Focus estimation in academic environments using Computer Vision*

Daniel Canedo, Alina Trifan, and António J. R. Neves

IEETA/DETI
University of Aveiro
3810-193 Aveiro, Portugal
{danielduartecanedo, alina.trifan, an}@ua.pt

Abstract. In this paper we propose a system capable of monitoring students' focus through cameras and using Computer Vision algorithms. Experimental results show that our system is capable of identifying students and tracking their focus during a class. At the end of the class, the system outputs graphical feedback to teachers regarding the average level of students' focus. Moreover, it can identify lecture periods in which students were less watchful and the corresponding topics that potentially need extra focus. In this paper we start by presenting the architecture of the system, followed by results obtained both during a small-group workshop and a classroom with a large number of attending students. The main goal of this work is to contribute to the transformation of the classroom as a sensing environment, providing information to both teachers and students about their engagement during the class.

Keywords: Class Monitoring · Face Detection · Face Recognition · Face Tracking · Focus Estimation.

1 Introduction

Student engagement is linked positively to desirable learning outcomes, such as critical thinking and grades obtained in a subject [1]. The student engagement and attention depend on several factors, being the teacher one of the most important [2]. Teachers' ability to connect well with students can be beneficial for students' attention. In this paper, we pretend to estimate the students' attention based on Computer Vision algorithms. With our approach, we can suggest that certain student is, in fact, looking to the teacher, to the board or to the projection.

We realize that estimating students' attention is an hard task and cannot be performed only relying on visual data. However, that visual data plays an essential role on determining the students' focus and eventually their behaviour, which then can be correlated with other kind of data in order to get a more

* Supported by the Integrated Programme of SR&TD SOCA (Ref. CENTRO-01-0145-FEDER-000010), co-funded by Centro 2020 program, Portugal 2020, European Union, through the European Regional Development Fund.

accurate estimation of their attention. For instance, that additional data could be a small quiz at the end of the class regarding the studied subjects [3]. However, if we only rely on visual data, we can still obtain satisfactory results about focus of a student. As the study in [4] showed, head orientation contributes 68.9% in the overall gaze direction and the authors achieved 88.7% accuracy at determining the focus. This conclusion implies that head orientation is a powerful method of measuring the students' focus.

In a preliminary work [5] we proposed a system that would in theory be capable of monitoring classrooms. This theoretical proposal was based on the analysis of relevant state of the art techniques in order to find out which methods were the most suitable for each of the blocks of the proposed architecture. The contribution of the current paper is the proof that the theoretical system is in practice capable of monitoring students' focus through cameras and Computer Vision algorithms.

This paper is organized as follows. Section 2 presents the methodology followed to implement our system. On Section 3 we present the System Architecture and the developed algorithms. Section 4 provides experimental results. Concluding remarks and the future work are featured in Section 5.

2 Methodology

The architecture of our system is made up of several blocks, from image acquisition during a classroom to the estimation of the students' focus.

2.1 Face Detection

As a first step of our methodology, we need to acquire the regions of interest in the classroom. Those regions of interest are the students' faces, which can be obtained through Face Detection algorithms. This is useful for extracting facial features which are used to estimate the head pose of the students and to identify them throughout the class. We must considerate the working distance in our scenario. Since the cameras need to capture the whole classroom, they need to be placed far away from the students, which leads to low resolutions images as input of the system. Therefore we must build a Face Detector that is efficient in long-range environments.

After the review of state of the art Face Detectors, we proposed in [5] one that fulfills the requirements of our system: MTCNN (Multi-task cascade convolution neural network) [6]. Depending on the used thresholds, this Face Detector is capable of detecting considerably low resolution faces in an image, at the cost of computer performance. Therefore, for different classroom sizes and different cameras, we can adjust the thresholds accordingly in order to detect every single student with a real-time performance. However, achieving the best threshold values for each situation is not straightforward, which may lead to false positives. Our system bypasses this problem by relying on a filter: it only assumes that a detected face is a real face if it detects the same face consecutively over 10 frames.

2.2 Identification and Tracking

Despite the low resolution faces, we must assure a good identification accuracy. This is for the sake of assigning the estimated focus levels to the respective students. It is highly undesirable to assign focus levels to the wrong students, since this could lead to the output of wrong data to both teachers and students regarding their performance and engagement in the class.

