DEVELOPMENT OF UNOBTRUSIVE AFFECTIVE COMPUTING FRAMEWORK FOR STUDENTS' ENGAGEMENT ANALYSIS IN CLASSROOM ENVIRONMENT

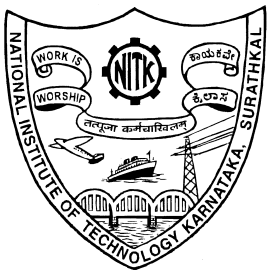
Thesis

Submitted in partial fulfilment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

by

Mr. Ashwin T S



DEPARTMENT OF INFORMATION TECHNOLOGY

NATIONAL INSTITUTE OF TECHNOLOGY KARNATAKA
SURATHKAL, MANGALORE - 575025

May 2020

Declaration

I hereby *declare* that the Research Thesis entitled ”Development of Unobtrusive Affective Computing Framework for Students' Engagement Analysis in Classroom Environment” which is being submitted to the National Institute of Technology Karnataka, Surathkal in partial fulfilment of the requirements for the award of the Degree of Doctor of Philosophy in Information Technology is a *bonafide report of the research work carried out by me*. The material contained in this thesis has not been submitted to any University or Institution for the award of any degree.

Place: NITK Surathkal
Date:

(148045IT14F04, Ashwin T S)
Department of Information Technology

Certificate

This is to *certify* that the Research Thesis entitled ”Development of Unobtrusive Affective Computing Framework for Students' Engagement Analysis in Classroom Environment” submitted by Mr. Ashwin T S (Register Number: 148045IT14F04) as the record of the research work carried out by him, is *accepted as the Research Thesis submission* in partial fulfilment of the requirements for the award of degree of Doctor of Philosophy.

Dr. G. Ram Mohana Reddy
Research Guide
Professor (HAG Scale), IT Department
NITK Surathkal, Mangalore - 575025

Chairman - DRPC
(Signature with Date and Seal)

Acknowledgements

I take the privilege to express my heartfelt gratitude to all the people who have been a constant support to me throughout my doctoral research. Foremost, my sincere thanks to my research guide Prof. G. Ram Mohana Reddy, Information Technology Department, NITK Surathkal for his valuable guidance, enthusiasm, inspiration, and dedication throughout my research. This research work is a result of his timely suggestions and firm decisions. I extend my gratitude to Mrs. Vijayalakshmi Ram Mohana Reddy for her affection, compassion, and continuous motivation, which was an added privilege.

My sincere thanks to the RPAC members, Prof. Shashidhar G Koolagudi, and Prof. Shyam S. Kamath, for their apt & timely suggestions and valuable reciprocations. I would also like to thank Prof. Gangadharan K V, for providing his constructive inputs and the data for the research work. I also thank all the teaching, technical, administrative, as well as non-teaching staff who have been very kind and helpful throughout my research work. Further, I extend my gratitude to Prof. Geetha Maiya, Director (Students' affairs), MIT Manipal, who motivated me to do research at NITK Surathkal.

I also thank my beloved parents, in-laws, and family members for their exhaustive encouragement and support. I extend my special thanks to my wife, Srigowri M P, for her cooperation, love, suggestions, and constant encouragement at every point of time.

I extend my gratitude to my dear ones Shridhar Domanal, Natesha, Ranjith, Karthik N, Manjunath, Gokul, Sanjay, Karthik K, my other co-researchers and also my brother Sachin T S who made my journey more enjoyable and fun.

Never to forget, the most exciting place for the doctoral research with updated technologies usable for research and with advanced resources made easily available, facilitating a convenient study, my heartfelt gratitude to NITK Surathkal for making this research memorable and possible.

(Mr. Ashwin T S)

Abstract

Pervasive intelligent learning environments can be made more personalized by adapting the teaching strategies according to the students' emotional and behavioral engagements. The students' engagement analysis helps to foster those emotions and behavioral patterns that are beneficial to learning, thus improving the effectiveness of the teaching-learning process. The students' emotional and behavioral patterns are to be recognized unobtrusively using learning-centered emotions (engaged, confused, frustrated, and so on), and engagement levels (looking away from the tutor or board, eyes completely closed, and so on).

Recognizing both the behavioral and emotional engagement from students' image data in the wild (obtained from classrooms) is a challenging task. The use of the multitude of modalities enhances the performance of affective state classification, but recognizing facial expressions, hand gestures, and body posture of each student in a classroom environment is another challenge. Here, the classification of affective states is not sufficient, object localization also plays a vital role. Both the classification and object localization should be robust enough to perform better for various image variants such as occlusion, background clutter, pose, illumination, cultural & regional background, intra-class variations, cropped images, multipoint view, and deformations.

The most popular and state-of-the-art classification and localization techniques are machine and deep learning techniques that depend on a database for the ground truth. A standard database that contains data from different learning environments with a multitude of modalities is also required. Hence, in the research work, different deep learning architectures are proposed to classify the students' affective states with object localization. A standard database with students' multimodal affective states is created and benchmarked. The students' affective states obtained from the proposed real-time affective state classification method is used as feedback to the teacher in order to enhance the teaching-learning process in four different learning environments, namely: e-learning, classrooms, webinars and flipped classrooms. More details on the contribution of this thesis are as follows.

A real-time students' emotional engagement analysis is proposed for both the individual and group of students based on their facial expressions, hand gestures, and body postures for e-learning, flipped classroom, classroom, and webinar environments. Both basic and learning-centered emotions are used in the study. Various CNN based architectures are proposed to predict the students' emotional engagement. The students' behavioral engagement analysis method is also proposed and implemented in the classroom and computer-enabled teaching laboratories. The proposed scale-invariant context assisted single-shot CNN architecture performed well for multiple students in a single image frame. A single group engagement level score for each frame is obtained using the proposed feature fusion technique.

The proposed model effectively classifies the students' affective states into teacher-centric attentive and in-attentive affective states. Inquiry interventions are proposed to address the negative impact of in-attentive affective states on the performance of students. Experimental results demonstrated a positive correlation between the students learning rate and their attentive affective state engagement score for both individual and group of students. Further, an affective state transition diagram and visualizations are proposed to help the students and the teachers to improve the teaching-learning process.

A multimodal database is created for both e-learning (single student in a single image frame) and classroom environments (multiple students in a single image frame) using the students' facial expressions, hand gestures, and body postures. Both posed and spontaneous expressions are collected to make the training set more robust. Also, various image variants are considered during the dataset creation. Annotations are performed using the gold standard study for eleven different affective states and four different engagement levels. Object localization is performed on each modality of every student, and the bounding box coordinates are stored along with the affective state/engagement level. This database is benchmarked with various popular classification algorithms and state-of-the-art deep learning architectures.

Keywords: Affective Computing; Affect Sensing and Analysis; Behavioral Patterns; Classroom Data in the Wild; Computer Vision; Multimodal Analysis; Student Engagement Analysis.

Contents

List of Tables

List of Figures

List of Abbreviations

Abbreviation	Meaning
AAM	Active Appearance Model
AFA	Automatic Face Analysis
AFEW	Acted Facial Expressions in the Wild
ANOVA	ANalysis Of VAriance
AR	Augmented Reality
AUC	Area Under the Curve
AUSSE	Australasian Survey of Student Engagement
AVEC	Audio/Visual Emotion and Depression Recognition
BROMP	Baker Rodrigo Ocumpaugh Monitoring Protocol
BU-3DFE	Binghamton University 3D Facial Expression
CCTV	Closed Circuit TV
CERT	Computer Expression Recognition Toolbox
CK+	Cohn-Kanade
CNN	Convolutional Neural Networks
CPM	Context-sensitive Predict Module
CPU	Central Processing Unit
DAGER	Deep Age, Gender and Emotion Recognition system
EL	Engagement Level
FABO	Face and Body Gesture
FACS	Facial Action Coding System
FDDB	Face Detection Data Set and Benchmark
FER	Facial Expression Recognition
GHQ	General Health Questionnaire
GPU	Graphics Processing Unit
HAPPEI	HAPpy PEople Images
HOG	Histogram of Oriented Gradients
IMFBD	Internet Movie Firearms Database
InIv	Inquiry Intervention
IOU	Intersection Over Union

ISE	Indian Spontaneous Expression
JAFFE	Japanese Female Facial Expression
JSON	JavaScript Object Notation
kNN	k-Nearest Neighbors
LBP	Local Binary Patterns
LFPL	Low-Level Feature Pyramid Layers
LFW	Labeled Faces in the Wild
LGBP	Local Gabor Binary Pattern
LSTM	Long Short-Term Memory
MCC	Matthews Correlation Coefficient
MCCN	Multitask Cascaded Convolutional Networks
MELD	Multimodal Emotionlines Dataset
MFCC	Mel-Frequency Cepstral Coefficients
MIT	Massachusetts Institute of Technology
MSE	Mean Square Error
NITK	National Institute of Technology Karnataka
NSSE	National Survey of Student Engagement
NYU	New York University
OS	Operating System
PA	Pyramid Anchors
PCA	Principle Component Analysis
PHOG	Pyramid Histogram of Gradients
RAM	Random Access Memory
RBM	Restricted Boltzmann Machine
RGB	Red Green and Blue
RNN	Recurrent Neural Networks
SAVEE	Surrey Audio-Visual Expressed Emotion
S3FD	Single-Shot Scale-invariant Face Detector
SD	Standard Deviation
SIFT	Scale-Invariant Feature Transform

SSD	Single-Shot MultiBox Detector
STAI	State Trait Anxiety Inventory
SVM	Support Vector Machine
TQ	Test Questionnaires
USA	United States of America
VR	Virtual Reality
WEKA	Waikato Environment for Knowledge Analysis
WITS	Wits Intelligent Teaching System
YFD	Yale Face Database
YOLO	You Only Look Once
ZCA	Zero Components Analysis

Chapter 1

Introduction

Learning environments play a vital role in the modern education system where the students learn and interact with the real-tutor (teacher) or auto-tutor to gain the required knowledge. In traditional learning, due to human to human interaction, teachers can interact with the students according to the students' visible behaviors, emotions, and so on. In this era of smart city, there are so many learning environments that are used in the teaching-learning process, such as social e-learning, collaborative e-learning, m-learning, and so on. Even there is the smart campus which not only uses traditional learning and e-learning but also uses webinars, and auto-tutors.

