Using Student Action Recognition to Enhance the Efficiency of Tele-education

Eleni Dimitriadou[1] and Andreas Lanitis[1,2]
[1]*Visual Media Computing Lab, Deptartment of Multimedia and Graphic Arts, Cyprus University of Technology, Cyprus*
[2]*CYENS - Centre of Excellence, Nicosia, Cyprus*

Keywords: Action Recognition, Deep Learning, Tele-education.

Abstract: Due to the COVID-19 pandemic, many schools worldwide are using tele-education for class delivery. However, this causes a problem related to students' active class participation. We propose to address the problem with a system that recognizes student's actions and informs the teacher accordingly, while preserving the privacy of students. In the proposed action recognition system, seven typical actions performed by students attending online courses, are recognized using Convolutional Neural Network (CNN) architectures. The actions considered were defined by considering the relevant literature and educator's views, and ensure that they provide information about the physical presence, active participation, and distraction of students, that constitute important pedagogical aspects of class delivery. The action recognition process is performed locally on the device of each student, thus it is imperative to use classification methods that require minimal computational load and memory requirements. Initial experimental results indicate that the proposed action recognition system provides promising classification results, when dealing with new instances of previously enrolled students or when dealing with previously unseen students.

1 INTRODUCTION

A major issue related to tele-education is the level of concentration of students during the course delivery process (Moubayed et al., 2020). This is especially true for primary school students, who find it more difficult to concentrate and attend a class delivered online (Putri et al., 2020). To make the situation worse, in many countries the use of cameras during the delivery of online courses to primary school students is forbidden as a means of preserving student privacy, and since the teacher has no optical contact with the students, it is very common that student's lose concentration during the course delivery process.

The purpose of this study is to provide a computer vision-based tool that can be used for examining the behavioral participation and behavioural disaffection of primary school students' during teleconferencing. Behavioral participation involves: effort, attention, initiation of action, involvement, intensity, persistence, and absorption, while behavioral disaffection involves: withdrawal, giving up, passivity, being inattentive, distracted, not engaged mentally and not prepared (Skinner et al., 2008, 2009). In line with the taxonomy listed above, we aim to develop a method for recognising student's actions related to student's attention to the lesson, their absence from the lesson, their active participation, and their distraction.

To deal with this problem, we propose to employ a machine vision-based approach that can be used for monitoring student's actions during the class delivery process and report to the instructor the status and level of concentration of the students. For this purpose, we have used different network architectures including faster R-CNN, SqueezeNet, GoogleNet, and Inception-v3. Figure 1 shows a block diagram of the proposed student monitoring system. An important aspect of this approach is that images of students captured during the process are processed locally and only information about the actions of the students is transmitted to the instructor so that privacy issues are not violated. When the proposed system is used in real classes, it is assumed that both parents and students should give their prior consent, ensuring that no ethical considerations related to privacy protection arise. Since the computation is done locally on the computers or smart devices of each student, it is imperative to adopt approaches that require minimum computational power, to ensure the smooth operation of the system regardless the specifications of the equipment that each student uses for the purpose of tele-education. Although the recognition of student's actions has been addressed in the literature before, to the best of our knowledge, the proposed formation of

Dimitriadou, E. and Lanitis, A.
Using Student Action Recognition to Enhance the Efficiency of Tele-education.
DOI: 10.5220/0010868200003124
In *Proceedings of the 17th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2022) - Volume 5: VISAPP*, pages 543-549
ISBN: 978-989-758-555-5; ISSN: 2184-4321

the application within the domain of tele-education has not been considered before.

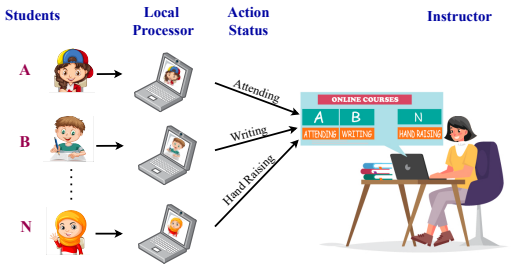

Figure 1: Schematic diagram of the proposed student monitoring system during online courses.

