Multimedia practice session - proposal (basic details)

1a. Project title: Engagement Detection in Online Teaching

1b. Project acronym: EDOT

1c. Applicants

Name: Fabio Fiori
E-mail: fabio.fiori@studenti.unitn.it
Student#: 231494
Institute: University of Trento

Name: Paloma Dominguez Sanchez
E-mail: p.dominguezsanchezi@studenti.unitn.it
Student#:238985
Institute: University of Trento

Name: Taylor Evan Lucero
E-mail:Taylor.Luceroi@studenti.unitn.it
Student#: 239001
Institute: University of Trento

2a. Summary of research proposal (max 250 words)

Through suffering the impact of a global pandemic we have pushed academia to utilize a new method of teaching online. However, without in person interaction teachers and students are suffering from low levels of engagement that can negatively impact a students level of education. Grasping the emotional understanding and feeling of students during these classes is imperative in making strides in online education.

Previous approaches identified the main emotions like Sadness, Happiness and Boredom and were able to use FERC Neural Network to achieve 96% accuracy of understanding the facial changes of a subject's face to match a specified emotion. Unfortunately, this is limited by its usage of only static images, any temporal changes will greatly decrease the accuracy of the model. Ensuring that the model overtime is able to accurately understand a subject's actions, we will pursue an ensemble between a FERC and LSTM model. This is due to the ability LSTM models have in learning and understanding a series of changes,that when learned, can help

classify student actions as well as OCR information that will be retrieved from the screen of the computer.
Through these student events, lectures and teachers will be able to identify when student engagement increases and decreases by looking at the engagement estimate, real-time feedback, and AI-enhanced post-lesson reviews.

2b. Abstract for laymen (max 250 words)

Our research goal is to create a new digital tool for engagement detection in online teaching. Currently, teachers can perceive students' struggle or disengagement just by observing their behavior or facial expressions in physical classrooms. However, detecting the engagement through face-to-face online communication is still missing.

We aim to create a digital tool that automatically estimates learners' engagement through computer vision-based techniques and provides teachers with real-time feedback and AI-enhanced post-lesson reviews.

Problems we have to deal with are:
- To recognize when a student is writing down and not looking at the camera, but he or she is attending the class
- when the student is distracted by external objects and he/she is not engaged.
- To provide to the professor a nice user-friendly interface that let him pivot on time the content without distraction

We propose to use an ensemble of FERC (Mehendale, 2020) and LSTM (Orozco, 2019) , as both have provided high quantitative results in facial emotion recognition and action recognition. We will use the classification categories of students as low-, normal- or high-engaged, or "off-screen" and apply them to the model.

Research proposal

4. Description of the proposed research

E-learning had a steep rise in popularity due to the flexibility and convenience it offered

educational agencies during the COVID-19 pandemic. Since 2020, more and more students have been opting to take courses online, which eventually has led to a proliferation of e-learning platforms and providers. The increased market competition resulted in a race for innovation and a growth in terms of software capabilities, with more platforms offering features such as live streaming, interactive content, and gamification.
For example, Zoom Video Communications Inc., one of the most prosperous and foremost leaders in internet teleconferencing, tried to meet the challenges posed by the upcoming digitalization of education by developing dedicated tools that mimic traditional classroom activities and dynamics. Whiteboarding, room polling, hand raising, breakout rooms and many other features were implemented over the past two years for the purpose of supporting remote lecturing.