All these different learning environments can be made more personalized by incorporating students' behavioral and emotional engagement analysis systems, which can automatically predict the students' engagement and accordingly adapt the teaching strategy to improve the teaching-learning process. Students' engagement using their non-verbal cues can be analyzed using Affective Computing. Affective computing is a separate domain that includes the emotional, cognitive, and behavioral aspects of humans. More details pertinent to different learning environments used in the teaching-learning process, students' affective content analysis, automatically measuring the students' engagement, and other related literature that uses affective computing in the education domain are given in the following sections. This chapter concludes with the motivation behind this research work and the outline of the thesis.

1.1 Affective Computing

Affective computing is the study and development of systems and devices that can recognize, interpret, process, and simulate human affects. Affective Computing combines engineering and computer science with psychology, cognitive science, neuroscience, sociology, education, psychophysiology, value-centered design, ethics, and more (Picard, 1997).

1.2 Learning Environments

Learning environments such as classroom, flipped classroom, e-learning, and webinar are widely used (Figure 1.1).

- **Classrooms:** In a traditional classroom learning, the teacher is the primary disseminator of information during the class period, and the students can practice and explore more post classroom instructional hours.
- **Flipped Classroom:** Flipped classroom instructional strategy is a reverse of classroom learning, where the students learn the concepts before coming to the class, and the classroom is used to explore topics in greater depth and create meaningful learning opportunities while students are initially introduced to new topics outside the classroom (Tucker, 2012). In flipped classrooms, students often use online sources to learn the concept outside the classroom.
- **E-learning:** eLearning is learning utilizing electronic technologies to access educational curriculum outside of a traditional classroom. In most cases, it refers to a course, program or degree delivered completely online. Students generally use e-learning platforms to learn the concepts, and few other options are also used like collaborative learning/ e-learning. Collaborative e-learning is defined as a process where a group of students learn or attempt to learn using e-learning platform.
- **Webinars:** Apart from e-learning and flipped classroom environments, the students can also learn from traditional classroom lectures in real-time using the online platforms such as web seminars referred to as webinars, where teacher responds to questions while students refer directly to the teacher for guidance and feedback.

All four learning environments fall under a wide classification of asynchronous and synchronous learning (Hrastinski, 2008). Learning is an affectively charged experience where the affective states continually occur throughout the learning process (D'Mello et al., 2007). An effective learning agent (human or auto-tutor) fosters the affective states which are beneficial to learning (Sidney et al., 2005). Synchronous and asynchronous learning are the two types of learning events based on students' engagement and medium of learning. Classroom-based learning is synchronous learning, which deals with a group of students taught by an instructor in an indoor/outdoor environment. On the other hand, most of the online education systems (e-learning) are asynchronous learning, where each student can learn the subject using electronic education

technologies like massive open online courses (MOOCs). In both synchronous and asynchronous learning environments, the students' affective states can be recognized using video-recorded data or image frames. The video recorded data with respect to (w.r.t.) classroom environment consists of a group of students (multiperson) in each image frame, whereas the e-learning scenario contains a single person in a single image frame.

1.3 Students' Engagement Analysis

Students' engagement is closely associated with their conceptual understanding, and it is broadly classified into four major categories, namely: emotional, behavioral, cognitive, and agentic engagements (Sinatra et al., 2015; Castellano et al., 2008).

- **Emotional Engagement:** Emotional engagement is defined as the students' emotional reactions to academic subject areas. Learning-centered and academic emotions are few popular categories used to measure emotional engagement (D'Mello et al., 2007; Bosch et al., 2016; Sinatra et al., 2015).
- **Behavioral Engagement:** The students' motivation to participate through their actions in learning is referred to as behavioral engagement. Behavioral aspects of attention, such as making eye contact and leaning forward during a discussion, self-directed academic behavior such as exhibiting resiliency in the face of obstacles, and so on, are used to measure the behavioral engagement (Whitehill et al., 2014; Sinatra et al., 2015; Yun et al., 2018).
- **Cognitive Engagement:** A widely used definition of cognitive engagement is a psychological investment where a student becomes psychologically invested when he(she) expands cognitive effort to understand, goes beyond the requirement of the activity, uses flexible problem solving, and chooses challenging tasks. The dimensions of cognitive engagement overlap with dimensions of both behavioral engagement and emotional engagement (Whitehill et al., 2014; Sinatra et al., 2015; Yun et al., 2018).
- **Agentic Engagement:** Agentic engagement is the fourth dimension of engagement, where students are proactive during instructions (Sinatra et al., 2015).

1.4 Analyzing Students' Engagement

The students' engagement is analyzed in various ways through self-reports, survey-based methods like NSSE (National Survey of Student Engagement), teacher introspective evaluations, checklists, speech/voice recognition techniques, physiological sensors

such as pulse rate, pressure sensors, learning environment's video content analysis, and others (Kuh, 2003; Kahu, 2013; Kuh et al., 2008; Zilvinskis et al., 2017; Calvo & D'Mello, 2010; D'mello & Graesser, 2012; D'Mello et al., 2010; Wang & Ji, 2015).

1.5 Unobtrusive Students' Engagement Analysis

Learning environments are classified as synchronous and asynchronous based on students' engagement and medium of learning. There are limited works on the students' engagement analysis performed in a synchronous learning environment like the classroom. The survey-based methods and the use of physiological sensors for each student present in a large classroom are both time-consuming and obtrusive (Kuh, 2003). Speech/voice recognition for students' engagement analysis in a large classroom is not feasible as each student may not get the opportunity to interact with the teacher all the time (Castellano et al., 2008). In a synchronized learning environment, the unobtrusive students' engagement can be effectively recognized using non-verbal cues such as facial expressions, hand gestures, and body postures captured from the video image frames of the classroom data (Whitehill et al., 2014; Zaletelj & Košir, 2017; Gupta et al., 2019; Ashwin & Guddeti, 2018).

Image frame based analysis deals with issues such as occlusion, background clutter, pose, illumination, cultural & regional background, intra-class variations, cropped images, multipoint view, and deformations. To address these issues, various techniques such as multiangle optimal pattern (Jain et al., 2017), video summarization (Muhammad et al., 2018), density estimation, and detection with scale-aware, context-aware, or multitask frameworks (Sindagi & Patel, 2018) were proposed. Multimodal analysis is another challenge, and various techniques such as Convolutional Neural Networks (CNN), Deep-CNN, Long Short-Term Memory (LSTM), and temporal CNNs were used for its analysis (Rahmani et al., 2018; Varol et al., 2018; Liu et al., 2018). But these techniques were not explored in different learning environments.

1.6 Affective State Classification

Knowledge and goals of the learner influence the students' affective states and vice versa (Bosch et al., 2016). Current research on the students' emotion recognition using their facial expressions include the classification of emotions into learning-centered emotions (boredom, confusion, frustration, eureka, flow/engagement) (D'mello & Graesser, 2012) and Ekman's basic emotions (happy, sad, fear, anger, disgust, and surprise) (Psaltis et al., 2017). Other classifications such as interested or not-interested (Klein & Celik, 2017); distracted or engaged were also considered (Thomas & Jayagopi, 2017). These classification techniques are confined to either e-learning or for the classroom, but not for both the environments. The existing literature focuses on student's video affective content analysis to predict their emotional engagement such as frustrated, confused, happiness and others (D'Mello et al., 2007; Ashwin et al., 2015), whereas a few other works considered mainly behavioral engagement such as looking away from the computer, eyes completely closed and the like from the students' video content (Whitehill et al., 2014; Silfver et al., 2018; Ashwin & Guddeti, 2018).

The students' emotional and behavioral engagement works are performed both in e-learning and classroom environments, but the students' engagement analysis is performed on a single person in a single image frame. However, in the real classroom scenario, there are multiple students in a single image frame with students' faces in the wild. Existing works for feature extraction like Histogram of Oriented Gradients (HOG), Scale-Invariant Feature Transform (SIFT), and tools like OpenFace are used for real-time students' emotion recognition. All these methods and tools fail to perform better for multiple students in a single image frame (Thomas & Jayagopi, 2017; Wang & Ji, 2015). Deep learning techniques for affective state classification are more accurate than existing machine learning techniques, but these are not well explored for students' identification and the affective state classification of multiple students in a single image frame.