In the remainder of the paper, we present a literature review on the topics of Computer vision in education and present the methodology adopted in our study. In Section 4, we present the results of the experimental evaluation followed by plans for future work and concluding comments.

2 LITERATURE REVIEW

Computer vision is often used to help educators to analyze videos and images of students as a means of acquiring important information regarding students. Within this context several researchers addressed the problem of automated student attendance registration. Chowdhury et al. (2020), proposed an automatic student attendance system based on face recognition using Convolutional Neural Networks (CNN). The proposed system can detect and recognize multiple students' faces from a real-time video stream obtained by a static camera located in a classroom. Mery et al. (2019), used deep networks to implement an automated student attendance system that identifies students in crowded classrooms using images captured by a smartphone camera, while Chintalapati and Raghunadh (2013) suggested an automated attendance management system based on face detection and recognition algorithms to detect the students who are entering the classroom. Instead of using a static camera, Gupta et al. (2018), used a rotating camera to capture images of students, followed by the detection of students' faces using an Inception-v3 CNN.

Computer vision methods have been used to help educators to keep control of all the students in physical classes (Radlak et al., 2015; Bebis et al., 2003; Rashmi et al., 2021). Ashwin and Guddeti (2020) proposed a system to analyze students' body postures, gestures and facial expressions to investigate engagement. Three states of student's engagement were examined: boredom, engaged and neutral. Ngoc Anh et al. (2019) present a system to monitor the behaviour of students in the classroom. Various computer vision techniques such as face embedding, gaze estimation, face detection and facial landmark detection were used for analyzing images. Similarly, Thomas and Jayagopi (2017) proposed a machine learning algorithm to determine students' engagement in a classroom by analyzing students' head position, direction of eye gaze and facial expressions. Furthermore, Yang and Chen (2011) presents a system that provides feedback regarding student actions that utilizes eye and face detection to determine if the students are active or not. Recently, Tran et al. (2021) developed a system regarding video face recognition in collaborative learning environments. Their proposed system could recognize multiple students' faces from different angles in a single image frame using static cameras in the classroom. Similarly, Teeparthi et al. (2021) developed a method for detecting multiple students' hands in a single image frame in a collaborative learning environment. Their proposed method included transfer learning and Faster R-CNN as their baseline method using time-projections, clustering and small region removal. In all approaches stated above fixed cameras were used for capturing images.

Instead of using a single camera, Li et al. (2019) describes the use of four cameras, which were fixed on the wall (front and back of classroom), to record the students' actions from various viewpoints. Furthermore, they studied 15 different types of student's actions, six of which are similar with the actions considered in our work. To evaluate the proposed model, four types of algorithms were used: IDT (Improved Dense Trajectory) used with Support Vector Machines, and CNN implemented with VGG-16 and Inception-v3 architectures. Unlike the work of Li et al. (2019), in our work we use images captured from a single laptop camera rather than employing multiple fixed cameras.

Recently, Rashmi et al. (2021), proposed a system, that recognizes multiple actions of students given an image frame captured by a CCTV camera. The dataset of student's actions included the actions of sleeping, eating, using phone, discussion and being engaged. The YOLOv3 framework was used for object detection and students' action recognition. While this study is related to our work, in our study we recognize the actions of a student in a single image frame during the online courses using a camera pointing at each student. As a result, the overall efficiency of the approach is optimized, as there is no need to include the computationally expensive YOLO algorithm that would have prevented the real-time operation of the system on less powerful machines.

The work of Bian et al. (2019), is closely related to our work since they consider the problem of facial expression recognition during the delivery of on-line courses. They use facial expression algorithms based on the VGG16 network and different CNN architectures. In our work, instead of focusing on the analysis of facial expressions, we consider the recognition of actions that relate to the physical presence (Raddon, 2006), active participation (Pratton and Hales, 1986), and distraction (Baron, 1986) as indicated in the relevant literature.

3 METHODOLOGY

The main steps of the proposed methodology include the steps of User Evaluation, Data Collection, and Action Recognition. In the following sections, we describe each step of the methodology.