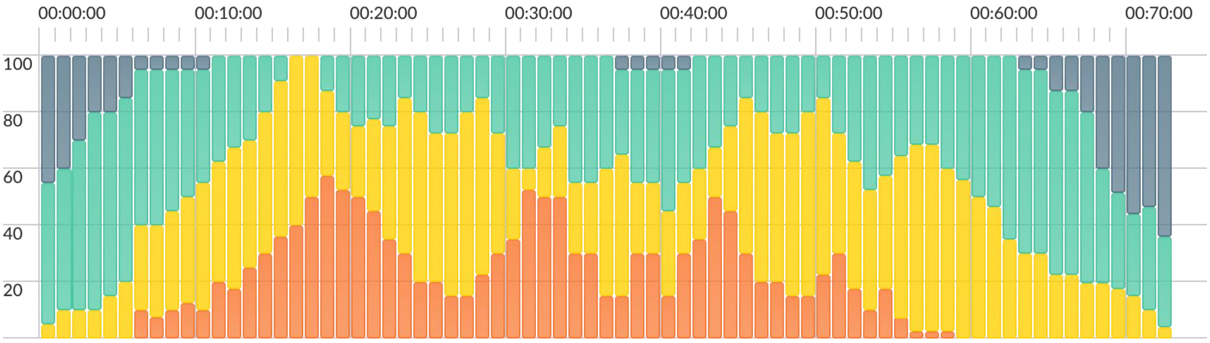
Despite the efforts of designers and engineers to simulate face-to-face communication, a crucial aspect of that experience is still missing. In physical classrooms, teachers can perceive students' struggle or disengagement just by observing their behavior or facial expressions. Such immediacy allows the teachers to quickly realize when communication has failed and possibly recover on the spot.
Unfortunately, this feature is not available in online environments, where students still have to express their engagement manually, resorting to verbal (e.g., written text) and nonverbal cues (e.g., emojis), at best, if solicited.

a. Research topic and envisaged results

We aim to create a digital tool that automatically estimates learners' engagement through computer vision-based techniques and provides teachers with real-time feedback and AI-enhanced post-lesson reviews.

To automate the detection of engagement we propose a deep learning approach, based on the analysis of users' facial expressions and behaviors. All necessary information is captured with PCs' built-in webcam and used as input for the algorithm to estimate the engagement of each learner in real-time. The algorithm can classify students as low-, normal- or high-engaged, or assign a fourth label named "off-screen" to take account of those who are not in front of the camera. Once all students are assigned with a label, the relative frequency of all categories are computed and then displayed on the teachers' screen in the shape of a stacked bar chart (on the right), and updated continuously. We expect this feature to work as a thermometer for engagement, helping teachers understand the overall engagement of their class at a single glance, without resorting to clumsy polling procedures.

At the end of the computation, the algorithm’s output doesn't get lost. The frequency of all categories at any given moment is stored and then used to provide post-lesson reviews. This feature would enable teachers to graphically visualize how the distribution of the four levels of engagement (3+1) varies throughout the duration of the lecture, and foresee possible disengagement patterns. A two-dimensional graph will be used for this purpose, plotting the time on the x-axis and the labels’ frequencies on the y-axis (below).



Since time may not be very informative itself, the teachers will be able to overlay the x-axis with the actual recording of the lesson, so as to identify connections between specific events and shifts in engagement. To support this activity, an optical character recognition system will be used on the slideshow record to extract information about the topics, and automatically highlight those that generated higher and lower engagement levels. We expect this feature to improve teachers’ self-assessment, by providing a better understanding of what takes hold of their students and what does not, both remotely and in person.

b. Approach

We propose to use an ensemble of FERC (Mehendale, 2020) and LSTM (Orozco, 2019) models, as both have provided high quantitative results in facial emotion recognition and action recognition. FERC is a two-part CNN, where the first section removes extraneous information, and the second concentrates on facial feature vector extraction. The secondary part of the ensemble would focus on the action recognition of the active user with an LSTM model. As action recognition mainly focuses on temporal data classification, the LSTM model can extract activity features and identify these actions with minimal parameters. Using Viola-Jones Face and Eye detection, we will identify Two ROIs of interest to monitor the students' face of engagement while learning. The collected software will, after set time intervals, take pictures of the located ROI using the laptop camera. These images will be assigned an engagement classification level consisting of low, normal, high, or off-screen. From here, by identifying the face, we can determine the movement of the location of the eyes. If the eyes are facing forward and looking at the screen, we can identify if the student is looking at the screen and engaged in some other activity. The model will also be able to put together specific action events that may take place on the screen, based on these specified ROIs that will be able to aid if a student is attentive to the less at hand. For instance, a student that is not engaged would be staring at an object away from the screen, have their head turned to interact with an individual or object next to them, or possibly, the individual left the view of the laptop camera. During any set online event, images will be taken of the screen by using a special program. This is for identification of an on-screen event that led to impact the level of engagement for subjects. Once the collected images are compiled and allocated into an appropriate category based on the level of engagement and action, they will pass through our ensemble model.