For the identification process, we propose a dynamic identification approach. In this approach, the Database is automatically filled during the class, while detecting the students. Through a face tracking algorithm, the system assigns an unique ID to each new detected student and stores several facial features during the class, based on which the Database is built-up. The main reason of implementing a tracking algorithm in our system is to bypass the computational cost of trying to recognize the students in every single frame. However, if for some reason the tracking is lost, the system recovers the students' identification through the Face Recognition, using the facial features stored in the Database. In order to increase the Face Recognition accuracy and avoid having identical identifications in the same frame, our system only compares a detected student with the ones who are not currently in the scene.

The Face Recognition implemented in our system is FaceNet [7] which uses a deep convolutional network and presents a 99,65% accuracy based on the Labeled Faces in the Wild Dataset (LFW [8]).

2.3 Focus Estimation

One way of measuring the students' focus that immediately comes to mind is by analyzing their eyes. However, as [9] mentioned, the accuracy of techniques like Eye Tracking tend to suffer when used on low resolution images.

Nevertheless, as already mentioned, we can rely on students' head pose to estimate their focus. To obtain the head pose we first need to acquire the five facial landmarks outputted from the MTCNN face detector: one in each eye, one in the nose and one in each mouth corner. These landmarks are usually stable, independently on the distance, so they can be trusted to estimate students' head pose. We created a sixth facial landmark in the chin, using the bounding boxes and the mouth landmarks' position, in order to improve the robustness of the head pose.

Lastly, we use an algorithm for head pose estimation [10] available on the OpenCV library [11] to estimate the head pose. This algorithm gives us a line that is drawn to the direction in which the head is oriented to. By comparing the direction of the line with a reference point that is marked by the teacher at the start of the class, it is possible to obtain a focus estimation of each student. This reference point dictates the direction in which the students should be looking at when they are focused.

3 System Architecture

The overview presented in the previous Section led to the design of the following system for monitoring classrooms. In Figure 1 we present the diagram of its architecture.

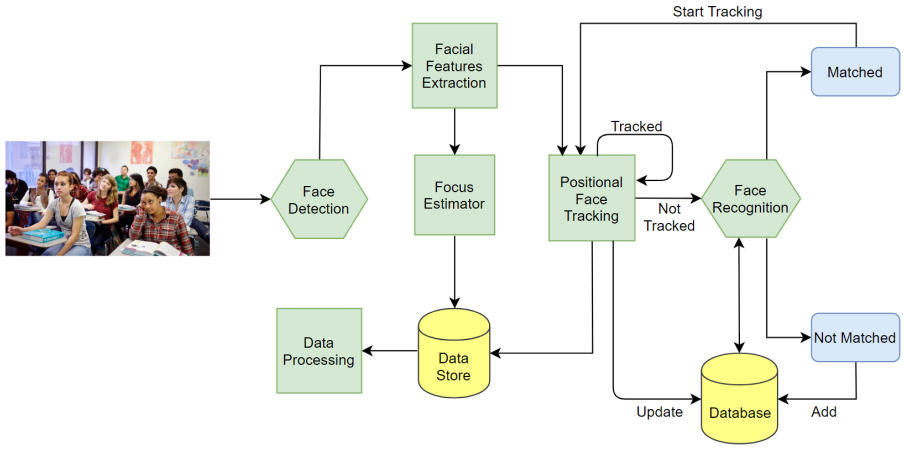

Fig. 1. Diagram of the system architecture, whose main tasks consist in Detection, Identification and Focus Estimation.

The system starts up with an input image captured by the camera. The camera should capture the whole classroom or part of it, depending on the students' acceptance of being recorded. This image is fed to our system, more specifically fed to a block called Face Detection. This block is responsible for detecting all the regions of interest in the image, which are the students' faces. After having these regions of interest, the system feeds them to the Facial Features Extraction block. This block is responsible for extracting the facial features of each student, which are used for the identification and for estimating their focus. Lastly, these facial features are fed to the Positional Face Tracking block and to the Focus Estimator block.

