8-2021

Biometric features modeling to measure students engagement.

Islam Mohamed Ahmed Mohamed Mahmoud Alkabbany
University of Louisville

BIOMETRIC FEATURES MODELING TO MEASURE STUDENTS ENGAGEMENT

Islam Mohamed Ahmed Mohamed Mahmoud Alkabbany
B.Sc., Electrical Engineering, Assuit University, 2007
M.Sc., Electrical Engineering, Assuit University, 2013

A Dissertation
Submitted to the Faculty of the
J. B. Speed School of Engineering of the University of Louisville
in Partial Fulfillment of the Requirements
for the Degree of

Doctor of Philosophy in Electrical Engineering

Department of Electrical and Computer Engineering
University of Louisville
Louisville, Kentucky

August, 2021

BIOMETRIC FEATURES MODELING TO MEASURE STUDENTS ENGAGEMENT

by

Islam Mohamed Ahmed Mohamed Mahmoud Alkabbany
B.Sc., Electrical Engineering, Assuit University, 2007
M.Sc., Electrical Engineering, Assuit University, 2013

A Dissertation Approved on
August 9, 2021
by the following Dissertation Committee:

Professor Aly Farag, Ph.D., Chair/Advisor

Asem Ali, Ph.D., Co-Advisor

Professor Thomas Tretter, Ph.D.

Chris Foreman, Ph.D.

Hongxiang Li, Ph.D.

Roman Yampolskiy, Ph.D.

Nicholas Hindy, Ph.D.

DEDICATION

To my parents, brothers and sisters, my beloved wife, and to my sons

ACKNOWLEDGEMENTS

First, I owe all my gratitude towards God, who bestowed his mercy and blessings upon me at all times. Then, I owe my appreciation to all the people who have made this dissertation possible, and because of whom my graduate experience has been one that I will cherish forever.

I would like to thank my advisor, Professor Aly Farag, for giving me an invaluable opportunity to work on challenging and exciting projects over the past years.

I would also like to thank my co-advisor, Dr. Asem Ali. Without his extraordinary support, this dissertation would not be possible. Thanks are due to Professor Thomas Tretter, Dr Chris Foreman, Dr. Dr Nicholas Hindy, Dr. Hongxiang Li, and Dr Roman Yampolskiy, for agreeing to serve on my dissertation committee and for the valuable feedback they gave me.

I would also like to thank my colleagues at the CVIP laboratory Mostafaa Mohamed and Mohammad Ghanoum for enriching my graduate life in many ways.

I owe my deepest gratitude to my family - my mother, father, brothers and sisters. Words cannot express the gratitude I owe them. Moreover, I have to mention my wife Esraa, who was suffering with me during this long journey while carrying many things on my behalf. Furthermore, special thanks to my little ones Mohamed,

Ahmed and Hamza, whom Allah blessed me with, and they kept playing around me all the time. I want to acknowledge financial support from the NSF.

It is impossible to remember all, and I apologize to those I have inadvertently left out. Lastly, thank you all, and thank God!

ABSTRACT

BIOMETRIC FEATURES MODELING TO MEASURE STUDENTS ENGAGEMENT

Islam Mohamed Ahmed Mohamed Mahmoud Alkabbany

August 9, 2021

The ability to measure students' engagement in an educational setting may improve student retention and academic success, revealing which students are disinterested, or which segments of a lesson are causing difficulties. This ability will facilitate timely intervention in both the learning and the teaching process in a variety of classroom settings. In this dissertation, an automatic students engagement measure is proposed through investigating three main engagement components of the engagement: the behavioural engagement, the emotional engagement and the cognitive engagement. The main goal of the proposed technology is to provide the instructors with a tool that could help them estimating both the average class engagement level and the individuals engagement levels while they give the lecture in real-time. Such system could help the instructors to take actions to improve students' engagement. Also, it can be used by the instructor to tailor the presentation of material in class, identify course material that engages and disengages with students, and identify students who are engaged or disengaged and at risk of failure. A biometric sensor network (BSN) is designed to capture data consist of individu-

als facial capture cameras, wall-mounted cameras and high performance computing machine to capture students head pose, eye gaze, body pose, body movements, and facial expressions. These low level features will be used to train a machine-learning model to estimate the behavioural and emotional engagements in either e-learning or in-class environment. A set of experiments is conducted to compare the proposed technology with the state-of-the-art frameworks in terms of performance. The proposed framework shows better accuracy in estimating both behavioral and emotional engagement. Also, it offers superior flexibility to work in any educational environment. Further, this approach allows quantitative comparison of teaching methods, such as lecture, flipped classrooms, classroom response systems, etc. such that an objective metric can be used for teaching evaluation with immediate closed-loop feedback to the instructor.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: INTRODUCTION

Despite the urgent demand for graduates from Science, Technology, Engineering, and Mathematics (STEM) disciplines, large numbers of U.S. university students drop out of engineering majors [6]. Nearly one-half of students fail to complete an engineering program at the University of Louisville, which is consistent with national retention rates at large, public institutions [7]. This number is even higher for at-risk women, racial and ethnic minorities, and first-generation college students [8]. The greatest dropout from engineering occurs after the first year, following standard gateway mathematics courses such as calculus [9] [10]. Dropout from the engineering major is strongly associated with performance in first-year mathematics courses [9]. Part of the difficulty, not limited to engineering, is the transition from secondary to college education in mathematics. Students often retain and apply only surface-level knowledge of mathematics [11].In addition, socio-psychological factors, such as perceptions of social belonging, motivation, and test anxiety, predict first-year retention [12] [9] [13] [14].

Thus, a plethora of research indicates that engagement at emotional, behavioral, and cognitive levels is a predictor and problem for retention in engineering. Student engagement contributes to higher grades, higher state assessment scores,

and better school conduct [15]. Suppose students are not engaged in the learning process inside the classroom. In that case, they are unlikely to obtain the skills necessary to successfully move on to the next level of education or into the global workforce [16]. The measurement of students' engagement in an educational setting may also provide essential information on how to improve student retention and academic success [17] [18] [19] [20].

Currently, feedback on student performance relies almost exclusively on graded assignments, with the in-class behavioral observation by the instructor a distant second. Performing the in-class observation of engagement by the instructor is problematic because he/she is primarily occupied with delivering the learning material. Indeed, adaptive learning environments allow free-form seating, and the instructor may not be able to have direct eye contact with the students. Even in traditional classroom seating, an instructor would not be able to observe a large number of students while lecturing. Therefore, it is practically impossible for the instructor to watch all students all the time while recording these observations per student and correlating with the associated material and delivery method. Moreover, these types of feedback are linked to the in-class environment. In an e-learning environment, the instructor may lose any feedback to sense student engagement. Performance on assignments can also be ambiguous. With some students deeply engaged yet struggling while other students are only minimally engaged, both groups end up with poor performance. Other students may manage good performance while lacking a deeper understanding and reflection of the material, e.g., merely studying to memorize an exam without engagement in the learning process.

One of the significant obstacles to assessing the effect of engagement in student learning is the difficulty of obtaining a reliable measurement of engagement. Using barometric sensors (such as cameras, microphones, heart rate wristbands sensors, and EEG devices) is more dynamic and objective approach for sensing. This dissertation focuses primarily on measuring the emotional and behavioral components of the engagement as well as on designing a biometric sensor network and technologies for modeling and validating engagement in various class setups.

1.1 The Dissertation Contribution

The main contribution of this study is to:

- Design a biometric sensor a biometric sesnor network and algorithms to be used to measure student engagement.
- Develop robust models of facial information for describing human engagement within an educational environment. In particular, at the behavioral level, where gross body, hand, and head movements, as well as eye-blinking and eye-gaze, are used as indicators of attention, and at the emotional level, where expressions correspond to muscle movements pertaining to attention.

Although our study has been conducted on a control set of students, this work has a broad impact. This research can be further extended to students with special needs. By detecting disengagement, future research may use this tool to develop an early-warning system to detect student anxiety and depression.

The remaining of the theses is organized as follow:

- Chapter two discusses Student Engagement problem, and the relation between the different component of engagement. It also reviews the research conducted in that area.

- Chapter three discusses the behaviour engagement in e-learning environment, it addresses the measured metrics to a classify the behaviour engagement.

- Chapter four discusses the behaviour engagement on the in-classes environment. It addresses the challenges on the in-class environment, and the extra metric that help in classifying the behaviour engagement.

- Chapter five discusses the emotional engagement, It addresses the measured metrics to a classify the emotional engagement.

- Chapter six introduces the proposed experiment and results on automatically measures both behaviour and emotional engagement frameworks.

- Chapter seven summarize the thesis conclusion and discuss the future works .

CHAPTER 2: STUDENT ENGAGEMENT IN LEARNING ENVIRONMENT

The three components that comprise student engagement are behavior, emotion, and cognition [21]. These components work together to fully encompass the student engagement construct, and each component has been found to contribute to positive academic outcomes (e.g., [21], [22]).

Behavioral engagement consists of the actions that students take to gain access to the curriculum. For example, behavioral engagement is measured in the classroom by self-directive behaviors, inattentive actions, and not participating cooperatively in class activities [23], [24]. Measures for behavioral engagement are correlated with school attendance and participation in extracurricular activities [25], and preparation for the class, including homework completion [26]. Although some manifestations of behavioral engagement include actions that are not physically observable within the classroom environment (e.g., completing homework), other behavioral actions as exhibited by particular postures (e.g., closed rather than open) [23], [24], [27] and fidgeting [28] can potentially be quantified. Once students engage behaviorally, they can be emotionally engaged with their learning.

Emotional engagement is broadly defined as how students feel about their

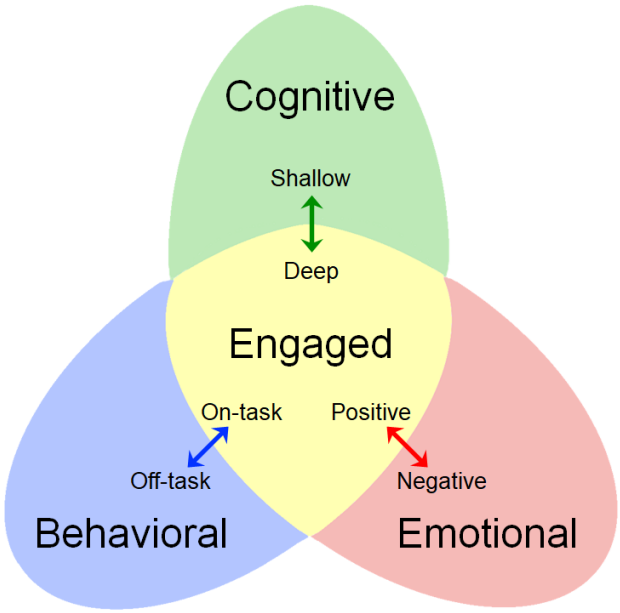


Figure 2.1: A conceptual framework linking on-task/off-task behavioral, positive/negative emotional, to deep/shallow cognitive engagement.

learning [29], learning environment (e.g., [30]), and instructors and classmates (e.g. [25], [21]). More specifically, measures of emotional engagement include expressing interest and enjoyment; reporting fun and excitement; reacting to failure and challenge; feeling safe; perceiving school as valuable, and expressing feelings of belonging [26]. Emotional engagement includes activities that display the "care" students have for their education and for the curriculum they have accessed [30].

Finally, the cognitive component is observed when students embrace the learning process, which leads to academic success outcomes (e.g., [21], [31]). In other words, Cognitive engagement is the mental investment in academic achievement, including the use of deep rather than superficial learning processes to self-regulate and persist in understanding the material (e.g., [32]).

Table 2.1 provides a summary of the psychological constructs[2, 3, 4, 5]for

Table 2.1: Psychological Constructs for the Three Types of Engagement.

TYPE OF ENGAGEMENT	**COGNITIVE**	**BEHAVIORAL**	**EMOTIONAL**
PSYCHOLOGICAL CONSTRUCT	Levels of processing [8] [7]	Targets of attention [9]	General activation systems [10]
ENGAGED STATE	Deep processing	On-task attention	Positive affect
DISENGAGED STATE	Shallow processing	Off-task attention	Negative affect

the three types of engagement, which would be used to devise a computational counterpart

Figure 2.1 describes the interrelationship between the three forms of engagement which will be quantified in this dissertation. and Table 1 provides a summary of the psychological constructs [8] [7] [9] [10] , for the three types of engagement.

The interrelationship between the three forms of engagement culd be modeled as stochastic process Fig.2.2. basically it could be summarized to three states :

- Not engaged
- Behaviorally engaged
- Emotionally engaged

All the students will start initial as not engaged and them state during the lecture will be tracked according to this model, (A,B,C,D,E) are the list of actions that be used by our automated engagement classifiers to detect an engagement state change. The students initial state is not engaged. The level of them behavioral engagement will be measured according to some metrics such as eye-gaze, head-pose , body-pose.

This behavioral engagement level will be used as a trigger to measure the emotion engagement.

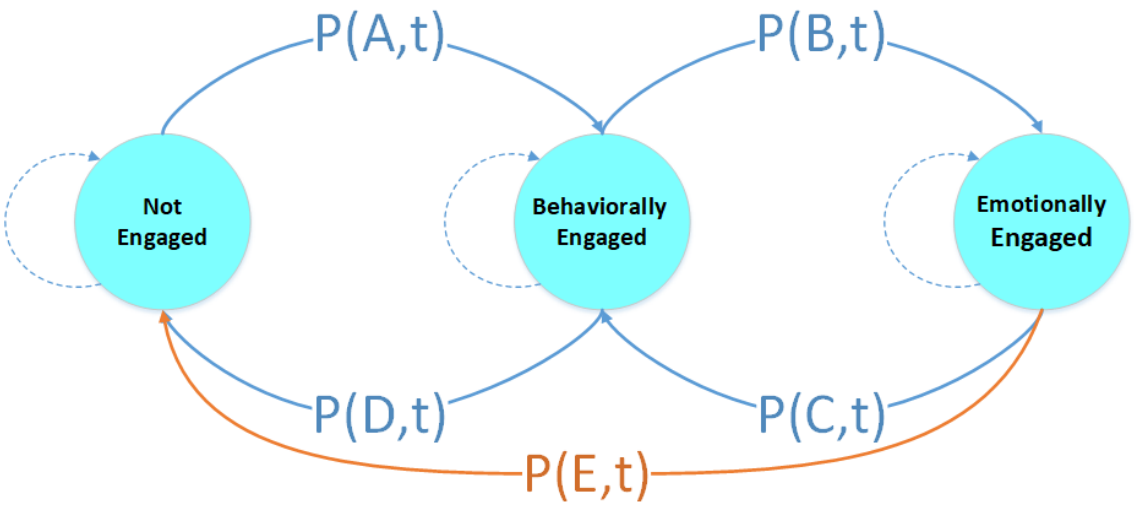


Figure 2.2: Engagement Model states diagram.

2.1 Literature Review

The education research community has developed various taxonomies describing student engagement. After analyzing many studies, Fredricks, et al. [21] organized engagement into three categories. Behavioral engagement represents the student's willingness to participate in the learning process. Emotional engagement refers to a student's emotional attitude towards learning. Cognitive engagement describes learning in a way that maximizes a person's cognitive abilities. The two former engagement categories can be easily sensed and measured.

Despite the advances in machine recognition of human emotion, there have been a small number of studies of facial expressions related to learning-centered

cognitive-affective states. Computer vision methodology can unobtrusively estimate a student's engagement from facial cues, e.g., [33–37] Such studies apply one or more of the following paradigms. Observation and annotation of affective behaviors, investigation of facial action units involved in learning-centered effect, and application of automated methods to detect affective states. Kapoor and Picard [33] used a camera equipped with IR LEDs to track pupils and to extract other facial features: head-nod, head-shake, eye blinks, eye and eyebrow shapes, and mouth activities. Also, a sensing chair is used to extract information about the postures. Moreover, they recorded the action that the subject is doing on the computer. Then a mixture of Gaussian processes combines all the information and predicts the current affective state. In their study, 8 Children (8 - 11 yrs) are enrolled. Children were asked to solve puzzles on a computer. For 20 minutes, the screen activity, side-view, and frontal view were recorded. From the collected videos, 136 clips are extracted (up to 8 secs long). Teachers were asked to observe and record the affective state at eight samples per second. The affective states under consideration are high, medium, and low interest, boredom, and "taking a break.". The recognition rates of an interest vs uninterest SVM classifier (for 65 interest samples and 71 uninterest 71 samples) are 69.84% (using upper face information) and are 57.06% (using lower face information). They got 86.55% recognition rate by combining all information, not only the facial features, using a mixture of Gaussian processes.

To detect the emotions that accompany deep-level learning, McDaniel et al. [34] investigated facial features. The affective states under consideration are boredom, confusion, delight, flow, frustration, and surprise. To perform their study,

they asked 28 undergraduate students to interact with AutoTutor. First, participants completed a pretest. Then, videos of the participants' faces were captured while interacting with the AutoTutor system for 32 minutes. Finally, they completed a posttest. After that, the affective states annotation was done by the learner, a peer, and two trained judges. The ground truth of the data is obtained from the trained judges, who have interjudge reliability Cohen's kappa (0.49). After that, the data was sampled to 212 emotion video clips (3-4 sec) with affective states: boredom, confusion, delight, frustration, and neutral. Finally, two trained coders coded participants' facial expressions using Ekman's Facial Action Coding System. They computed correlations to determine the extent to which each of the AUs was diagnostic of the affective states of boredom, confusion, delight, frustration, and neutral. Their analyses indicated that specific AU's could classify confusion, delight, and frustration from neutral, but boredom was indistinguishable from neutral.

In order to study the learning-centered effect, Grafsgaard et al. [36] used an automated facial expression recognition tool to analyze videos of computer-mediated human tutoring. They collected a dataset of 67 undergraduate students who are learned an introductory engineering course using JavaTutor software. Participants took six sessions of 45 min. Each session started with a pretest, then the teaching session, post-session surveys, and finally posttest. During the teaching session, database logs, webcam facial video, skin conductance, and Kinect depth video were collected. Two trained coders coded participants' facial expressions using Ekman's Facial Action Coding System to annotate the data. They recorded the five most frequently occurring AUs (1, 2, 4, 7, and 14). The authors used the CERT toolbox

[38] to extract these 5 AU's automatically. Also, they computed the normalized learning gain from the posttest and the pretest scores. They claimed the following conclusions: outer brow raise (AU2) was negatively correlated with learning gain. Brow lowering (AU4) was positively correlated with frustration. Mouth dimpling (AU14) was positively correlated with both frustration and learning gain. Also, facial actions during the first five minutes were significantly predictive of frustration and learning at the end of the tutoring session.

Recently, Whitehill et al. [37] introduced an approach for automatic recognition of engagement from students' facial expressions. They claimed that human observers reliably agree when discriminating low versus high degrees of engagement (Cohen's k = 0.96). This reliability decreases to (k = 0.56) for 4 distinct levels of engagement. Also, they claimed that static expressions contain the bulk of the information used by observers, not the dynamic expressions. This claim means that engagement labels of 10-second video clips can be reliably predicted from the average labels of their constituent frames (Pearson r = 0.85). They collected a dataset of 34 undergraduate students who trained using cognitive skills training software. Each session started with an explaining video (3 min), then a pretest (3 min), a training video (35 min), and finally a posttest. The participant's face was recorded during the training. To annotate the data, the video frames are coded by seven labelers using a scale to rate the engagement: 1: Not engaged, 2: Nominally engaged, 3: Engaged in the task, 4: Very engaged, and X: unclear frame. Then 24285 frames were selected such that the difference between any two labelers doesn't exceed one, and no labeler assigned X to the frame. The "ground truth" label of a frame is

the integer average of all labels. Gabor features were extracted from the detect face to generate a 40x48x48 feature vector. Then four binary SVM classifiers were used to detect a level out of the four levels of engagement. Finally, a multinomial logistic regressor was used to combine the output of the four binary classifiers. They claimed that automated engagement detectors perform with comparable accuracy to humans.

Li and Hung [39] report enhancement of student engagement by the fusion of facial expressions and body features. Fusion of more disparate data can also enhance engagement measures, such as video facial expression with wristband heart rate data [40] by Monkaresi et al., and posture with electrodermal activity data fusion [41]. The use of context was explored by Dhamija and Boult [42] in the area of online trauma recovery, and they and others have found significant evidence [43] [44] [40] [37] that facial expression estimation of engagement was nearly universal. Additional work by Svati and Boult [45] explored the influences of mood awareness on engagement classification, where the mood is the prevailing state of emotion independent of the current task, e.g., classroom learning. Emotion affects the domain in which facial expressions and other biometrics are collected, and the understanding of how emotion affects engagement serves to fine-tune the use of these biometrics.

