ISSN 2278-3091
Volume 9, No.5, September - October 2020
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse68952020.pdf
https://doi.org/10.30534/ijatcse/2020/68952020

A Novel Architecture for Student's attention detection in classroom based on Facial and Body expressions

Tarik Hachad[1], Abdelalim Sadiq[1], Fadoua Ghanimi[2], Lamiae Hachad[3]
[1] Laboratory of Information Modelling and Communication Systems, Ibn Toufail University, Kenitra, Morocco, Tarik.hachad@uit.ac.ma
[2] Laboratory of Technological Information and Modeling, University Hassan II, Casablanca, Morocco, ghanimi_fadoua@yahoo.fr
[3] Laboratory of Signals Systems and Components, Sidi Mohamed Ben Abdellah University, Fez, Morocco, lamiae.hachad@usmba.ac.ma

ABSTRACT

Student attention has become a key topic for the educational system. Automatic and efficient systems are needed for monitoring classrooms and providing feedback to the teacher. In this paper, we have presented a novel architecture for student's attention detection. Our goal is to estimate the student's state of attention at any time during the lecture based on the analysis of the student's facial and body expressions. Moreover, we have provided a detailed comparison of recent systems existing in literature. Several features are considered to be analyzed and that will afford us a good solution for automatic students' attention estimation during lectures.

Key words: Attention detection, Computer vision, Human behavior analysis, Machine learning.

1. INTRODUCTION

Attention is the first step in the learning process. Indeed, it is the ability to constantly maintain mental effort in order to be able to select and focus on what is important. This must be done while resisting distraction. Many authors affirm that at the beginning of the lecture, the students pay particular attention to teaching but for most of them, they end up losing this attention after about 10 minutes [1]. This leads educational institutions to consider the students' attention not only as a tool to improve learning but as an essential component that must be measured and analyzed.

For many researchers, attention is synonymous with engagement. Fredricks et al [2] analyzed 44 studies and proposed that there are three different forms of engagement: behavioral, emotional, and cognitive.

Analysis of student attention becomes a key topic for the educational system[3], [4], which needs automatic and efficient systems to monitor the variation of student's attention and provide feedback to the teacher.

In this work, the main objective is to achieve the detection and the analysis of the student's attention through the use of different technologies of detection of facial and body expressions. The ultimate goal is to have accurate information about the student's attention at any given time. Thus, this paper proposes a multimodal architecture for the detection and measurement of the attention. This architecture is enough flexible to allow a possible evolution such as the integration of new components. On the other hand, it was designed to be adaptable to function according to the variable conditions of a classroom.

Contributions: In this article, we review existing attention detection systems. This review will be concluded by a comparison of the various techniques proposed. We propose a multimodal framework for automatic detection of the student's level of attention. This system is based on the use of computer vision techniques. Then we discussed the pros and cons of the different existing systems and our proposal to make an improvement.

Conceptualization of attention: Our goal is to estimate the attention of students. In other words, detect the facial and body expressions that characterize the attention of the student and this in accordance with the observations made by the experts. Thus, teachers can have the necessary information in order to adapt their educational behavior.

Our article is structured as follows. First, existing systems are presented. Next, we explain in detail the architecture of our system. Then, we discuss the main strengths and weaknesses of existing systems and the solutions that our system offers. We end our document with some conclusions.

2. EXISTING SYSTEMS

The literature on student attention detection shows a variety of approaches,
Table ***1*** summarizes the different approaches used by some existing systems.

Krithika et al [5] have proposed a system for monitoring student concentration in an e-learning environment. The objective of their work is to detect the student's concentration from two important measurements: the rotation of the head and the eyes' movement. The analysis of these components provides one of the three levels of concentration defined by the authors: high, medium and low level of concentration.

In [6], the authors aim to integrate affective calculus to help teachers assess the quality of their instructions, by analyzing student engagement through the classification of facial features. They also consider that the student's engagement has three components: behavioral, cognitive and emotional, and they try to establish a link between facial expressions and learning ability.

For Zaletelj et al [7] The basic idea of their system is to use the advanced capabilities of the Kinect One sensor to discreetly collect behavioral data from students during lectures. They proposed a methodology to compute features from the Kinect data corresponding to visually observable behaviors. Then, they apply machine learning methods to build models to predict the attentive state of each student. For them the main issue was to establish a correlation between the student's attention and the observations made by the teacher. They analyze attention scores provided by human observers and match them with the observable behaviors of the students (activities, gestures, etc.). This makes it possible to discriminate the different behaviors associated to a level of attention. Then, a machine learning algorithm is used to generate the model for predicting the student's state of attention from the features provided by the Kinect One sensor.

Canedo et al [8] have proposed an autonomous agent theoretically capable of tracking the students' attention and providing output data of this component for each student. Based on [9] which reveals that the orientation of the head contributes 68.9% in defining the direction of gaze and achieves an accuracy of 88.7% at determining the focus of attention. According to this, they chose the head orientation method as a powerful measure of student attention. As they consider that 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. [8], [10]. To improve the results of their system, they decided to pair the head orientation technique with a second technique based on the body gesture.