1.7 Multitude of Modalities

Existing works analyzed a student's emotion in a controlled environment for a single student in a single image frame. Some works induced frustration, curiosity, and other emotions of the students, and analyzed their emotional relationship with their performance (D'Mello, 2012; Craig et al., 2004). But there are no works on faces in the wild data of a classroom. Though facial expressions are used in the majority of affective state recognition techniques, many research studies demonstrate that there is a significant contribution of body posture and hand gesture for the recognition of the students' affective states (D'Mello et al., 2007; Patwardhan & Knapp, 2014). For example, results in (Patwardhan & Knapp, 2014) demonstrated that more than 70% of the recognition accuracy is due to the hand gesture and body posture concerning the anger affective state.

1.8 Unobtrusive Students' Engagement Database

The human affective states can be recognized through facial expressions, voice, hand gestures, and body postures. There are many research studies on the human affective state recognition (Picard, 1997; Glowinski et al., 2011; Wang & Ji, 2015). There exist several state-of-the-art machine & deep learning and computer vision methods for affective state recognition and its analysis (Szegedy et al., 2016; Krizhevsky et al., 2012; Kleinsmith & Bianchi-Berthouze, 2013). These methods require a proper database to train, test, and validate. The accuracy of affective state recognition is dependent on the training data. This accuracy is affected by various parameters of training data such as the pose, background clutter, intra-class variation, illumination, deformation, occlusion, posed expression, natural expressions, cultural, and regional variations. A training data which covers all the said parameters with sufficient data will give better results (Setty et al., 2013). Currently, there exists no standard database for e-learning and classroom environments to recognize the affective state of single and multiperson using facial expressions, hand gestures, and body postures.

Some intelligent tutoring systems not only include student's affective state prediction but also include their identification to personalize and improve the teaching-

learning process with the help of object localization. Object localization is the prediction of the objects within an image along with its boundaries. Object localization can be used in the classroom response system for accurate identification of students and so on. Though there are many standard databases for object localization (Huang et al., 2007; Jain & Learned-Miller, 2010; Zhang & Deng, 2016), none of them contain the data which includes the face, hand gestures and body postures, specifically for both e-learning and classroom response systems.

Creating a database for both affective state prediction and object localization with all the required parameters is a challenging and time-consuming task. There are several databases for face detection, posture recognition, segmentation, human pose, object localization, and for affective state classification using facial expressions, but these are not used in both e-learning and classroom environments as these databases do not contain expressions related to learning-centered emotions nor contain multiple people in a single image frame. There are a few databases that contain multi-face in a single image frame with profile and tilted faces (Tarrés & Rama, 2012; Martinez, 1998), but they are not explored for affective state recognition and its analysis. This is the motivation to create a database for affective state recognition with object localization using three components, namely: facial expression, hand gestures, and body postures.

1.9 Students' Affective States as Feedback

With the advancement in technology, personalized (one to one tutoring) learning emerged as an effective tool, and the automation in personalized tutoring led to the development of intelligent tutoring systems (Kulik & Fletcher, 2016). Intelligent tutoring systems promote engagement and learning by dynamically detecting and responding to a student's affective states (Calvo & D'Mello, 2010; D'Mello et al., 2010, 2007). These personalized tutoring systems are generally categorized under asynchronous learning, which is a student-centered tutoring system using online learning resources. On the contrary, a classroom is a well known synchronous learning environment where a group of students involve in learning directly through human to human interaction. An effective teaching agent (teacher or auto-tutor) should possess various teaching strategies to

carry out a better teaching-learning process with the students. An effective teaching strategy includes inquiry-based instruction, cooperative learning, utilizing technology in the classroom, behavioral management, and a few other strategies (Ahlfeldt et al., 2005; Hu & Li, 2017; Walker et al., 2008; Chi et al., 2011).

From the existing literature, it is observed that inquiry-based instruction is one of the most effective teaching strategies in the classroom environment (Ku et al., 2014). In inquiry-based instructions, teachers pose thought-provoking questions that inspire the students to think for themselves and become more independent learners. Inquiry-based instructions are based on the context and the affective states of the students (Eison, 2010). There exist several adaptive tutoring systems such as Wayang and Crystal Island tutors, which use virtual agents to provide hints and express empathy to the students based on their affective states in the classrooms (Arroyo et al., 2014; Rowe et al., 2009). Other works include providing hints, motivational messages, adapting curriculum to individual needs, auto-generating hints based on the students' behavior, and so on (Silva et al., 2019; Moore & Stamper, 2019; Rajendran et al., 2018). These existing works use self-reports, agents, and text-based analysis to recognize the students' engagement. But, the unobtrusive students' engagement analysis in real-time as feedback to enhance the teaching-learning process is not explored in the literature.

In order to perform inquiry intervention, accurate recognition of student's affective states is necessary (Arroyo et al., 2009; D'Mello et al., 2007; Arroyo et al., 2014). The existing literature considers the student's emotion/affective state recognition from computer vision techniques in e-learning environment, but they do not use this as feedback for any automatic interventions. Affective image content analysis is widely used for unobtrusive students' affective state prediction in the e-learning environment. The requirement of unobtrusive affective content analysis methods extended even to the universities after the introduction of IoT and Smart Campus, where the inbuilt cameras like surveillance cameras present in the computer-enabled teaching laboratories are used to analyze the students' engagement (Ashwin & Guddeti, 2018). This led to the use of affective video content analysis methods where image frames are used to predict the multiple students' facial expressions and behavioral patterns in a single image frame.

Recent advancement in deep learning showed better classification and object localization in real-time (Szegedy et al., 2016; Redmon et al., 2016). There are few works which use deep learning to recognize the students' engagement in real-time (Klein & Celik, 2017; Zaletelj & Košir, 2017; Burnik et al., 2017). A common goal of these studies includes adaptive teaching strategies based on students' engagement for classroom environments.

1.10 Motivation

The existing literature focuses on the students' affective content analysis for predicting their emotional engagement (boredom, confusion, flow/engagement, and so on) (D'Mello et al., 2007; Bosch et al., 2016). Whereas a few other works have mainly considered behavioral engagement (looking away from the computer, eyes completely closed, etc.) from the students' video content (Whitehill et al., 2014), to aid the prediction. Though these works are performed in both e-learning and classroom environments, the students' engagement analysis is performed on a single person in a single image frame. However, in the real classroom scenario, there will be multiple students in a single image frame, and this data contain students' faces in the wild. Further, there exists no single robust technique to predict the students' engagement in the classroom environment, and also, there are no standard datasets available for the same. This is the motivation to address the following issues and challenges:

(a) Recognition of the emotional/behavioral engagement of each student.

(b) Type of engagement analysis to be performed for better accuracy.

(c) Recognition of every student present in the single frame with localization to predict the engagement patterns in the wild.

(d) Use of multimodal[1] analysis for better performance.

(e) Prediction of a single group engagement level value for each frame.

(f) Use of a single robust technique to predict the students' engagement in both the environments.

[1]The word 'Multimodal' used in the proposed methodologies of the entire thesis refers to intra-image multimodality where the features of the facial expressions, hand gestures, and body postures of each student present within that image frame are considered.

(g) Test for the possibility of replacing or substituting the robust unobtrusive technique for the popular/state-ot-the-art students' engagement analysis methods.

(h) Real-time performance of the students' engagement analysis.

(i) Use of the results of (h) as immediate feedback to enhance the teaching-learning process.

1.11 Summary of Research Contributions

Figure 1.1 shows the complete research framework for unobtrusive students' engagement analysis in learning environments. As shown in Figure 1.1, the students' engagement analysis is performed using the nonverbal cues, and the required data is obtained from image frames. The image frames obtained from the video data make this an unobtrusive method. Both emotional and behavioral engagements are analyzed using nonverbal cues. The nonverbal cues include intra-image multimodalities such as facial expressions, hand gestures, and body postures of each student present in an image frame. These multi modalities are used to classify the students' affective states. These classified affective states, along with the object localization, helps in analyzing both the emotional and behavioral engagement of students. The data are obtained from various learning environments, such as classrooms, webinars, e-learning, and flipped classrooms. Generally, e-learning and flipped classroom image frame data contain a single student in a single image frame. In contrast, the classroom and webinar image frame data contains multiple students in a single image frame. Each multiple student image frame data is used to analyze the group engagement of that image frame. Students' affective state classification and object localization techniques are used for the engagement analysis. Real-time classification of affective state classification is used as feedback to enhance the teaching-learning process. To address the issues and challenges mentioned in the previous section, the following are the key contributions of this thesis.