Ahuja and et al. introduced a framework to sense a set of engagement-related features (EduSense) [3]. They extract facial landmarks and use them to find facial features such as head pose and smile detection. They also perform body segmentation and body keypoints extraction. Then use this to extract features such as

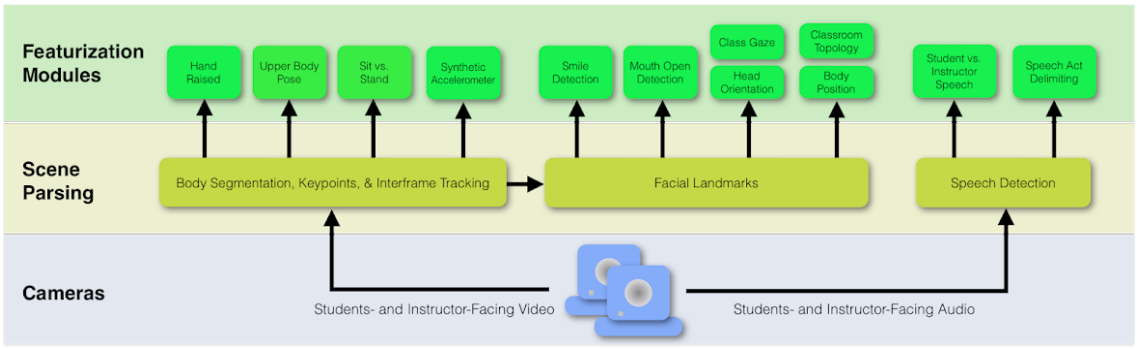


Figure 2.3: Edusense frameworks and feature [3].

detection of hand raise and sit vs. stand detection. Furthermore, they perform speech detection to find the ratio between instructor speech time to student speech time. Fig. 2.3 show the introduced framework and the extracted features

In [46] Ahuja, and et al. used two RGB cameras to extract student and instructor head pose, then they use these features to estimate heatmap of where students gaze.

CHAPTER 3: BEHAVIORAL ENGAGEMENT IN E-LEARNING ENVIRONMENT

Behavioral engagement consists of the actions that students take to gain access to the curriculum. These actions include self-directive behaviors outside of class, such as doing homework and studying, as well as other activities, such as shifting in the seat, hand, body, or other subs/conscious movements while observing lectures. Also, participating cooperatively in-class activities [23] [24].

Head pose and eye gaze are the main metrics to measure the students' behavioral engagement. By estimating the student point of gaze, it could be told if they are looking to be engaged with the lecture or not. If the student looks to his laptop, he is probably highly behaviourally engaged. While if he looks to other points, he is probably not engaged.

3.1 Facial Metric

3.1.1 Facial Landmark

The first step to obtaining the proposed facial metric is to extract the facial landmark. Facial landmarks are mainly located around facial components such as

eyes, nose, and mouth. Facial landmarks allow us to align faces for various tasks; they also help finding the head pose, eye gaze, and facial expression.

The facial landmarks detector [47] combines a part-based model and holistic face information. OpenFace 2.0 [48] uses a Convolutional Experts Constrained Local Model (CE-CLM) [49] for facial landmark detection and tracking. This module consists of two main components: 1- Point Distribution Model (PDM), which captures landmark shape variations. 2- patch experts which model local appearance variations of each landmark.

Figure 3.1 shows the extracted 68 facial landmark.

Figure 3.1: 68 Facial landmark.

3.1.2 Head Pose

Head pose estimation is to find both the head in-plane and out off-plane rotations, see Fig3.2 This estimation could be formulated as a perspective n point problem (PnP) [50]. After obtaining the 68 facial landmark points. And given a 3D face model with its 3D landmark known. Then this problem can be solved as Perspective-n-Point problem.

$$sp_c = K[R|T]P_w \tag{3.1}$$

Where p_c is the image 2D point, P_w is the world 3D point, K is the matrix of intrinsic camera parameters, while s is the scale R is the 3D rotation matrix, and T is the 3D translation matrix which represents the extrinsic camera parameters.

OpenFace 2.0 [48] take advantage of using CE-CLM, which uses a 3D representation of facial landmarks and projects them to the image using orthographic camera projection, which allows the framework to estimate the head pose accurately once the landmarks are detected. The resulting head pose could be represented in 6 degrees of freedom (DOF) (3 degrees of freedom of head rotation [R] - yaw, pitch and roll - and 3 degrees of translation [T] - X, Y, and Z)

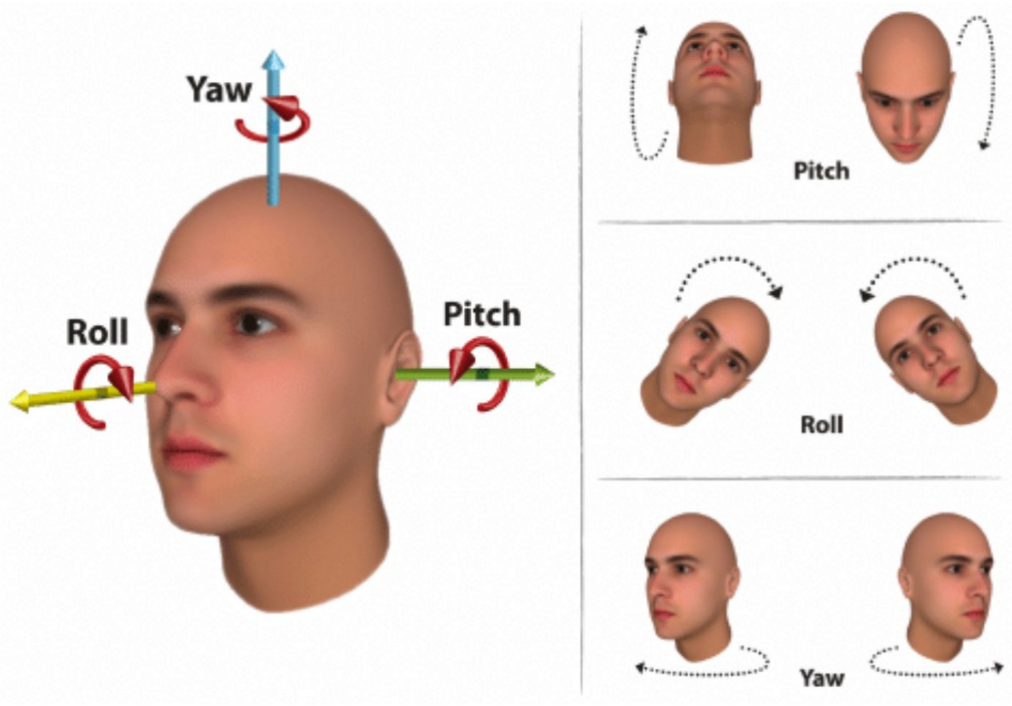


Figure 3.2: Head pose.

3.1.3 Eye Gaze

Eye gaze tracking is the process of measuring either the point of gaze or the motion of an eye relative to the head. The eye gaze could be represented as the vector from the 3D eyeball center to the pupil. Various works use eye tracking to find out student behavioral engagement [51] by either using either special hardware devices or regular RGB cameras with the help of software algorithms. Hardware devices mainly use Near-infrared light, which is directed towards the eyes pupil, causing detectable reflections in both the pupil and the cornea. These reflections – the vectors between the cornea and the pupil – are tracked by an infrared camera. This optical tracking of corneal reflections, known as pupil center corneal reflection (PCCR), is shown in Fig. 3.3.

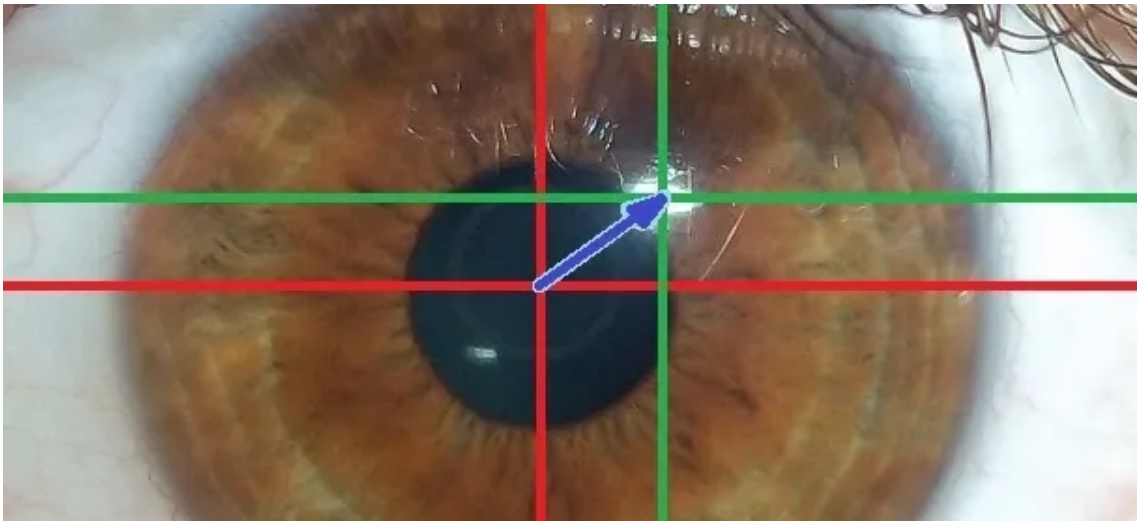


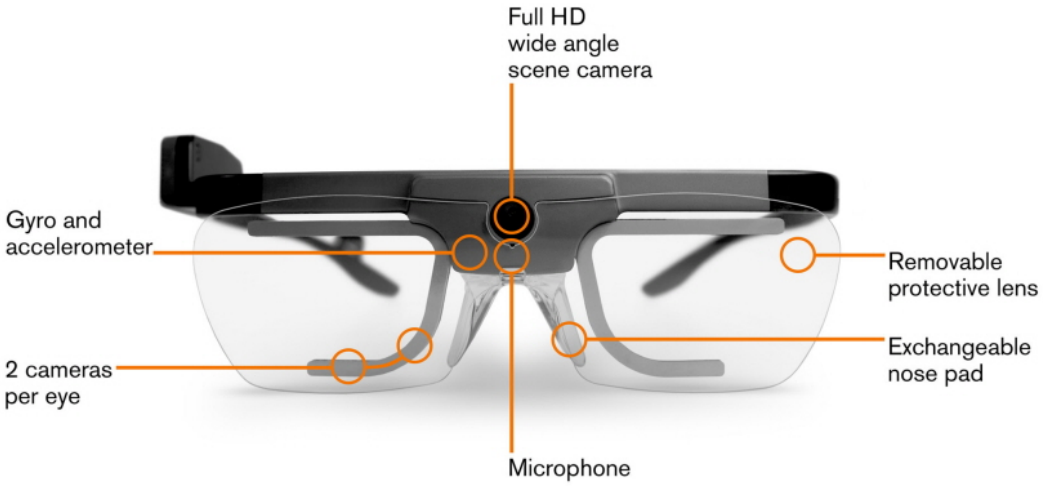
Figure 3.3: Pupil center corneal reflection (PCCR).

There are two types of hardware eye tracker, screen-based eye tracker, which is usually a bar attached to the screen containing IR source and camera.This type is used in stationary setup Fig. 3.4a. The other type is eye-tracking glasses, in which the IR camera and sensor are attached to the glasses frame, and it allows the subject to move freely Fig. 3.4b.

In order to estimate the eye gaze using the software approach, the eyelids, iris, and the pupil are detected using [52]. The detected pupil and eye location are used to compute the eye gaze vector for each eye.A vector from the camera origin to the center of the pupil in the image plane is drawn, and its intersection with the eye-ball sphere was calculated to get the 3D pupil location in world coordinates.

Openface 2.0 [48] estimate eye gaze individually for each eye by using a Constrained Local Neural Field (CLNF) landmark detector [53] [54] to detect eyelids, iris, and the pupil. They obtain the pupil location in 3D camera coordinates by firing a ray from the camera origin toward the center of the pupil in the image

(a) Screen-based eye trackers



(b) Eye tracking glasses

Figure 3.4: Type of eye tracking hardware.

plane and compute its intersection with the eyeball sphere. The vector from the 3D eyeball center to the pupil location is the estimated gaze vector.

3.2 The Proposed Behavioral Engagement Framework For E-learning Environment

In this section, Novel framework for automatic measurement of the behavioral engagement level of students in the e-learning environment is proposed. The proposed frameworks capture the user's video using a regular webcam; it tracks their faces through the video's frames. Different features are extracted from the user's face, e.g., facial landmark points, head pose, eye gaze, as shown in Fig. 3.5.

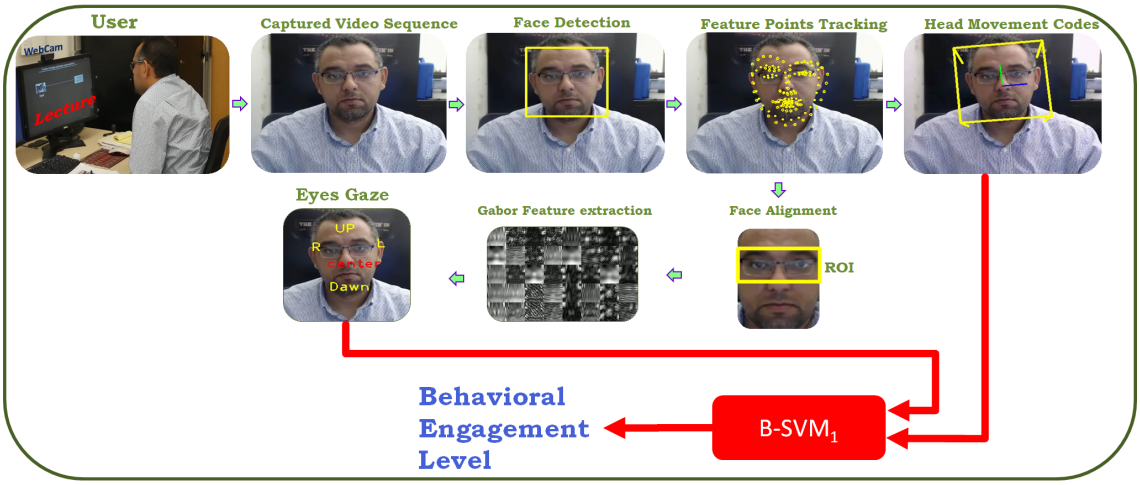


Figure 3.5: The proposed e-learning behavioral engagement framework.

To extract and track the ROI through frames, pipeline of cascade algorithms are applied. Fig. 3.6 illustrates the block diagram of this framework. First, a face is detected using the face detection algorithm based on the Viola-Jones face detector and its implementation in the OpenCV library. Since this algorithm detects many face candidates, largest detected candidate is selected . This selection is appropriate to the camera setup where a single client is in front of a web camera. A skin detector

is used To reduce the false-positive faces. It measures the skin ratio in the detected candidate face.

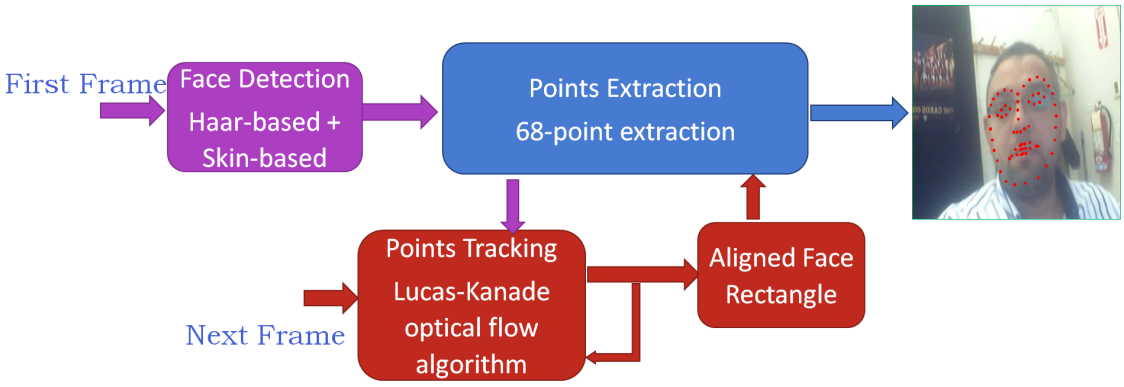


Figure 3.6: Face and facial points tracking framework.

After the face is detected, 68 facial feature points are extracted using the approach in [47]. This approach's performance depends on a well-trained model. The current model is trained on the multiview faces 300 Faces In-the-Wild database [55]. Then, the facial image is aligned by transforming these landmarks to a common space to eliminate the in-plane rotation. next, a region of interest (eye) is cropped to 100×32. A bank of 40 Gabor filters is applied to the ROI to extract the feature, which is used to train an RBF–based multivariate SVM classifier. The classifier gives the probabilities of the eye pupil is looking at the frontal, up, down, left, and right.Columbia Gaze Dataset [56] and RaDF [57] are used to train the eye gaze SVM. A cross-validation experiment is conducted using 123 training and 123 testing sets constructed from 1888 images of these databases, then the recognition results are obtained as shown in Table 3.1 .

The 68 facial landmarks are used along with a 3D face model from [56] to

Table 3.1: Eye gaze classification confusion Matrix

	frontal	Right	Left	Up	Down
frontal	**83.66**	2.27	3.84	4.26	5.97
Right	4.17	**92.44**	0.0	1.23	2.16
Left	4.78	0.15	**91.51**	1.54	2.01
Up	27.98	3.57	4.76	**57.14**	6.55
Down	26.79	5.95	4.17	1.79	**61.31**

estimate the head pose by solving this Perspective-n-Point problem. To evaluate the face tracking and the head pose estimation approaches, the proposed pipeline was applied , which consists of these two approaches, on different datasets:

(a) Head Pose Database [58], which has 120 videos for 10 subjects. Each video has 300 frames

(b) Boston Univ. Head-Tracking dataset [59], which has 72 sequences and the sequence is 200 frame

(c) Head Pose and Eye Gaze (HPEG) Dataset [60], which has 10 subjects, are captured at two sessions. Each has 20 video sequences of 200-400 frames.

(d) Pointing '04 Head Pose Image Database [56]. 15 image galleries related to 15 different persons. Each gallery contains two sequences of 93 face images.

(e) Columbia Gaze Data Set [61]. Contains 5,880 images of 56 subjects,21 subjects wore prescription glasses.

The pose estimation's Mean Absolute Error (MAE) and its standard deviation are computed w.r.t. ground truth as shown in Table 3.2.

Table 3.2: Mean Absolute Error (MAE) and the error's deviation in degrees of the three rotations angels for different datasets

Database	Roll	Yaw	Pitch
(a)Head Pose [58]	0.6 ± 0.8	2.2 ± 2.6	1.9 ± 2.1
(b)Boston Univ. Head-Tracking [59]	2.2 ± 2.2	4.8 ± 4.5	3.3 ± 3.4
(c) HPEG [60]	-	5.6 ± 4.5	4.2 ± 3.8
(d)Pointing '04 Head Pose Image [56]	-	8.9 ± 7.8	15.3 ± 10.5
(e)Columbia Gaze [61]	1.3 ± 1.1	6.8 ± 5.4	1.3 ± 1.2

The extracted features; head pose, and eye gaze are used to fed a support vector machine (SVM) to classify the students behavioral engagement level.

As improvement for this framework, OpenFace 2.0 [48] platform could be used to extract the head pose and eye gaze. The eye gaze is provided for each eye as a 3D vector from the eyeball center to the pupil center. And the head pose is provided as 3D translation and 3D Rotation. Then this extracted head pose and eyes gaze could be fed to the behavioural engagement SVM classifier.

3.3 Conclusion

This chapter proposed a framework to measure the students behavioural engagement level,The proposed framework could be implemented to run either offline or at the Client/Server model. In some settings, such as a student watching tutorial or online lecture, the framework modules are implemented on one machine. In contrast, in the case of e-learning, this framework is implemented as a Client/Server program. The algorithm will track the face and extract the eye pose and head gaze at the client-side and send this small feature vector to the server-side. The server

will receive thus features from multiple students simultaneously, then calculate each student's behavioral engagement level. It also could find the average level for the whole class.

CHAPTER 4: BEHAVIORAL ENGAGEMENT IN CLASS ENVIRONMENT

Estimating the behavioral engagement in the class environment is more complicated than in the e-learning environment. Rather than presence of only one target of interest (laptop screen) in the case of e-learning, there are multiple targets of interest in the class environment. The student may look at the instructor, the whiteboard, projector screen, or even one of his/her peers. Therefore the framework should track where each student gaze and also where their peer gaze. Then relates them together to estimate the student behavioral engagement level

Other metrics such as students body pose and body actions also must be taken into consideration while estimating the behavioral engagement level.