3.1 User Evaluation

Prior to the application development, the proposed system and its operation was presented to a group of teachers, students and parents, who provided feedback related to the class monitoring tool (Dimitriadou and Lanitis, 2021). Data were collected through semi-structured interviews since this method constitutes the most appropriate tool for recording detailed opinions (Barriball and While, 1994).

The analysis of the feedback received shows that the participants were enthusiastic about the system of recognizing students' actions in online lessons. With regards to the disadvantages, the educators and parents characterized the application as excellent in contrast with the students, where the majority declared that they would have difficulties in accepting the application of the system as it will prevent them from engaging in activities not related to the class delivery. At the end of the interviews, most of the participants found the proposed application complete, while they were given the chance to suggest ways to enrich the system with additional features. More details related to the analysis of results of the user evaluation are presented in (Dimitriadou and Lanitis, 2021).

Based on the feedback received from the stakeholders, the development process was carried out, as exemplified in the following subsections.

3.2 Data Collection

Data have been collected using a laptop camera, from 15 primary school students while they were using a laptop in a style that is usually adopted during the attendance of on-line courses from their homes. Students participating in the experiment were asked to perform the actions presented in Table 1. All actions considered are related to the three categories of "Physical Presence", "Active Participation" and "Distraction". Furthermore, the definition of the exact actions considered were based on suggestions of several active educators, engaged in tele-education.

Table 1: Taxonomy of student actions considered.

Categories	Classes	Description
Physical Presence	Absent	Not being present in the camera point of view.
Active Participation	Attending	Looking towards the camera.
	Hand Raising	Raising the left or right hand to request permission to talk.
	Writing	Student writing notes on a pad.
Distraction	Telephone Call	Using a smartphone for making a call.
	Using Phone	Using a smartphone for texting or playing games.
	Looking Elsewhere	Looking at different directions rather than looking towards the screen of the laptop.

For each action stated in table 1, a 10 second video was recorded depicting each student performing the action. For each student, two set of videos have been recorded (Session A and Session B), in which the students wear different clothes while the background was different for each session and student. An example of typical images depicting different actions considered are shown in Figure 2. Currently, the dataset contains 194 videos with a total of 9700 image frames, where Session A has 5200 frames and Session B 4500 frames. We are still in the process of collecting additional data, and once the data collection is completed, we plan to make the dataset publicly available.

Figure 2: Typical students' images regarding the seven actions.(a) Telephone Call (b) Attending (c) Hand Raising (d) Looking Elsewhere (e) Writing (f) Using Phone (g) Absent.

3.3 Action Recognition

Action recognition is performed using deep networks. We trained and evaluated the performance of several deep network architectures that include a GoogleNet (Szegedy et al., 2015), SqueezeNet (Iandola et al., 2016), Inception-v3 (Szegedy et al., 2016) and faster R-CNN (Ren et al., 2015).

To solve the problem of limited data, we used data augmentation to augment images in the training dataset. In the present work, augmentation algo-

rithm included reflection, rotation, scale and translation of all training images. During the training procedure, transfer learning was adopted where starting from a pre-trained version of the networks trained using images from the ImageNet database (www.image-net.org), the network weights of certain layers are adapted based on the train data and the classification task in question. To train the networks, we used the Adam optimizer (Kingma and Ba, 2014) with a momentum value of 0.9 and weight decay of 0.0001 to minimize a cross-entropy loss function. In the training options, we set the learning rate 0.0001, defined the size of batch 11 while training involved 15 epochs.

4 EXPERIMENTAL INVESTIGATION

From the initial stages of the experimental evaluation, it was evident that the faster R-CNN architecture was the least appropriate network for the proposed application framework, due to the computational load required during the classification stage, that would prevent the real time operation of the system on student's machines. Therefore, the experimental evaluation was focused on the comparative evaluation of the GoogleNet, SqueezeNet and Inception-v3 architectures.

Based on the taxonomy in Table 1, results are reported both for the seven-class classification problem, where the exact student action is recognized, and for the three-class classification problem where student's actions are classified into the states of "Physical Presence", "Active Participation" and "Distraction". Two experiments were performed :

4.1 Experiment 1: Using New Instances of Previously Enrolled Students

For this experiment image frames from Session A were used for training the system and image frames from Session B (see section 3.2) for testing the system, allowing in that way the assessment of the ability of the proposed system to recognize actions in new videos of students who provided training data.

