Monitoring Students' Attention in a Classroom Through 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. Monitoring classrooms using cameras is a non-invasive approach of digitizing students' behaviour. Understanding students' attention span and what type of behaviours may indicate a lack of attention is fundamental for understanding and consequently improving the dynamics of a lecture. Recent studies show useful information regarding classrooms and their students' behaviour throughout the lecture. In this paper we start by presenting an overview about the state of the art on this topic, presenting what we consider to be the most robust and efficient Computer Vision techniques for monitoring classrooms. After the analysis of relevant state of the art, we propose an agent that is theoretically capable of tracking the students' attention and output that data. The main goal of this paper is to contribute to the development of an autonomous agent able to provide information to both teachers and students and we present preliminary results on this topic. We believe this autonomous agent features the best solution for monitoring classrooms since it uses the most suited state of the art approaches for each individual role.

Keywords: Class monitoring · Learning environment
Face Detection · Face Recognition · Face Tracking · Pose estimation

1 Introduction

The studies presented on [1] show that the student engagement is linked positively to desirable learning outcomes, such as critical thinking and grades obtained in a subject. The student engagement and attention depend on several factors. One important factor that influences the preservation of the student engagement is the teacher [2]. Teachers' positive emotions are likely to induce students' positive emotions, referred to as "emotional contagion". Therefore, positive teacher emotions may not only be essential for the wellbeing of teachers but they may also affect students' wellbeing and, in turn, learning in class. This suggests that teachers' ability to connect well with students can be beneficial regarding students' attention.

Additionally, [3] complements that the size of a classroom influences the students' attention. In large classes, the teacher will have to use up more time to

J. Bajo et al. (Eds.): PAAMS 2018 Workshops, CCIS 887, pp. 371–378, 2018.
https://doi.org/10.1007/978-3-319-94779-2_32

draw students' attention, which is emotionally exhausting. In contrast, smaller classes seem to allow an environment in which students are less likely to receive corrective talk from their teachers, since they are naturally more engaged to the class. This appears to be a more productive educational environment.

However, [4] showed that only 46% to 67% of the students pay attention during a class. This means that up to half of the students could not be productive in their learning. It is important to recognize the potential factors that may lead to this scenario and in which situations of the class the students tend to lose their focus more than others. With this information in hands, the teachers can search for possible problems during their classes and try to correct them, which may benefit the learning efficiency of their students.

The first contribution of this paper is a thorough revision of the state of the art on monitoring classrooms. There is already some work done in this area, such as apps and social networks for monitoring the students activities as mentioned in [5] or Teacher Assistance apps as shown in [6]. However, as the title suggests, our main focus is a non-invasive approach using cameras placed in convenient spots of the classroom.

We first review state of the art techniques on Computer Vision that are useful in a classroom environment. This review is focused on Face Detection, Face Recognition, Facial Features and Pose Estimation. Firstly 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. After obtaining them, we can extract their Facial Features. This component is not only important for measuring the students' attention, but also for Face Recognition. We need to make sure that the identification process is as accurate as possible, since we obviously don't want to assign information to the wrong students. As for the Pose Estimation, we can use it to measure certain behaviours and relate them to the attention level of the students.

After studying the state of the art, we propose an autonomous agent concatenating the most suited techniques for monitoring a classroom, showing some preliminary results. 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 inputs low resolutions faces for our proposed agent.

Lastly, we make a brief conclusion of what was presented in this paper.

2 Understanding Attention in Large Classrooms

In this section we present the state of the art techniques on Computer Vision that are useful in a classroom, divided into several subsections, corresponding to useful approaches we have identified for monitoring students' attention.

2.1 Camera Self-calibration

The classroom environment is not uniform, its luminosity changes as the day goes by. Opening or closing the windows, turning on or turning off the lights

are also causes for an irregular luminosity during a day of classes. All of these situations induce different image intensities, therefore the camera needs to be calibrated to capture analyzable and uniform images. Maintaining the uniformity of the input images increases the Face Recognition accuracy. Comparing faces that are within the same conditions is obviously more accurate than comparing faces that are within different conditions. A study of [7] proves this point: an image that is not calibrated would potentially deteriorate the accuracy of Face Detection and Face Recognition.