4.1 Body Metric

4.1.1 Body Pose

Body pose estimation is the process of identifying the body posture of a person by estimating the human body's Key-Points (joints) such as shoulders, elbows, and wrists in videos or images. Then it can indicate and track a person's various postures

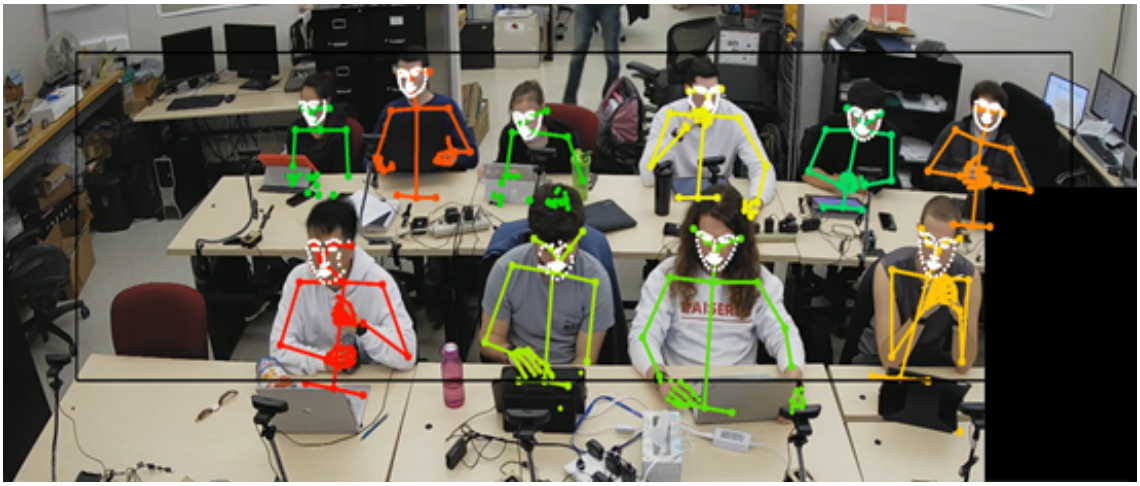


Figure 4.1: Body Pose

by connecting the related points,see Fig.4.1. The human pose estimation problem is not only to find the human body joints but also is to register them correctly to assemble the human skeleton. This process may be challenging, especially in the case of the crowd or when part of the body is occluded. Human body keypoints estimation has been an interesting point of research for decades. recent research investigated extracting human pose for single and multi-person in-the-wild which include body, foot [62] [63] [64] [65], and hand keypoints [66] [67].

The estimation of Human body pose could be performed using special hardware or by using a regular RGB camera and algorithms. Azure Kinect (Fig. 4.2) has a 12-megapixel RGB camera supplemented by one megapixel-depth camera for body tracking. It uses the Bottom-up approach on IR images to estimate the body pose. It obtains the body joints heat map, part affinity field, and part segmentation map. Then it uses them to estimate 3D skeletons.

Cao and et al. introduced Openpose [4] framework to extract whole body pose

Figure 4.2: Azure Kinect

in-the-wild. It detects the skeleton (body, face, hand, and foot keypoints) in 2D for multi-person. It also can estimate the 3D skeleton in the case of a single person. In order to detect the parts, Openpose [4] extract the part confidence map, then it uses art Affinity Fields (PAFs) to perform part association to form the full-body poses. In the case of multi-person, it performs non-maximum suppression on the detection confidence maps to obtain a discrete set of part candidate locations. The pipeline of the Openpose [4] framework is shown in Fig. 4.3.

4.1.2 Body Action

Land and Harris [68] conducted unstructured observations of several large classes to study the patterns of student action that would define behavioral engagement. They defined a set of actions to represent engaged behavior such as listening

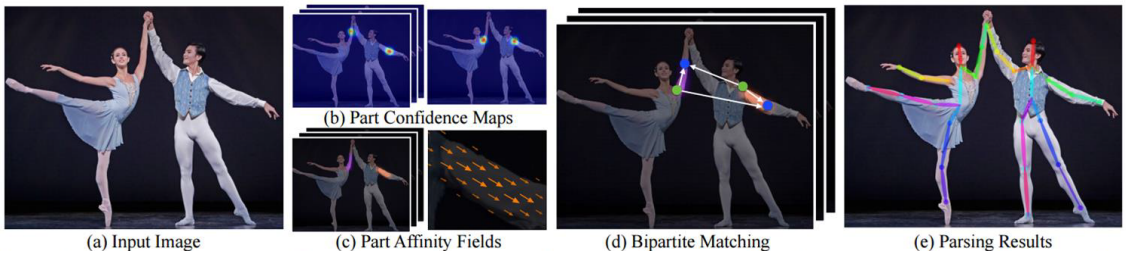


Figure 4.3: Openpose [4] pipeline

while eye contact focuses on instructor or activity, writing notes, read material related to the class, engaged laptop usings like take notes or read class material, students interaction with instructor to ask or answer question, and student interact with peers in a discussion relates to class material.

They also defined a set of actions to represent non-engaged behavior such as settling in the lecture (finding a seat, download material, organize notes), packing up (pack the notes), unresponsive like an eye is closed, off-task like working in homework or study other courses, a disengaged laptop useings such as web browsing or watching a video or playing a game, student interact with peers in discussion not related to course material, and distraction by other students. Tabassum and et al. [69] also made a similar study and also ended up with a very similar set of student patterns that define behavioral engagement.

4.1.3 Proposed Set of Actions

A study has been conducted to find the patterns which define behavioral engagement in undergraduate classes. Table 4.1 shows the proposed metrics to define

low behavioral engagement patterns and the token that could be used to measure this pattern, while Table 4.2 shows the proposed metrics to determine high behavioral engagement patterns and the token that could be used to measure this pattern

Table 4.1: Low behavioural engagement patterns.

Pattern	**Measurable token**
Eyes frequently moving from place-to-place (e.g. 5 or more in 2-minute frame)	Eye gaze/head pose changes $\geq$ 3 times/min
Hands frequently moving from place-to-place (e.g. 5 or more in 2-minute frame)	Arms pose changes $\geq$ 3 times/min
Engaging with phone	Eye gaze/head
Looking at tablet on off-task	Eye gaze/head with hand pose
Engaged with peers in discussion not related to course material	Eye gaze/head towards peer
Staring in a direction without instruction	Head pose/Eye gaze away from target

Table 4.2: High behavioural engagement patterns.

Pattern	**Measurable token**
Eyes consistently focused on instruction	Eye gaze/head pose
Writing notes	Arms pose
Looking at tablet and make instructed task	Eye gaze/head with hand pose
Engaged with peers in discussion related to course material	Eye gaze/head towards speaker

4.2 The Proposed Behavioral Engagement Framework For in-class Environment

On this framework, two sources of video streams were used. The first source is a wall mount camera that captures the whole class, while the second source is dedicated webcams in front of each student. The proposed pipline is shown on Fig. 4.4 The first step in the framework is to apply OpenFace [48] algorithm on each stream to extract the pose/gaze features. The wall-mounted camera provides the head pose only as the faces size is too small to get accurate eye gaze from it, while the students' cameras provide us with both head poses and eye gazes. Each camera provides the output in its world coordinate. Therefore the second step is to align all the camera's coordinates to get all students' head poses and eye gazes in a common world coordinate. Given a well known class setup, the target planes could be found as a one-time pre-calibration for the class. The intersections of the students' head pose/eye gaze rays and the target planes are calculated. To eliminate noise, the feature was combined within a window of time of size T. Then the mean point of gaze could be found on each plane in addition to the standard deviation for each window of time. The plane of interest in each window of time is the one with the least standard deviation of students gaze. For each student, the student pose/gaze index could by calculated as the deviation of his gaze points from the mean gaze point in each window if time. This index is used to classify the average student behavioral engagement within a window of time.

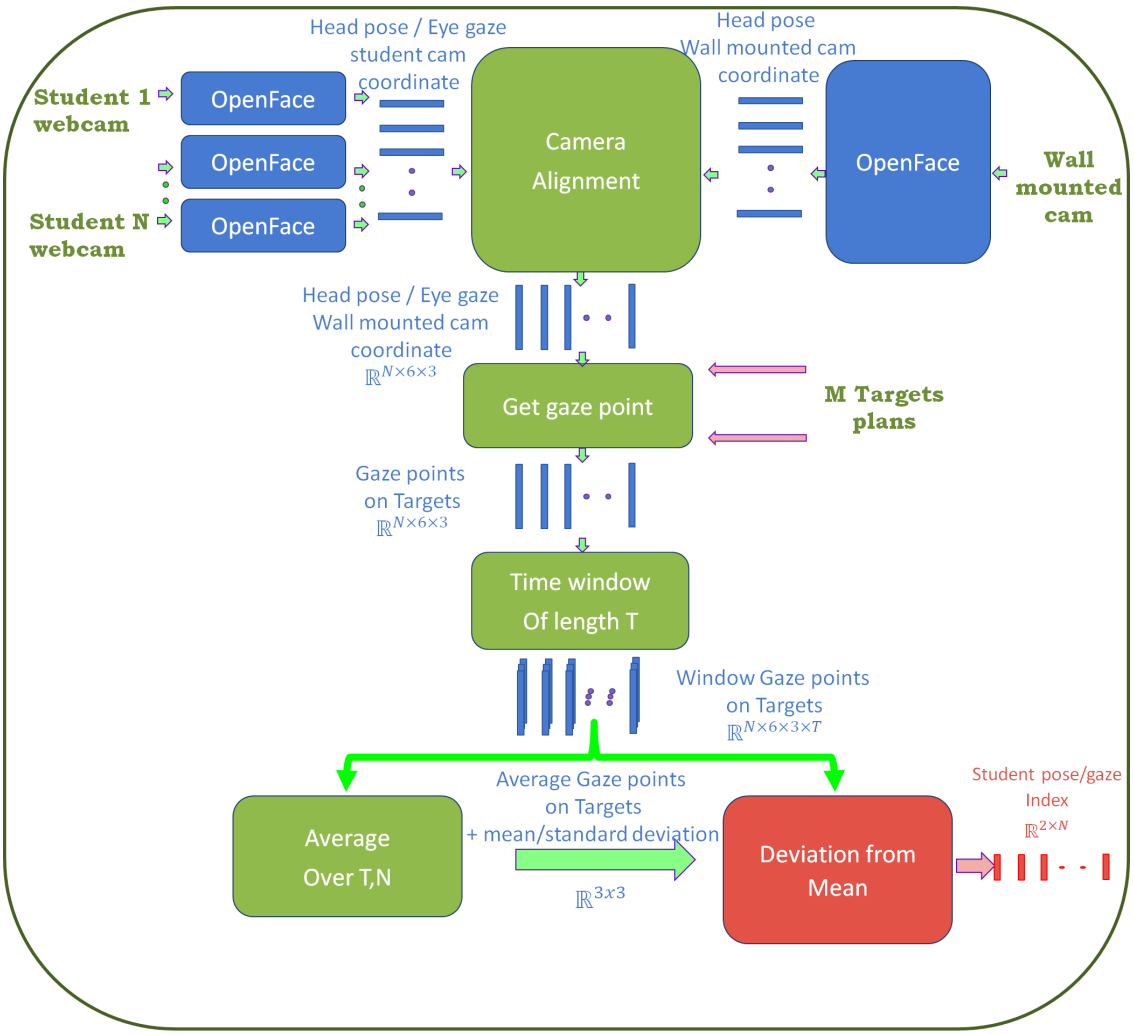


Figure 4.4: In the class behavioral engagement framework.

In order to enhance the behavioral engagement estimation, the framework was extended to have a body action-based feature. Figure 4.5 shows the body actions detection module. The stream from wall mounted camera is used to estimate the student pose by using OpenPose [4] framework. OpenPose provides 21 body joints + 70 facial landmarks + 25 joints per hand. The proposed modules neglect the lower body and foot joints, as in the class environment the students lower body parts are occluded by the benches ,see Fig. 4.1. Therefore, The total length of the

used feature per student is 121 (9 body joints + 70 facial landmarks + 21 x2 hand joints). Next, each student skeleton is aligned and normalized so that the relative location of the student does not affect the module decision. Thereafter, the module combines the feature within a window of time of size T, and a long short term memory (LSTM) network is trained to classify the desired actions. To train that LSTM network NTU RGB+D [70] dataset are used. The dataset [70] contains 60 different action classes, including daily, mutual, and health-related actions. Among them, ten actions related to in the class behavioral engagement was chosen for the proposed framework. The dataset was collected from 40 subjects performing the action in two different trials in front of 3 cameras; these cameras were located at the same height but from three different horizontal angles $(-45°, 0°, +45°)$. The dataset is collected in 17 different setups; in each, the height of the cameras is changed. The data set provides RGB stream, depth map from IR camera, and 3D/2D skeleton. To be consistent with the class environment,the proposed body action detection module pipeline was run to extract the body pose features. 30 subjects were used for training process and 10 for validation. The proposed module obtain an accuracy of %84 in body action classification.

Figure 4.6 show sample of the selected actions. Some actions such as eating, drinking, and play with phone imply that the student is not behaviorally engaged . On the other hand, Some actions such as typing on keyboard, writing notes, and pointing imply that the student is behaviorally engaged

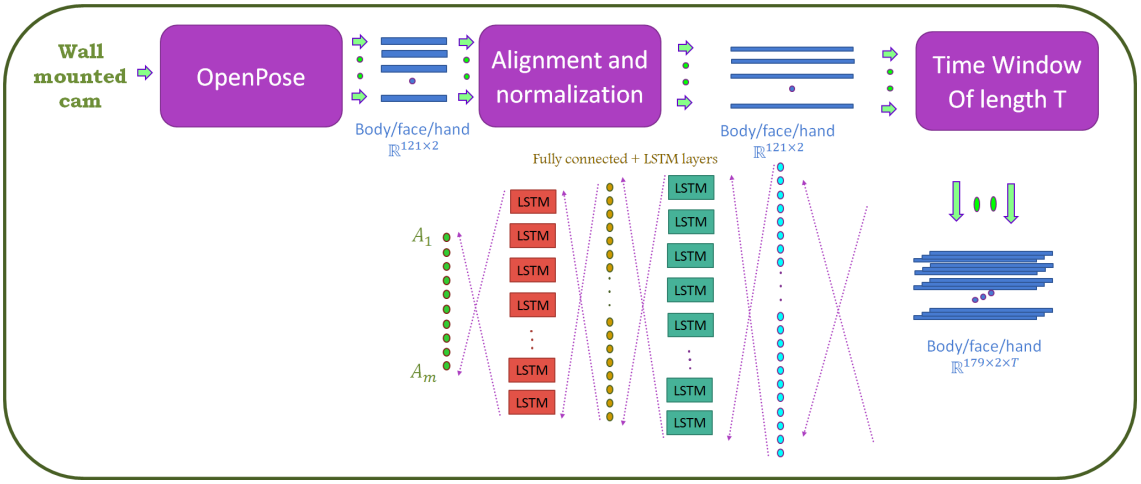


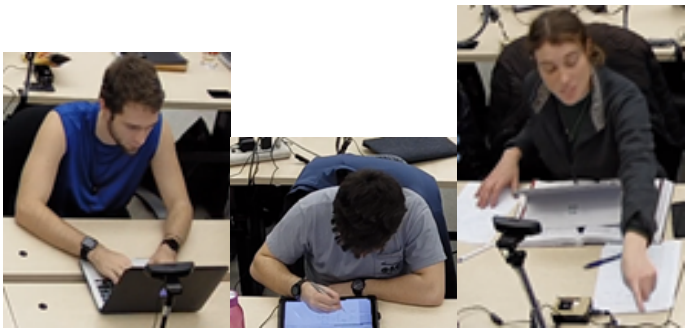
Figure 4.5: Body action detection module.

4.3 Conclusion

This chapter proposed a framework to measure the students behavioural engagement level on the in-class environment,The proposed framework has multiple source of video streams. Beside track the face and extract the eye pose and head gaze for all students, it will find the mean point of interest on multiple target and evaluate the attention of each student separately depend of his/here divergent from that mean. It also detects the students body pose and use a series of the change in body pose to recognize the body actions.

Checking time Drink Eat Play with phone

(a) Actions imply lower behavioral engagement level



Type on keyboard Taking notes Pointing

(b) Actions imply lower behavioral engagement level

Figure 4.6: Sample of the selected Action

CHAPTER 5: EMOTIONAL ENGAGEMENT IN LEARNING ENVIRONMENT

Emotional engagement is broadly defined as how students feel about their learning [29], learning environment [71] and instructors and classmates. Emotions include happiness or excitement about learning, boredom or disinterest in the material, or frustration and struggling to understand [25]. This chapter discuss the metrics to classify facial emotions, the main metric facial muscle movements. The chapter discuss the effect of facial muscle moment on facial geometry. It also discuss how to detect this muscle movements using stream of video for the students faces, and how this could be used to classify the students emotional engagement.

5.1 Effect of Muscle Contractions on The Facial Geometry

The human face has many muscles whose contractions constitute facial expressions. Figure 5.1, which is generated using ARTNATOMY tool [72], illustrates different muscles and the facial action due to each muscle contraction. To highlight the effect of these muscle contractions on the facial geometry, a mesh is fit on an image of the expressed face and compare it with respect to a neutral mesh. Then changes in the mesh's triangles areas are computed .

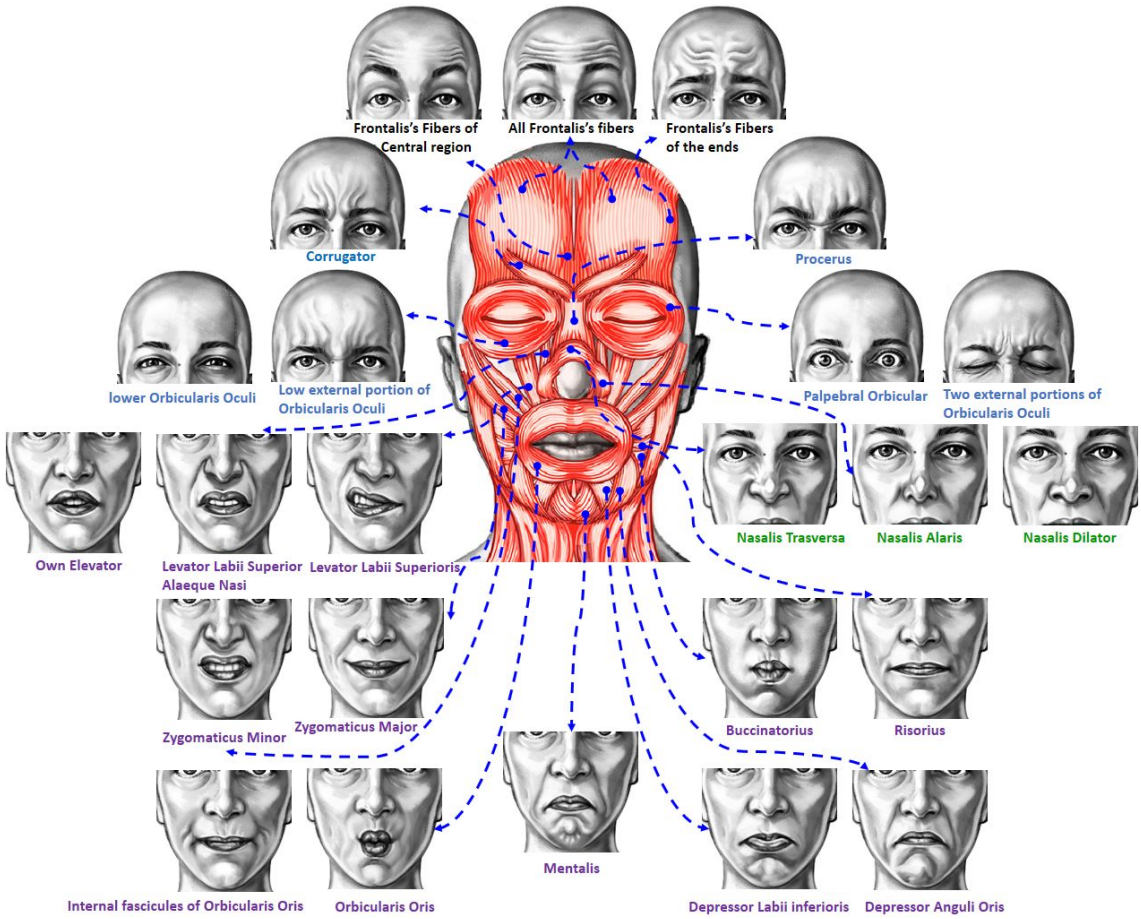
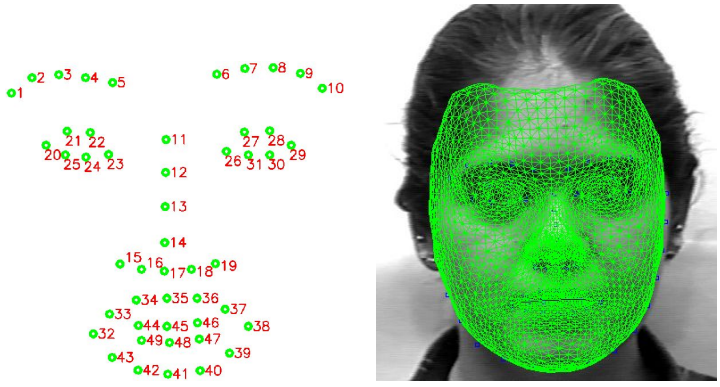


Figure 5.1: Facial muscles and their related actions.