The paper of Whitehill et al [11] explored the development of real-time automated recognition of engagement from student's facial expressions. The first step of their work was collecting experimental data for the engagement study of 34 subjects who participated in the Spring 2011 version of the Cognitive Skills Training study at the Historically Black College /University (HBCU) and the Summer 2011 version of the Cognitive Skills Training study at a University in California. The students were taken on video while they are performing cognitive tasks on iPad. Different scenarios for image labelling were explored in order to choose the most reliable and feasible method. Four binary classifiers of engagement were trained, one for each of the four levels previously set by the authors. Next evaluating the accuracy of their model, they proceeded to reverse engineering to understand how the human labelers formed their judgments. Finally, they investigate the correlation between human and automatic perceptions of engagement with student test performance and learning.

3. GENERAL DESCRIPTION OF OUR ARCHITECTURE FOR STUDENT ATTENTION DETECTION

The basic schema in Figure 1 shows the essential processing nodes of the framework. The data to be processed by this system is a stream of high definition images encoded in video format provided by a camera. In such a system the computing power is very expensive. Consequently, the various functionalities must be performed in an appropriate manner to acquire real-time execution. Accordingly, the proposed architecture for the detection and estimation of attention is divided into 4 main modules (nodes) that must run as a distributed processing. Each node is responsible to perform a number of treatments in an independent way. As can be seen in Figure 1, the nodes are responsible for the tasks associated with the detection of facial expressions, body gesture classification, eye gaze estimation, and head pose estimation. The outputs of these various nodes provide heterogeneous results and must be merged (Fusion) in order to generate an understandable representation of the level of attention.

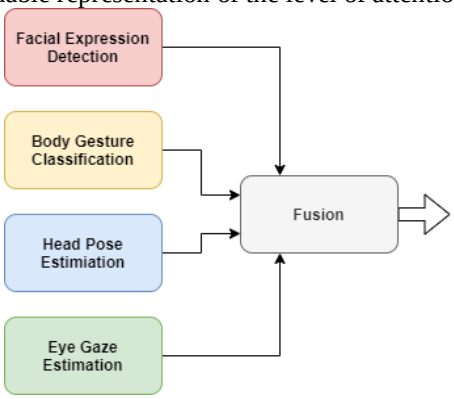

Figure 1: Basic diagram of the distributed architecture for the student's attention detection and estimation.

The system relies essentially on components formed of a combination of sensor and processing unit. These must be executed in a distributed manner. They must also be autonomous and have a perception of their environment through a set of input and output interfaces in order to establish an orchestrated hierarchy of operations. A central module would coordinate all the components to correctly perform the algorithms of attention evaluation. For example, each component of our system (facial expression detection, body gesture classification, eye gaze estimation, and face gaze estimation) has a module for collecting information from its sensor, which is in our case the camera. Then, it will execute its own processing to obtain the output data in a format understandable by the central component that ensures the tasks of coordination and fusion of the results. To address these issues, we have opted for an architecture called multi-agent systems (MAS).

Table 1: Features summary of existing attention detection systems.

Ref	Hypothesis/ Idea	methodologies	technologies	Data set
[5]	In an e-learning environment where the student is facing the screen of his computer, his gaze and his head pose provide crucial information on his concentration level [5].	The system (SERS) proposes an approach allowing to detect the level of concentration of the student by continuously monitoring the rotation of the head and the eyes' movement.	MatLab functions for face detection and features detection using Viola Jones and LBP	
[6]	Effective classification of student engagement would only be identified by actual teachers in a classroom setting[12]. The authors would like to prove that the initial dataset labelled by teachers would be effective as a basis in the classification of student engagement by the Predictive Classification Model[6].	The framework of study presented in this work is composed of four major components: Input, Face Detection and Facial Features Extraction, Predictive Classifier and Output or Feedback to user. The training dataset is labelled as engaged or not with the aid of human experts (two teachers) Test data are composed of three data sets the first and the second include images captured from two classes with 31 and 53 students, respectively and the last consists of combined instances from the first two data sets.	Microsoft Cognitive Toolkit (CNTK)[13]. OpenFace [14]	Extended Cohn-Kanade (CK+) AU-Coded facial expression database [12]
[9]	The use of advanced capabilities of the Kinect One sensor to discreetly collect behavioral data from students during traditional classroom lectures, in order to analyze it and detect the level of attention.	The authors proposed a methodology to calculate the features from Kinect data corresponding to visually observable behaviors. These later were labelled as describing a level of attention. Machine-learning methods have been applied to these features in order to build models to predict the attentive state of each student.	Kinect One sensor and its Toolbox	The dataset was obtained during four lecturing sessions in a classroom, where students were asked to perform some learning tasks like take notes, answer questions... The Kinect One sensor was placed frontally to students from a distance of 1.8 m, 15 Colored frames per second were captured with full HD resolution (1920 by 1080).
[8]	Students that have their motion synchronized with the majority are assumed to be paying attention. The study of Stiefelhagen et al [10] implies that head orientation is a powerful method of measuring the student's attention.	The prototype assumes that the students looking towards the camera are paying attention.	MTCNN [15] LFW [16]	MPII Multi-Person Data set
[11]	This study explores the development of real-time automated recognition of engagement from students' facial expressions.	The study compares the observations made by the participants in the student engagement estimation operation based on their facial expressions. When there is an agreement between the observers, the image is labelled which gives more reliability to this work. Then, machine learning methods are used to automatically predict engagement.	MLR(CERT)[17]s	The dataset is composed of images and clips containing facial expressions of 34 students. Data are collected while students are performing cognitive tasks.