A work presented in [8] showed a calibration of the most important parameters of the vision system which makes algorithms for object detection efficient. They propose a self-calibration method based on the image luminance histogram. The histogram of the luminance shows how many times each intensity value appears in an image. This can easily indicate if the image is underexposed or overexposed to light. Thus, they propose the application of a Proportional-Integral controller (PI) to handle this underexposure or overexposure to light. With this method they managed to preserve the intensity uniformity of the captured images.

2.2 Face Detection and Recognition

Face Detection and Face Recognition are the most important techniques of monitoring a classroom. Assigning attention levels to the wrong student is something undesirable, so the priority is to make a proper recognition before assigning data. Despite the inevitable low resolution of the input faces because of the distance between the camera and the students, we must assure a good identification accuracy.

A work made by [9] proposed a system that uses a multi-task cascade convolution neural network (MTCNN [10]) for Face Detection and uses the ResNet-101 layers convolution neural network for Face Recognition. The MTCNN has 3 stages to output a proper Face Detection, in each stage the face that is being analyzed goes through a convolutional neural network (CNN). The first stage obtains the candidate windows and their bounding box regression vectors, merging highly overlapped candidates [11]. The second stage feeds those candidates to another CNN, which rejects a large number of false candidates. The third stage is similar to the second one, but it also outputs five facial landmarks' positions. Paired with the proposed ResNet-101 layers CNN trained with 65 million samples, they claim to achieve 98.87% accuracy rate for the Face Recognition based on the Labeled Faces in the Wild (LFW [12]) Face Recognition benchmark.

Their proposed method was developed for classrooms, so they also faced the same problems we mentioned above. They showed that their system can detect low resolution faces while preserving the Face Recognition accuracy rate and real time performance. This is the ideal state of the art solution for detecting and recognizing students in a classroom environment, therefore we will consider it in the following section of this paper.

2.3 Features and Pose Extraction

One way of measuring the students' attention that immediately comes to mind is by analyzing their eyes. However, as [13] mentioned, the accuracy of techniques like Eye Tracking tend to suffer from low resolution images. Knowing this, we can't rely on that technique for extracting data from the students. Nevertheless, there are other methods to measure the students' attention which work around the distance problem.

As a study of [14] showed, head orientation contributes 68.9% in the overall gaze direction and achieved 88.7% accuracy at determining the focus of attention. This conclusion implies that head orientation is a powerful method of measuring the students' attention.

A work presented in [15] proposed a head pose estimation algorithm by associating a few facial landmarks with 3D world coordinates. Calculating the rotation and translation of those landmarks, it's possible to transform the 3D points in world coordinates to 3D points in camera coordinates and project them onto the image plane. Consequently, a resulting line starts in the nose landmark and it is drawn to the direction in which the head is oriented to.

Although the head pose alone already has a high accuracy determining the focus of attention, paired with another technique could additionally improve the results. Students that are paying attention normally react to a stimulus in the same way. In other words, students that have their motion synchronized with the majority are assumed to be paying attention. An example of this synchronization is when the class has to write down something important if the teacher tells them to [16].

The work described in [17] proposes a real time multi-person 2D pose estimation (OpenPose). The image to be analyzed is fed to a CNN, predicting a set of 2D confidence maps of body part locations and a set of 2D vector fields of part affinities, encoding the degree of association between parts. Using non-maximum suppression they are able to obtain the body part candidates. The set of 2D vector fields of part affinities has the role of eliminating wrong associations in a multi-person scenario (where the classroom is included).

The OpenPose average precision on detecting the body parts (head, shoulders, elbows, wrists, hip, knees and ankles) is claimed to be between 75% and 80% based on the MPII Multi-Person Dataset test, beating the precision of the other state of the art solutions.

With this body pose estimation, students' motion can be tracked. Concatenating this output with the head pose estimation's output, we might have an approach of calculating the attention of a classroom that presents a satisfactorily high accuracy.

3 Proposed Agent and Preliminary Results

The overview presented in the previous section led to the proposal of the following autonomous agent for monitoring classrooms. In Fig. 1 we present its workflow diagram.