(a) (b)

Figure 5.2: An example of the mesh alignment. a) 2D landmarks extracted from the facial image. b) Aligned mesh on the 2D image.

5.1.1 Mesh Alignment

To generate a mesh reflecting the query facial expression, a 3D mesh is estimated using a linear combination of expression blend-shapes [73]. The blend-shape models are based on the six basic facial expressions. The 3D mesh estimation is a regression process that is formulated as a function mapping an initial shape into a regressed shape. Using the extracted 49-2D landmarks and their 3D correspondences, which are extracted from a 3D mesh, a least-squares approximation of a camera matrix could easily be found as follows. First, similarity transforms T_v and T_u are used, in homogeneous coordinates, to normalize the extracted 2D landmark point $x_i \in R^3$ and the corresponding 3D model point $X_i \in R^4$. These similarity transforms translate the mean to the origin and scale the points so that the Root-Mean-Square (RMS) distance from their origin is $\sqrt{2}$ for x_i and $\sqrt{3}$ for X_i, respectively: $\tilde{x_i} = T_v x_i$ with $T_v \in R^{3\times3}$, and $\tilde{X}_i = T_u X_i$ with $T_u \in R^{4\times4}$. Then the required camera matrix M is computed from the normalized camera matrix $\tilde{M} \in R^{3\times4}$, $\tilde{x_i} = \tilde{M}\tilde{X}_i$ as follows: $M = T_v^{-1}\tilde{M}T_u$.

Finally, the camera matrix is used to project the regressed 3D mesh to a 2D mesh that reflects the query facial expression. An example of the fitted mesh is shown in Fig. 5.2-b.

5.1.2 Muscle Contraction Map

To estimate the effect of muscles contraction on the facial mesh, the fitted mesh is compared with respect to a neutral one. Due to a muscle contraction, the muscle

can be shortening or lengthening. This leads to changes in the mesh's triangles near this muscle. The areas of these triangles increased or decreased according to the movement. This assumption is used to generate a map that represents the facial muscles' contraction according to a specific expression. Examples of maps for the six basic expressions are shown in Fig. 5.3. This subject belongs to CK+ dataset [74].

The generated mesh consists of 3448 vertices and 6736 triangles. Using this mesh in extracting geometric features leads to a very high dimensional features vector. Therefore, a sampled version of this mesh is used in the following framework.

Algorithm 1 Muscle movement map algorithm

```
1:  procedure GET_MUSCLE_MAP(samples_list, neutral_areas)
2:      mesh_model = load_model(model_files)
3:      for all sample in samples_list do
4:          pts = extract_Landmarks()
5:          mesh = mesh_Alignment(mesh_model, pts)
6:          areas = calculate_mesh_faces_area(mesh)
7:          areas = areas/(areas)                    ▷ normalize the area
8:          area_ratio = areas/neutral_areas
9:          area_ratio = clip(area_ratio, 0.5, 2)    ▷ limit the ratio
10:         feature = log2(area_ratio)
11:         features_list.add(feature)
12:     return features_list
```

5.1.3 Facial Action Coding System

The face is an essential tool for nonverbal social communication. Thus analysis of facial movement is an active research topic for behavioral scientists since the work of Darwin in 1872 [75]. The Facial Action Coding System (FACS), which was developed by Ekman and Friesen [76], is an index of facial expressions. Each Action

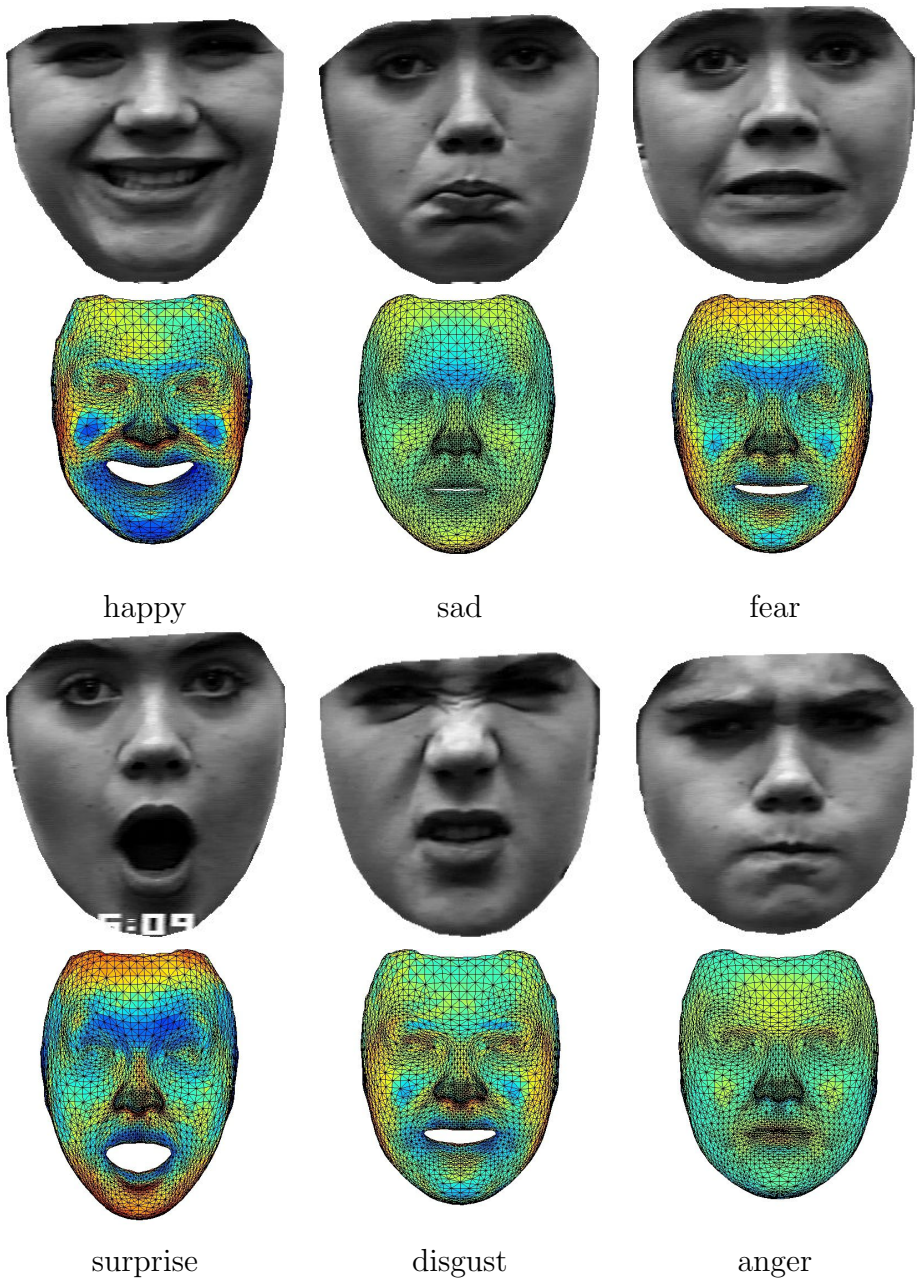


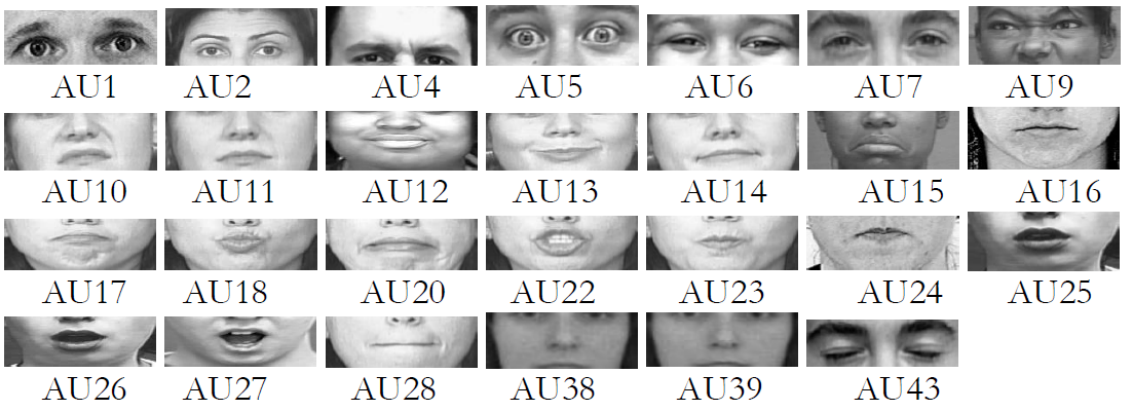
Figure 5.3: Mesh movement maps according to different facial expressions. Muscle contraction is scaled from muscle shortening (blue areas) to muscle lengthening (red areas) and it is represented as a heat map.

unit represents a facial muscle or group of facial muscle movements. FACS decomposes facial expressions in terms of action units (AU's). AU's are the fundamental actions of individual muscles or groups of muscles.

- Main Codes is the set of 46 AU's related to facial muscles. Examples of these AU's are shown in Fig.5.4a.
- Head Movement Codes is the set of AU's related to head movements. Examples of these AU's are shown in Fig.5.4b.
- Eye Movement Codes is the set of AU's related to eyes movements. Examples of these AU's are shown in Fig. 5.4c.

. Facial Action Coding System (FACS) became the most used method for measuring these facial movements, i.e., Action Units (AUs). Action units have a broad impact on several facial expression-based applications such as human-computer interaction [77] and measuring student's engagement [37].

According to a study by G. Duchenne [78], who electrically stimulated facial muscles, movement of the muscles around the mouth, nose, and eyes constitute the facial expressions. This reveals the sparse nature of the dominant AUs regions. Therefore, the performance of AUs detectors can be enhanced using region-based signatures. These signatures can be extracted from uniform patches (e.g., [79–82]) or from patches centered around facial landmarks (e.g., [83,84]). Instead of directly defining these patches, Li et al. [85] introduced a deep learning-based approach to find important areas and crop these regions of interest. From a psychological point of view, recently, Liu et al. [82] investigated the effect of each facial region on various



(a) Example of codes Main AUś

(b) Example of head movement AUś

(c) Example of eye movement AUś

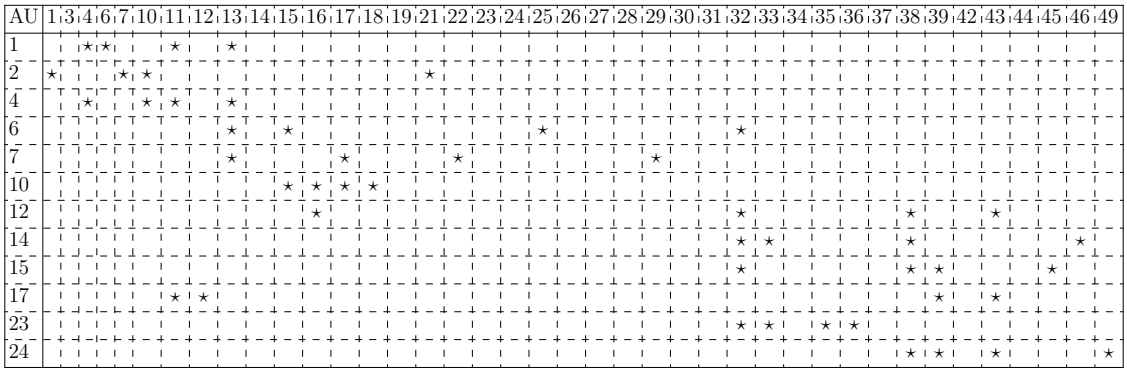
facial expressions. Similarly, Zhong et al. [80] identified the active facial patches of each facial expression. In the Joint Patch and Multi-label Learning (JPML) approach [83], 49 patches are chosen around facial landmarks. Then these sparse facial patches are used to learn a multi-label classifier. For each action unit, the authors identified the most effective set of those patches. A single person's emotion activates a set of AUs [86]. As an example, the smile expression simultaneously activates “Lip Corner Puller” and “Cheek Raiser” action units. Therefore, detecting AUs individually (i.e., one-vs-all classification such as SVM [87] and ADABoost [88]) does not exploit these semantic relationships. On the other hand, many researchers

(e.g., [79, 83, 84, 89, 90]) investigated the correlations among different action units. To learn these relationships, Tong and Ji [91] used a Bayesian network model, and Wang et al. [90] used a restricted Boltzmann machine. In the JPML approach [83], Zhao et al. proposed a multi-label classifier to identify AUs that co-occur frequently and others that unlikely co-occur.

Features that are used in AUs detection can be categorized into appearance-based features (e.g., SIFT, histogram of gradient (HOG), and Local Binary Pattern (LBP)) [84, 92], geometric-based features [93] or both [94]. The appearance-based features (e.g., a 6272-D SIFT feature vector representing each patch in JPML [83]) are histogram descriptors without any shape information. On the other hand, the geometric-based features ignore any visual information. Recently, features that are learned by deep learning approaches replace these hand-crafted features. As an example, a Convolutional Neural Network (CNN) model [95] was proposed to jointly learn dynamic appearance and shape features for facial AUs detection. A Deep Region and Multi-Label learning (DRML) network [79] was proposed to capture local appearance changes for facial regions. A recent CNN-based facial action unit detection approach is EAC-Net [85], which enhances a pre-trained CNN model to learn both features enhancing and region cropping functions.

5.2 The Proposed Facial Action Unit Classifier Under Pose Variation

This section exploit both the sparse nature of the dominant AUs regions and semantic relationships among AUs for action units detection. First, to handle pose



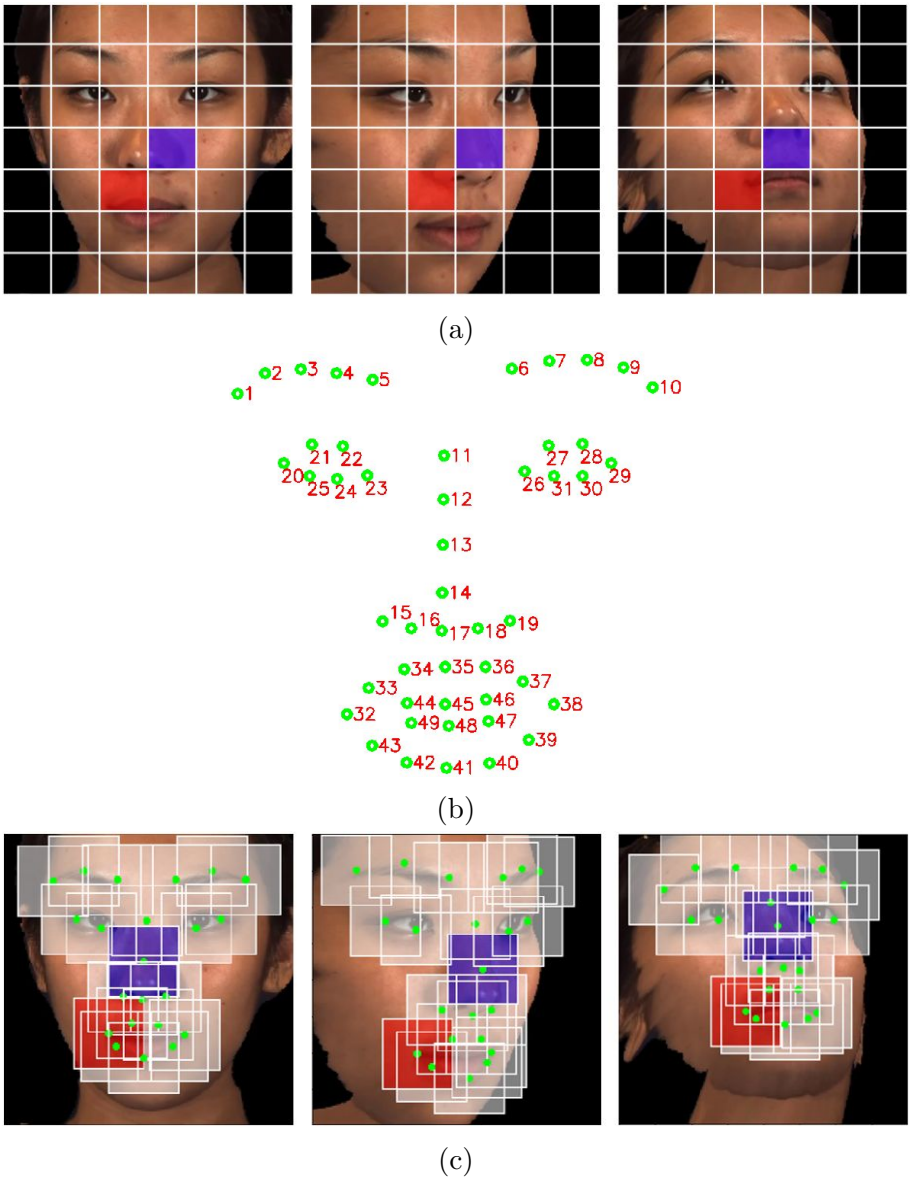
AU	1	3	4	6	7	10	11	12	13	14	15	16	17	18	19	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	42	43	44	45	46	49
1			⋆	⋆			⋆		⋆																															
2	⋆				⋆	⋆										⋆																							
4			⋆			⋆	⋆		⋆																														
6									⋆		⋆									⋆							⋆												
7									⋆				⋆				⋆							⋆															
10											⋆	⋆	⋆	⋆																								
12												⋆															⋆							⋆			⋆			
14																											⋆	⋆						⋆						
15																											⋆							⋆	⋆			⋆		
17							⋆	⋆																										⋆			⋆		
23																											⋆	⋆		⋆	⋆							
24																																	⋆	⋆		⋆				⋆

Table 5.1: A chart illustrates the most four significant patches (marked with ⋆) for each AUs. Rows represent AUs and Columns represent patches.

variations, patches around facial landmarks are defined instead of using a uniform grid which suffers from displacement and occlusion problems as shown in Fig. 5.5. Then,a new deep region-based neural networks architecture in a multi-label setting is proposed to learn both the required features as well as the semantic relationships of AUs. Moreover, a weighted loss function is used to overcome the imbalance problem in multi-label learning.

Face alignment is the first step in any facial system. First, the pipeline start by detecting 68 facial landmarks (see Fig. 5.5(b)) using the detector in [47]. Then, the facial image is aligned by transforming these landmarks to a common space to eliminate the in-plane rotation. Finally, a region of interest is cropped to 200×200 such that the left corner of the right eye becomes the origin of the common space.

Recently, Convolutional Neural Network (CNN) has been presented as an end-to-end framework that performs both feature extraction and classifier training. However, the convolutional layers treat image pixels equally. This spatial stationarity does not hold in faces i.e., structured objects. On the other hand, locally connected



(a)

(b)

(c)

Figure 5.5: In the presence of pose, the uniform grid (a) suffers from lack of correspondences (red and blue rectangles) due to displacement and occlusion. To minimizes this lack of correspondence, facial landmarks (b) are used to define a sparse set of patches (c).

layers treat each image pixel differently. But this needs a huge number of parameters to be tuned. To treat each region differently, a recent region-based layer was proposed by Zhao et al., [79]. However, regions are defined using a uniform grid, which is prone to lack of correspondence in the presence of pose, as shown in Fig. 5.5.

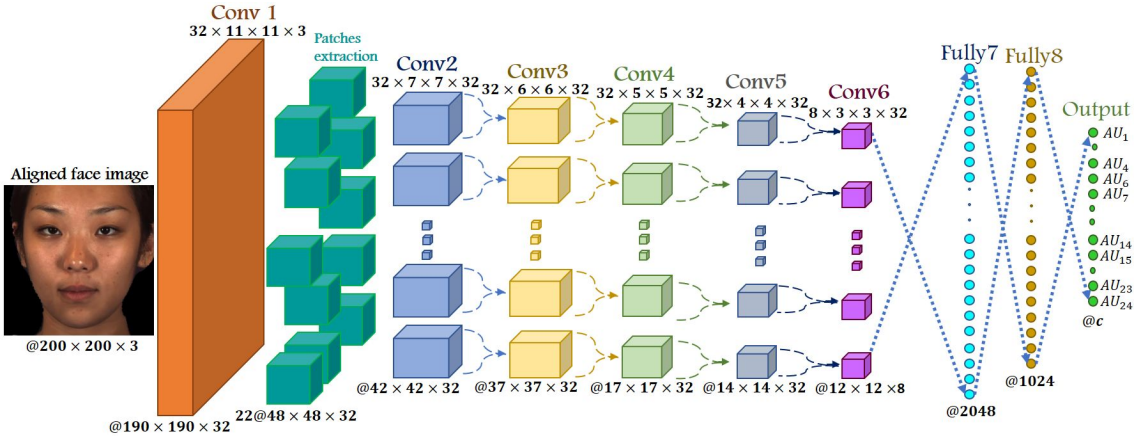


Figure 5.6: The proposed deep region learning architecture. Low level features are extracted from an aligned RGB facial image by a convolutional layer (Conv1). Then 22 overlapped patches of sizes 48×48 are extracted from the convolutional layer output. Each patch is processed by a different cascaded of five convolutional layers (Conv2-Conv6). The filter size of each layer is written on the top, and the dimensions of the layer's output are written on the bottom. The 22 feature vectors extracted by Conv6 are concatenated and fed to three consecutive fully connected layers to detect c AUs.