The complete architecture of the system is illustrated in Figure 2. It comprises 4 main parts, namely "facial expressions detection", "body gesture classification", "face gaze estimation" and "eye gaze estimation" they process the students' images captured by the camera and provide their results to the "Multimodal Fusion" unit that merge them to return an estimation of the students state of attention.

3.1 Image acquisition

The student image acquisition process is performed by a high-resolution color camera located above the projection to capture the entire audience in the classroom. In this way, all students' faces must be present as part of the image captured accurately. A series of image snapshots are taken at well-defined time intervals (15 frames / second).

Face detection is the first step in our biometric analysis system, its accuracy significantly affects the performance of subsequent operations, hence the importance of using a high-performance face detection algorithm. Each face present on the image is then extracted and saved in a cropped format required for the analysis tasks that will succeed. Indeed, it will be transmitted to the following units: Facial expressions detection, Head pose estimation and eye gaze estimation units.

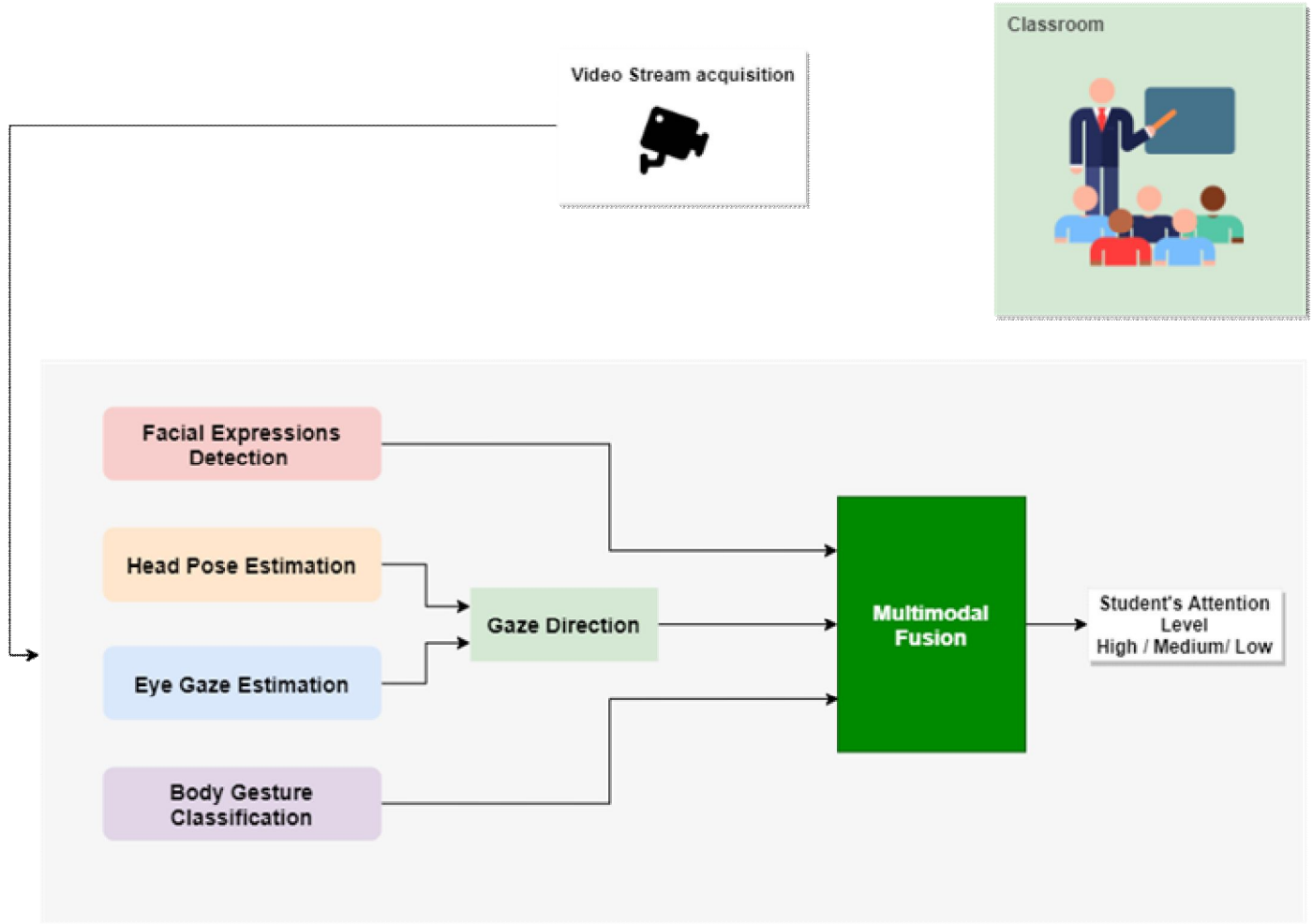

Figure 2: The distributed architecture of the student attention detection system

3.2 Facial expressions detection

The ability of the human being to carry out the daily tasks of life, starting from what is serious and requires concentration to entertainment is greatly dependent on his emotions. According to Immordino et al, the relationship between learning, emotion, and body state is much deeper than many educators realize and that the original purpose for which our brain evolved was to change our physiology, to optimize our survival and allow us to flourish [18].