Since the system starts with no information from the past, it is implied that this block can't successfully track the students in the first frame. However, it assigns an unique ID to each detected student and saves their current positions for the next iterations. At the same time, the Database is updated, as it also starts up empty.

When the system already knows the position of each student, it can successfully track them. By comparing all the current positions of the bounding boxes retrieved from the Face Detection block with the positions of the bounding boxes from the previous frame, the system assigns the previous IDs to the closest students in the current frame. If the Positional Face Tracking fails to assign the

correct IDs for reasons such as occlusions, the system has a way to reassign the correct ID through a refresh algorithm. This algorithm refreshes all the students' IDs through the Face Recognition block each 9 frames, by comparing the actual facial features with the facial features that are assigned to the respective ID. For instance, if a student takes the ID of other student because of occlusions, the refresh algorithm is going to realize that he/she has the wrong ID because the student who "stole" the ID is only going to be compared with the student who got his/her ID "stolen". In this case, the refresh algorithm outputs a great distance value in the comparison, reassigning him/her the correct ID.

As mentioned in Subsection 2.2, since the Face Recognition block only compares the students with the ones who are not in the scene for a better accuracy, the student who gets his/her ID "stolen" by other one is not going to be compared with himself/herself, since his/her ID is still in the scene. In order to bypass this problem we set the counter for the refresh algorithm (9 frames) to be lower than the counter for the Face Detection block (10 frames). This will lead to a refresh before the system attributes a new ID to the student who lost his/her ID for assuming he/she is a new student in the scene. Therefore, in this unfavorable scenario, the system is able to reassign the correct ID to the "robber", and the "stolen" student has his/her own ID out of the scene and available for reassignment.

In a scenario where the Positional Face Tracking fails to track certain student, the system goes through the Face Recognition block in order to reassign the respective ID. If the student is correctly matched, the system starts tracking him/her again. If the student is not matched, it means he/she was not detected in the previous frame, therefore the system adds him/her to the Database and assigns a new ID.

While the students are being tracked, the system is constantly retrieving information from their facial features in order to estimate their focus. This estimation is calculated through the algorithm mentioned in Subsection 2.3. The focus values are stored in the Data Store and, at the end of the class, the Data Processing block is responsible for calculating an average focus for each individual ID. Afterwards, the system outputs this information through graphical feedback regarding the students' focus during the class, which ideally will be consulted by both teachers and students.

4 Experimental Results

In this Section we present the experimental results obtained while testing the developed system in real world scenarios. It is important to mention that, to our knowledge, currently there are no public datasets in academic environments available for testing.

4.1 Ground Truth

Initially, a small-group experiment was conducted with the aim of reaching a Ground Truth of our focus estimator. This served as well as a test to the system

as a whole and allowed us to identify potential flaws. Figure 2 shows how the test was conducted.

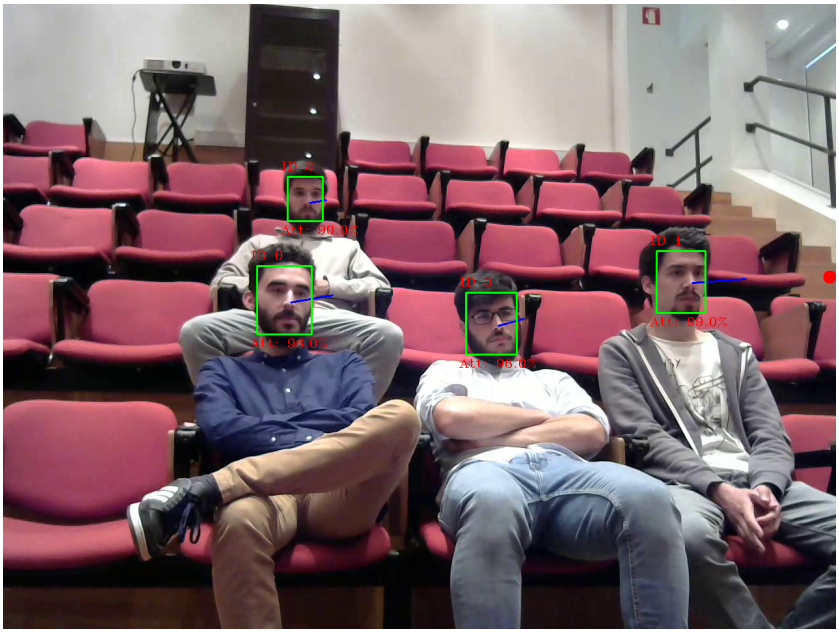

Fig. 2. Small-group experiment for reaching a Ground Truth of our focus estimator. On top of each bounding box the corresponding person ID is displayed. Below the bounding box the person's focus value is displayed.