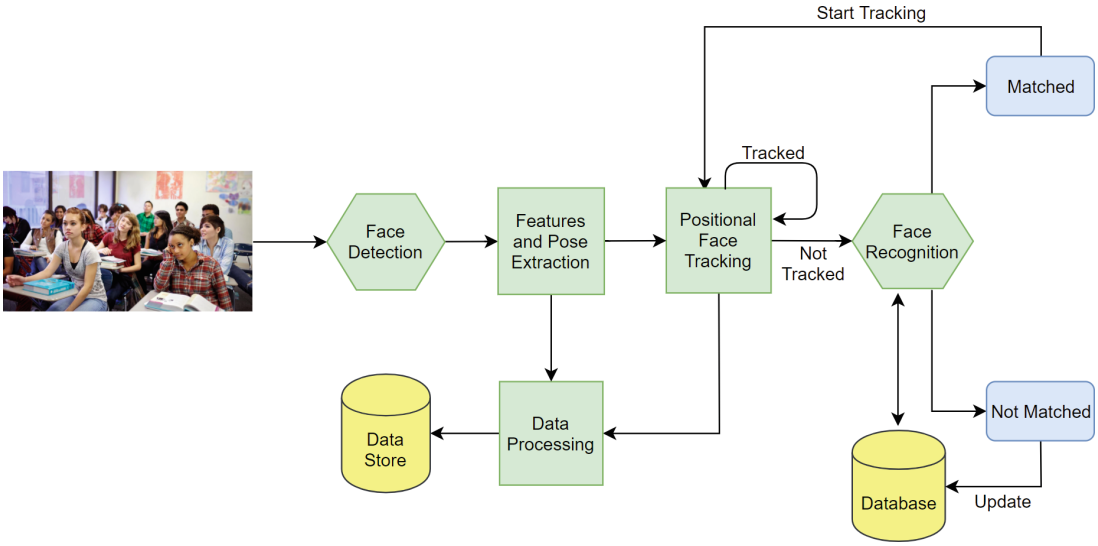

Fig. 1. Workflow diagram of the proposed autonomous agent.

The proposed agent receives an image from an acquisition module that captures the images of the classroom. This module analyzes the image and calibrates the camera parameters (Fig. 2) using the camera self-calibration approach mentioned in the previous section.

Fig. 2. From left to right, an image captured with some of the parameters of the camera set to higher values and the corresponding image obtained after the calibration.

In Fig. 2 the importance of camera calibration is explained. In the left image, it would be difficult to detect the face since the image has high intensity values. The right image is the resulting image obtained after the calibration, which homogenized the intensity values of the face, making it simpler to be detected.

Afterwards we propose the use of the MTCNN mentioned above for the Face Detection block. This block is responsible to feed the Features and Pose Extraction block with images of the students' faces, which has the role of retrieving their facial features and body pose. We propose the adjustment of the head pose estimation approach mentioned previously for estimating the head orientation and the use of OpenPose for the body pose estimation (Fig. 3).

Fig. 3. From left to right, preliminary results of the head pose estimation and the body pose estimation.

At this point, there is the need of identifying the students before assigning them the data regarding their attention. We propose a block called Positional Face Tracking to assist the Face Recognition block, which is responsible for tracking the students' faces by comparing their actual position with their previous position. Since the students usually are sat down on their chairs, we can avoid that the proposed agent would keep on trying to recognize them (which has a significant computational cost and reduces the frames per second) by adding this type of tracking. However, if certain student moves out of place, the tracking may fail. If this is the case, the proposed agent would advance to the Face Recognition block and try to recognize the respective student. We propose the ResNet-101 layers CNN mentioned in the previous section for the Face Recognition block.

In Fig. 4 we show a preliminary result of our proposed autonomous agent prototype, obtained from a real environment. The students were participating in a workshop about Computer Vision organized at our University.

Fig. 4. Preliminary result of Positional Face Tracking, Face Recognition and Attention Levels in a real scenario during a workshop at our University.

In the beginning of the referred workshop, we registered the students that agreed to participate in the experiment through a NFC Reader, which automatically turned on the camera, storing images of their faces. We trained their aligned faces using the ResNet-101 layers CNN and turned on the proposed agent prototype. When the students were identified through the Face Recognition block, the agent started tracking their faces and their attention. Since the prototype assumes that the students looking towards the camera are paying attention, the students in Fig. 4 have low attention levels despite their obvious focus on the lecture. However, the target position can be configured beforehand and will be considered in future experiments. Although we used a static Database in this experiment, we propose a dynamic Database that would assign a unique ID to each student and store their faces automatically, providing more autonomy to the agent.