Emotions are important in nonverbal communication, and emotions influence cognition in many ways: how we process information, our attention, and our biases towards information [19].

In order to approach the process of detecting emotions, we were inspired in the circumplex model of affect [20]. The circumplex model of affect suggests that emotions are distributed in a two-dimensional circular space, with dimensions of excitement (activation) and valence (pleasure).

The different states of arousal are represented on the vertical axis while the valence occupies the horizontal axis, the center of the circle signifies a neutral valence and an average level of excitation. In this model, each emotion can be represented by a level of valence and excitement, or at a neutral level of one or the other of these factors as illustrated in the Figure 3.

The camera provides a continuous, non-intrusive way to capture images of students' faces. facial information can be used to understand certain facets of the student's current state of mind, and there are several techniques to automate this measurement process [21], [22].Knowing the affective state of the student can lead to deduce his level of attention or at least gives partial information that can be combined with others to calculate his level of attention.

The information extracted the student's face are decisive for obtaining an emotional interpretation: positive promoting attention and learning or negative decreasing these two components. Figure 3 shows the different emotional states that are provided by our system. In this regard, it will offer one of the basic emotions as output: happiness, sadness, anger, fear, disgust, neutral, and surprise.

Ekman and Friesen (1978) propose a coding system for facial actions (FACS) which highlights the expressive aspects of emotions. This system describes the specific facial behaviors, based on muscles and their correspondence with facial expressions (basic emotions). Basically, each movement in the face is called an action unit AU. There are approximately 58 action units. These facial models have been used to identify the emotions of happiness, sadness, surprise, disgust, anger and fear [23].

In the field of facial expression analysis, there are two main approaches: geometric-based and appearance-based approaches [24]. The geometric facial features present the shape and locations of facial components (including the mouth, eyes, brows, and nose). The facial components or facial features points are extracted to form a features vector that represents the face geometry. With appearance-based methods, image filters, such as Gabor wavelets, are applied to either the whole face or specific regions in a face image to extract a feature vector [25].

There are strengths and weaknesses in both the geometric based and appearance-based approaches. Geometric based methods typically track the position of a number of facial points in time. With this approach, some features of facial appearances (e.g., shape of mouth, position of eyebrows) can be extracted, while features related to texture of the face (e.g., furrows and wrinkles) cannot be extracted. In contrast, appearance-based methods may be more sensitive to changes in illumination (e.g., brightness and shadows), head motions and differences between shapes of the faces [21], [26].

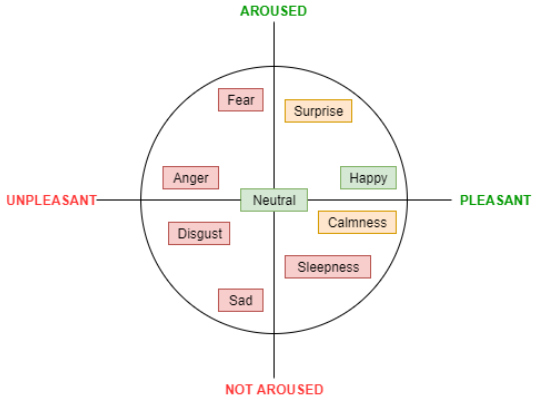

Figure 3: Circumplex model of affect

3.3 Head pose estimation

Head pose estimation is required for tracking where each student is looking in the classroom. Multi-person head pose estimation presents technical challenges in a large classroom setting [27]. Our only data source is a single camera that continuously provides 2D images. For that, we are interested in the study of 2D images-based methods to estimate the student's head position. These approaches have shown a certain unreliability, especially in cases where a person's eyes, nose and mouth are occluded. Indeed, it has been observed that students adopt common behaviors during a lesson scene such as supporting the head with the hand or scratching their hair. These partial facial occlusions are common and can interrupt the tracking of head pose for periods. In the practical case, a student can focus multiple points in the classroom. An attentive student: will face the slideshow or the professor, this indicates that he is following the explanations. He can also be in a state of attention and have the gaze towards his notepad, but he must manifest the posture of transcribing notes. The situations where the student reveals a lack of interest or a distraction is when he looks around him, fixes the ceiling, or looks at his notepad for a long time without writing anything.