- The thesis extensively reviewed the methods available for the emotional engagement of multiple students in a single image frame and found that there is no work done in this area. Hence, we proposed students' multi-facial emotion recognition for Ekman's basic emotions. We further enhanced the method to predict all the learning-centered emotions in both e-learning and classroom environments. The proposed method predicts the students' learning-centered emotions for all

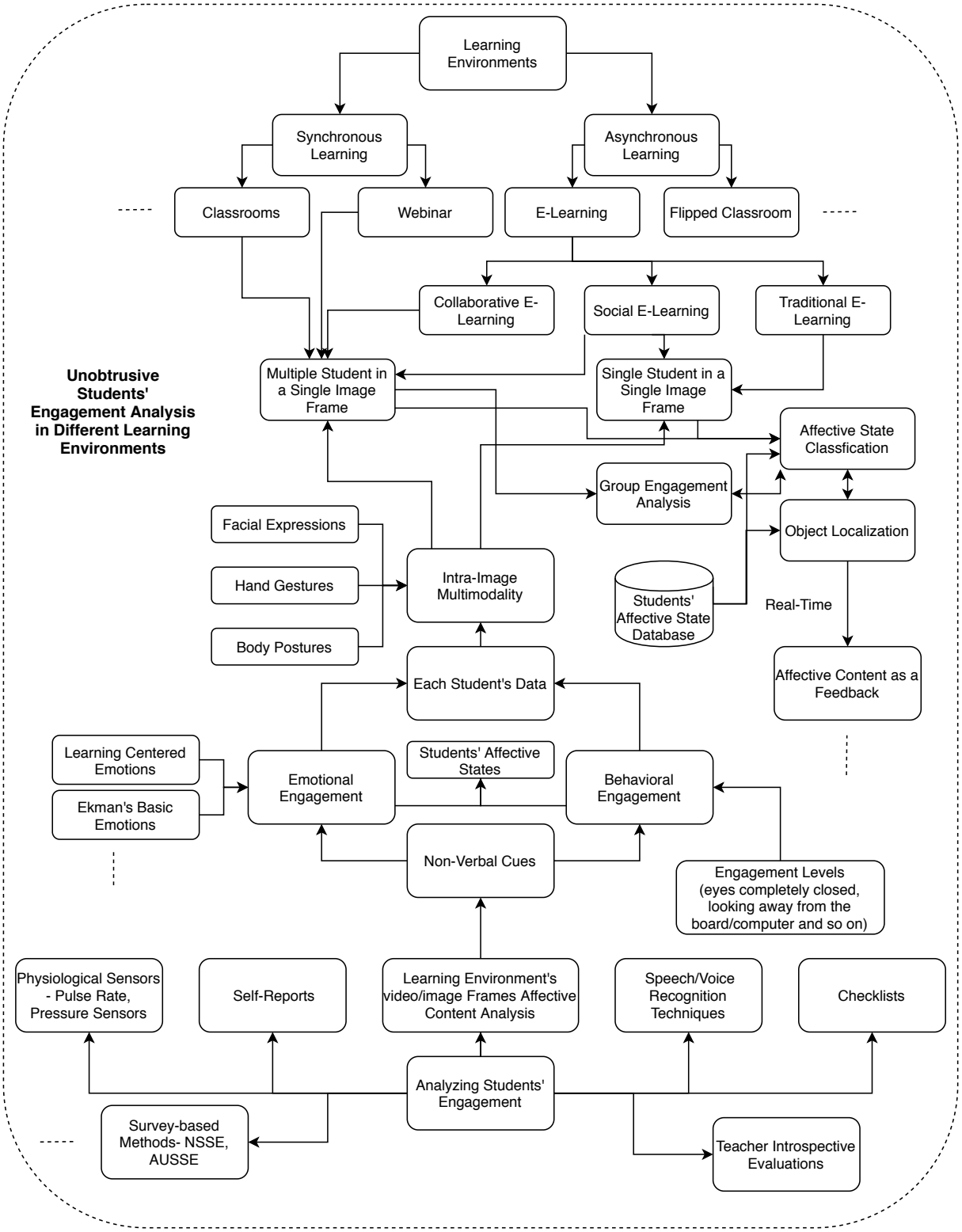


Fig. 1.1. Unobtrusive students' engagement analysis in learning environments.

the students present in the image frame and uses bounding boxes to localize the student present in the image frame. Group engagement analysis is also performed on each image frame.

- An architecture is proposed to analyze the behavioral engagement analysis of students in both the classrooms and computer-enabled teaching laboratories. The multitude of modalities is used to classify the students' engagement levels. Various statistical analysis is performed to observe the correlation among students' engagement levels and their performance. Group engagement analysis is also performed on each student present in an image frame.
- An automatic inquiry intervention is proposed in different learning environments, which uses the proposed real-time affective state classification and localization method as feedback. The proposed automatic inquiry intervention uses purposeful questioning to automatically pose the questions to the student based on their affective states. Statistical analysis proved that there is a significant impact of the proposed automatic intervention method.
- The final contribution of this thesis is the creation of an affective database with students' affective states. The created database contains various image variants, data from different learning environments, a multitude of modalities, and object localization. Annotations are performed using gold-standard study, and the database follows the rules provided by the institutional ethics committee. The created database is benchmarked with various state-of-the-art classification and localization techniques.

1.12 Organization of the Thesis

The remainder of this thesis is organized as follows. Chapter 2 reviews the related work on students' engagement analysis, different works on affective state detection, classification, & localization techniques, real-time affective feedback systems, and the available databases for students' affective content analysis using their non-verbal cues in different learning environments. Based on the outcome of the literature survey, the problem statement and research objectives are defined.

Chapter 3 describes the emotional engagement analysis. An ensemble of Haar cascades, RBM, SVM, and the proposed modified affine transformations are used to analyze the basic emotions for multiple faces. Students' affective states such as boredom and engaged are detected using the proposed CNN model. All the dominant learning-centered emotions are detected and classified along with the object localization using the proposed CNN model, and it is tested with various state-of-the-art classification and localization techniques.

Chapter 4 focuses on behavioral engagement analysis. Scale-invariant context-

assisted single-shot CNN is proposed to analyze the students' behavioral engagement in the classroom environment. Various statistical methods are used to analyze the correlation between the obtained students' behavioral engagement data with their performance in the class. Further, the behavioral engagement analysis is also analyzed for computer-enabled teaching laboratories, and the results of classification and localization are compared with state-of-the-art classification and localization architectures.

Chapter 5 focuses on using the students' engagement analysis in real-time as feedback to enhance the teaching-learning process. Separate models are built in four different learning environments, such as e-learning, flipped classrooms, classroom and webinar environments. To perform the engagement analysis is real-time CNN based architecture is proposed. Based on the students' engagement, automatic inquiry interventions are used to enhance the teaching-learning process. Various statistical methods are used to analyze the impact of the proposed method.

Chapter 6 describes the created students' affective state database. The created database contains students from different learning environments, such as e-learning and classroom. The details like camera setup, annotation using gold-standard study, variants used in the database, storing the annotated data, and so on are clearly explained in this chapter. The dataset is trained and tested with various state-of-the-art architectures, and it is also benchmarked for various feature extraction methods.

Finally, Chapter 7 summarizes the contributions of the research work and highlights the possible future directions in enhancing the teaching-learning process.

1.13 Summary

The chapter introduces affective computing, various learning environments, different types of students' engagement analysis methods, students' affective states and its classification, unobtrusive students' engagement analysis and its use as feedback to enhance the teaching learning process. Finally, the motivation behind this research work is discussed along with the outline of the thesis. The next chapter provides the details on the exisitng literature.

Chapter 2

Literature Survey

In this chapter, a complete review of existing students' engagement analysis is discussed. Students' engagement analysis includes various class labels used in different studies to classify the students' emotions and behavioral patterns, various ways to measure those class labels, and different state-of-the-art techniques to classify the observed students' emotional and behavioral patterns. Further, we also discuss the real-time student' engagement analysis techniques and the use of those real-time analysis systems as feedback to improve the teaching-learning process. Also, there are several databases for emotion recognition and engagement analysis; we discuss in detail about the existing unobtrusive student engagement analysis databases. Finally, we discuss the outcome of the literature review, followed by the problem statement and the objectives of the research work.

2.1 Students' Engagement Analysis

Affective State Class Labels: There are a limited number of systems which explored multimodal affect detection in the learning environment (Psaltis et al., 2017; Castellanos et al., 2017). Existing intelligent tutoring systems, auto-tutors, and humanoid robots predict students' engagement using their facial expressions and other body parts related to behavioral aspects. Learning-centered emotions like anger, boredom, confusion, contempt, curiosity, disgust, eureka, and frustrations are used to analyze the students' engagement in the Emote-aloud study (Sidney et al., 2005; D'Mello, 2012). Constructive and destructive learning emotions like frustration, confused, happiness, and hopeless are considered for the students' engagement analysis in humanoid robot tutors (Singh et al., 2013). Whitehill et al. (2014) proposed four behavioral engagement levels to analyze the students' engagement in a laboratory environment. Holmes et al. (2018) proposed a real-time non-verbal behavior recognition of e-learners. They considered 37 behavioral patterns and analyzed the student's engagement as positive or negative. They did not consider multiple faces for affective state recognition and

feedback mechanism to improve their auto-tutor. Further, D'Mello et al. (2007) showed that only confusion, boredom, frustration, delight, and flow emotions are sufficient to analyze the student's emotional engagement in any learning environment.

2.2 Affective State Detection and Classification:

There are several existing works on the recognition of students' affective states performed using text, audio, and video data. Most of the unobtrusive computer vision based students' affective state analysis techniques only use the facial expressions. Though there are limited works on exploring the students' affective state analysis using state-of-the-art techniques, there are some cutting edge techniques on exploring the recognition of basic emotions' in various other domains. Those are techniques for emotion recognition, multimodality, multiperson in a single frame affective content recognition, and object localization (Klein & Celik, 2017; Wang & Ji, 2015). Duncan et al. (2016) proposed a real-time facial emotion recognition system using convolutional neural networks to classify the image frames into seven basic emotions. They performed transfer learning and used CK+ & JAFFE datasets for training the model. The overall training accuracy of 90.7%, but the test accuracy was 57.1% only. These methods considered only facial expressions for emotion classification.