The face ROIs are standardized to a size of 96 × 96 pixels. After generating the LDP map and dividing the face image into blocks, a histogram of 32 bins would be created from each block. After concatenating the histogram bins for all the blocks, a histogram of LDP with $(32\times9\times3+32\times18)=1440$ dimension is extracted from each face ROI and reduced into 200 using PCA and KPCA projection.

We will use the classification categories of students as low-, normal- or high-engaged, or "off-screen" and apply them to the model. For the two-level and three-level engagement detection models, the output layer would consist of 4 nodes, respectively, indicating the training of the 4 learning activities. During training, once the weights of the first RBM are trained, h1 is fixed. Then the weights of the second RBM are trained using h2. The third RBM is trained using the weights in the previous RBM. Finally, the output of the third RBM is used as input for the BP network for supervised learning.

In the experiment, we would randomize the dataset and then apply k-fold cross-validation with 70% face images for the training and 30% face images for the classifier testing in each round.

After the model is trained, we will attempt to replicate Dewan et al. (2018)’s experiment(Gupta, 2016) using the DAiSEE dataset, compiled to gauge engagement in online courses. This dataset includes 112 individuals — 32 females and 80 males. A total of 9068 video snippets are collected, each approximately 10 seconds long. The dataset was collected in an unconstrained environment, such as dorm rooms, crowded lab spaces, and libraries, with three different illumination settings — light, dark, and neutral. The dataset is labeled four differing affective states - engaged, confused, frustrated, and bored. The following labeling scheme was used to help identify the level of engagement students felt while retaining the student's emotional state portrayed in the image. As this is a multi-classification problem, the engaged face images with a neutral intensity that portray little to no emotion are labeled as normally engaged. The engaged face images with intensity values higher or positive emotions greater than two are given the label highly engaged. The face images with identifiable negative emotions are given the label not-engaged.

From this, we would further implement the network on the authentic expression database(Li, X, 2020)and the Geneva Multimodal Expression database (Bänziger et al, 2012). The Authentic expression database would focus on how the network is able to understand and classify non-posed facial emotions as learners' interactions and contents in the wild are made subconsciously depending on the scenario at hand. This database monitored viewers' faces while they watched various display segments from then-recent movie trailers. After the images were collected, subjects were interviewed to determine their emotional states. The third database would be the Geneva Multimodal Expression Database. This set is a collection of audio and video recordings featuring ten actors portraying 18 affective states with different verbal contents and modes of expression. Not only does this database have the most significant number of affective states it also determines the intensity, plausibility, and authenticity of the emotion portrayed by the actors.

c. Scientific or economic relevance

Our system would boost research in the educational field and it consists of a combination of different techniques with a low cost camera, to make video identification possible in order to recognise engagement through facial expressions. The results of this research could be used by teachers to improve their lessons both in presence and remotely. The goal of this paper is to demonstrate how a low-cost video monitoring system can be used to verify a student's performance in presence or remotely.

Traditional methods of education as we know it are becoming a thing of the past. They are becoming increasingly digitized, and being driven by technology innovations. In fact, the so-called EdTech, the education technology industry, is expected to reach $680.1 billion by 2027, growing at an annual rate of 17.9%.

Edtech startups across the globe are riding a wave of investor optimism to push forward advances in just about every corner of education. Globally, startups in the education space raised a record $13.3B in venture investment—up nearly 50% from the previous high mark—while U.S.-based startups took in a record $2.2B in funding.

However, how do these technologies affect student’s engagement? That’s what we are researching for.

In order to identify and localize humans from video, a variety of methods have been proposed in the literature. For example, Dewan et al. (2018) trained their model on dataset called Affective States in E-Environments (Gupta, 2016) to demonstrate the effectiveness of the proposed method, where the two-level engagement identification obtains a better accuracy (90.89%) than the three-level engagement detection (87.25%), demonstrating the usefulness of the suggested technique. Within mobile settings, human activity recognition from videos, or more specifically, human activity classification, is a popular research topic. A typical approach is to segment the video into blocks and analyze each block individually. Each block is then assigned a label.

In our research we presented a camera management system that consists of a low cost camera to capture images in the classroom, a PC with a camera driver software, a user-friendly graphical interface, and a database. This paper aims to present the camera system components that have been improved and optimized for efficiency, to make the system as cost-effective as possible.

From an economical point of view the research could be very interesting. There are many public and private schools, online teaching platforms and governments all over the world that would very much be interested in such a system when it is fully operational. This means we have a whole education ecosystem as our potential customers. In that sense the research does not only include academic interests, but also the more economic aspects, where we could imagine some real uses of this system.