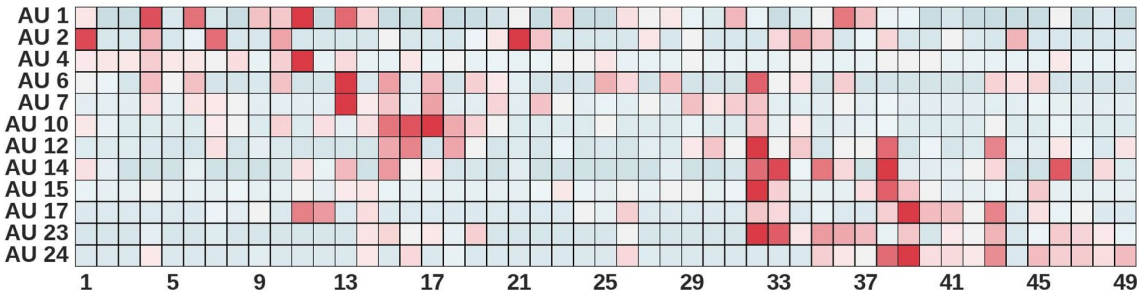


Figure 5.7: An image illustrates the patches significance (ordered from dark red to dark blue) for each AUs.

The proposed network architecture overcomes these drawbacks by treating each region differently. Moreover, patches are defined around facial landmarks instead of a uniform grid. 22 patches are defined to be 48×48 pixels centered around 22 landmarks out of the 49 landmarks. The 22 overlapped patches were chosen to cover the area of interest in the face as shown in Fig. 5.5(c). The proposed network architecture, which is inspired by architectures presented in [79] and [96], is shown in Fig. 5.6. The input to the proposed network is the aligned RGB facial image and its 22 landmarks. First, the image is filtered using 32 filters of size $11 \times 11 \times 3$. This convolutional layer "Conv1" is used to extract a set of low-level features.

Subsequently, 22 patches are extracted from the 32 feature maps (i.e., outputs of "Conv1") around the specified landmarks (which are justified to fit the new size i.e., 190×190) with size of $48 \times 48 \times 32$. Then local features are extracted from each patch by applying five consecutive sets of filters (i.e., "Conv2" - "Conv6") as shown in Fig. 5.6. In each layer, the number and the size of the filters are the same for each patch but with different weights. As an example, in the convolutional layer "Conv2", there are 22 sets of filters. Each set has 32 filters of the same size $7 \times 7 \times 32$ but different wights. To guarantee the non-linearity in this cascade, an activation function is applied after each layer. Rectified Linear Unit (ReLU) [97] is selected to be the activation function due to its sparse features output. This sparsity is an encouraged behavior for the deep network layer because it acts as a regularization factor.

Finally, the 22 feature vectors extracted by the cascade of convolutional networks (i.e., the features of size $22 \times 12 \times 12 \times 8$) are concatenated and are fed to

two fully connected layers ("Fully7" - "Fully8"). ReLU is used as an activation function for these two fully connected layers. Also, these fully connected layers are mainly used to capture the correlations among these features and compress them into a smaller vector (i.e., 1024-D). After representing the input facial image by a 1024-D features vector, the multi-labels classification is performed by another fully connected layer with c outputs. Sigmoid function is used as an activation function in this "Output" layer to make each value in the c-D outputs vector representing the prediction $x_j \in [0,\ \ 1]$ of the j^{th} AU of interest.

This setting of AUs classification is a multi-label learning problem. $L(Y, X)$ is serves as the weighted cross-entropy to be minimized. This function measures the probability error in AUs classification.

$$L(Y, X) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{c}\alpha_j y_{ij}\log x_{ij} + (1 - y_{ij})\log(1 - x_{ij}),$$

where $X \in \mathbb{R}^{N\times c}$ is the matrix of the output layer responses for N samples. $Y \in \{0, 1\}$ is the matrix of the ground truth labels where each element y_{ij} is the ground-truth label of i^{th} sample for j^{th} AU. The weight α_j is multiplied by the first term to up-weight the cost of a positive error relative to a negative error for j^{th} AU. These weights are used to overcome the well-known imbalanced data problem, i.e., the number of positive samples of AUs is less than the negative ones. Finally, two regularization methods are used to prevent overfitting during the training process: the dropout and the ℓ_2 norm of the weights, which is added to the loss function.

5.2.1 Patch Significance

As shown in Fig. 5.5(c), 22 overlapped patches out of the 49 patches are chosen. To show the significance of the selected patches and how these patches affect AUs classification, a similar method to the occlusion sensitivity maps approach [98] is applied as follows: the proposed model shown in Fig. 5.6 is using all 49 patches. Using $30,000$ samples, the score of each AU is calculated, but to occlude a certain patch effect the "Conv6" output of this patch is fed as zeroes to "Fully7". This is sequentially repeated for all 49 patches, and the patch significance is calculated as its average effect on the score of a certain AU. The significance of patches for each AU is shown in Fig. 5.1. Note that the numbers of the patches correspond to the numbers of the landmarks shown in Fig. 5.5(b). From Fig. 5.1, the following facts about the semantic relationships among AUs could be inferred, which are similar to what has been illustrated in the state-of-the-art e.g., [79, 83]: patches around inner eyebrow 4 and 6 as well as patches in between 11 and 13 are the most significant for "Inner Brow Raiser" AU1; the set of most significant patches for "Outer Brow Raiser" AU2 contains outer brow patches 1 and 10; the high significance of patches 7 for AU2 confirms the positive correlation between AU1 and AU2; and the high significance of patch 32 for "Cheek Raiser" AU6 and "Lip Corner Puller" AU12 highlights the positive correlation between these two AUs. Fig. 5.7 highlights that lips-related AUs, i.e., 12, 14, 15, 23 and 24, have their most significant patches around the lips. These correctly learned correlations among different AUs confirm the effectiveness of the proposed architecture in detecting different action units. The

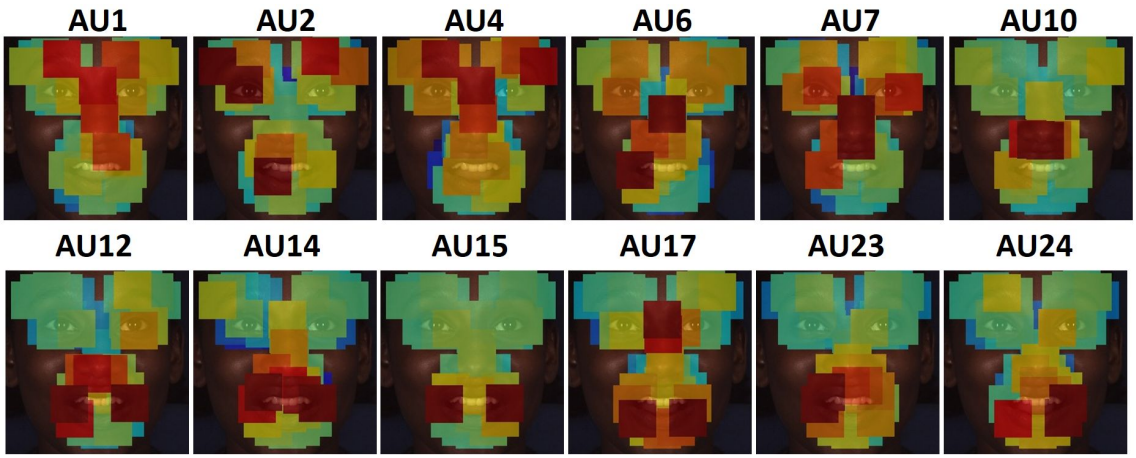


Figure 5.8: Maps illustrate the significance of different patches for each AUs. Heat color coding is used as a measure of the significance: from the dark red (i.e., the most significant) to the dark blue (i.e., the least significant).

selected 22 overlapped patches include all the regions of interest of these AUs.

Pose variance is one of the leading causes that degrade the performance of AUs detectors even using state-of-the-art deep learning approaches. Therefore, the majority of the presented CNN-based models addressed AUs detection for frontal or near-frontal faces. To detect AUs in non-frontal faces, Tosér et al. [99] proposed a deep learning model that tracks the facial fiducial landmark of the individuals and uses them to obtain a normalized face. Another cause for the performance degradation is that negative samples predominate the positive ones. This is a common problem for imbalanced large-scale multi-label learning frameworks [100, 101]. To overcome this limitation, Zhang et al. [102] proposed a class-imbalance aware algorithm.

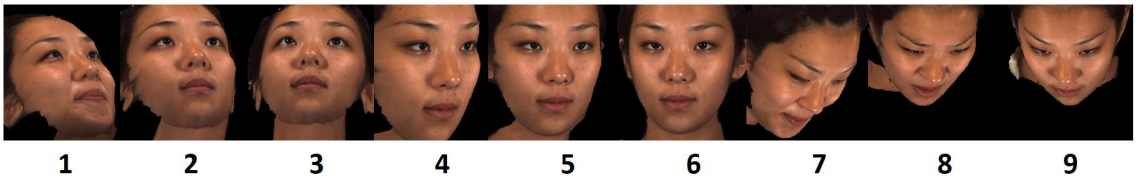


Figure 5.9: Nine different poses in FERA17 dataset [2].

5.2.2 Evaluate The Proposed Facial Action units Classifier

To evaluate the performance of the proposed network, two datasets were used: BP4D-Spontaneous dataset [1] and the recently released FERA17 dataset [2].

BP4D-Spontaneous dataset [1]: This dataset consists of 328 videos that were captured during a series of eight emotional expressions for 23 females and 18 males. The dataset has a frame-based action units coding. experiment was conducted using videos of 31 subjects as training/validation data and videos of the remaining 10 subjects as testing data. The huge number of frames in these videos are sampled to obtain valid aligned facial images. This sampling reduces the dataset to approximately $110,000$ valid frames.

FERA17 dataset: The range of head movements in the BP4D dataset is moderate. So, recently, the FERA17 dataset [2] was released with 9 different poses. FERA17 has 328 3D sequences for 41 subjects of the BP4D [1]. These are used as a train/validation set. Another 159 3D sequences for 20 subjects that were derived from a subset of BP4D+ database [103] are used as a test-set (This is the development partition of FERA17, which is publicly available). These 3D sequences in BP4D and BP4D+ are rotated by pitch angles ($-40°$, $-20°$, and $0°$) and yaw angles ($-40°$, $0°$, and $40°$). Then nine videos were created, see Fig. 5.9. Also, the

dataset has a frame-based action units coding. Approximately $300,000$ valid frames out of 3896 videos were sampled. Videos with poses 1 and 7, shown in Fig. 5.9, are excluded because the preprocessing step does not generate many valid frames from these videos due to the occlusion in the left eye.

To illustrate the imbalance in these datasets, skew, i.e., the ratio of the number of negative samples to the number of positive samples, in these train-sets for each AU is shown in Table 5.2.

Learning: To train the proposed network, an adaptive learning rate optimization method [104] was used. The initial learning rate is 1.0, and the momentum is 0.95. The weight α_j in the loss function is chosen to be the skew of the corresponding AU in the data. The dropout rate is 50%. The batch size is 128. The weight decay is 0.0005. All experiments were performed on one NVIDIA Titan X GPU.

Performance Measures: Three metrics were used as a performance measure: Area under the Precision Recall curve (APR), Area under the ROC curve (AROC), and f_1 score. However, the APR and f_1 score are attenuated by the skewed distributions [101]. Thus, the normalized versions nAPR and nf_1 score of these metrics are calculated.

As a first evaluation, the proposed model was trained using the BP4D [1] train-set. To illustrate the capability of the proposed network in learning the semantic relationships among AUs, the relation matrix of the ground truth AUs and the relation matrix of the proposed network predictions are computed. Each relation matrix contains the correlation coefficients between pairwise AUs. These matrices are shown in Fig. 5.10. The element-wise Euclidean distance 0.004 between the

Table 5.2: Skew of different AUs within BP4D [1] and FERA17 [2] train sets

AU	1	2	4	6	7	10	12	14	15	17	23	24
BP4D	2.2	3.4	2.4	1.2	0.8	0.7	0.8	1.0	3.1	1.3	3.4	3.6
FERA17	1.7	-	2.4	1.0	0.6	0.5	0.6	0.9	1.6	0.9	1.4	-

Table 5.3: Results on the BP4D test-set using different performance measures: f_1 score, nf_1 score, APR, nAPR, and AROC.

AU	1	2	4	6	7	10	12	14	15	17	23	24	Avg
skew	2.4	2.3	2.4	0.9	0.6	0.6	1.0	0.9	3.1	1.3	2.7	3.5	-
f_1	0.51	0.56	0.65	0.78	0.81	0.83	0.81	0.74	0.60	0.68	0.59	0.60	0.68
nf_1	0.68	0.70	0.78	0.77	0.75	0.77	0.81	0.72	0.77	0.72	0.72	0.79	0.75
APR	0.52	0.58	0.72	0.83	0.85	0.85	0.86	0.75	0.61	0.68	0.63	0.51	0.70
nAPR	0.71	0.74	0.84	0.82	0.78	0.76	0.86	0.73	0.82	0.73	0.80	0.77	0.78
AROC	0.69	0.73	0.84	0.84	0.81	0.81	0.87	0.75	0.83	0.75	0.78	0.82	0.79

two matrices confirms the ability of the proposed network to learn the semantic relationships of AUs. The trained model is then used to predict the presence of the action units in BP4D test set. The different performance metrics: f_1 score, nf_1 score, APR, nAPR, and AROC of this experiment are shown in Table 5.3.

Also, saliency map approach [105] was used to visualize the significant region

Table 5.4: Different performance measures for the proposed model when is tested on BP4D test-set and trained on FERA17 train-set

AU	1	4	6	7	10	12	14	15	17	23	Avg
skew	2.4	2.4	0.9	0.6	0.6	1.0	0.9	3.1	1.3	2.7	-
f_1	0.45	0.48	0.64	0.78	0.82	0.83	0.81	0.73	0.48	0.68	0.67
nf_1	0.67	0.67	0.75	0.77	0.77	0.77	0.80	0.70	0.72	0.72	0.73
APR	0.35	0.41	0.71	0.79	0.84	0.87	0.88	0.72	0.44	0.66	0.67
nAPR	0.53	0.60	0.83	0.78	0.76	0.79	0.88	0.69	0.70	0.72	0.73
AROC	0.47	0.60	0.81	0.81	0.80	0.82	0.88	0.72	0.72	0.74	0.74

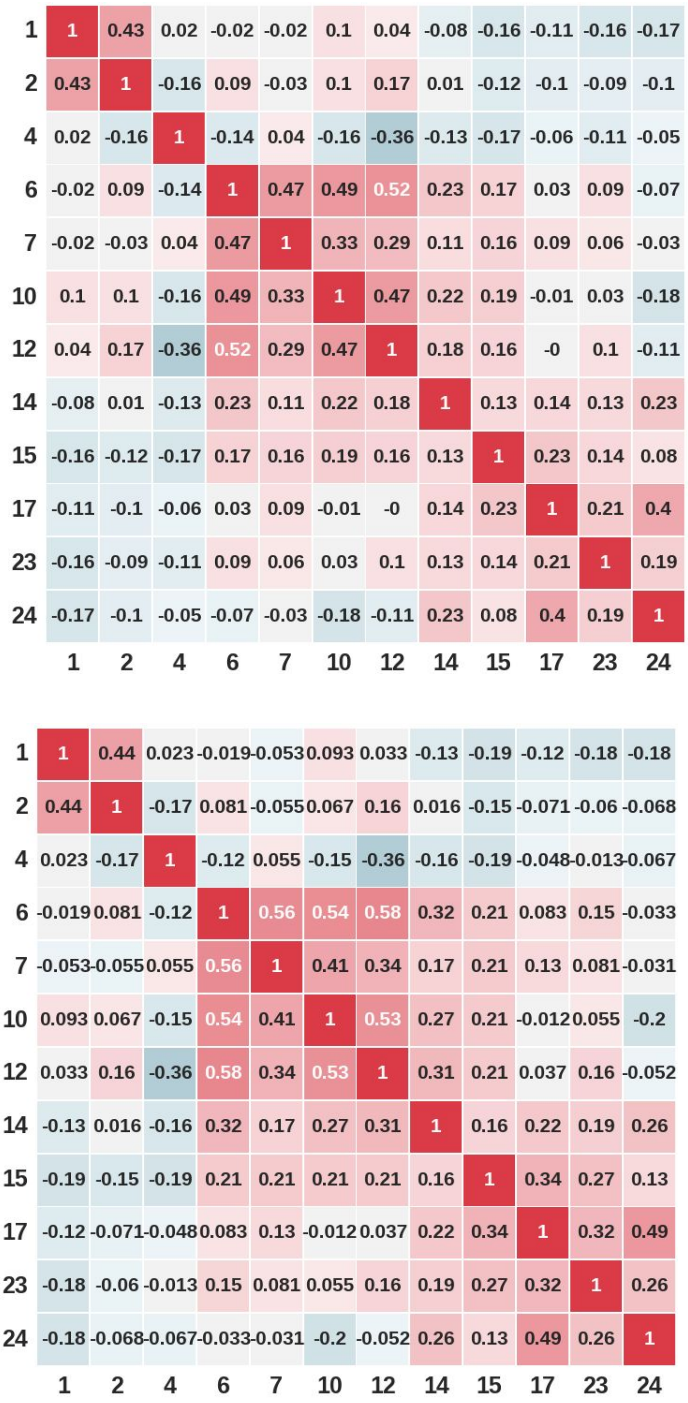


Figure 5.10: The ground truth relation matrix of BP4D dataset (top) and the corresponding relation matrix computed by predictions of proposed approach (bottom).

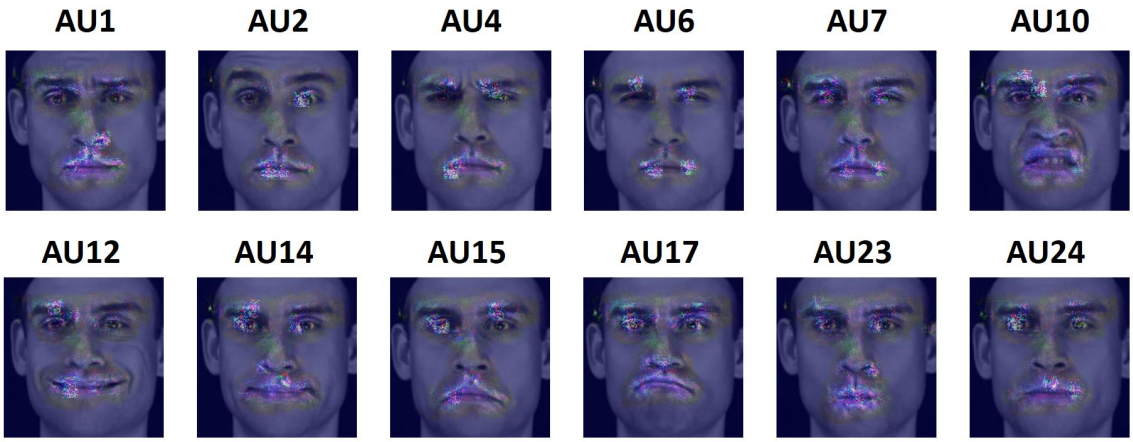


Figure 5.11: Visualization of saliency maps for different AUs.

pixels selected by the trained model. The model is trained using the BP4D [1] train set, then the saliency maps, shown in Fig. 5.11, are generated for different AUs samples from the Max-Planck-Institut (MPI) dataset [106]. Note that the region significance illustrated in Fig. 5.8 is computed as an average over the $25,000$ used images. However, the saliency map visualizes the response of the model for a specific sample. The samples from Max-Planck-Institut (MPI) dataset [106] was used, hence the subject tries to activate a single action unit at a time, which enhances the appearance of the AUs. Similar to the conclusions from the patches significance, the saliency maps highlight how the proposed model identifies sparse discriminative regions for e for AUs detection.