This experiment was a small workshop, in which the presenter was substituted by a 10 minutes video. The circle presented in the right side of the Figure 2, as mentioned previously, is the reference point to where the participants should be looking at when they are focused.

The system outputted graphical feedback regarding the focus of each participant obtained through the Data Processing block. Afterwards, through manual annotation, it was noted down if each participant was focused or not for each frame of the experiment, over 6522 frames. The annotator marked "1" if the participant was looking to the reference point, and "0" if the participant was not looking to the reference point for each frame. This led to a slight discrepancy between the experimental graphics and the manual graphics. This is due to the fact that the system uses the head pose for calculating the direction of the gaze and the observer used the eyes to understand if the participants were focused or not. As an example of discrepancy, the head pose estimation may be directed towards the reference point, however the gaze is not. Figure 3 shows the obtained results.

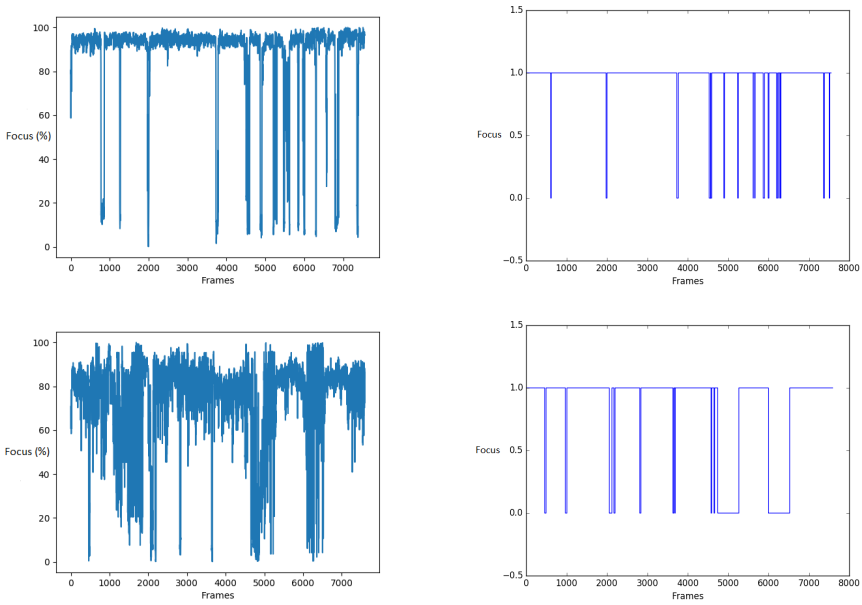

Fig. 3. From left to right, the experimental graphic and the manual graphic obtained for ID 0 and ID 2 (Focus values in the Y axis, number of the frame in the X axis). From top to bottom, the belonging IDs for each graphics: 0 and 2.

The left images are the experimental results and the right images are the results obtained through manual annotation. The Ground Truth was 93.07%, 98.23%, 85.69% and 93.56% from ID 0 to ID 3, making an average Ground Truth of 94.13% for this experiment. While making manual annotations, we deduced that the Ground Truth can never be 100% due to the two dimensional approach limitations in the focus estimator. However, the accuracy is satisfactorily high considering the computational cost.

4.2 Classroom Experiment

A formal experiment was conducted on a real classroom scenario at the University of Aveiro. The class duration was 38 minutes and 17 seconds and 60 students attended it. The results were quite satisfactory: the identification block presented an accuracy of 100% during the whole class and the system successfully provided graphics of the estimated focus for each individual ID. Figure 4 shows how the experiment was organized and Figure 5 shows 2 examples of different focus levels during this experiment.

Fig. 4. Formal experiment during a class in University of Aveiro.