This could eventually open up many doors for us and lead us to a much more exciting ride, and we hope it will. However, as stated before, a huge challenge is to create a system that is not just a copy but a whole new approach.

d. International development

To gauge how engaged online learners are, online learning environments can make use of sensing technology and affective computing approaches. In this sense, researchers go in two directions: Create tools that increase engagement in real time or measure engagement and provide the results to take the pertinent measures.

Dewan et al. (2018) provided a deep learning-based method for assessing the level of involvement of online students by observing their facial expressions. They employ Kernel Principal Component Analysis (KPCA) to capture the nonlinear correlations among the collected features and Local Directional Pattern (LDP) to obtain person-independent edge features for the various face expressions. The findings of the experiment demonstrate that the suggested strategy is highly accurate in classifying the various degrees of involvement that students may exhibit when engaging in online learning activities.

Other researchers, such as Karimah & Hasegawa (2022), presented a literature review of current advancements in engagement definitions, datasets, and machine learning-based estimate approaches for automated engagement. To answer the RQs and explain engagement definitions, datasets, and techniques, 47 papers were chosen.

Finally there are other investigations focused on specific areas such as language learning. Xu et al. (2022) looked at an automated system driven by artificial intelligence (AI) that employs voice and face recognition to track interactions, facial expressions, and speech in real-time between teachers and students. The outcomes showed that in this one-to-one online learning environment, young students were very engaged. English proficiency levels and learner frontal face exposure, which shows how attentive they are in class, are important and beneficial determinants of learner engagement. The amount of time teachers spent speaking overall and during teaching appeared to be significant indicators of student engagement.

Equipment

6. Expected use of instrumentation

Cameras - *Logitech C920* **cost 55 euros**

GPU - *RTX 3080* **704 euros**

Reasons: CNN learning needs a minimum of 10 GB

DataBase - *AWS S3*

The main differences between HDFS and AWS S3 are: #1: S3 is more scalable than HDFS. Difference #2: When it comes to durability, S3 has the edge over HDFS. Difference #3: Data in S3 is always persistent, unlike data in HDFS. Amazon S3 pricing ; S3 Standard - General purpose storage for any type of data, typically used for frequently accessed data ; First 50 TB / Month, $0.023 per GB. **- 20Gb a month * .023 (36) = 16.56**

Software - *OpenCV*

OpenCV: Open Source Computer Vision Library. OpenCV was designed for computational efficiency and with a strong focus on real-time applications. Written in optimized C/C++, the library can take advantage of multi-core processing. Enabled with OpenCL, it can take advantage of the hardware acceleration of the underlying heterogeneous computing platforms.

(Open Source software)

Computers- CYBERPOWERPC

Best PC under $ 1k. Has Intel processors, suitable RAM size, fair expandability, and RTX GPUs.

750 euros

Specs:

Processor: Intel Core i5–11400F up to 4.5 GHz.

Memory: 8 GB DDR4.

Hard Drives: 500 GB NVMe SSD.

GPU: NVIDIA GeForce RTX 2060 6 GB.

Computing Power: 7.5

Ports: 1x HDMI 2.0, 1x USB 3.1 Type-C, 2x USB 3.1, 1x USB 2.0.

OS: Windows 11 Home.

Connectivity: WiFi 802.11ax, Gigabit LAN (Ethernet), Bluetooth.

7. Work programme

Goal	Y1 Q1	Y1 Q2	Y1 Q3	Y1 Q4	Y2 Q1	Y2 Q2	Y2 Q3	Y2 Q4	Y3 Q1	Y3 Q2	Y3 Q3	Y3 Q4
Planning the user research	🟦											
Partnership with 2-3 public / private schools		🟦										
Test Lab Construction		🟦	🟦									
Qualitative Data collection			🟦	🟦								
Milestone: Qualitate Data Collected					■							
Data Cleaning					🟦	🟦						
Data Analysis						🟦	🟦					
Starting model development							🟦	🟦				
Software Development								🟦	🟦			
Milestone:MVP									■			
Planning testing research										🟦		
Partnership with public and private schools										🟦	🟦	
Testing Development										🟦	🟦	🟦
New Model Training											🟦	🟦
Milestone: First operative model												■