The data regarding the students' attention was saved into vectors during the lecture. When the agent was turned off, the Data Processing block calculated the averages of the attention levels for each student and stored them into the Data Store block. For this block, we propose an online database that is updated at the end of the class, being accessible to both teachers and students.

4 Conclusion

In this paper we proposed an agent for monitoring classrooms. We made a research about the potentially best state of the art on Computer Vision techniques that may be applied in a classroom environment. By putting them together in a workflow that theoretically handles problems like false positives or losing track of a student, we proposed an autonomous agent showing some of the preliminary results of its prototype. We believe that this agent has the potential of transforming the classroom in a sensing environment, by providing guidance and feedback not only to the teachers on how to improve their teaching role during a class, as well as to the students on how to improve their behaviour in the classroom and consequently increase their academic performance. The long term goal of our research is to provide visual feedback to the teachers regarding the average level of students' attention, and provide counseling to the students regarding their behaviour during the class. Such counseling can be, for example, the identification of lecture periods in which students were less watchful and the corresponding topics that potentially need extra attention.

Acknowledgments. This work was 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.

References

1. Carini, R.M., Kuh, G.D., Klein, S.P.: Student engagement and student learning: testing the linkages. Res. High. Educ. **47**(1), 1–32 (2006)
2. Hagenauer, G., Tina, H., Volet, S.E.: Teacher emotions in the classroom: associations with students' engagement, classroom discipline and the interpersonal teacher-student relationship. Eur. J. Psychol. Educ. **30**(4), 385–403 (2015)
3. Blatchford, P., Bassett, P., Brown, P.: Examining the effect of class size on classroom engagement and teacher-pupil interaction: differences in relation to pupil prior attainment and primary vs. secondary schools. Learn. Instr. **21**(6), 715–730 (2011)
4. Raca, M., Kidzinski, L., Dillenbourg, P.: Translating head motion into attention-towards processing of student's body-language. In: Proceedings of the 8th International Conference on Educational Data Mining. No. EPFL-CONF-207803 (2015). APA
5. Carchiolo, V., et al.: Monitoring students activities in CS courses. In: 2016 15th RoEduNet Conference: Networking in Education and Research. IEEE (2016)
6. Gutierrez-Santos, S., et al.: Scalable monitoring of student interaction indicators in exploratory learning environments. In: Proceedings of the 25th International Conference Companion on World Wide Web. International World Wide Web Conferences Steering Committee (2016)
7. Zhao, W., et al.: Face recognition: a literature survey. ACM Comput. Surv. (CSUR) **35**(4), 399–458 (2003)
8. Neves, A.J.R., Trifan, A., Cunha, B.: Self-calibration of colormetric parameters in vision systems for autonomous soccer robots. In: Behnke, S., Veloso, M., Visser, A., Xiong, R. (eds.) RoboCup 2013. LNCS (LNAI), vol. 8371, pp. 183–194. Springer, Heidelberg (2014). https://doi.org/10.1007/978-3-662-44468-9_17
9. Fu, R., et al.: University classroom attendance based on deep learning. In: 2017 10th International Conference on Intelligent Computation Technology and Automation (ICICTA). IEEE (2017)
10. Zhang, K., et al.: Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Sig. Process. Lett. **23**(10), 1499–1503 (2016)
11. Farfade, S.S., Saberian, M.J., Li, L.-J.: Multi-view face detection using deep convolutional neural networks. In: Proceedings of the 5th ACM on International Conference on Multimedia Retrieval. ACM (2015)
12. Huang, G.B., et al.: Labeled faces in the wild: a database for studying face recognition in unconstrained environments. Technical report, vol. 1, no. 2, pp. 07–49, University of Massachusetts, Amherst (2007)
13. Krafka, K., et al.: Eye tracking for everyone. arXiv preprint arXiv:1606.05814 (2016)
14. Stiefelhagen, R., Zhu, J.: Head orientation and gaze direction in meetings. In: CHI 2002 Extended Abstracts on Human Factors in Computing Systems. ACM (2002)
15. Head Pose Estimation using OpenCV and Dlib. https://www.learnopencv.com/head-pose-estimation-using-opencv-and-dlib/
16. Raca, M., Dillenbourg, P.: System for assessing classroom attention. In: Proceedings of the Third International Conference on Learning Analytics and Knowledge. ACM (2013)
17. Cao, Z., et al.: Realtime multi-person 2d pose estimation using part affinity fields. In: CVPR, vol. 1, no. 2 (2017)