3.4 Eye gaze estimation

While engagement/attention seems to be a difficult concept to model mathematically, there exist informative cues that can be used to infer the attention level of an audience. One very natural cue is a person's gaze. In particular, the location of a person's gaze focus and the duration for which they maintain their focus are useful indicators of attention [28]. Just et al indicate that the duration of a person's eye fixations is directly related to the amount of neural processing power they are devoting to a particular task, which can then be directly related to a person's attention level[28], [29].

In the literature, there are a large number of techniques that accurately estimate the position of the gaze. Sharma et al [30] propose a very good comparative survey of eye gaze estimation techniques.

The eye gaze estimation could only be measured in situations when the eyes are detected. Actually, several factors can prevent the detection of the eyes such as the presence of occlusive objects or an over brightness of the image, etc. The eye gaze estimation will be used in conjunction with the head pose to complete the information. Indeed, when the eyes are visible, head pose becomes a requirement to accurately predict gaze direction. Physiological investigations have demonstrated that a person's prediction of gaze comes from a combination of both head pose and eye direction [31], [32]. In Figure 4, two views of a head are presented at different orientations, but the eyes are drawn in an identical configuration in both. Glancing at this image, it is quite clear that the perceived direction of gaze is highly influenced by the pose of the head. If the head is removed entirely and only the eyes remain, the perceived direction is similar to that in which the head is in a frontal configuration[32].

Figure 4: Wollaston illusion: Although the eyes are the same in both images, the perceived gaze direction is dictated by the orientation of the head [32], [33]

3.5 Body gesture classification

Detecting attention from body gestures is a very challenging task. In order to handle this issue, we studied the behavior of students during the lecture. In our case we are interested in the upper part of the body. Since this latter is the visible part in a scenario of a student seated behind a table. Each behavior is linked to a level of attention such as writing notes or standing in a position like supporting the head with the hand and have a gaze outside the slideshow.
The authors in [34] give very good surveys on body posture detection. The survey paper reviews the various electronic devices on the market that allow performing body postures detection tasks. It establishes a benchmark to compare the automatic recognition systems and the body posture data sets necessary for training of the detection model. As mentioned previously, in our approach, the objective of this module is to detect postures revealing a level of attention. The output of the body posture detection node will take the following values: upright sitting, hand supporting the head, lean back, and writing notes

3.6 Multimodal fusion

Our system is intended to help teachers detect possible loss of students' attention during the lectures. Such a system must handle perfectly real-world cases. Indeed, the sources of information on which it must rely have to be multiple and analyzed in a combined manner. A fusion step is necessary in order to build a multidimensional image of the student's attention and to reduce the inaccuracies related to the extracted features.

The different analysis units of our system receive the students' images in order to extract features. These are heterogeneous, hence the need to combine them to obtain complete information. This operation is known as multimodal fusion.

Multimodal Fusion is achieved using an artificial neuronal network (ANN). The ANN is used to predict the student's state of attention from the different facial and body features extracted by all units of the system. It will therefore have to learn with great precision the dataset of student attention states that we have partially created and which will be continuously fed with the teachers' observations, Table 2 gives a small overview of this dataset.

The architecture of our ANN is made up of three layers: an input layer (i), which receives information from the facial and body features, a hidden layer (h), responsible for processing information from the input layer, and an output layer (o) that admits one of three possible outcomes: high, medium and low level of attention. The architecture of our neural network is shown in the Figure 5. Once the dataset of facial and body features has been fixed by the experts, we can perform the training which will allow us to estimate the weights of the neurons in each layer of the network.

Table 2: The facial and body features for student states of attention

	Emotion	**Gaze direction**		**Body gesture**	**Attention level**
		Head Pose	**Eye gaze**		
1	Neutral	Slide		Upright sitting	High
2	-	Fixes the ceiling		Lean back	Low
3	Sad	Slide		Upright sitting	Medium
4	-	-		Writing notes	High