8. Costs

Budget Items			**Amount (€)**
Salary	x2 Researchers	paid monthly, for 3 year	201,600.00
	x1 Front-end developer	paid per project	25,000.00
	x1 Back-end developer	paid per project	28,000.00
		Sub-total	254,000.00
Camera Logitech C920			55.00
GPU RTX 3080			720.00
CyberPowerPC			750.00
Database AWS 3			16.56.00
Unexpected costs		20% of the total amount	51,228.31
		Total	307,369.87

References

Bänziger, T., Mortillaro, M., & Scherer, K.R. (2012). Introducing the Geneva Multimodal Expression corpus for experimental research on emotion perception. *Emotion, 12*(5), 1161-1179.

https://www.unige.ch/cisa/gemep/coreset/

Dewan, M. A. A., Lin, F., Wen, D., Murshed, M., & Uddin, Z. (2018). A Deep Learning Approach to Detecting Engagement of Online Learners. *2018 IEEE SmartWorld, Ubiquitous Intelligence &Amp; Computing, Advanced &Amp; Trusted Computing, Scalable Computing &Amp; Communications, Cloud &Amp; Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/ SCALCOM/ UIC/ ATC/ CBDCom/ IOP/ SCI)*. https://doi.org/10.1109/smartworld.2018.00318

Dewan, A. A. M., Murshed, M., Lin, F. (2019). *Engagement detection in online learning: a review - Smart Learning Environments*. SpringerOpen. https://slejournal.springeropen.com/articles/10.1186/s40561-018-0080-z

Drobac, S., & Lindén, K. (2020). Optical character recognition with neural networks and post-correction with finite state methods. *International Journal on Document Analysis and Recognition (IJDAR), 23*, 279 - 295.

https://link.springer.com/article/10.1007/s10032-020-00359-9

Gupta, A., Jaiswal, R., Adhikari, S., & Balasubramanian, V.N. (2016). DAISEE: Dataset for Affective States in E-Learning Environments. *ArXiv, abs/1609.01885*

https://www.semanticscholar.org/paper/DAISEE%3A-Dataset-for-Affective-States-in-E-Learning-Gupta-Jaiswal/6fdc0bc13f2517061eaa1364dcf853f36e1ea5ae

Karimah, S. N., & Hasegawa, S. (2022). Automatic engagement estimation in smart education/learning settings: a systematic review of engagement definitions, datasets, and methods. *Smart Learning Environments*, *9*(1). https://doi.org/10.1186/s40561-022-00212-y

Li, X., Zhang, X., Yang, H., Duan, W., Dai, W., & Yin, L. (2020). An EEG-Based Multi-Modal Emotion Database with Both Posed and Authentic Facial Actions for. Emotion Analysis. *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, 336-343. https://ieeexplore.ieee.org/abstract/document/9320173?casa_token=-Cg_Ax-NfNwAAAAA:vdou6tSKkAZQ5Ks44m3AQntvRIsqBY2o756Dpv6i4rbzx0NcCR9ijodTnjJqYE2m8bYg4wi2fA

Mehendale, N.D. (2020). Facial emotion recognition using convolutional neural networks (FERC). SN Applied Sciences, 2, 1-8.

https://www.semanticscholar.org/paper/Facial-emotion-recognition-using-convolutional-Mehendale/910706991687078e4314e3b06f4da64eae88f477

Orozco, C.I., Buemi, M.E., & Berlles, J.J. (2019). CNN-LSTM Architecture for Action Recognition in Videos. https://www.semanticscholar.org/paper/CNN-LSTM-Architecture-for-Action-Recognition-in-Orozco-Buemi/df8beecc6c0d16e9b75675c46b99aee80aaa83d5

Xu, X., Dugdale, D. M., Wei, X., & Mi, W. (2022). Leveraging Artificial Intelligence to Predict Young Learner Online Learning Engagement. *American Journal of Distance Education*, 1–14.

https://doi.org/10.1080/08923647.2022.2044663

Yu, Z., & Yan, W.Q. (2020). Human Action Recognition Using Deep Learning Methods. *2020 35th International Conference on Image and Vision Computing New Zealand (IVCNZ)*, 1-6.

https://openrepository.aut.ac.nz/bitstream/handle/10292/14076/YuZ.pdf?sequence=3&isAllowed=y