The results obtained from the proposed CNN architectures for the performance evaluation of the seven-class classification problem are presented in Table 2. The errors of the GoogleNet, SqueezeNet and Inception-v3 are 0.022, 0.059 and 0.374 respectively.

GoogleNet and SqueezeNet are proved as the most suitable networks in the classification of the students' actions in Experiment 1, in contrast to Inception-v3. However, the SqueezeNet needs less computational time and memory requirements in comparison with the other networks. The low performance measures in Inception-v3 are due to the fact that more data are required for training.

Table 2: Accuracy metrics for experiment 1.

Performance Measures	GoogleNet	SqueezeNet	Inception-v3
Accuracy(%)	97.81	94.15	62.61
Recall	0.964	0.910	0.601
Precision	0.964	0.926	0.510
F1-score	0.963	0.912	0.542
MCC*	0.960	0.907	0.479
Cohen's k	0.910	0.761	0.345
Classification time (s)	0.089	0.039	0.258
Memory requirements (MB)	21.3	2.60	77.1

**MCC=Matthews Correlation Coefficient*

Based on the results it is evident that the algorithm can classify students' actions almost with perfect accuracy for GoogleNet. From the confusion matrix for the seven-class problem, all test images were recognized correctly in class 1 (Absent), class 4 (Looking Elsewhere) and class 6 (Using phone), while for class 2 (Attending), class 3 (Hand Raising), class 5 (Telephone Call) and class 7 (Writing) the accuracy is over 90%. Figure 3 represents the ROC curve for each class for the GoogleNet.

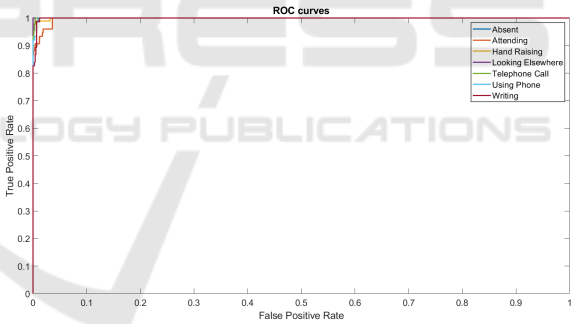

Figure 3: ROC curves for GoogleNet for the seven-class classification problem for experiment 1.

The impact of the action recognition system is negative when networks misclassify images of active participation as distraction with high error rates. Based on the confusion matrix for the seven-class problem, the networks GoogleNet, SqueezeNet, and Inception-v3 mostly misclassify the class "Writing" as "Using Phone", with percentages of 0.6%, 1.8%, and 2.7%, respectively. Furthermore, SqueezeNet misclassifies 1.8% of the "Attending" images as "Using Phone". Overall, the errors rates of the algorithms are minimal. However, more training data and network tuning are required to minimize the errors that could give to the teacher a wrong impression about a student's behaviour.

By aggregating the seven actions into three more general categories the GoogleNet, SqueezeNet and Inception-v3 acheive accuracies of 98.66%, 95.74%, and 73.45% respectively. Figure 4, illustrates the confusion matrix in the GoogleNet network which classifies the three general categories related to student status. The class related to student physical presence is perfectly classified while the active participation and distraction categories have an accuracy of 96.3% and 99.6% respectively. Based on the results, we conclude that the GoogleNet can recognize with increased accuracy the behavioural disaffection instead of behavioural participation of primary school students', which is important for educators to know during teleconferencing.

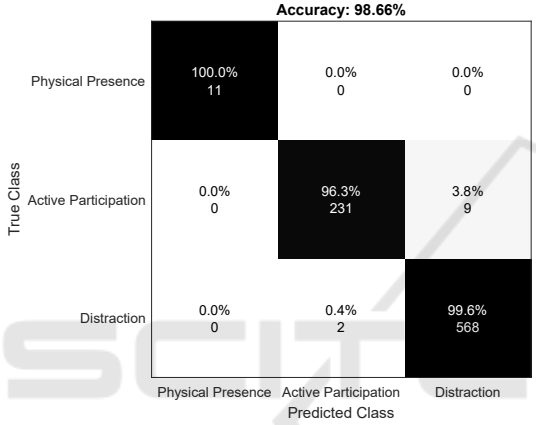

Figure 4: Confusion matrix for GoogleNet for the three-class classification problem for experiment 1.