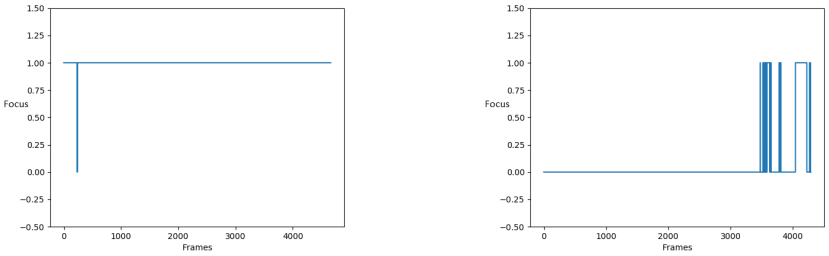

Fig. 5. Different focus levels for 2 students during the experiment (Focus values in the Y axis, number of the frame in the X axis).

Figure 5 shows 2 different types of behaviours during the experiment. For this experiment we decided to normalize the focus levels for better analyzable graphics: focus levels less than 50% are converted to "0" and focus levels greater than or equal to 50% are converted to "1". In the left image, we have a student that was focused throughout most of the class, since most focus levels were "1". In the right image, we have a student that was distracted throughout most of the class, only staying focused a few moments in the end of the class.

This graphical feedback provides identification of lecture periods in which students were less watchful and the corresponding topics that need extra focus. This has the goal of improving academic performance. As future work, we intend to link focus to engagement and classify different types of student behaviours during the class.

5 Conclusion

In this paper we presented a system for monitoring classrooms. After researching about the potentially best state of the art approaches for each required

techniques, we built a system capable of transforming the classroom in a sensing environment. This system is capable of running automatically. No registration of the students at the beginning of the class is required, the only requirement being the intervention of the lecturer to instruct the system about the location of the reference point. This is simply done through a mouse click. The results for the Ground Truth of our focus estimator and for our real scenario experiment are satisfactory and encouraged us to follow up with the use of the system in a real classroom environment. As for future work, we intend to add a body pose estimator [12] to calculate the teacher's mood during the class and how it affects his/her performance. Moreover, we are working towards creating a classification of students' behaviour based on their focus and class engagement.

References

1. Carini, Robert M., George D. Kuh, and Stephen P. Klein. "Student engagement and student learning: Testing the linkages." Research in higher education 47.1 (2006): 1-32.
2. Hagenauer, Gerda, Tina Hascher, and Simone E. Volet. "Teacher emotions in the classroom: associations with students engagement, classroom discipline and the interpersonal teacher-student relationship." European Journal of Psychology of Education 30.4 (2015): 385-403.
3. Nguyen, Khuyen, and Mark A. McDaniel. "Using quizzing to assist student learning in the classroom: the good, the bad, and the ugly." Teaching of psychology 42.1 (2015): 87-92.
4. Stiefelhagen, Rainer, and Jie Zhu. "Head orientation and gaze direction in meetings." CHI'02 Extended Abstracts on Human Factors in Computing Systems. ACM, 2002.
5. Canedo, Daniel, Alina Trifan, and Antnio JR Neves. "Monitoring Students Attention in a Classroom Through Computer Vision." International Conference on Practical Applications of Agents and Multi-Agent Systems. Springer, Cham, 2018.
6. Zhang, Kaipeng, et al. "Joint face detection and alignment using multitask cascaded convolutional networks." IEEE Signal Processing Letters 23.10 (2016): 1499-1503.
7. Schroff, Florian, Dmitry Kalenichenko, and James Philbin. "Facenet: A unified embedding for face recognition and clustering." Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.
8. Huang, Gary B., et al. "Labeled faces in the wild: A database forstudying face recognition in unconstrained environments." Workshop on faces in'Real-Life'Images: detection, alignment, and recognition. 2008.
9. Krafka, Kyle, et al. "Eye tracking for everyone." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.
10. Head Pose Estimation using OpenCV and Dlib. `https://www.learnopencv.com/head-pose-estimation-using-opencv-and-dlib/`
11. OpenCV library. `https://opencv.org/`
12. Cao, Zhe, et al. "Realtime multi-person 2d pose estimation using part affinity fields." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017.