 115–130. [CrossRef]
17. Mazer, J.P. Associations among Teacher Communication Behaviors, Student Interest, and Engagement: A Validity Test. *Commun. Educ.* **2013**, *62*, 86–96. [CrossRef]
18. Christophel, D.M. The relationships among teacher immediacy behaviors, student motivation, and learning. *Commun. Educ.* **1990**, *39*, 323–340. [CrossRef]
19. Raes, A.; Detienne, L.; Windey, I.; Depaepe, F. A systematic literature review on synchronous hybrid learning: Gaps identified. *Learn. Environ. Res.* **2019**, *23*, 269–290. [CrossRef]
20. Lee, O.; Anderson, C.W. Task engagement and conceptual change in middle school science classrooms. *Am. Educ. Res. J.* **1993**, *30*, 585–610. [CrossRef]
21. Whitehill, J.; Serpell, Z.; Lin, Y.-C.; Foster, A.; Movellan, J.R. The Faces of Engagement: Automatic Recognition of Student Engagement from Facial Expressions. *IEEE Trans. Affect. Comput.* **2014**, *5*, 86–98. [CrossRef]
22. Spanjers, D.M.; Burns, M.K.; Wagner, A.R. Systematic Direct Observation of Time on Task as a Measure of Student Engagement. *Assess. Eff. Interv.* **2008**, *33*, 120–126. [CrossRef]
23. Revere, L.; Kovach, J.V. Online technologies for engaged learning, a meaningful synthesis for educators. *Q. Rev. Distance Educ.* **2011**, *12*, 113–124.
24. Pirsoul, T.; Parmentier, M.; Nils, F. The rocky road to emotion measurement in learning and career development: On the use of self-reports. In Proceedings of the 18th Biennial EARLI Conference for Research on Learning and Instruction, Aachen, Germany, 12–16 August 2019; Available online: https://hdl.handle.net/2078.1/218801 (accessed on 23 November 2020).
25. Brown, G.T.; Andrade, H.L.; Chen, F. Accuracy in student self-assessment: Directions and cautions for research. *Assess. Educ. Princ. Policy Pr.* **2015**, *22*, 444–457. [CrossRef]
26. Skinner, E.A.; Kindermann, T.A.; Furrer, C.J. A Motivational Perspective on Engagement and Disaffection. Conceptualization and assessment of children's behavioral and emotional participation in academic activities in the classroom. *Educ. Psychol. Meas.* **2008**, *69*, 493–525. [CrossRef]
27. Liao, W.; Xu, W.; Kong, S.; Ahmad, F.; Liu, W. A Two-stage Method for Hand-Raising Gesture Recognition in Classroom. In Proceedings of the 2019 8th International Conference on Educational and Information Technology (ICEIT 2019), Cambridge, UK, 2–4 March 2019; pp. 38–44. [CrossRef]
28. Lin, J.; Jiang, F.; Shen, R. Hand-Raising Gesture Detection in Real Classroom. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 6453–6457.
29. Böheim, R.; Knogler, M.; Kosel, C.; Seidel, T. Exploring student hand-raising across two school subjects using mixed methods: An investigation of an everyday classroom behavior from a motivational perspective. *Learn. Instr.* **2020**, *65*, 101250. [CrossRef]
30. Chen, Y.-L.E.; Kraklow, D. Taiwanese College Students' Motivation and Engagement for English Learning in the Context of Internationalization at Home. *J. Stud. Int. Educ.* **2014**, *19*, 46–64. [CrossRef]
31. Barbadekar, A.; Gaikwad, V.; Patil, S.; Chaudhari, T.; Deshpande, S.; Burad, S.; Godbole, R. Engagement Index for Classroom Lecture using Computer Vision. In Proceedings of the 2019 Global Conference for Advancement in Technology (GCAT), Bangalore, India, 18–20 October 2019; pp. 1–5.