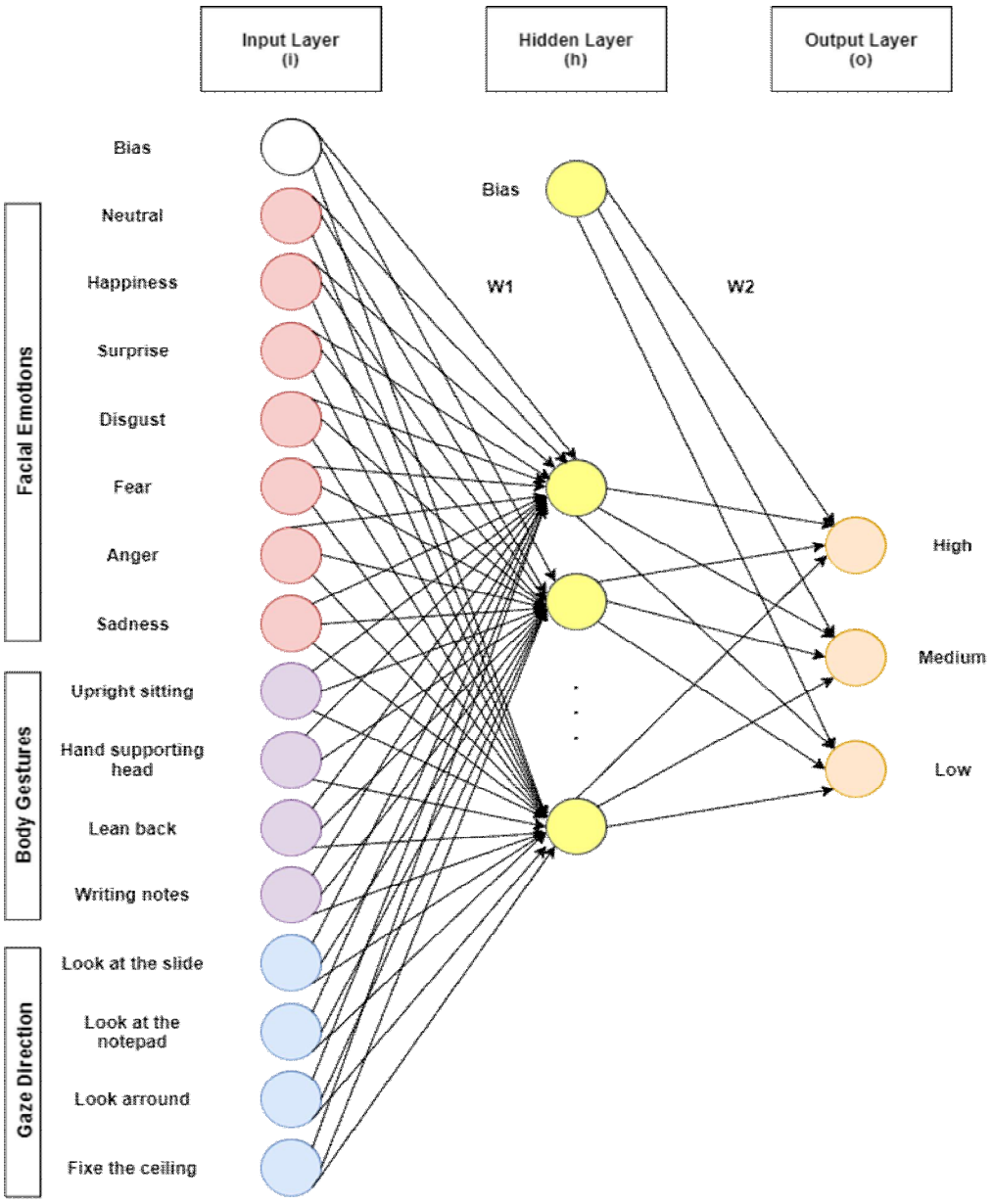

Figure 5: ANN ensuring Multimodal Fusion

4. DISCUSSION

Student attention detection has become one of the current active research topics in computer vision. In this paper we propose a comparison study of several existing systems to detect student's attention. Each system is based on one or more axes of analysis such as the analysis of emotions or the gaze direction, used independently or combined to build an image of what attention is. In an e-learning context where a student is placed in front of his screen reading a text or answering questions, these approaches can be convincing. However, in the scenario of a classroom, the development of a fully automated attention recognition system is a non-trivial task because of the large number of students that a class can contain, the occlusions that may exist, the complexities of human gestures and data acquisition problems. Table 3 provides a complete comparison of the methods used by existing systems and lists the pros and cons of each system.

The first step in building an attention recognition system is to acquire a dataset on the student's attention states or, in the absence of a such resource, we must compose our own labelled images and video sequences dataset. This dataset will be used to generate the attention recognition model. It should contain a sufficient variety of student's behaviors and correspond to real-world scenarios. One of the most important things to consider is the quality of the media that forms the dataset. A good dataset should take into account all technical and functional constraints:

- The input medium must include images and video sequence;
- There must be enough data to overcome the over-fitting issues;

Table 3: The detailed comparison of existing attention detection systems

Reference	Eye gaze	Head pose	body gesture	Facial expressions	Pro and cons
[5]	non	yes	no	no	- The system uses very few measures to detect the level of concentration of the student. - The detection of the student's eyes cannot provide the real direction of the gaze.
[6]	no	no	no	yes	- The use of a single measure (emotions detection) is not enough to assess the level of engagement of students
[7]	yes	yes	yes	yes	- Data from human observations is not entirely reliable; - The training dataset is rather limited, and its total length for 18 persons is 122 min; - The variations of human behavior which are present within the dataset is limited and is not covering all possible student behaviors; - Inter-personal differences in behavior during lecture were clearly visible and influenced the accuracy of a person-independent classifier; - Detection of the faces of all students at a given moment is not always guaranteed; - The reliability of gaze detection depends on the orientation of the face (frontal or not) and presence of obstructing objects such as hands; - The similar issue is with person's skeleton which is not accurate due to obstructed view of the person sitting behind a table.
[8]	yes	yes	no	no	- Student can have low attention levels despite his obvious focus on the lecture
[11]	no	no	yes	yes	- They followed a rigorous method for the collection and labeling of images which reinforces the precision of this system

- The quality of the input medium (resolution, grey-scale or colour);
- Great variety of the same gesture,
- Large intraclass variations (i.e. variations in the pose of the subjects);
- Shooting under partial occlusion.