Alizadeh & Fazel (2017) used grayscale images obtained from Kaggle for emotion recognition using CNN. Further, they combined raw pixel data and HOG features to train the CNN. Their results demonstrated that the hybrid model with raw pixel and HOGs has no impact on the results obtained from the CNN. The overall accuracy observed using the shallow model was 48.77%, and the deep model was 60.92%. Fan et al. (2016) proposed a video based emotion recognition system using CNN-RNN (Recurrent Neural Networks) and C3D (Deep 3-Dimensional Convolutional Networks) hybrid networks. They used FER2013 database to train and obtained an accuracy of 42.82% for Face-AFEW dataset.

Ng et al. (2015) proposed a transfer learning based deep learning architecture for emotion recognition on small datasets. They used Emotiw and FER datasets for training and validation. Their results demonstrated an overall accuracy of 48.5% for validation

and 55.65% for testing. Ranganathan et al. (2016) proposed a multimodal emotion recognition system using a deep learning architecture for emoFBVP database. This database consists of four multimodal components, namely: face, body gesture, voice, and physiological signals, and the components are used for classifying 23 different emotions. They used a convolutional deep belief model for emotion recognition and obtained an accuracy of 58.5% for MAHNOB-HCI dataset and 97.3% for the Cohn Kanade database, but these works are not explored in learning environments.

Tzirakis et al. (2017) proposed a deep neural network framework for end to end multimodal emotion recognition. The accuracy analysis was performed for both spontaneous and natural emotions using RECOLA and AVEC 2016 database with an accuracy of 62% and 71%, respectively. Dehghan et al. (2017) proposed DAGER: Deep Age, Gender, and Emotion Recognition system using CNN for Sighthound dataset and obtained an accuracy of 76.1% for emotion recognition. They used LFW dataset for face recognition during the training phase, and the same dataset was fine tuned using Sighthound's emotion data. Kahou et al. (2016) proposed EmoNets, which used multimodal deep learning approaches for emotion recognition based on the video data. They used activity, audio, bag-of-mouth, and convnet for the emotion prediction and obtained an overall accuracy of 41.03% and 47.67% for AFEW2 and AFEW4 datasets, respectively. Zhao & Itti (2017) proposed an improved deep learning technique for object categorization using the pose information. They used what/where CNN (2W-CNN) on iLab-20M dataset and obtained an accuracy of 84.8%. Though the methods mentioned in Tzirakis et al. (2017); Dehghan et al. (2017); Mollahosseini et al. (2016); Burkert et al. (2015); Kahou et al. (2016); Zhao & Itti (2017) carried out multimodal emotion recognition with state-of-the-art techniques, these were performed only for basic emotions. Further, these techniques were not used for the students' affective content analysis and hence did not consider object localization and affective state analysis with respect to (w.r.t.) students' data.

Facial expressions are recognized using Gabor features, Haar cascades, pyramid histogram of gradients, local binary patterns, and others (Wang & Ji, 2015). But there are very few works on facial expression recognition using deep learning techniques

in e-learning and classroom environments (Klein & Celik, 2017; Ashwin & Guddeti, 2019b,a, 2018; Gupta et al., 2019). Existing works for the feature extraction such as Histogram of Oriented Gradients (HOG), Scale-Invariant Feature Transform (SIFT), and tools like OpenFace are used for real-time student's emotion recognition (Wang & Ji, 2015; Monkaresi et al., 2017; Thomas & Jayagopi, 2017). All these methods and tools fail to perform well for multiple students in a single image frame. Deep learning techniques for affective state classification are more accurate than existing machine learning techniques. Long Short-Term Memory (LSTM) is a widely used deep learning technique for emotion recognition from videos (Booth et al., 2017; Xia et al., 2018; Dhamija, 2017; Sun et al., 2018; Dhamija & Boult, 2017), but these are not explored much for the students' identification and also for the real-time affective state classification in learning environments.

Classification of students' affective states in a classroom environment is a challenging task. Group happiness intensity level is calculated using the weighted group expression model, which uses global and local contexts (relevant position of people in the image) (Dhall et al., 2015). This model is further optimized for multimodal group affect analysis using hand, body, and scene details (Huang et al., 2018). But this work is limited to types of happiness, and also, the use of local contexts are not relevant to the classroom scenarios. LSTM (with AlexNet as base) is used to recognize multimodal group emotion recognition using feature fusion (Sun et al., 2016), and Group behaviour analysis is performed using sociometric badges and surveillance cameras in poster presentation and cocktail party effects (Alameda-Pineda et al., 2016), but these are not explored in learning environments. Castellanos et al. (2017) used Atkinson's index with students' text data to calculate group behavioral engagement scores in the classroom environment.

State-of-the-art Recognition and Localization Techniques

There are several state-of-the-art multimodal, multiperson detection, classification, and localization techniques. Feature pyramid networks are popular in recognizing the high-level feature semantic maps at all scales (Lin et al., 2017). But these are memory

intensive, and to address this issue, scale-invariant single-shot detectors were used (Zhang et al., 2017). These techniques failed to recognize small faces or objects. Tang et al. (2018) proposed PyramidBox, which uses low-level feature pyramids as well as context-sensitive prediction modules to recognize small objects present in the image. But all these techniques are neither used for recognition of students' multimodality (face, hand gesture, and body posture) nor localization of students' faces in the wild.

Students' Affective State Detection and Classification

There are a limited number of systems that explore multimodal students' engagement in the learning environment. Existing intelligent tutoring systems, auto-tutor, and humanoid robots use the students' engagement analysis based on their facial expressions and other body parts related to behavioral aspects. In the existing literature, non-verbal cues are used to classify either emotional or behavioral engagement.

Emotional Engagement

Learning-centered emotions like anger, boredom, confusion, contempt, curiosity, disgust, eureka, and frustration were used to analyze the students' engagement in the Emote-aloud study (Sidney et al., 2005). Constructive and destructive learning emotions such as happiness, confusion, frustration, and hopelessness were considered for the students' engagement analysis in humanoid robot tutors (Singh et al., 2013). Bosch et al. (2016) considered the facial expressions and aggregate body movements of 137 subjects and thus classified them into different affective states, namely: boredom, confusion, delight, engagement, frustration, and off-task behavior. They used CERT (Computer Expression Recognition Toolbox) computer vision software to classify the obtained facet features using WEKA (Waikato Environment for Knowledge Analysis) tool with 14 different classifiers and obtained an overall AUC (area under the curve) of 0.708. But, these works were performed in controlled environments.

Behavioral Engagement

Whitehill et al. (2014) captured the student's facial expressions and body postures using iPad's web camera and classified students' engagement into four different engagement levels (using their behavioral patterns). They considered 34 subjects and obtained the AUC of 0.729 using Gabor features. Zaletelj & Košir (2017) used a Kinect sensor for classifying the students' attention using facial expression, eye gaze, and body posture. They considered writing, yawning, supporting the head, leaning back, and a person's gaze to classify the students' attention into high-level, mid-level, and low-level attentions. The authors considered 18 subjects and used the classifiers such as decision trees and k-NN to obtain an accuracy of 0.753. Kahu (2013) analyzed the students' engagement using their facial expressions, body movement, and gaze patterns; and classified the student engagement into engaging and non-engaging parts. They considered multiple deep instance learning based frameworks to obtain an average mean square error (MSE) of 0.10 on 78 subjects of the e-learning environment. But, the above mentioned methods were not explored for multiple students present in large classrooms.

Students' Multimodal Engagement Analysis

A few techniques used hand gestures and body postures along with the head movement for the analysis of the students' engagement. Emotion intensity models were applied on the obtained web frames (used webcams or Kinect to obtain the image frames), and tools like WEKA were used to obtain the classification results (Calvo & D'Mello, 2010; Patwardhan & Knapp, 2014; Grafsgaard et al., 2013). But all these were performed on a single person in a single image frame.

Students' Engagement Analysis in Classrooms

Existing works used a Kinect device to capture multiple students present in a classroom. k-Nearest Neighbors, decision trees, support vector machines, Haar cascades, and CNNs (AlexNet) were used to classify the students' behavioral patterns (Zaletelj & Košir, 2017; Burnik et al., 2017; Klein & Celik, 2017; Thomas & Jayagopi, 2017). The capturing range of Kinect was low when applied to large classrooms, and the techniques used were not robust enough to classify the students' expressions in the wild.

2.3 Affective Content as Feedback

The use of the analyzed students' engagement data to improve the teaching-learning process is another challenge. There are a few works where the student's engagement is analyzed by the dwelling time (amount of time student spent on the video), dwelling rate (how much content in the video, was seen by the student), and accordingly the complexity of the video lecture was analyzed and optimized (Van der Sluis et al., 2016). State diagram (Coffrin et al., 2014), competency map (Grann & Bushway, 2014) and various other learning analytic dashboards (Bodily & Verbert, 2017) are used as feedback to both the students and teachers which serves as a recommender system to improve the teaching-learning process and thus aiding in the visualization of the patterns among student's engagement and the corresponding marks obtained by them. But these works are limited to Massive Open Online Courses (MOOCs) or e-learning environments.

Existing smart/digital classrooms provide feedback after the completion of each class or course. Some feedback systems include the details about the instructional and social responses from the students (Yu, 2017). There are several works on game-based learning which dynamically adapts the teaching strategy or responds to a single student. Prime Climb is a game based learning test-bed which dynamically adapts the teaching strategy for each student using user-adaptive hints (Conati, 2002). Crystal Island recognizes the students' affect expressions and shapes their affective trajectories using narrative-centered learning interactions (Rowe et al., 2009). These are confined only to game-based learning, and there is no affective feedback. Other diverse works on student's affective content analysis and engagement analysis performed in various learning environments are summarized in Tables 2.1 and 2.2.