32. Canedo, D.; Trifan, A.; Neves, A.J.R. Monitoring Students' Attention in a Classroom through Computer Vision. In *Highlights of Practical Applications of Agents, Multi-Agent Systems, and Complexity: The PAAMS Collection. PAAMS 2018. Communications in Computer and Information Science*; Springer: Cham, Switzerland, 2018; Volume 887, pp. 371–378.
33. Li, W.; Jiang, F.; Shen, R. Sleep Gesture Detection in Classroom Monitor System. In Proceedings of the (ICASSP 2019) 2019 IEEE International Conference on Acoustics, Speech and Signal Processing, Brighton, UK, 12–17 May 2019; pp. 7640–7644.
34. Nezami, O.M.; Dras, M.; Hamey, L.; Richards, D.; Wan, S.; Paris, C. Automatic Recognition of Student Engagement Using Deep Learning and Facial Expression. In Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Dublin, Ireland, 10–14 September 2018; pp. 273–289.
35. MacHardy, Z.; Syharath, K.; Dewan, P. Engagement Analysis through Computer Vision. In Proceedings of the 7th International Conference on Collaborative Computing: Networking, Applications and Worksharing (CollaborateCom 2011), Orlando, FL, USA, 15–18 October 2012; pp. 535–539.
36. Bosch, N.; D'Mello, S.; Ocumpaugh, J.; Baker, R.S.; Shute, V. Using Video to Automatically Detect Learner Affect in Computer-Enabled Classrooms. *ACM Trans. Interact. Intell. Syst.* **2016**, *6*, 1–26. [CrossRef]
37. Raca, M.; Tormey, R.; Dillenbourg, P. Sleepers' lag—study on motion and attention. In Proceedings of the 4th International Conference on Learning Analytics and Knowledge (LAK '14), Indianapolis, IN, USA, 24–28 March 2014; pp. 36–43.
38. Matthews, G.; De Winter, J.; Hancock, P.A. What do subjective workload scales really measure? Operational and representational solutions to divergence of workload measures. *Theor. Issues Ergon. Sci.* **2020**, *21*, 369–396. [CrossRef]
39. Carreira, J.; Zisserman, A. Quo vadis, action recognition? A new model and the kinetics dataset. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 6299–6308.
40. Wang, Z.; Jiang, F.; Shen, R. An Effective Yawn Behavior Detection Method in Classroom. In *Mining Data for Financial Applications*; Springer: Cham, Switzerland, 2019; Volume 11953, pp. 430–441.

41. Macey, W.H.; Schneider, B. The Meaning of Employee Engagement. *Ind. Organ. Psychol.* **2008**, *1*, 3–30. [CrossRef]
42. Organisation for Economic Co-Operation and Development. *Student Engagement at School: A Sense of Belonging and Participation*; Organisation for Economic Co-Operation and Development (OECD): Paris, France, 2003.
43. Finn, J.D.; Rock, D.A. Academic success among students at risk for school failure. *J. Appl. Psychol.* **1997**, *82*, 221–234. [CrossRef]
44. Gendron, M.; Roberson, D.; Van Der Vyver, J.M.; Barrett, L.F. Perceptions of emotion from facial expressions are not culturally universal: Evidence from a remote culture. *Emotion* **2014**, *14*, 251–262. [CrossRef]
45. Golann, J.W.; Darling-Aduana, J. Toward a multifaceted understanding of Lareau's "sense of entitlement": Bridging sociological and psychological constructs. *Sociol. Compass* **2020**, *14*, e12798. [CrossRef]
46. Derry, S.J.; Pea, R.D.; Barron, B.; Engle, R.A.; Erickson, F.; Goldman, R.; Hall, R.; Koschmann, T.; Lemke, J.L.; Sherin, M.G.; et al. Conducting Video Research in the Learning Sciences: Guidance on Selection, Analysis, Technology, and Ethics. *J. Learn. Sci.* **2010**, *19*, 3–53. [CrossRef]
47. Luckin, R. AI is coming: Use it or lose to it. *Times Educ. Suppl.* **2018**, 5306.