4.2 Experiment 2: Recognizing Actions of Previously Unseen Students

All images of eight students (from both session A and Session B) were used for training the system and images of the remaining seven students were used for testing the system, so that recognition performance is tested on previously unseen students.

Table 3 presents the performance measures obtained from the suggested CNN architectures for the seven-class classification problem. For Experiment 2, the errors of GoogleNet, SqueezeNet, and Inception-v3 are 0.057, 0.127 and 0.551, respectively.

We observe from Table 3 that in Experiment 2, GoogleNet and SqueezeNet have better performance than Inception-v3 which requires more data in order to be efficient. The computational time for SqueezeNet network is infinitesimal and half the time required for GoogleNet.

GoogleNet achieved the highest correct recognition rates among all networks considered. From the confusion matrix for GoogleNet for the seven-class

Table 3: Accuracy metrics for experiment 2.

Performance Measures	Googlenet	Squeezenet	Inception-v3
Accuracy(%)	94.32	87.32	44.91
Recall	0.913	0.820	0.540
Precision	0.939	0.812	0.369
F1-score	0.911	0.804	0.361
MCC*	0.911	0.790	0.325
Cohen's k	0.768	0.482	0.555
Classification time (s)	0.082	0.038	0.234
Memory requirements (MB)	21.3	2.60	77.1

**MCC=Matthews Correlation Coefficient*

problem the images were recognized correctly in class 1 (Absent), class 4 (Looking Elsewhere), class 5 (Telephone Call) and class 6 (Using Phone). For the remaining classes, class 3 (Hand Raising), class 2 (Attending) and class 7 (Writing) have accuracy 98.8%, 78.6% and 61.4% respectively. Figure 5 represents the ROC curve for Experiment 2 for each class for the GoogleNet.

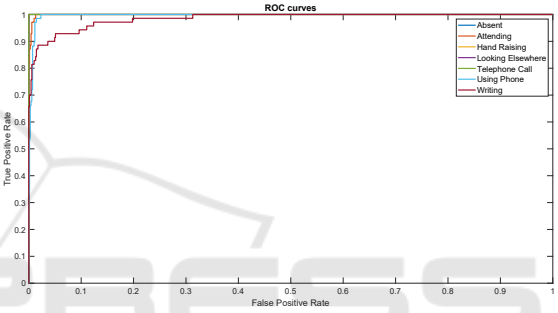

Figure 5: ROC curves for GoogleNet for the seven-class classification problem for experiment 2.

For the three-class classification problem the GoogleNet, SqueezeNet and Inception-v3 achieve correct recognition rates of 94.32%, 92.46% and 65.13% respectively. GoogleNet proved to be the network with the highest classification accuracy, like in the case of Experiment 1. Based on the confusion matrix (Figure 6) of GoogleNet, physical presence and distraction are perfectly classified while the active participation category has an accuracy of 80.8%. Overall, the classification rates of GoogleNet and SqueezeNet are satisfactory, despite the fact that we deal with images of previously unseen students.

Our results suggest that the GoogleNet mostly misclassifies the classes "Writing" and "Attending" with phone usage by 3.4% and 1.9%, respectively. We observe that in both Experiments, GoogleNet confuses the aforementioned classes with minimal errors. Otherwise, SqueezeNet mostly overestimates the phone usage of students during active participation by 3.4%. On the other hand, Inception-v3 overestimates phone usage and class "Looking elsewhere" during active participation by 6.4% and 13%, respectively. Error percentages for the three networks are higher in Experiment 2 than in Experiment 1, since

the size of the train set in Experiment 1 is larger than the size in Experiments 2. In the future, in order to minimize the error rates, the overall experimental evaluation will be implemented using several recorded videos of online lessons of larger duration.