To summarize, there are several tools to recognize the users' affective states, such as affectiva[1], emotinent[2], imotions[3] and so on, but they are not explored in the learning environments. The existing students' affective state recognition techniques are limited to single modality and considered for e-learning environments.

[1]https://www.affectiva.com
[2]http://emotient.com
[3]https://imotions.com

Table 2.1
Related Work Summary

Authors	Methodology	Engagement Analysis	F/I	Environment
D'Mello et al. (2007)	Affect sensitive auto-tutor	Student's affective states	No	Auto-Tutor
Kim et al. (2015)	Smartphone response system	Student's marks	No	Classroom
Castellanos et al. (2017)	Group engagement score for virtual learning environment	Engagement level (Balanced, un- even, unengaged)	No	Virtual Learning
Balaam et al. (2010)	Subtle Stone is used for engagement analysis	Student's affective states	No	Classroom
Liu et al. (2015)	Learning analytic system called tracer is used	Point and Intensity based engagement analysis	No	Online Learning
Yousuf & Conlan (2017)	Visual narrative framework VisEN is used	Self Reports (average, good or excellent)	No	Online Learning
Burnik et al. (2017)	Vision based observed attention estimator	Behavior patterns	No	Classroom
Klein & Celik (2017)	The WITS intelligent tutoring system	Students' affective states	No	Classroom
Maneeratana et al. (2017)	Class-Wide course feedback	Class Survey	No	Seminar
Zaletelj & Košir (2017)	Student engagement monitoring	Students' behavior	No	Classroom
Bosch et al. (2016)	FACET, CERT Computer Vision	Students' affective states	No	Classroom
Whitehill et al. (2014)	iPad based engagement monitoring	Behavioral patterns	No	e-learning
Thomas & Jayagopi (2017)	Classroom monitoring	Facial Behavioral Cues	No	Classroom
Tiam-Lee & Sumi (2019)	WEKA and OpenFace	Student's affective states	No	e-learning
Yun et al. (2018)	Children engagement monitoring	Behavioral patterns	No	Classroom Laboratory
Psaltis et al. (2017)	Multimodal Student engagement	Basic Emotions	No	Virtual Reality (VR) Game

F/I: Feedback/ Intervention

Table 2.2
Related Work Merits and Limitations

Authors	Merits	Potential Limitations
D'Mello et al. (2007)	Classified student engagement into boredom, confusion, delight, flow and frustration	Different sensors are used to recognize different affective states
Kim et al. (2015)	Used twitter to enable effective interaction	Rapid feedback system is required for real-time scenarios
Castellanos et al. (2017)	Both individual activity and similarity of participation are considered	Engagement metrics are available only for the instructors
Balaam et al. (2010)	Tangible technology designed to support student's active emotional communication	Use of Subtle Stone makes it obtrusive
Liu et al. (2015)	Analyzed the behavior pattren of student's writing on cloud based applications	Impact of visualization on learning is not evaluated
Yousuf & Conlan (2017)	Analyzed the behavior pattern of student's writing on cloud based applications	Impact of visual narratives in online learning is not performed
Burnik et al. (2017)	Student's attention monitoring during a lecture using gaze and behavior cues	No feedback is given to teacher or student for further improvements
Klein & Celik (2017)	Detecting students' engagement during lectures using CNN	Considered only behavioural engagement
Maneeratana et al. (2017)	Student's engagement based feedback system	Manual engagement analysis is performed
Zaletelj & Košir (2017)	Kinect based engagement analysis	Kinect capture range is small, it cannot cover a big classroom
Bosch et al. (2016)	Considered facial expressions and gross body movements	Works for single person in a single image frame only
Whitehill et al. (2014)	Face and head movement, eye tracking and posture is considered	Works when the student is close to the camera
Thomas & Jayagopi (2017)	Classroom data is considered where each student's face is cropped and processed for emotion recognition	Recognizes only facial features of each student
Tiam-Lee & Sumi (2019)	Both pose and face are considered for the analysis	Fails for scale invariant and tilted face images
Yun et al. (2018)	CNN is used for both pose and scale invariant face images	Tested on only children's data
Psaltis et al. (2017)	Multi-modal fusion methods are used	Considered only basic emotions

From Tables 2.1 and 2.2, it is observed that there are several works on students' engagement analysis, but none of them predicts the students' affective states in real-time, and hence, those recognized students' affective states are not used as real-time feedback. A few existing works recognize the students' affective state in classroom environments (Thomas & Jayagopi, 2017; Yun et al., 2018), but real-time affective state predictions are not performed. There are a few tools in the game-based learning environment which recognizes the students' affective state in real-time (Conati, 2002; Psaltis et al., 2017). Here they recognize a single student who is standing or playing in front of the camera or using a VR headset. But these tools fail to recognize sitting and multiple students present in other learning environments like e-learning and classrooms. Hence, there is no existing work which predicts the students' affective state in real-time and use it as feedback to improve the teaching-learning process. Also, there are no existing works that recognize the students' affective states in all four learning environments (for both single and multiple students in a single image frame). Therefore in this work, we propose a method to recognize the students' affective state in real-time and use this data as immediate feedback to improve the teaching-learning process.

2.4 Multi-Modal and Faces in the Wild Data Analysis

Multi-task convolutional neural networks were proposed to address pose-invariant unconstrained face recognition (Yin & Liu, 2018). Weighted Mixture Deep Neural Network was used to recognize the basic emotions for CK+ database (which include occluded faces) with 97% accuracy. Further, the authors proved that CNN based architectures outperformed all the handcrafted feature based methods (Yang et al., 2018). Deep Fusion CNN was used for multi-modal expression recognition to classify the basic emotions (Li et al., 2017). Blur-aware bi-channel deep neural network was proposed for face recognition in the wild (Ding & Tao, 2018). Several other works were proposed for face recognition, emotion classification, and multi-modal analysis for the data of faces in the wild (Xie & Hu, 2019; Zhang et al., 2019; Li et al., 2017; Ding & Tao, 2015; Wu et al., 2018). The base network for most of the above mentioned methods use one among the following architectures: AlexNet (Krizhevsky et al., 2012), VGGNet (Simonyan & Zisserman, 2014), ResNet (He et al., 2016) or GoogleNet (Szegedy et al., 2016).

But all these above methods were used for crowd data analysis in the entertainment domain or for sports data analysis but not for students' affective state recognition and classification. Hence, we proposed a novel hybrid CNN architecture for the students' affective state analysis in the classroom environment. Though there were several works related to affective state recognition of students, very few works were done for the recognition of learning-centered affective states using both emotional and behavioral patterns with some cognitive aspects involved in it. Further, it is observed that there exists no standard dataset for the verification and validation of the students' affective states in a classroom environment. This motivated us to create our dataset for the students' affective state analysis.

Table 2.3 shows the summary of recent works on affective content analysis performed in various domains. From the existing literature it is clearly evident that the traditional machine learning techniques such as ANNs, SVMs, etc. were outperformed by deep learning techniques such as CNN, RNN and LSTM (Long Short-Term Memory) for image frame based emotion recognition (Baveye et al., 2017). Table 2.3 shows existing works on multi-face emotions recognition for different types of smiles related to happiness. Similarly, there were other works in the field of entertainment which were tested on MELD (Multimodal Emotionlines Dataset) and EmotiW dataset using InceptionV3 and VGG16 deep learning architectures (Poria et al., 2018; Guo et al., 2018). But, these architectures were not sufficient to test on learning-centered emotions such as engaged and bored where the students' expressions and behavioural patterns vary significantly when compared to the facial expressions and behavioural patterns of humans obtained from the entertainment filed (DeFalco et al., 2018).

2.5 Students' Affective State Databases

Most of the standard affective state classification databases use facial expressions for the classification of affective states. Generally, the classification is based on Ekman's basic emotions. There are a few research studies on affective state recognition and classification of students in a learning environment. Happy et al. (2015) created an Indian Spontaneous Expression (ISE) database with 428 segmented video clips, including 29

Table 2.3
Summary of Works on Image Based Affective Content Analysis

Authors	Method	Tested on	Emotional Descriptors	D	IFC	M	GA
Gupta et al. (2016)	InceptionV3 Model	DAiSEE Dataset	Boredom, Confusion, Engagement, Frustration	Ed	SP	No	No
Lopes et al. (2017)	Modified CNN	CK+, JAFFE and BU-3DFE	Ekman's basic emotions	Ed	SP	No	No
Huang et al. (2019)	Long Short Term Memory	DAiSEE Dataset	Boredom, Confusion, Engagement, Frustration	Ed	SP	No	No
Hayashi (2019)	Facial Action Coding System	21 Japanese University Students	Ekman's basic emotions	Ed	SP	No	No
Ramirez L (2019)	Decision Trees, Data obtained from Kinectv2	16 Undergraduate Students	Engagement, Frustration	Ed	MP	No	Yes
Tiam-Lee & Sumi (2019)	WEKA and OpenFace	73 Students	Boredom, Confusion, Engagement, Frustration	Ed	SP	YES	No
Subramanian et al. (2018)	Linear SVMs	ASCERTAIN	Engagement, Liking, Familiarity	En	SP	YES	No
Liu & Jiang (2019)	Particle Swarm Optimization, kNNs	10 Adults from shooting team	Happiness, Sadness, Anger, Fear	S	SP	YES	No
Huang et al. (2018)	Information Aggregation, decision fusion	HAPPEI and GAFF	Neutral, Small smile, Large smile, Small laugh, Large laugh and Thrilled	R	MP	YES	Yes

IFC - Image Frame Content; M - Multimodality; GA - Group Analysis; D: Domian;
Ed: Education; En Entertainment; S: Sports; R: Real-life activities
SP - Single Person in a Single Image Frame; MP - Multiple Person in a Single Image Frame

male and 21 female participants. They classified the facial expressions into four basic emotions, namely: happy, sadness, surprise, and disgust. This database also includes mild to strong spontaneous facial expressions but does not contain posed expressions.