To measure the generalizability of the proposed model in a cross-dataset scenario, the proposed model was trained using the FERA17 train set. This model is then used to predict the presence of the action units in the BP4D test set. The performance measures are reported in Table 5.4. The results in Tables 5.3 and 5.4 are very close to each other. This confirms that cross-dataset protocol is successfully

applied to the proposed model.

The other set of experiments are conducted to evaluate the performance of the proposed network under pose variations. The proposed model, which is trained using the FERA17 train-set, is used to predict the presence of the action units in the FERA17 test-set. For each pose, the different performance metrics: nf_1 score, nAPR, and AROC are shown in Fig. 5.12. As shown in Fig. 5.13(a), the low standard deviations of the different metrics for the seven poses highlights the pose invariant capability of the proposed model to detect different action units.

Another experiment is conducted to illustrate the significance of the patch-based model. A similar model (named "convnet") to the proposed shown in Fig. 5.6 was build. In this "convnet" model, the region-based layers (i.e.,"Conv2" - "Conv6") are replaced by standard convolutional layers, while keeping the sizes of filters as in the proposed model. Similarly, the "convnet" model is trained using the FERA17 train-set with the same settings of the proposed trained model. Then, the proposed model and the "convnet" model are used to predict the presence of the action units in the FERA17 test-set. The average overall posses of the different performance metrics: f_1, nf_1 , APR, nAPR, and AROC are shown in Table 5.5. These performance measures illustrate the following: the "convnet" model has a slight enhancement in only AU6; other action units are more accurately detected using the proposed model than the standard 'convnet". This enhancement is up to 15% in ROC and nAPR of AU23. Moreover, the proposed model has lower standard deviations (see Fig. 5.13(a)) than the "convnet" model for the different metrics (see Fig. 5.13(b)). This confirms that the proposed patch-based layer is more effective in capturing the

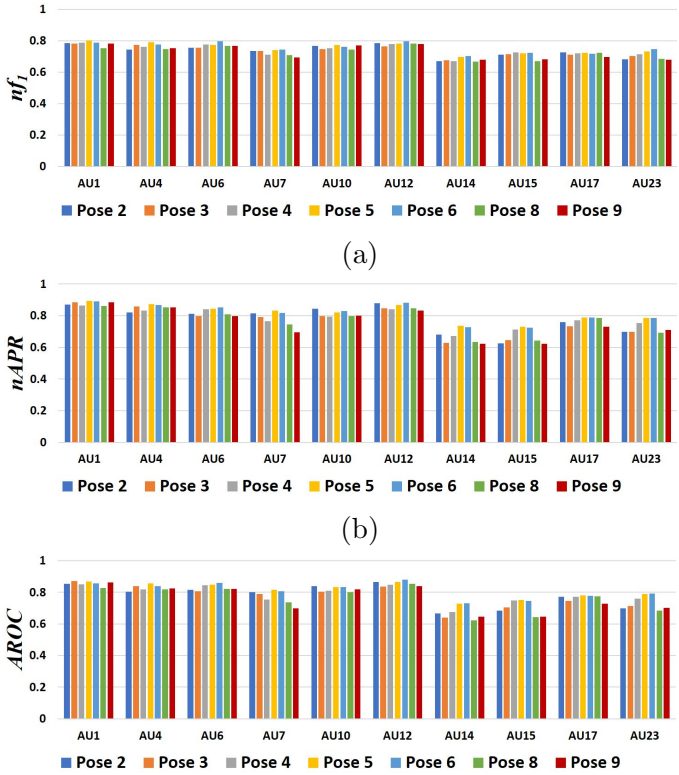
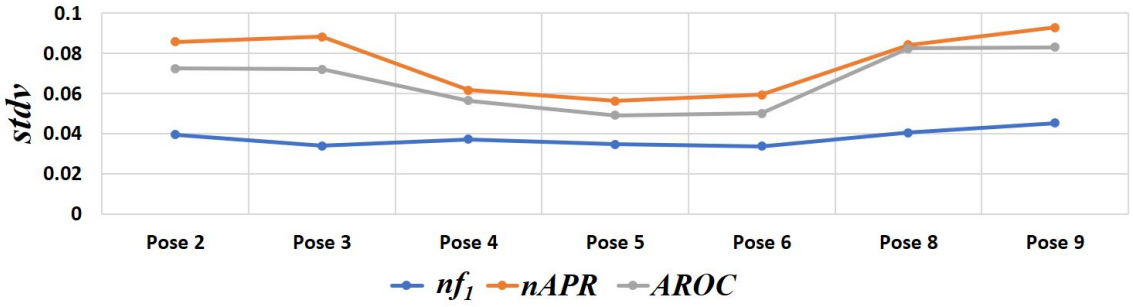


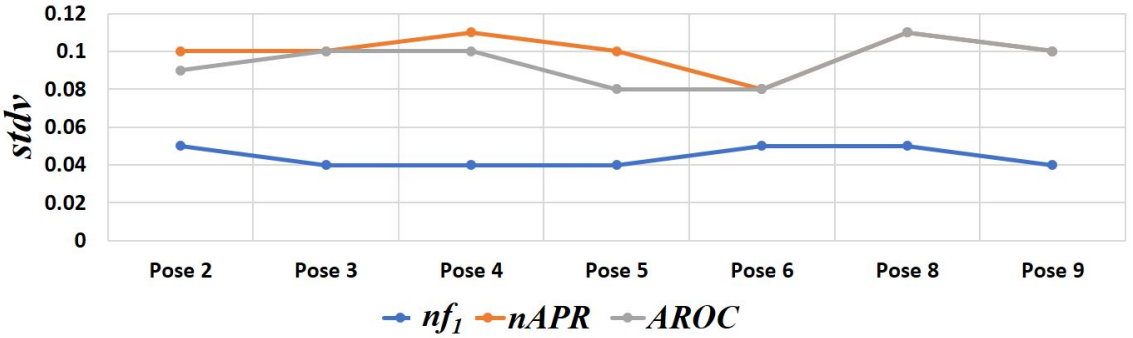
Figure 5.12: Different performance measures for the proposed AUs detection approach under different poses using FERA17 [2] test-set F1-score (a), Area under PR curve (b), and Area under ROC curve (c).

required structural features of the face and learns the correlations among AUs under pose variations than the standard convolutional layer.

Comparison with the-state-of-the-art: The closest work to the proposed one is recently introduced by Zhao et al. [79]. They conducted experiments using the BP4D dataset [1]. However, unlike the proposed model sampling, they sampled 100 positives and 200 negative frames for each sequence, and they adopted a 3-fold partition instead of the proposed model partitioning. The authors reported (see Table 2 in [79]) the performance of different related work such as the classical linear support vector machine classification, patch-learning method [80], JPML [83] and other deep network-based methods (e.g., locally connected network, AlexNet [107],



(a)



(b)

Figure 5.13: Standard deviations of the different metrics for the 7 poses.

and their DRML model [79]). Comparing these reported performance measures to the proposed model result (i.e., APR Avg= 70% in Table 5.3 and DRML Avg= 56% [79]), confirms the high performance of the proposed approach, but using a different setting as explained. Also, the element-wise Euclidean distance 0.004 between the two relation matrices is smaller than what was reported for AlexNet and DRML models in [79]. This confirms that the proposed approach outperforms the state-of-the-art approaches.

It is worth mentioning that Tosér et al. [99] recently conducted a similar experiment for action units detection under pose variations. The authors used an old version of the FERA17 dataset (i.e., FERA15). The proposed model nf_1 scores are better than what has been reported in [99] for the multi-label model. However, since

Table 5.5: AUs detection performance of the proposed model, the "convnet" model, and baseline results [2] using the FERA17 test-set. Different performance measures: APR, nAPR, AROC, f_1 score and nf_1 score are used.

AU		1	4	6	7	10	12	14	15	17	23
skew		9.6	10.5	1.6	0.5	0.6	1.0	0.4	4.1	2.9	1.9
f_1	convnet	0.50	0.39	0.73	0.82	0.80	0.77	0.82	0.36	0.49	0.54
	FERA17 [2]	0.15	0.17	0.56	0.73	0.69	0.65	0.62	0.15	0.22	0.21
	Proposed model	0.57	0.49	0.70	0.82	0.81	0.77	0.82	0.42	0.54	0.59
nf_1	convnet	0.75	0.73	0.78	0.72	0.74	0.78	0.67	0.67	0.70	0.68
	Proposed model	0.78	0.76	0.77	0.72	0.76	0.78	0.68	0.71	0.71	0.71
APR	convnet	0.51	0.38	0.79	0.87	0.86	0.82	0.78	0.27	0.48	0.48
	Proposed model	0.59	0.50	0.75	0.87	0.87	0.85	0.82	0.34	0.55	0.60
nAPR	convnet	0.85	0.80	0.85	0.78	0.79	0.83	0.62	0.60	0.71	0.63
	Proposed model	0.88	0.85	0.82	0.78	0.81	0.85	0.67	0.68	0.7	0.73
AROC	convnet	0.82	0.79	0.85	0.77	0.79	0.84	0.63	0.62	0.71	0.65
	Proposed model	0.86	0.83	0.83	0.77	0.82	0.85	0.67	0.71	0.76	0.74

the datasets are different, this cannot be considered as a fair comparison. Finally, as shown in Table 5.5 action units are more accurately detected using the proposed model than the FERA17 baseline results [2]. The enhancement in the f_1 scores are from 9% to 42%.

5.3 The Proposed Emotional Engagement Framework

In this section, novel framework for automatic measurement of the emotional engagement level of students in both of e-learning environment or in-class environment is proposed. The proposed frameworks capture the user's video using a regular webcam; it tracks their faces through the video's frames. Different features are extracted from the user's face, e.g., facial landmark points, and facial action units, as shown in Fig. 5.14.

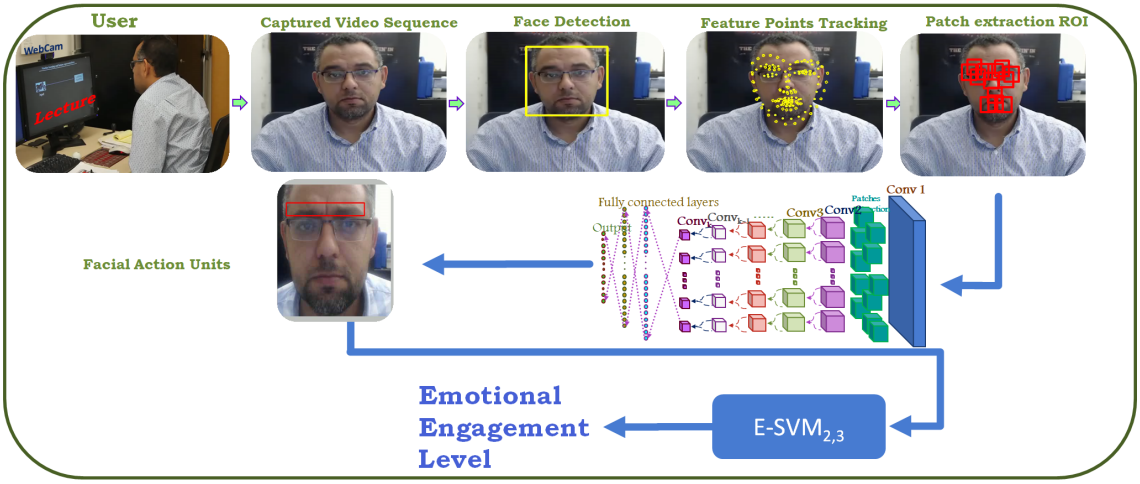


Figure 5.14: The proposed emotional engagement framework.

Similar to the proposed behavioral engagement framework for e-learning environment section 3.2. To extract and track the ROI through frames,the same pipeline are applied. The same face detector Fig. 3.6 is used to extract and track the face region. Also the 68 facial feature points are extracted using the approach in [47]. Next the 22 patches to be used for the proposed facial action unit detection under pose variation are extracted as discussed in 5.2.1. The extracted facial action units are used to fed a support vector machine (SVM) to classify the students emotional engagement level.

5.4 Conclusion

This chapter introduced the proposed facial action unit detection that work under pose variation. It also presented a framework to measure the students emotional engagement level. The proposed framework could be implemented to run either offline or at the Client/Server model. In some settings, such as a student

watching tutorial or online lecture, the framework modules are implemented on one machine. In contrast, in the case of e-learning, this framework is implemented as a Client/Server program. This module will track the face and extract facial point, and the facial action units at the client-side and send this small feature vector to the server-side. The server will receive these features from multiple students simultaneously, then calculate each student's emotional engagement level.

CHAPTER 6: AUTOMATIC MEASUREMENT OF BEHAVIORAL AND EMOTIONAL ENGAGEMENT

In this chapter, novel frameworks for automatic measurement of the engagement level of students are proposed in either an in-class environment or an e-learning environment. Also, a biometric sensor network is proposed to be used in collecting data. This chapter also compares the proposed modules with the state-of-the-art.

6.1 The Proposed E-learning Student Engagement Automatic Measurement Framework

The proposed frameworks capture the user's video and track their faces through the video's frames. The proposed framework is built as a client/server application. The client-side uses the modules introduced in section 3.2 to extract the head pose and eye gaze for behavioral engagement measurement, and the module in section 5.3 to extract the facial action units for the emotional engagement measurement. It sends this feature vector over the internet after attaching student identification and timestamp. The client designed to have low dimensional features vector as payload

for the internet to guarantee that it is reliable even with a low bandwidth networks. The server received this feature vector along with other student vectors. It classifies each student engagement level individually, as well as summarizes the whole class engagement levels, and displays the results graph to the instructor dashboard, see Fig. 6.1.

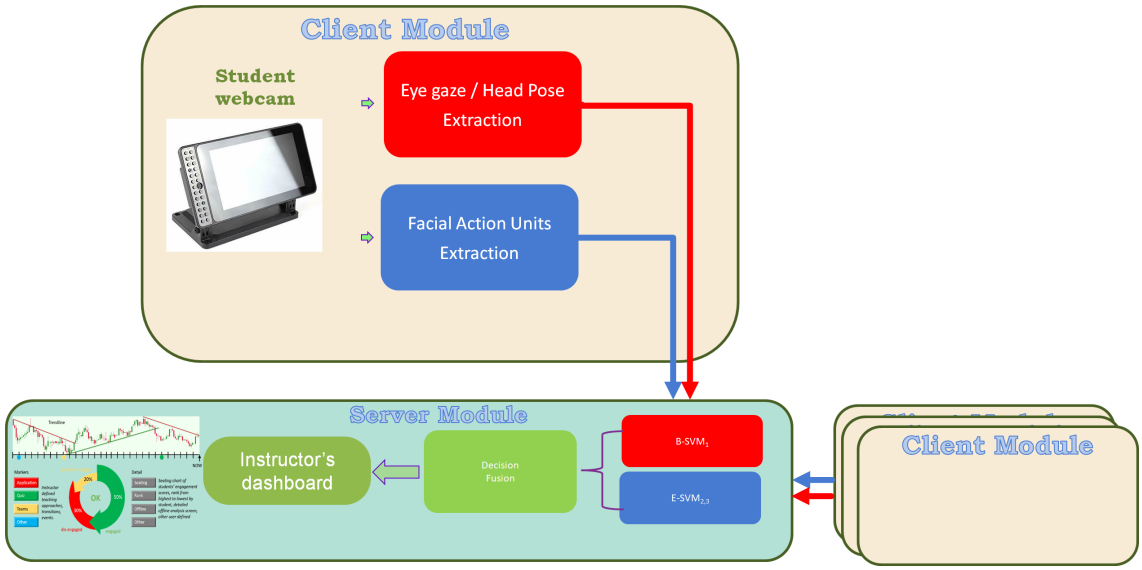


Figure 6.1: The proposed e-learning student engagement automatic measurement framework.

6.1.1 Hardware Setup

Using student webcams and machines to run the proposed client module raises many issues, especially with the huge variety that students have in terms of hardware and software. The camera quality cannot be guaranteed, and multiple versions of the software are needed to ensure that it runs on each operating system. Also, the student may fold his/her laptop and use it to take notes which leads to the impossibility of capturing the student's face. Therefore, a special hardware unit is

designed and sent to the student to be used as our client module. This module is composed of a Raspberry Pi micro-controller connected to a webcam and touch display, see Fig. 6.2. The student has to perform a one-time setup to ensure internet connectivity.

The Raspberry Pi micro-controller runs program that connects to the server, captures the video stream, applies the introduced pipelines to extract the feature vector, and sends that vector to the server. The program allow the students to adjust the webcam to ensure that the video has a good perspective for the face. This module was also used in the data collection phase. It recorded a video stream of the student's face during the lecture and uploaded it to a cloud server by the end of the lecture. The The Raspberry Pi uses a TLS encrypted connection to ensure the student data security and privacy.

Figure 6.2: The student hardware module.

We also send a wristband (Samsung galaxy watch) to measure other biometric features such as heart pulse rate. This feature should be used in the future to enhance the emotional engagement measurement.

6.1.2 Data Collection

The Hardware described in the previous section is used to capture subjects' facial videos while attending a lecture. The facial videos were recorded during the lecture. The collected dataset consists of 13 students.

To annotate the dataset, four annotators coded the video frames using a four engagement level.

- 0) No face
- 1) Not engaged
- 2) Look to be engaged, which mean behavioural engaged while emotionally not engaged.
- 3) Engaged, which mean both behavioural and emotionally engaged.

After excluding the "0" category, we collected over 100,000 frames. We next applied the following voting process: a frame is included in the dataset if at least three annotators used the same label for the frame. Therefore, the total number of selected frames is 70,000 reduced to with the following distribution: (12%) of them in category "1", (56.5%) for category "2", (31.5%)for category "3". The pairwise inter-coder agreement for this set is computed using Pearson correlations. The

average pairwise agreement value was found to be 0.66.

6.1.3 Evaluation and Comparison

Experiments were conducted to compare the proposed modules with the state-of-the-art engagement measurement systems. The collected data are used to evaluate the performance using the area under the Receiver Operating Characteristics (ROC) curve A' statistics and the accuracy for the comparison.

We compared the proposed e-learning student engagement automatic measurement with Affdex SDK [5] as it was almost designed for the same environment. McDuff et al. [5] introduced a facial expression analysis module that measures nine emotions and 32 facial micro-expressions; among them, they measure engagement and attention. Affdex SDK [5] also detects and tracks the facial landmarks, then it extracts a histogram of gradient (HOG) features from the aligned faces. These features are used as an input to support vector machine (SVM) classifiers to detect the facial action units (AUs). Then they use an emotional facial action coding system (EMFACS) [108] to express a set of emotions, which include engagement and attention, as combinations of facial actions. While The Affdex SDK [5] is not designed to work in a client/server architecture, it uses a single video stream to capture the students face and classify both behavioral engagement (attention) and emotional engagement level. To use in e-learning environment, the whole system is installed on the student's machine, which comes at the cost of processing at client-side or streams the webcam video to the server machine the process the video which comes

at the cost of used bandwidth.

For the proposed Modules Leave-one-out cross-validation (LOOCV) experiment is conducted using 13 training and testing sets. And the collected data was also used to evaluate the Affdex SDK [5]. Table. 6.1 show the comparison of the behavioral engagement performance for Affdex SDK [5] and the proposed framework. It also shows the enhancement on the proposed framework if OpenFace 2.0 [48] is used to obtain the eye gaze and head pose. While Table. 6.2 show the comparison of the emotional engagement performance for Affdex SDK [5] and the proposed framework. It also shows the enhancement of the proposed framework if a higher dimensional feature vector is used. This feature vector could be obtained by applying a bank of 40 Gabor filters (4 orientations, 5 spatial frequencies,and 2 scales) after resizing the aligned face (ROI) .to 32x32, which leads to obtaining a 40960-D feature vector.

Table 6.1: Comparison of recognition rates for the behavioral classifiers.

	A'	Accuracy r
Affdex SDK [5]	69%	78%
Proposed framework	66%	80%
Proposed framework with OpenFace [48]	84%	83%

Table 6.2: Comparison of recognition rates for the emotional classifiers.

	A'	Accuracy r
Affdex SDK [5]	71%	52%
Proposed framework	74%	70%
Proposed framework with high dimensional feature	82%	80%

The results from the two systems were too noisy and had a lot of high frequencies changes. This means that the two systems are very sensitive for high-frequency movements such as eye blinking. Therefore, the results are smoothed using a moving average filter of span equal to 1 minute. Figure. 6.3 shows samples of behavioral engagement result of the proposed model and attention of Affdex sdk [5], and Fig. 6.4 shows samples of emotional engagement result of the proposed model and engagement of Affdex sdk [5] for the same videos.