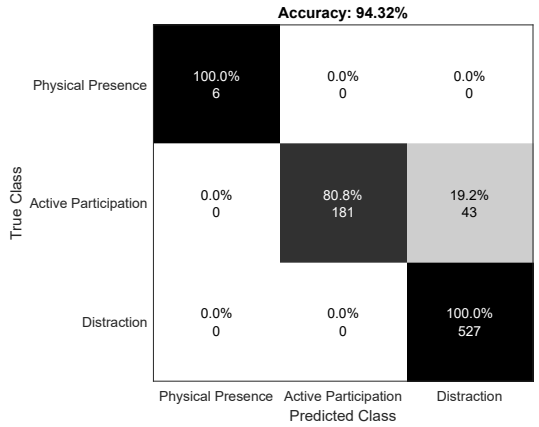

Figure 6: Confusion matrix for GoogleNet for the three-class classification problem for experiment 2.

5 CONCLUSION AND FUTURE WORK

In our study, we developed a deep network model which can provide to educators real-time video analysis of young students, providing in that way information about the behavioral participation and behavioral disaffection of their students during the delivery of online courses.

Initial results demonstrate high classification rates for the three class and seven class classification problems, both when dealing with new images of previously enrolled students or when dealing with previously unseen students. The results indicate the feasibility of this approach that can be used as the basis for implementing an integrated system that helps educators to monitor actions of participants during remote class delivery. An important aspect of the system is the fact that the privacy of students is protected since teachers receive only real-time information about students' actions rather than receiving images of the students. To eliminate any additional ethical considerations regarding student privacy, in the final system students and parents will be thoroughly informed about the operation of the action recognition system, and they will be required to approve its use during the delivery of on-line courses.

Since in the proposed system, the action recognition process will take place locally, on the device of each student, it is imperative to use classification methods that require minimal computational load. In the final system implementation, students who possess devices with low computational capabilities will be using the SqueezeNet network for classification, as in that case the computational and memory requirements are minimized, allowing real-time operation, and a reasonable recognition accuracy. Otherwise, if the students have a powerful device, then it is more appropriate to use the GoogleNet network so that we take advantage of the increased recognition accuracy achieved by this architecture.

In future work, we aim to explore different network architectures, extend the dataset to include more students and additional actions, perform an exhaustive user evaluation with all stakeholders, and investigate practical issues of integrating the proposed action recognition system in tele-education teaching sessions.

ACKNOWLEDGEMENTS

We would like to thank all volunteers who participated in the interviews and all student who provided training data. This project was partially funded by the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement No 739578 and the Government of the Republic of Cyprus through the Directorate General for European Programmes, Coordination and Development.

REFERENCES

Ashwin, T. and Guddeti, R. M. R. (2020). Impact of inquiry interventions on students in e-learning and classroom environments using affective computing framework. *User Modeling and User-Adapted Interaction*, 30(5):759–801.

Baron, R. S. (1986). Distraction-conflict theory: Progress and problems. *Advances in experimental social psychology*, 19:1–40.

Barriball, K. L. and While, A. (1994). Collecting data using a semi-structured interview: a discussion paper. *Journal of Advanced Nursing-Institutional Subscription*, 19(2):328–335.

Bebis, G., Egbert, D., and Shah, M. (2003). Review of computer vision education. *IEEE Transactions on Education*, 46(1):2–21.

Bian, C., Zhang, Y., Yang, F., Bi, W., and Lu, W. (2019). Spontaneous facial expression database for academic emotion inference in online learning. *IET Computer Vision*, 13(3):329–337.

Chintalapati, S. and Raghunadh, M. (2013). Automated attendance management system based on face recognition algorithms. In *2013 IEEE International Confer-*

ence on Computational Intelligence and Computing Research, pages 1–5. IEEE.

Chowdhury, S., Nath, S., Dey, A., and Das, A. (2020). Development of an automatic class attendance system using cnn-based face recognition. In *2020 Emerging Technology in Computing, Communication and Electronics (ETCCE)*, pages 1–5. IEEE.

Dimitriadou, E. and Lanitis, A. (2021). Student action recognition as a means of supporting remote learning. In *ICERI2021 Proceedings*, 14th annual International Conference of Education, Research and Innovation, pages 8175–8180. IATED.