Riaz & Mushtaq (2016) analyzed the six basic categories of aesthetic emotions with 12 female and 15 male respondents, but this database was created only for the e-learning environment. Whitehill et al. (2014) proposed four-level engagement classification (highly engaged, engaged in the task or stay on task, sleepy or not in the task and not at all engaged or not thinking about the task) of students in an e-learning environment. Here, behavioral engagement was considered more than the emotional engagement of students for engagement level classification. There are several other research studies for the e-learning system, but those are not based on hand gestures or body postures for affective state recognition (Ashwin et al., 2015).

Patwardhan & Knapp (2014) proposed a single-face affective state recognition system using affective intensity estimation with a range of 0 to 1. They considered 5 participants with six basic emotions using face, body, hand, speech, and the head. They demonstrated that the contribution of face for affective state recognition for a few basic emotions like anger is less than 40%. But the experimental results were analyzed only for anger affective state. Sidney et al. (2005) integrated an affective sensor for the analysis of frustration, confusion, and boredom using face and body postures. They experimented with 20 college students; further, they used a sensor for posture pattern recognition embedded on both the seat and the back of the chair. The entire affective state analysis was based on single-face recognition in a single image frame and was performed along with body pressure measurement system. But, they did not carry out multi-face affective state analysis. Grafsgaard et al. (2013) proposed a multimodal feature-based method for predicting engagement and frustration during the tutoring phase using face and hand gestures. This method was tested and analyzed for 67 students and tutors. They also proposed a model (Grafsgaard et al., 2014) that gives a relationship between self-efficacy and the effectiveness of tutorial interactions. This was tested for 66 students and classified for two different affective states, namely: frustration and engagement. But this work was designed only for the e-learning scenario.

Table 2.4
Summary of Affective Databases in Computer Vision.

Authors	Database Name	Database Content	Recognition Type	Affective States	Merits	Demerits
Andriluka et al. (2014)	MPII Human Pose Dataset	25K Images	Entire body	Everyday Human Activities	Richer annotations with occluded images, 3D torso and head orientations	No annotations for affective state recognition
Valstar & Pantic (2010)	MMI Facial Expression Database	25 Participants	Face	Six Basic Emotions	Facial emotions with side, profile, and tilted face.	Posed expressions only
Martinez (1998)	AR Face Database	126 People with 4000 Frames	Face	Smile, Anger, Scream and Neutral	Different facial expressions, illumination conditions and occlusions.	Posed expressions only
Georghiades et al. (1997)	Yale Face Database	165 Images of 15 Persons	Face	Sad, Sleepy, Surprised, and Wink	It contains occluded images	Contains only frontal face images
Lyons et al. (1998)	JAFFE	213 Images of 7 Different Expressions	Face	Basic Emotions	Coded facial expression images with multi-orientation, multi resolution	Only female expressions were considered

Continuation of Table 2.4						
Authors	Database Name	Database Content	Recognition Type	Affective States	Merits	Demerits
Phillips (2004)	FERET	1564 Sets of Images	Face	No	Detection and verification algorithms were used	No annotation for facial expressions
Marin et al. (2014)	Microsoft Kinect and Leap Motion Dataset	10 Different Gestures performed by 14 Persons	Hand	No	Gesture detection and annotations were performed	This dataset was not designed for affective state recognition
Tompson et al. (2014)	NYU Hand Pose Dataset	6736 Image Frames	Hand	No	It is a RGB depth data with ground-truth of hand-pose information	It contains only one person's hand gestures
Tarrés & Rama (2012)	Gtav Dataset	44 Persons with 27 Pictures	Face	No	It contains occluded images with different illuminations	Posed expressions only
Zhang et al. (2014)	BU-3DFE Database	100 Subjects with 2500 Images	Face	Basic Emotions	Spontaneous expressions were stored (2D and 3D)	All the expressions and images contain only frontal face expressions

Continuation of Table 2.4						
Authors	Database Name	Database Content	Recognition Type	Affective States	Merits	Demerits
Setty et al. (2013)	IMFBD Dataset	100 Film Video	Face	Basic Emotions	Movie image frames were considered for classification of six basic emotions	Posed expressions only
Dhall et al. (2011)	AFEW Database	957 clips	Face	Basic Emotions	Temporal data with 7 basic emotions	All the expressions were extracted from movie scenes
Happy et al. (2015)	HAPPEI Database	428 video clips of 50 participants	Face	Basic Emotions	Images were collected from wild	Contains only frontal face
Filntisis et al. (2019)	Child-Robot Interaction Database	30 Students with 315 emotion sequences	Face and body posture	Basic Emotions	Multilabel training from multiple visual cues	Explored only on robotic interactions
Kaur et al. (2018)	Student Engagement Database	78 subjects with 5 mins video	Head Pose and Eye Gaze	Behavioral Patterns	Both prediction and localization were performed	Only behavioral patterns were considered

Continuation of Table 2.4						
Authors	Database Name	Database Content	Recognition Type	Affective States	Merits	Demerits
Sapiński et al. (2018)	Multimodal Database	16 subjects with 560 annotated images	Face and posture	Basic Emotions	Subjects were of diverse age groups	Learning-centered emotions were not considered
Bian et al. (2018)	Spontaneous Expression Database	82 students and 30184 images	Face	Learning-Centered Emotions	Data Augmentation was also used	Only online learning subjects were considered
Gavrilescu (2015)	LIRIS-ACCEDE Database	64 Subjects with 8 - 12 second videos	Expressions, gestures and postures	Basic Emotions	Stochastic context-free grammars were used	Only frontal face and pose of person were considered
Gupta et al. (2016)	DAiSEE Database	9068 video snippets captured from 112 users	Face	Learning -Centered Emotions	Engagement Recognition in the Wild	Multi-modality was not considered

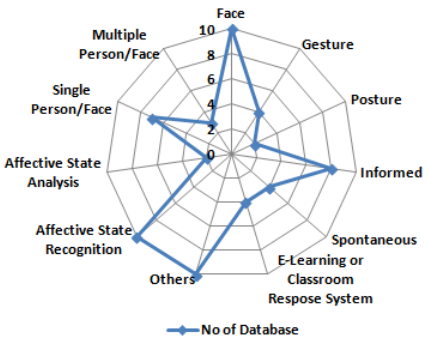


Fig. 2.1. Spider chart of affective databases in computer vision.

There exists affective based research work for the classification of multi-person in a single image frame. Dhall et al. (2015) experimented on 417 images with 1756 faces and analyzed three affective states in the group image frames, namely: positive, neutral, and negative. Happy et al. (2015) created HAPPEI database for classification of happy affective state for multi-person in a single image frame. They used only the happy affective state for the creation of their database. They also considered neutral, small smile, large smile, small laugh, large laugh, and thrilled, happy affective state variants for affective state classification. These were not suited for learning environments.

From the works Calvo & D'Mello (2010); D'Mello et al. (2007); Ekman (1992); Picard (1997), it was observed that students often tend to express only a few affective states in e-learning and classroom environments. This included seven basic emotions and a few learning-centered emotions. Table 2.4 summarizes the affective databases in computer vision. In summary, there are a few databases which are used only for face affective state recognition (Tarrés & Rama, 2012; Lyons et al., 1998; Georghiades et al., 1997; Valstar & Pantic, 2010; Setty et al., 2013; Zhang et al., 2014). There are a few standard databases which are used only for face detection (Zhang & Deng, 2016; Huang et al., 2007; Jain & Learned-Miller, 2010). There are some databases which are used only for hand gesture recognition (Marin et al., 2014; Tompson et al., 2014; Matar et al., 2016; Radeta & Maiocchi, 2013; Garber-Barron & Si, 2012). There are a few works that have given insights about the difference in facial expressions of Caucasian and Asian subjects (Happy et al., 2015; Lyons et al., 1998). Further, some research studies have used face, hand gesture, and body posture for affective state recognition of single-person in a single image frame (Calvo & D'Mello, 2010; D'Mello et al., 2007;

Grafsgaard et al., 2013; Ezen-Can et al., 2015; Filntisis et al., 2019; Sapiński et al., 2018; Noroozi et al., 2018; Gavrilescu, 2015).

Figure 2.1 shows the spider chart of affective databases in computer vision w.r.t. visual cues recognition that includes affective state recognition using facial expressions, hand gestures, body postures; spontaneous and posed expressions; single and multi-face affective state recognition in a single image frame; and other datasets for affective state analysis specifically designed for e-learning and classroom environments. It is observed from Figure 2.1 that there are only a few databases that consider all the three visual cues recognition components of single-person in a single image frame, namely: face, hand gesture, and body postures. Further, there are a few databases that consider multi-person in a single image frame with spontaneous facial expressions only. It is also observed from Figure 2.1 that there are a few databases available for affective state analysis for e-learning and crowd affective state recognition. Further, only a few databases are available for Indian or Asian people with annotated facial affective state recognition. But there exists no standard database for analyzing students' emotional and behavioral engagement using students' affective state analysis based on facial expressions, hand gestures, and body postures for both e-learning and classroom environment. Hence, one of the objectives of this is to create an affective database for both e-learning and classroom environments, thus overcoming the limitations of the existing works.