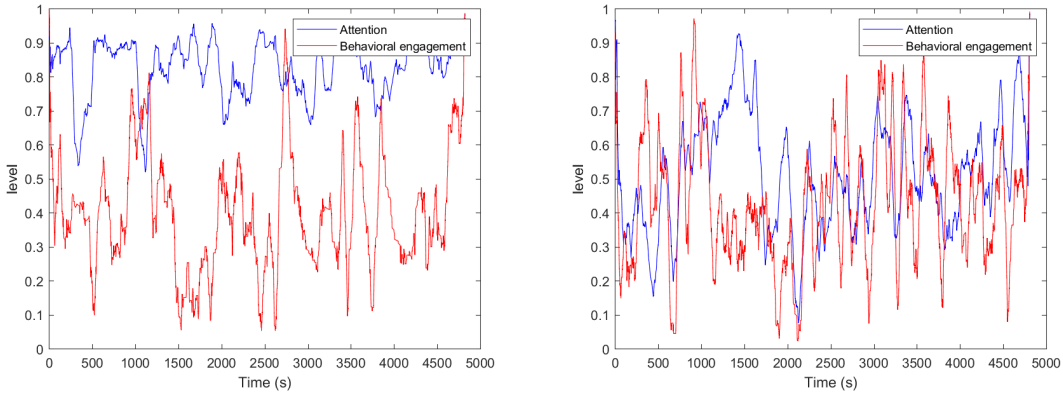


Figure 6.3: Samples of behavioural engagement result of proposed model and attention of Affdex sdk [5].

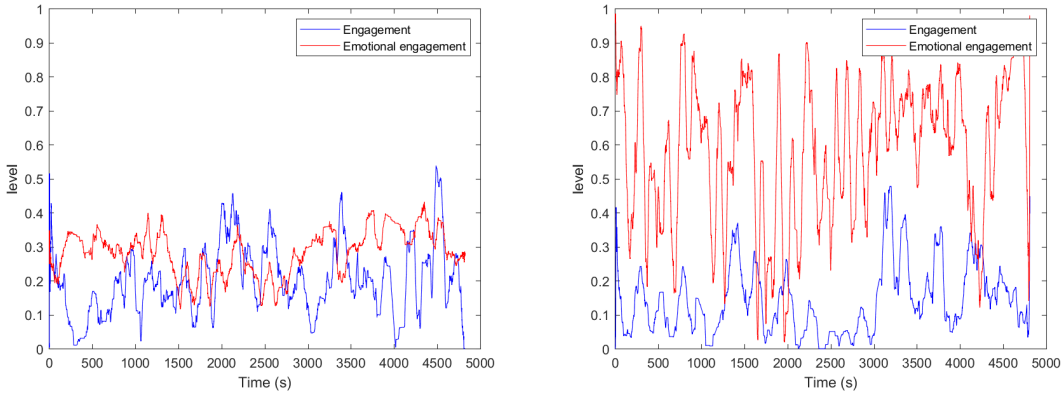


Figure 6.4: Samples of emotional engagement result of the proposed model and engagement of Affdex sdk [5].

The proposed framework extracts a 41-D features vector from each frame. These features include 33 AUs code, 3 pose angles, and 5 eye gaze codes. This 41-D features vector is extracted in less than 200 ms. This makes the system is applicable for online processing (i.e., 5 fps).

6.2 The Proposed In-class Student Engagement Automatic Measurement Framework

As discussed in chapter 4, the in-class student engagement is more complex, especially for behavioral engagement. Unlike e-learning, where there is only one target for student gaze. In the class, the student may give attention to the whiteboard, instructor, projector screen, or his/her peers. Therefore, This section introduces the framework to measure student engagement for the in-class environment. The proposed framework uses the module proposed in section 4.2 to extract the head pose, eye gaze, and body actions, then use them to classify. For each window of time, it detects the point of interest that attracts most student gazes and individually calculates the deviation of each student gaze. It uses the same emotional engagement in section 5.3 to extract the facial action units for the emotional engagement measurement. Using this feature, The behavioral and emotional levels are estimated at local high-performance computing machine on the same network. It classifies each student engagement levels individually, as well as summarizes the whole class engagement levels and displays the results graph to the instructor dashboard, see Fig. 6.5.

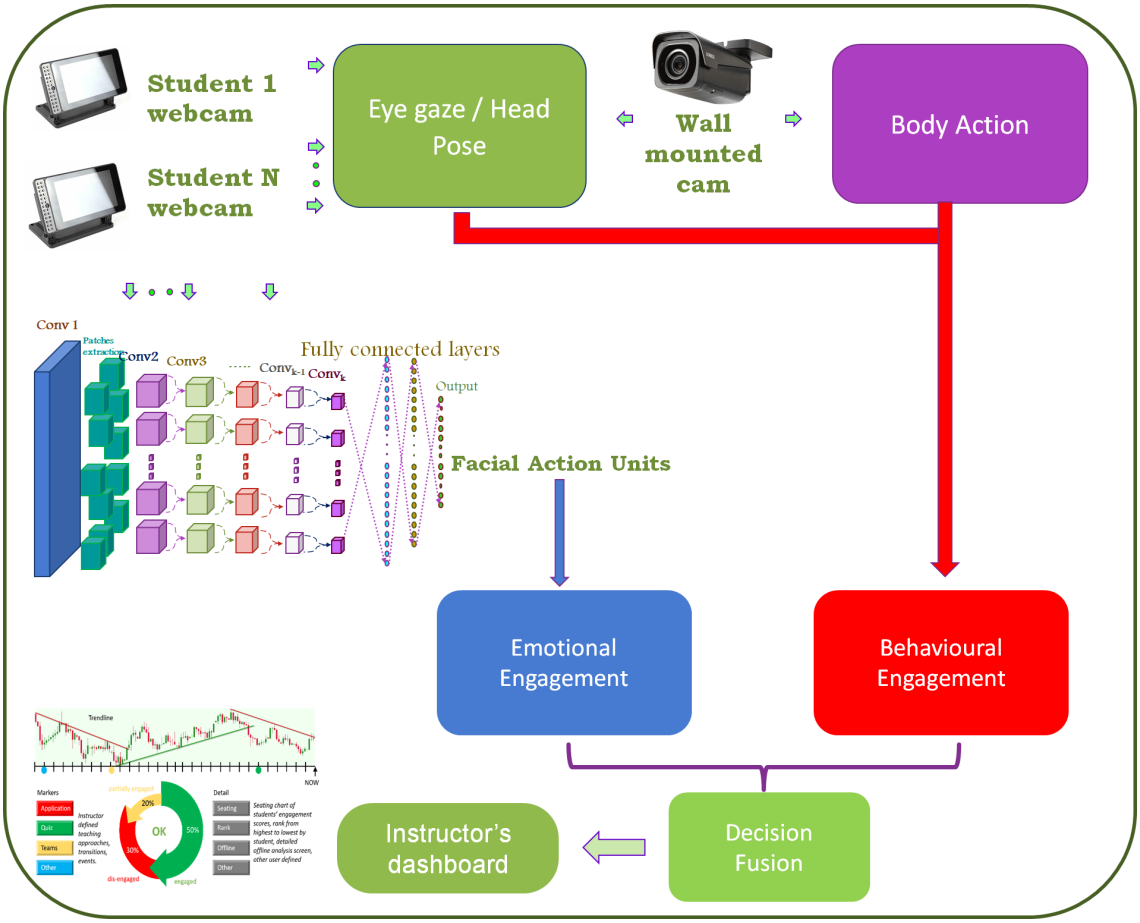


Figure 6.5: The proposed e-learning student engagement automatic measurement framework.

6.2.1 Hardware setup

The proposed class setup uses the same student hardware module used in the e-learning setup, but in this setup, this module is connected to a high-performance computing machine in the same local network. This high-performance computing machine collects all the streams/features and classifies the engagement level in real-time. The setup also includes 4K wall-mounted cameras to capture a stream of student bodies to be used in the body pose and body actions extraction. Also,

the configuration provides high bandwidth network equipment for both wire and wireless connections. Also, wristbands (Samsung galaxy watch) could be used to measure other biometric features such as heart pulse rate. That could be integrated with the system in the future. Figure 6.6 shows the biometric sensor network used in the in-class environment

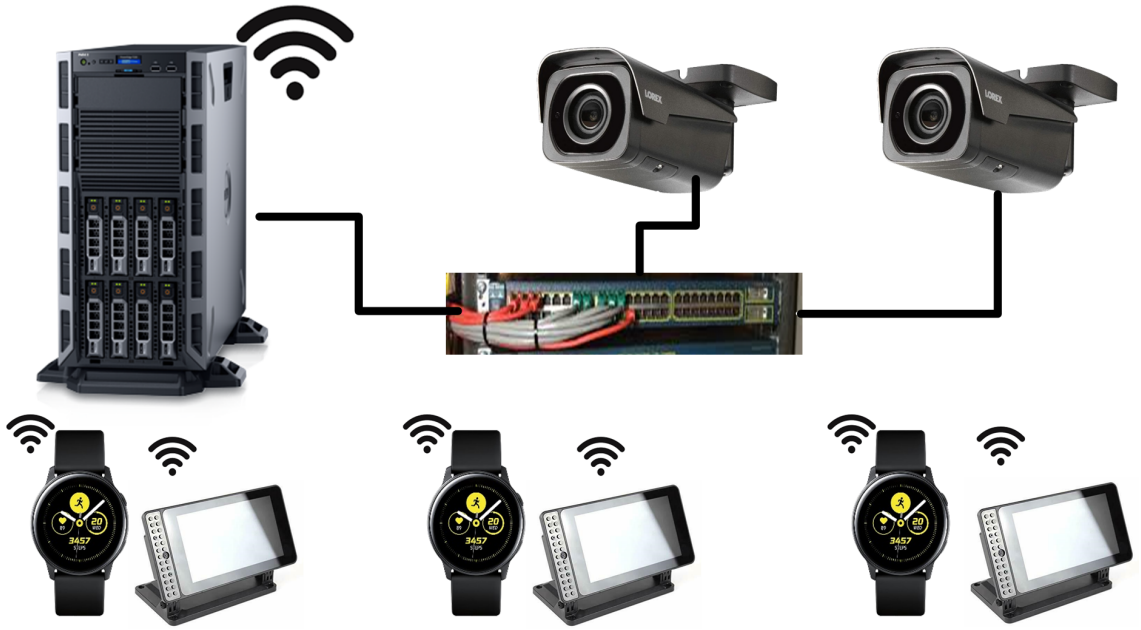


Figure 6.6: The proposed biometric sensor network.

6.2.2 Data Collection

The Hardware described in the previous section is used to capture subjects' facial videos and body videos while attending four lectures. The facial videos and body videos were recorded during the lecture. The collected dataset consists of 10 students who are learned an engineering course. This data is annotated by professorial educators. Each lecture is 75 min in length and divided into 2-minute

windows, which result in 1360 samples.

6.2.3 Evaluation and Comparison

The proposed framework uses the same emotional engagement classifiers used in the e-learning framework. Therefore this section focus on evaluating the behavioral engagement module. It is hard to compare the performance of the proposed module as there is no other frameworks that handle the same environment. The frameworks introduced in [46] [3] are very close in purpose to the proposed framework. However, they focus on extracting the features without any contribution in automatically measure the engagement level. As shown, the agreement ratio for the disengaged and engaged in terms of behavioral engagement are 83% and 88%, respectively.

Figure 6.7 shows the confusion matrix of the proposed behavioural engagement classification

6.3 Conclusion

This chapter presented two frameworks to automatically measure student engagement either in e-learning or in-class environment. It also introduces the hardware setup needed for these frameworks. The e-learning framework shows better performance than the state-of-the-art. This chapter also shows the performance of the in-class environment framework. On the other hand, there is a lack of framework

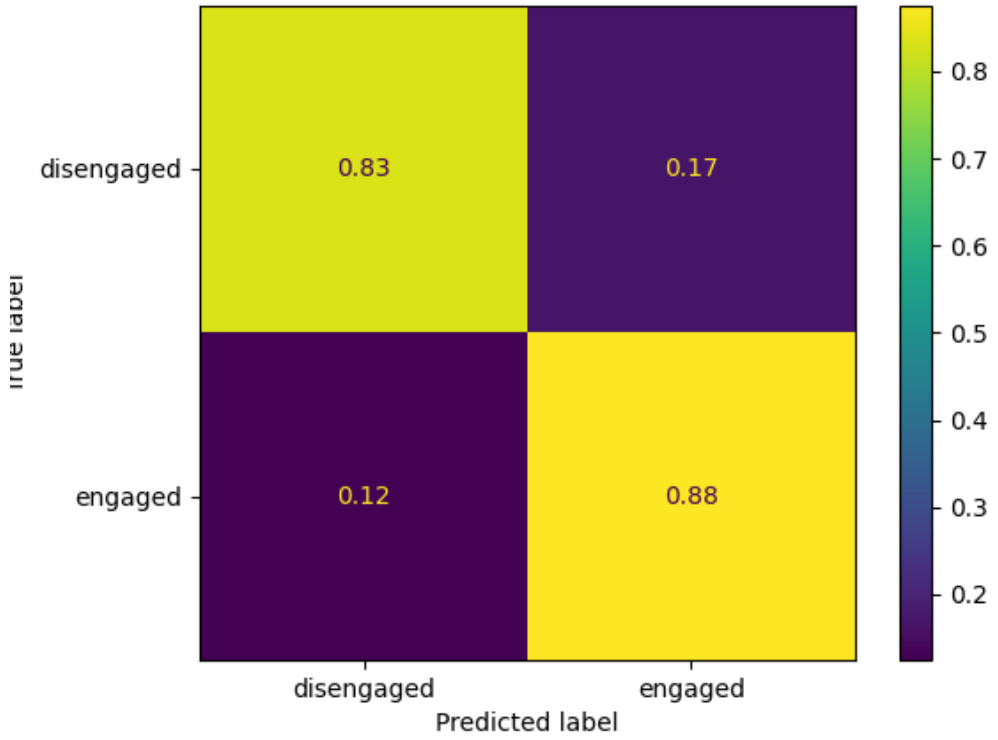


Figure 6.7: The confusion matrix of the proposed behavioural engagement classification.

for the in-class environments to be used on the comparison, and more experiments need to be conducted on larger data sets (Data from multiple courses during the whole semester).

CHAPTER 7: CONCLUSION & FUTURE WORK

7.1 Conclusion

In this dissertation, novel frameworks to automatically measure the student's engagement levels in either e-learning or in-class environments are proposed. These frameworks provide the instructors with real-time estimation for both the average class engagement level and the individual's engagement levels, which will help the instructor make decisions and plans for the lecturer.

The dissertation discussed the drawbacks of the traditional methods to estimate the student's engagement and the necessity to have an automated system for that problem. It presented the three main components that form the student's engagement; behavioral, emotional, and cognitive engagement. Also, it reviewed the previous research done in that area. Various types of biometric features are discussed to be used in engagement estimation. Biometric sensor network was designed to capture student's gesture and expression. The behavioral engagement level depends on the head pose, the eye gaze, the body pose, and the body action are presented. While the facial action coding system is introduced for the measurement of student's emotional engagement. The differences between the various environment,

i.e., e-learning and in-class, and the choice of the suitable metric for each environment are discussed. In an e-learning environment, the student head pose and eye gaze are tracked to ensure that the student gaze at his/her laptop screen. On the other hand, in the in-class environment, the student head pose and eye gaze are also tracked against various targets such as instructor, whiteboard, etc. Also, student's body pose are tracked to extract their body actions. Machine learning algorithms are used to train suitable classifiers for each environment. Finally, the performance of each module is discussed.

A comparison with the state-of-the-art framework is held. The e-learning framework shows better performance than the state-of-the-art. In contrast, this type of comparison was difficult to perform in the in-class environment because of the lack of frameworks that work for the in-class environments. For this environment, the state-of-the-art frameworks focused on extracting the metrics without any evaluation or classification for the engagement.

This dissertation showed that the students behavioral and emotional engagement could be measured using biometric features which are from head, eye, and body of the students in either e-learning or in-class environment with an average accuracy of %86.

7.2 Limitations and Future Work

Due to the Pandemic, the dataset collected for the in-class environment was relatively small. It only contains ten students attending four sessions for a single

course. A large-scale dataset should be collected, for more students, who attending multiple courses during the entire semester. This will help in the process of training and evaluating both behavioral and emotional engagement measurement modules. It will also allow the emotional engagement measurement module to become more complicated by classifying chunks of video (time window) rather than individual frames.

The proposed framework uses only one metric to estimate emotional engagement. Using more biometric features such as heart rate and galvanic skin response (GSR) will improve the estimation of emotional engagement. Also, this work did not discuss the estimation of the third component of engagement which is the cognitive engagement. Measuring this component is too complicated, and using a sensor such as an electroencephalogram (EEG) headset is very intrusive. A study to relate the measured behavioral and emotional engagement levels to the third component needs to be performed.

REFERENCES

[1] Xing Zhang, Lijun Yin, Jeffrey F. Cohn, Shaun J. Canavan, Michael Reale, Andy Horowitz, Peng Liu, and Jeffrey M. Girard. BP4D-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database. Image Vision Comput., 32(10):692–706, 2014.

[2] Michel F. Valstar, Enrique Sánchez-Lozano, Jeffrey F. Cohn, László A. Jeni, Jeffrey M. Girard, Zheng Zhang, Lijun Yin, and Maja Pantic. FERA 2017 - addressing head pose in the third facial expression recognition and analysis challenge. In IEEE Int. Conf. on Automatic Face & Gesture Recognition, 2017.

[3] Karan Ahuja, Dohyun Kim, Franceska Xhakaj, Virag Varga, Anne Xie, Stanley Zhang, Jay Eric Townsend, Chris Harrison, Amy Ogan, and Yuvraj Agarwal. Edusense: Practical classroom sensing at scale. Proc. ACM Interact. Mob. Wearable Ubiquitous Technol., 3(3):71:1–71:26, September 2019.

[4] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019.

[5] Daniel McDuff, Mohammad Mavadati, May Amr, Jay Turcot, and Rana Kaliouby. Affdex sdk: A cross-platform real-time multi-face expression recognition toolkit. pages 3723–3726, 05 2016.

[6] Brandi N Geisinger and D Raj Raman. Why they leave: Understanding student attrition from engineering majors. International Journal of Engineering Education, 29(4):914, 2013.

[7] Guili Zhang, Timothy J Anderson, Matthew W Ohland, and Brian R Thorndyke. Identifying factors influencing engineering student graduation: A longitudinal and cross-institutional study. Journal of Engineering education, 93(4):313–320, 2004.

[8] Catherine Good, Aneeta Rattan, and Carol S Dweck. Why do women opt out? sense of belonging and women's representation in mathematics. Journal of personality and social psychology, 102(4):700, 2012.

[9] Jeffrey L Hieb, Keith B Lyle, Patricia AS Ralston, and Julia Chariker. Predicting performance in a first engineering calculus course: Implications for interventions. International Journal of Mathematical Education in Science and Technology, 46(1):40–55, 2015.

[10] Campbell Rightmyer Bego, Il Young Barrow, and Patricia A Ralston. Identifying bottlenecks in undergraduate engineering mathematics: Calculus i through differential equations. In 2017 ASEE Annual Conference & Exposition, 2017.

[11] Ann Kajander* and Miroslav Lovric. Transition from secondary to tertiary mathematics: Mcmaster university experience. International Journal of Mathematical Education in Science and Technology, 36(2-3):149–160, 2005.

[12] David B Bellinger, Marci S DeCaro, and Patricia AS Ralston. Mindfulness, anxiety, and high-stakes mathematics performance in the laboratory and classroom. Consciousness and Cognition, 37:123–132, 2015.

[13] Gregory M Walton, Christine Logel, Jennifer M Peach, Steven J Spencer, and Mark P Zanna. Two brief interventions to mitigate a "chilly climate" transform women's experience, relationships, and achievement in engineering. Journal of Educational Psychology, 107(2):468, 2015.

[14] Joanna Perry Weaver, Marci S DeCaro, Jeffrey Lloyd Hieb, and Patricia A Ralston. Social belonging and first-year engineering mathematics: A collaborative learning intervention. In 2016 ASEE Annual Conference & Exposition, 2016.

[15] Shui-fong Lam, Shane Jimerson, Bernard PH Wong, Eve Kikas, Hyeonsook Shin, Feliciano H Veiga, Chryse Hatzichristou, Fotini Polychroni, Carmel Cefai, Valeria Negovan, et al. Understanding and measuring student engagement in school: The results of an international study from 12 countries. School Psychology Quarterly, 29(2):213, 2014.

[16] Alex J Bowers. Grades and graduation: A longitudinal risk perspective to identify student dropouts. The Journal of Educational Research, 103(3):191–207, 2010.

[17] Joseph Gasper, Stefanie DeLuca, and Angela Estacion. Switching schools: Revisiting the relationship between school mobility and high school dropout. American Educational Research Journal, 49(3):487–519, 2012.

[18] James J Appleton, Sandra L Christenson, and Michael J Furlong. Student engagement with school: Critical conceptual and methodological issues of the construct. Psychology in the Schools, 45(5):369–386, 2008.