Gupta, S. K., Ashwin, T., and Guddeti, R. M. R. (2018). Cvucams: Computer vision based unobtrusive classroom attendance management system. In *2018 IEEE 18th International Conference on Advanced Learning Technologies (ICALT)*, pages 101–102. IEEE.

Iandola, F. N., Han, S., Moskewicz, M. W., Ashraf, K., Dally, W. J., and Keutzer, K. (2016). Squeezenet: Alexnet-level accuracy with 50x fewer parameters and¡ 0.5 mb model size. *arXiv preprint arXiv:1602.07360*.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Li, X., Wang, M., Zeng, W., and Lu, W. (2019). A students' action recognition database in smart classroom. In *2019 14th International Conference on Computer Science & Education (ICCSE)*, pages 523–527. IEEE.

Mery, D., Mackenney, I., and Villalobos, E. (2019). Student attendance system in crowded classrooms using a smartphone camera. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 857–866. IEEE.

Moubayed, A., Injadat, M., Shami, A., and Lutfiyya, H. (2020). Student engagement level in an e-learning environment: Clustering using k-means. *American Journal of Distance Education*, 34(2):137–156.

Ngoc Anh, B., Tung Son, N., Truong Lam, P., Le Chi, P., Huu Tuan, N., Cong Dat, N., Huu Trung, N., Umar Aftab, M., Van Dinh, T., et al. (2019). A computer-vision based application for student behavior monitoring in classroom. *Applied Sciences*, 9(22):4729.

Pratton, J. and Hales, L. W. (1986). The effects of active participation on student learning. *The Journal of Educational Research*, 79(4):210–215.

Putri, R. S., Purwanto, A., Pramono, R., Asbari, M., Wijayanti, L. M., and Hyun, C. C. (2020). Impact of the covid-19 pandemic on online home learning: An explorative study of primary schools in indonesia. *International Journal of Advanced Science and Technology*, 29(5):4809–4818.

Raddon, A. (2006). Absence as opportunity: Learning outside the institutional space and time. *Journal of Further and Higher Education*, 30(02):157–167.

Radlak, K., Frackiewicz, M., Szczepanski, M., Kawulok, M., and Czardybon, M. (2015). Adaptive vision studio—educational tool for image processing learning. In *2015 IEEE Frontiers in Education Conference (FIE)*, pages 1–8. IEEE.

Rashmi, M., Ashwin, T., and Guddeti, R. M. R. (2021). Surveillance video analysis for student action recognition and localization inside computer laboratories of a smart campus. *Multimedia Tools and Applications*, 80(2):2907–2929.

Ren, S., He, K., Girshick, R., and Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28:91–99.

Skinner, E., Furrer, C., Marchand, G., and Kindermann, T. (2008). Engagement and disaffection in the classroom: Part of a larger motivational dynamic? *Journal of educational psychology*, 100(4):765.

Skinner, E. A., Kindermann, T. A., and Furrer, C. J. (2009). A motivational perspective on engagement and disaffection: Conceptualization and assessment of children's behavioral and emotional participation in academic activities in the classroom. *Educational and psychological measurement*, 69(3):493–525.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826.

Teeparthi, S., Jatla, V., Pattichis, M., Pattichis, S.-C., and Lopez Leiva, C. (2021). Fast hand detection in collaborative learning environments? *International Conference on Computer Analysis of Images and Patterns.*

Thomas, C. and Jayagopi, D. B. (2017). Predicting student engagement in classrooms using facial behavioral cues. In *Proceedings of the 1st ACM SIGCHI international workshop on multimodal interaction for education*, pages 33–40.

Tran, P., Pattichis, M., Celedon-Pattichis, S., and Lopez Leiva, C. (2021). Facial recognition in collaborative learning videos. *International Conference on Computer Analysis of Images and Patterns.*

Yang, S. and Chen, L. (2011). A face and eye detection based feedback system for smart classroom. In *Proceedings of 2011 International Conference on Electronic & Mechanical Engineering and Information Technology*, volume 2, pages 571–574. IEEE.