Some reviewed systems use standard data sets like CK+ or MPII Multi-Person Dataset for training and testing purposes, others generate their own labelled datasets based on observations made by experts in the education sector. The standard datasets cover a specific area, such as the emotions CK+ or human pose (sport) MPII. However, these limitations constitute an unrealistic scenario that does not represent real-world situations and does not meet the specifications of a dataset for student's activity classroom. Moreover, the datasets labelled by education experts do not contain a large amount and variety of data. Labelling methodologies are not often explained or reveal a lack of rigor and objectivity.

A teacher can easily detect that a student is not interested or distracted, this is part of the skills he has acquired during the ply of his profession. A human can easily learn to classify objects or phenomena he encounters. This task is not as simple to perform for a machine. Indeed, it needs specific learning to perform this operation with a relatively correct accuracy compared to that of a human being. In order to simulate the human approach, we must ask the following questions:

1. What are the visual features that allow a teacher to determine that a student is attentive or not?
2. Are the visual features evaluated separately to conclude the state of attention of the student or must be combined or merged to provide an output result?

Choosing the appropriate features for attention detection is a problem that needs to be addressed before the classification phase. In our study, the possible features to extract from a camera stream are: facial and body expressions and gaze direction.

The reviewed systems propose unimodal and multimodal approaches with features coupling. But what they are criticized for is the absence of the features' fusion. In fact, they use body and facial expressions independently to decide on the student's state of attention or in the case of the combination of several features, they are used in a sequential manner which is in opposition to the principle of complementarity. So, we have proposed a fusion of features based on multimodal data sources. The combination of multimodal features that we offer retains the coupling of body features and gaze orientation to produce a complete image of attention.

An automated system must operate effectively in different situations. Indeed, all the stages of attention modelling and analysis must be performed automatically. This requires rigor in carrying out the following tasks: modelling (choice and extraction of features) which provides the information necessary for recognition of attention; and the classification phase which will classify a feature or a set of features in a defined level of attention. Several factors can prevent proper functioning of the system in a classroom: the variation in lighting, the partial occlusion of the face or body and the number of attention situations that the system must handle and recognize.

5. CONCLUSION

This paper has described a distributed architecture for student attention detection in a classroom based on the analysis of facial and body expressions. The ultimate goal is to be able to keep the attention of the students throughout the lecture. To accomplish this, we have chosen a number of features to analyze for our study. The choice of these features was not arbitrary but, conditioned by the data source that we use, namely the camera. We have also drawn up a comparison of existing systems in order to propose a generic architecture that will overcome the limitations of existing systems.

The main limitation of this study is the limited access to real-world data, so, the size of the samples available is insufficient for statistical measurement which we will take into consideration in future work. The next step of our work is to collect more samples in order to accomplish our student attention detection dataset and to test the performance of our architecture.

REFERENCES

[1] N. A. Bradbury, "**Attention span during lectures: 8 seconds, 10 minutes, or more?**" American Physiological Society Bethesda, MD, 2016.

[2] J. A. Fredricks, P. C. Blumenfeld, and A. H. Paris, "**School Engagement: Potential of the Concept, State of the Evidence**," *Rev. Educ. Res.*, vol. 74, no. 1, pp. 59–109, Mar. 2004.

[3] N. Sabri, N. H. Musa, N. N. A. Mangshor, S. Ibrahim, and H. H. M. Hamzah, "**Student emotion estimation based on facial application in E-learning during COVID-19 pandemic**," *Int. J. Adv. Trends Comput. Sci. Eng.*, vol. 9, no. 1.4 Special Issue, 2020.

[4] J. Zhang, E. Kamioka, and P. X. Tan, "**Emotions detection of user experience (Ux) for mobile augmented reality (mar) applications**," *Int. J. Adv. Trends Comput. Sci. Eng.*, vol. 8, no. 1.4 S1, pp. 63–67, 2019.

[5] Krithika L.B and Lakshmi Priya GG, "**Student Emotion Recognition System (SERS) for e-learning Improvement Based on Learner Concentration Metric**," *Procedia Comput. Sci.*, vol. 85, pp. 767–776, 2016.

[6] R. Manseras, T. Palaoag, and A. Malicdem, "**Class Engagement Analyzer using Facial Feature Classification**," no. November, pp. 1052–1056, 2017.

[7] J. Zaletelj and A. Košir, "**Predicting students'**

attention in the classroom from Kinect facial and body features," *EURASIP J. Image Video Process.*, vol. 2017, no. 1, p. 80, 2017.

[8] D. Canedo, A. Trifan, and A. J. R. Neves, "**Monitoring Students' Attention in a Classroom Through Computer Vision**," 2018, pp. 371–378.

[9] R. Stiefelhagen and J. Zhu, "**Head orientation and gaze direction in meetings**," in *CHI'02 Extended Abstracts on Human Factors in Computing Systems*, 2002, pp. 858–859.