[19] Chandra P Carter, Amy L Reschly, Matthew D Lovelace, James J Appleton, and Dianne Thompson. Measuring student engagement among elementary students: Pilot of the student engagement instrument—elementary version. School Psychology Quarterly, 27(2):61, 2012.

[20] Jennifer A Fredricks and Wendy McColskey. The measurement of student engagement: A comparative analysis of various methods and student self-report instruments. In Handbook of research on student engagement, pages 763–782. Springer, 2012.

[21] Jennifer A Fredricks, Phyllis C Blumenfeld, and Alison H Paris. School engagement: Potential of the concept, state of the evidence. Review of educational research, 74(1):59–109, 2004.

[22] Mary Sinclair, Sandra Christenson, Camilla Lehr, and Amy Anderson. Facilitating student engagement: Lessons learned from check & connect longitudinal studies. The California School Psychologist, 8:29–41, 01 2014.

[23] Ning Hao, Hua Xue, Huan Yuan, Qing Wang, and Mark Runco. Enhancing creativity: Proper body posture meets proper emotion. Acta psychologica, 173:32–40, 12 2017.

[24] Valentina Andolfi, Chiara di nuzzo, and Alessandro Antonietti. Opening the mind through the body: The effects of posture on creative processes. Thinking Skills and Creativity, 24, 02 2017.

[25] James Appleton, Sandra Christenson, Dongjin Kim, and Amy Reschly. Measuring cognitive and psychological engagement: Validation of the student engagement instrument. Journal of School Psychology, 44:427–445, 10 2006.

[26] Jennifer Fredricks and Wendy Mccolskey. The Measurement of Student Engagement: A Comparative Analysis of Various Methods and Student Self-report Instruments, pages 763–782. Springer Science & Business Media, 01 2012.

[27] Ning Hao, Huan Yuan, Yi Hu, and Roland Grabner. Interaction effect of body position and arm posture on creative thinking. Learning and Individual Differences, 32, 05 2014.

[28] Paul Seli, Jonathan Carriere, David Thomson, James Cheyne, Kaylena Ehgoetz Martens, and Daniel Smilek. Restless mind, restless body. Journal of experimental psychology. Learning, memory, and cognition, 40, 12 2013.

[29] Ellen Skinner and Michael Belmont. Motivation in the classroom: Reciprocal effects of teacher behavior and student engagement across the school year. Journal of Educational Psychology, 85:571–581, 12 1993.

[30] Kristin E. Voelkl. Identification with school. American Journal of Education, 105(3):294–318, 1997.

[31] Judith L. Meece, Phyllis Blumenfeld, and Rick H. Hoyle. Students' goal orientations and cognitive engagement in classroom activities. 1988.

[32] Michelene T. H. Chi and Ruth Wylie. The icap framework: Linking cognitive engagement to active learning outcomes. Educational Psychologist, 49:219 – 243, 2014.

[33] Ashish Kapoor and Rosalind W. Picard. Multimodal affect recognition in learning environments. In Proceedings of the 13th Annual ACM International Conference on Multimedia, pages 677–682, 2005.

[34] Bethany McDaniel, Sidney D'Mello, Brandon King, Patrick Chipman, Kristy Tapp, and Art Graesser. Facial features for affective state detection in learning environments. In Proceedings of the Annual Meeting of the Cognitive Science Society, volume 29, 2007.

[35] Sidney K. D'Mello and Arthur Graesser. Multimodal semi-automated affect detection from conversational cues, gross body language, and facial features. User Modeling and User-Adapted Interaction, 20(2):147–187, 2010.

[36] Joseph F. Grafsgaard, Joseph B. Wiggins, Kristy Elizabeth Boyer, Eric N. Wiebe, and James C. Lester. Automatically recognizing facial indicators of frustration: A learning-centric analysis. 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction, 2013.

[37] Jacob Whitehill, Zewelanji Serpell, Yi-Ching Lin, Aysha Foster, and Javier R. Movellan. The faces of engagement: Automatic recognition of student engagement from facial expressions. IEEE Trans. Affective Computing, 5(1):86–98, 2014.

[38] Gwen Littlewort, Jacob Whitehill, Tingfan Wu, Ian Fasel, Mark Frank, Javier Movellan, and Marian Bartlett. The computer expression recognition toolbox (cert). In 2011 IEEE International Conference on Automatic Face & Gesture Recognition (FG), pages 298–305. IEEE, 2011.

[39] Yan-Ying Li and Yi-Ping Hung. Feature fusion of face and body for engagement intensity detection. In 2019 IEEE International Conference on Image Processing (ICIP), pages 3312–3316. IEEE, 2019.

[40] Hamed Monkaresi, Nigel Bosch, Rafael A Calvo, and Sidney K D'Mello. Automated detection of engagement using video-based estimation of facial expressions and heart rate. IEEE Transactions on Affective Computing, 8(1):15–28, 2016.

[41] Nathan L Henderson, Jonathan P Rowe, Bradford W Mott, and James C Lester. Sensor-based data fusion for multimodal affect detection in game-based learning environments. In EDM (Workshops), pages 44–50, 2019.

[42] Svati Dhamija and Terrance E Boult. Exploring contextual engagement for trauma recovery. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pages 19–29, 2017.

[43] Debora Duarte Macea, Krzysztof Gajos, Yasser Armynd Daglia Calil, and Felipe Fregni. The efficacy of web-based cognitive behavioral interventions for chronic pain: a systematic review and meta-analysis. The Journal of Pain, 11(10):917–929, 2010.

[44] Susan Unok Marks and Russell Gersten. Engagement and disengagement between special and general educators: An application of miles and huberman's cross-case analysis. Learning Disability Quarterly, 21(1):34–56, 1998.

[45] Svati Dhamija and Terranee E Boult. Automated mood-aware engagement prediction. In 2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII), pages 1–8. IEEE, 2017.

[46] Karan Ahuja, Deval Shah, Sujeath Pareddy, Franceska Xhakaj, Amy Ogan, Yuvraj Agarwal, and Chris Harrison. Classroom digital twins with instrumentation-free gaze tracking. CHI '21, New York, NY, USA, 2021. Association for Computing Machinery.

[47] Eslam Mostafa, Asem A. Ali, Ahmed Shalaby, and Aly Farag. A facial features detector integrating holistic facial information and part-based model. In CVPR-Workshops, 2015.

[48] Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. Openface 2.0: Facial behavior analysis toolkit. In 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), pages 59–66. IEEE, 2018.

[49] Amir Zadeh, Yao Chong Lim, Tadas Baltrusaitis, and Louis-Philippe Morency. Convolutional experts constrained local model for 3d facial landmark detection. In Proceedings of the IEEE International Conference on Computer Vision Workshops, pages 2519–2528, 2017.

[50] Joel A Hesch and Stergios I Roumeliotis. A direct least-squares (dls) method for pnp. In 2011 International Conference on Computer Vision, pages 383–390. IEEE, 2011.

[51] Narayanan Veliyath, Pradipta De, Andrew A Allen, Charles B Hodges, and Aniruddha Mitra. Modeling students' attention in the classroom using eyetrackers. In Proceedings of the 2019 ACM Southeast Conference, pages 2–9, 2019.

[52] Erroll Wood, Tadas Baltrusaitis, Xucong Zhang, Yusuke Sugano, Peter Robinson, and Andreas Bulling. Rendering of eyes for eye-shape registration and gaze estimation. In Proceedings of the IEEE International Conference on Computer Vision, pages 3756–3764, 2015.

[53] Tadas Baltrusaitis, Peter Robinson, and Louis-Philippe Morency. Constrained local neural fields for robust facial landmark detection in the wild.

In Proceedings of the IEEE international conference on computer vision workshops, pages 354–361, 2013.

[54] Erroll Wood, Tadas Baltrusaitis, Xucong Zhang, Yusuke Sugano, Peter Robinson, and Andreas Bulling. Rendering of eyes for eye-shape registration and gaze estimation. In Proceedings of the IEEE International Conference on Computer Vision, pages 3756–3764, 2015.

[55] Christos Sagonas, Epameinondas Antonakos, Georgios Tzimiropoulos, Stefanos Zafeiriou, and Maja Pantic. 300 faces in-the-wild challenge: database and results. Image and Vision Computing, 47:3–18, 2016.

[56] Tal Hassner, Shai Harel, Eran Paz, and Roee Enbar. Effective face frontalization in unconstrained images. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015.

[57] Oliver Langner, Ron Dotsch, Gijsbert Bijlstra, Daniel H. J. Wigboldus, Skyler T. Hawk, and Ad Van Knippenberg. Presentation and validation of the radboud faces database. Cognition & Emotion, 24(8):1377–1388, 2010.

[58] Mikel Ariz, José J Bengoechea, Arantxa Villanueva, and Rafael Cabeza. A novel 2d/3d database with automatic face annotation for head tracking and pose estimation. Computer Vision and Image Understanding, 148:201–210, 2016.

[59] Gianluca Fadda, Gian Luca Marcialis, Fabio Roli, and Luca Ghiani. Exploiting the golden ratio on human faces for head-pose estimation. In International Conference on Image Analysis and Processing, pages 280–289. Springer, 2013.

[60] Stylianos Asteriadis, Dimitris Soufleros, Kostas Karpouzis, and Stefanos Kollias. A natural head pose and eye gaze dataset. In Proceedings of the International Workshop on Affective-Aware Virtual Agents and Social Robots, pages 1–4, 2009.

[61] Brian A. Smith, Qi Yin, Steven K. Feiner, and Shree K. Nayar. Gaze locking: Passive eye contact detection for human-object interaction. In Proceedings of the 26th Annual ACM Symposium on User Interface Software and Technology, pages 271–280, 2013.

[62] Mykhaylo Andriluka, Stefan Roth, and Bernt Schiele. Monocular 3d pose estimation and tracking by detection. In 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pages 623–630. IEEE, 2010.

[63] Vasileios Belagiannis and Andrew Zisserman. Recurrent human pose estimation. In 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017), pages 468–475. IEEE, 2017.

[64] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 7291–7299, 2017.

[65] Yilun Chen, Zhicheng Wang, Yuxiang Peng, Zhiqiang Zhang, Gang Yu, and Jian Sun. Cascaded pyramid network for multi-person pose estimation. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 7103–7112, 2018.

[66] Tomas Simon, Hanbyul Joo, Iain Matthews, and Yaser Sheikh. Hand keypoint detection in single images using multiview bootstrapping. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, pages 1145–1153, 2017.

[67] Christian Zimmermann and Thomas Brox. Learning to estimate 3d hand pose from single rgb images. In Proceedings of the IEEE international conference on computer vision, pages 4903–4911, 2017.

[68] Erin S Lane and Sara E Harris. A new tool for measuring student behavioral engagement in large university classes. Journal of College Science Teaching, 44(6):83–91, 2015.

[69] Tasnia Tabassum, Andrew A Allen, and Pradipta De. Non-intrusive identification of student attentiveness and finding their correlation with detectable facial emotions. In Proceedings of the 2020 ACM Southeast Conference, pages 127–134, 2020.

[70] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. Ntu rgb+d: A large scale dataset for 3d human activity analysis. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 1010–1019, 2016.

[71] Kristin E Voelkl. Identification with school. American Journal of Education, 105(3):294–318, 1997.

[72] Victoria Contreras. ARTNATOMYA. www.artnatomia.net, 2006.

[73] Patrik Huber, Guosheng Hu, Jose Rafael Tena, Pouria Mortazavian, Willem P. Koppen, William J. Christmas, Matthias Rätsch, and Josef Kittler. A multiresolution 3d morphable face model and fitting framework. In Proceedings of the 11th Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2016) - Volume 4: VISAPP, Rome, Italy, February 27-29, 2016., pages 79–86, 2016.

[74] Patrick Lucey, Jeffrey F. Cohn, Takeo Kanade, Jason Saragih, Zara Ambadar, and Iain Matthews. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression, 2010.

[75] C. Darwin. The Expression of Emotions in Man and Animals. John Murray 1872, reprinted by University of Chicago Press.

[76] P. Ekman and W. Friesen. Facial Action Coding System: A Technique for the Measurement of Facial Movement. Consulting Psychologists Press, 1978.

[77] Christine L. Lisetti and Diane J. Schiano. Automatic facial expression interpretation: Where human-computer interaction, artificial intelligence and cognitive science intersect. Pragmatics & Cognition, 8(1):185–235, 2000.

[78] G. Duchenne. Mecanisme de la physionomie humaine. Paris, France: Renouard, 1862.

[79] Kaili Zhao, Wen-Sheng Chu, and Honggang Zhang. Deep region and multi-label learning for facial action unit detection. In CVPR, 2016.

[80] Lin Zhong, Qingshan Liu, Peng Yang, Junzhou Huang, and Dimitris N. Metaxas. Learning multiscale active facial patches for expression analysis. IEEE Trans. Cybernetics, 45(8):1499–1510, 2015.

[81] Shizhong Han Ping Liu. Facial expression recognition via a boosted deep belief network. In CVPR, 2014.

[82] Ping Liu, Joey Tianyi Zhou, Ivor Wai-Hung Tsang, Zibo Meng, Shizhong Han, and Yan Tong. Feature disentangling machine - A novel approach of feature selection and disentangling in facial expression analysis. In ECCV, 2014.

[83] Kaili Zhao, Wen-Sheng Chu, Fernando De la Torre, Jeffrey F. Cohn, and Honggang Zhang. Joint patch and multi-label learning for facial action unit detection. In CVPR, 2015.

[84] Stefanos Eleftheriadis, Ognjen Rudovic, and Maja Pantic. Multi-conditional latent variable model for joint facial action unit detection. In ICCV, 2015.

[85] Wei Li, Farnaz Abtahi, Zhigang Zhu, and Lijun Yin. EAC-Net: A region-based deep enhancing and cropping approach for facial action unit detection. In IEEE Int. Conf. on Automatic Face & Gesture Recognition, 2017.

[86] P. Ekman, W.V. Friesen, and J.C. Hager. Facial Action Coding System (FACS): Manual. A Human Face, 2002.

[87] Patrick Lucey, Jeffrey F. Cohn, Takeo Kanade, Jason Saragih, Zara Ambadar, and Iain Matthews. The extended cohn-kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression. In CVPR-Workshops, 2010.

[88] Gwen Littlewort, Marian Stewart Bartlett, Ian R. Fasel, Joshua Susskind, and Javier R. Movellan. Dynamics of facial expression extracted automatically from video. Image Vision Comput., 24(6):615–625, 2006.

[89] Xiao Zhang and Mohammad H. Mahoor. Task-dependent multi-task multiple kernel learning for facial action unit detection. Pattern Recogn., 51(C):187–196, 2016.

[90] Ziheng Wang, Yongqiang Li, Shangfei Wang, and Qiang Ji. Capturing global semantic relationships for facial action unit recognition. In ICCV, 2013.

[91] Yan Tong and Qiang Ji. Learning bayesian networks with qualitative constraints. In CVPR, 2008.

[92] Wen-Sheng Chu, Fernando De la Torre, and Jeffery F. Cohn. Selective transfer machine for personalized facial action unit detection. In CVPR, 2013.

[93] Simon Lucey, Iain A. Matth ews, Changbo Hu, Zara Ambadar, Fernando De la Torre, and Jeffrey F. Cohn. AAM derived face representations for robust facial action recognition. In Int. Conf. on Automatic Face & Gesture Recognition, 2006.

[94] Carlos Fabian Benitez-Quiroz, Ramprakash Srinivasan, and Aleix M. Martínez. Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. In CVPR, 2016.

[95] Shashank Jaiswal and Michel F. Valstar. Deep learning the dynamic appearance and shape of facial action units. In IEEE Winter Conference on Applications of Computer Vision, 2016.

[96] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. Web-scale training for face identification. In CVPR, pages 2746–2754, 2015.

[97] Vinod Nair and Geoffrey E. Hinton. Rectified linear units improve restricted boltzmann machines. In Proceedings of the 27th International Conference on Machine Learning (ICML-10), pages 807–814, 2010.

[98] Matthew D. Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In ECCV, 2014.

[99] Zoltán Tosér, László A. Jeni, András Lörincz, and Jeffrey F. Cohn. Deep learning for facial action unit detection under large head poses. In ECCV-Workshops, 2016.

[100] Min-Ling Zhang and Zhi-Hua Zhou. A review on multi-label learning algorithms. IEEE Trans. Knowl. Data Eng., 26(8):1819–1837, 2014.

[101] László A. Jeni, Jeffrey F. Cohn, and Fernando De La Torre. Facing imbalanced data–recommendations for the use of performance metrics. In IEEE Humaine Association Conference on Affective Computing and Intelligent Interaction, 2013.

[102] Min-Ling Zhang, Yu-Kun Li, and Xu-Ying Liu. Towards class-imbalance aware multi-label learning. In Int. Joint Conf. on Artificial Intelligence, 2015.

[103] Zheng Zhang, Jeff M. Girard, Yue Wu, Xing Zhang, Peng Liu, Umur Ciftci, Shaun Canavan, Michael Reale, Andy Horowitz, Huiyuan Yang, Jeffrey F. Cohn, Qiang Ji, and Lijun Yin. Multimodal spontaneous emotion corpus for human behavior analysis. In CVPR, June 2016.

[104] Matthew D. Zeiler. Adadelta: An adaptive learning rate method. CoRR, abs/1212.5701, 2012.

[105] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. CoRR, abs/1312.6034, 2013.

[106] Kathrin Kaulard, Douglas W. Cunningham, Heinrich H. Bülthoff, and Christian Wallraven. The mpi facial expression database — a validated database of emotional and conversational facial expressions. PLoS ONE, 7(3), 2012.

[107] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In Advances in Neural Information Processing Systems 25, pages 1097–1105. 2012.

[108] Wallace V Friesen, Paul Ekman, et al. Emfacs-7: Emotional facial action coding system. Unpublished manuscript, University of California at San Francisco, 2(36):1, 1983.

CURRICULUM VITAE

Islam Mohamed Ahmed Mohamed Mahmoud Alkabbany
Louisville, Kentucky 40217
Islam.Alkabbany@Louisville.edu

Education

- PhD in ECE, Speed School of Engineering, University of Louisville

  Working on modeling students' Emotinal Engagment

  Current GPA: 4.0 Expected graduation date: 2021

- MSc. Electrcial Engineering, Faculty of Engineering, Assiut University, 2013

  Thesis Title: Energy Efficient Connectivity Techniques for Mobile Wireless Sensor Network.

- BSc. Electrcial Engineering, Faculty of Engineering, Assiut University, 2007

  Total grade: Excellent with degree of honor

  Graduation Project: Real time system in CNC operation.

Current Position

- Research Assistant, CVIP Lab, Speed School of Engineering, University of Louisville, 6/2020 - now.

Work Experience

- Teaching Assistant, Electrcial and Computer Engineering, Speed School of Engineering, University of Louisville, 1/2019 - 5/2020
- Research Assistant, CVIP Lab, Speed School of Engineering, University of Louisville, 1/2016 - 12/2018
- Lecturer Assistant, Electrical Engineering Department, Faculty of Engineering, Assiut University, 5/2013 - now
- Demonstrator, Electrical Engineering Department, Faculty of Engineering, Assiut University, 11/2007 - 5/2013

Selected Publications
Full list of publications could be found on
https://scholar.google.com/citations?user=FFjQRw0AAAAJ&hl=en

- Aly Farag, Asem Ali, **Islam Alkabbany**, Foreman, James Christopher, Thomas Tretter, Marci S. DeCaro, and Nicholas Carl Hindy. "TToward a Quantitative Engagement Monitor for STEM Education." In 2021 ASEE Virtual Annual Conference Content Access. 2021.

- Foreman, James Christopher, Aly Farag, Asem Ali, **Islam Alkabbany**, Marci S. DeCaro, and Thomas Tretter. "Toward a Multi-dimensional Biometric Approach to Quantifying Student Engagement in the STEM Classroom." In 2020 ASEE Virtual Annual Conference Content Access. 2020.

- **Islam Alkabbany**, Asem Ali, Amal Farag, Ian Bennett, Mohamad Ghanoum, and Aly Farag. "Measuring student engagement level using facial information." In 2019 IEEE International Conference on Image Processing (ICIP), pp. 3337-3341. IEEE, 2019.

- Ali, Asem M., **Islam Alkabbany**, Amal Farag, Ian Bennett, and Aly Farag. "Facial action units detection under pose variations using deep regions learning." In 2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII), pp. 395-400. IEEE Computer Society, 2017.

Linguistic skills
First language: Arabic (Native speaker)
Second language: English (good)
Personal Information
Current Address: 776 David Fairliegh Ct. Apt. 3 Louisville, KY, 40217
Cell-Phone: +1 (502) 956-9496
E-Mails: Islam.Alkabbany@louisville.edu
Date of birth: 01/26/1986 Nationality: Egyptian
Marital Status: Married Religion: Muslim