[10] M. Raca and P. Dillenbourg, "**System for assessing classroom attention**," in *Proceedings of 3rd International Learning Analytics & Knowledge Conference*, 2013, no. CONF.

[11] J. Whitehill, Z. Serpell, Y.-C. Lin, A. Foster, and J. R. Movellan, "**The faces of engagement: Automatic recognition of student engagementfrom facial expressions**," *IEEE Trans. Affect. Comput.*, vol. 5, no. 1, pp. 86–98, 2014.

[12] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "**The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression**," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops*, 2010, pp. 94–101.

[13] Microsoft, "**The Microsoft Cognitive Toolkit - Microsoft Research.**" .

[14] T. Baltrušaitis, P. Robinson, and L.-P. Morency, "**Openface: an open source facial behavior analysis toolkit**," in *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2016, pp. 1–10.

[15] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "**Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks**," *IEEE Signal Process. Lett.*, vol. 23, no. 10, pp. 1499–1503, Oct. 2016.

[16] G. B. Huang, M. Mattar, T. Berg, and E. Learned-Miller, "**Labeled Faces in the Wild: A Database forStudying Face Recognition in Unconstrained Environments**," in *Workshop on Faces in "Real-Life" Images: Detection, Alignment, and Recognition*, 2008.

[17] G. Littlewort *et al.*, "**The computer expression recognition toolbox (CERT)**," in *Face and gesture 2011*, 2011, pp. 298–305.

[18] M. H. Immordino-Yang and A. Damasio, "**We feel, therefore we learn: The relevance of affective and social neuroscience to education**," *Mind, brain, Educ.*, vol. 1, no. 1, pp. 3–10, 2007.

[19] J. Broekens, D. Degroot, and W. A. Kosters, "**Formal models of appraisal: Theory, specification, and computational model**," *Cogn. Syst. Res.*, vol. 9, no. 3, pp. 173–197, 2008.

[20] J. A. Russell, "**A circumplex model of affect**.," *J. Pers. Soc. Psychol.*, vol. 39, no. 6, p. 1161, 1980.

[21] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "**A survey of affect recognition methods: Audio, visual, and spontaneous expressions**," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 1, pp. 39–58, 2008.

[22] T. S. C.P. Sumathi and M.Mahadevi, "**Automatic facial expression analysis a survey**," *Int. J. Comput. Sci. Eng. Surv.*, vol. 4, no. 1, pp. 424–435, 2012.

[23] B. McDaniel, S. D'Mello, B. King, P. Chipman, K. Tapp, and A. Graesser, "**Facial features for affective state detection in learning environments**," in *Proceedings of the Annual Meeting of the Cognitive Science Society*, 2007, vol. 29, no. 29.

[24] B. Fasel and J. Luettin, "**Automatic facial expression analysis: a survey**," *Pattern Recognit.*, vol. 36, no. 1, pp. 259–275, Jan. 2003.

[25] T. Kanade, J. F. Cohn, and Y. Tian, "**Comprehensive database for facial expression analysis**," in *Proceedings Fourth IEEE International Conference on Automatic Face and Gesture Recognition (Cat. No. PR00580)*, 2000, pp. 46–53.

[26] H. Monkaresi, N. Bosch, R. A. Calvo, and S. K. D'Mello, "**Automated Detection of Engagement Using Video-Based Estimation of Facial Expressions and Heart Rate**," *IEEE Trans. Affect. Comput.*, vol. 8, no. 1, pp. 15–28, Jan. 2017.

[27] J. Bidwell and H. Fuchs, "**Classroom analytics: Measuring student engagement with automated gaze tracking**," *Behav Res Methods*, vol. 49, p. 113, 2011.

[28] P. Khorrami, V. Le, J. C. Hart, and T. S. Huang, "**A system for monitoring the engagement of remote online students using eye gaze estimation**," in *2014 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, 2014, pp. 1–6.

[29] M. A. Just and P. A. Carpenter, "**Eye fixations and cognitive processes**," *Cogn. Psychol.*, vol. 8, no. 4, pp. 441–480, Oct. 1976.

[30] A. Sharma and P. Abrol, "**Eye gaze techniques for human computer interaction: A research survey**," *Int. J. Comput. Appl.*, vol. 71, no. 9, 2013.

[31] S. R. H. Langton, H. Honeyman, and E. Tessler, "**The influence of head contour and nose angle on the perception of eye-gaze direction**," *Percept. Psychophys.*, vol. 66, no. 5, pp. 752–771, 2004.

[32] E. Murphy-Chutorian and M. M. Trivedi, "**Head pose estimation in computer vision: A survey**," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 4, pp. 607–626, 2008.

[33] W. H. Wollaston, "**Xiii. on the apparent direction of eyes in a portrait**," *Philos. Trans. R. Soc. London*, no. 114, pp. 247–256, 1824.

[34] Z. Liu, J. Zhu, J. Bu, and C. Chen, "**A survey of human pose estimation: The body parts parsing based methods**," *J. Vis. Commun. Image Represent.*, vol. 32, pp. 10–19, 2015.