2.6 Major Gaps of the Existing Literature

Tables 2.1, 2.2, and 2.4 summarizes the key existing works on the students' engagement analysis in different learning environments. It is observed that the existing works are extensively explored for students' emotional engagement analysis using their facial expressions. Multimodality in emotion recognition, such as the use of postures and gestures along with the facial expressions, are considered in existing studies which are confined only to the e-learning environment. Also, the use of deep learning techniques for the students' engagement analysis is limited, and this reduces the robustness of the method to recognize scale-variant faces within the image. The existing works which use deep learning techniques are not explored much for multiple students in a single image

frame. There are works on analyzing the students' engagement in classrooms where multiple students in a single image are considered, but these works use the Kinect sensor as a capturing device that limits its data capturing capacity to a certain range and hence cannot be used for large classrooms. All these works did not analyze the group engagement level of students within the classroom. Even if the group engagement analysis is performed, it is done by using text data for analysis (Castellanos et al., 2017). To summarize, the following are the major gaps in the existing literature:

- The existing literature did not consider the data from large classrooms, webinars, and flipped classroom environments.
- Though there is multimodal emotion recognition for a single student in a single image frame (e-learning environment) in the existing studies, it is not explored for multiple students in a single image frame (classroom environment).
- There exist no works on image/video frame based group engagement analysis or group level score prediction using multiple students in a single image frame. The existing group level score prediction algorithms use text data to analyze group level engagement predictions.
- Most of the existing works predict only the students' emotional engagement using Ekman's basic, learning-centered, or academic emotions. Behavioral engagement of the students is studied in e-learning and game-based learning environments, but not explored in the classroom environment.
- The students' engagement analysis is not explored in computer-enabled teaching laboratories.
- There is minimal existing work on adapting the teaching strategy based on the students' affective states. Also, popular and effective teaching strategies like inquiry interventions are not automated in the existing study.
- There exists no such database which contains the students' affective state data obtained from the image frame of different learning environments. Further, there are limited works on the multitude of modalities for each student present in the image frame along with the object localization.

2.7 Motivating Examples

- How can a intelligent tutoring systems use the student's visual data to predict their engagement in real-time?
- How can a smart classroom environment unobtrusively predict the students' engagement?
- How can a pervasive intelligent tutoring system use the predicted students' emotional and behavioral patterns to adapt the teaching strategy/style in real-time?

2.8 Problem Statement

The goal of this research work is to propose an unobtrusive students' emotional and behavioral engagement analysis method using their non-verbal cues, which works in both e-learning and classroom environments. Accordingly, the research problem is stated as follows: **"To design and develop an unobtrusive affective computing framework for students' engagement analysis in the classroom environment."**

2.9 Research Objectives

The research objectives are defined as:

- To develop an effective method for multimodal students' affective content analysis.
- To develop an effective deep learning architecture for the students' behavioral engagement analysis.
- To develop a personalized intelligent tutoring system with automatic inquiry intervention based on the students' affective states in learning environments.
- Creation of a database with students' affective states and behavioral patterns using their facial expressions, hand gestures, and body postures.

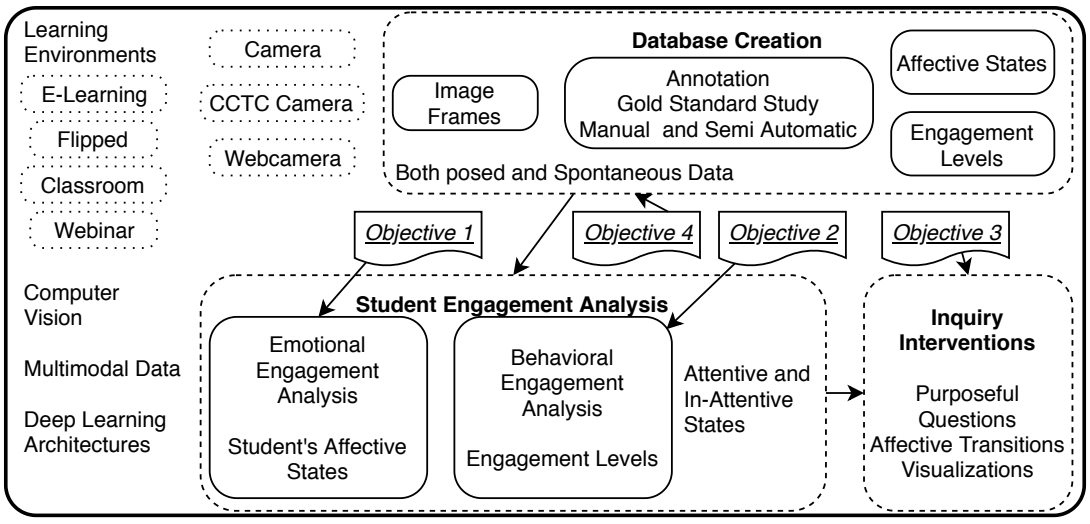


Fig. 2.2. Overview of the research contributions.

The overall flow of the research contributions w.r.t. the above mentioned objectives is shown in Figure 2.2.

2.10 Summary

In this chapter, we presented the existing state-of-the-art techniques related to students' emotional and behavioral engagement analysis. Further, various architecture to classify the students' affective states along with the object localization, the use of multitude in modalities, existing students' affective states databases, and existing models that use students' affective states as feedback in real-time are discussed. The challenging issues based on the outcome of the literature review are clearly discussed, along with problem statement and research objectives. In the following chapters, the issues mentioned in this chapter are addressed with suitable solutions which work in different learning environments. The next chapter emphasis the students' emotional engagement analysis.

Chapter 3

Emotional Engagement Analysis

Recognition of students' affective states for analyzing their emotional and behavioral engagement is a challenging task. Recent advancements in computer vision have broadened the scope for the affective content analysis of students in e-learning and classroom environments. Multi-person emotion recognition is one of the significant challenges in affective computing. Further, there are limited works on object localization based personalized affective state recognition systems for e-learning and classroom scenarios. The students' emotional engagement is analyzed using the learning-centered emotions, and there are no existing works that use a multitude of modalities to recognize the students' learning-centered emotions in both e-learning and classroom environments.

Hence, in this chapter, novel deep learning methods are proposed to recognize and analyze the students' emotional engagement. Two different methodologies are proposed in this chapter, along with the preliminary work on multifacial emotion recognition. The first part contains the methodology for detecting the multifacial emotion recognition of the user. The second part contains the proposed hybrid CNN architecture for students' emotional engagement prediction using two different affective states along with the neutral. The third part includes another proposed method that can recognize all dominant learning-centered emotions of students in four different learning environments, namely: e-learning, flipped classroom, classroom, and webinar environments.

The key contributions of students' emotional engagement analysis are as follows:

- A novel method for multifacial emotion recognition to classify Ekman's basic emotions, face detection of tiled & occluded faces in streaming data, and video affective content analysis.
- A novel hybrid CNN model for affective state analysis of students in the classroom environment using:
  - students' facial expression, hand gestures and body posture,
  - group engagement (class) score (for each frame with spontaneous expression (multi-person in a single frame image)), and
  - classification of engaged and boredom affective states along with neutral.

- A novel CNN-based architecture for recognizing the students' affective states using:
  - facial expressions, hand gestures & body postures in e-learning & flipped classroom environments (single person in a single image frame) and in classroom & webinar environments (multi-person in a single image frame).
  - a novel object localization technique for detecting students' faces, hand gestures, and body postures.

3.1 Proposed Methodology for Multifacial Emotion Recognition

Nowadays, there are several types of popular e-learning methods, such as collaborative learning and social learning, where multiple students sit together and use e-learning tools. Here the emotion recognition needs to be performed on various students present in that single image frame obtained from the camera. In this study, Ekman's basic emotions are predicted for the students present in the input image data using the proposed multifacial emotion recognition. It consists of two phases: The training phase and the execution phase.

(a) Training Phase

Preprocessing: The given images are converted to grayscale image; the resolution is diminished to 1024*768. For some images, cropping of the facial images is also done to increase the accuracy of feature extraction.

Face Detection using CPU: Haar feature-based classifier is used to detect the frontal face in the data set. The detector's parameters are varied such that the dimension of the object to detected (face) will have minimal effect on the detection accuracy.

Feature Extraction Using AAM: The landmark points or the feature points are identified using an AAM (Active Appearance Model) which gather 100 such points of which some key distances and measures such as the distance between the eye lids, mouths, the width of the mouth, etc. are extracted.

SVM Training: The SVM is trained using a combination of Radial Basis Function Kernel and polynomial kernel with the input as the feature points. Each set of feature points is mapped to a particular emotional expression out of 6 basic emotions and the neutral. The trained SVM data is stored in a serialized format.

(b) Execution Phase:

Face Detection using GPU: A GPU based object detector based on Local Binary Patterns is used in this step. Though the accuracy of this detection is not impressive but much faster than other detectors. It can be used to localize the search or minimize the computational load of detecting the object in the upcoming phase.

Face Verification using CPU: Similar to the detection stage in the training phase, this module uses a CPU based accurate detector based on Haar Cascades, but it is slow. Hence the regions so detected as a face by the previous stage are applied to the execution stage where it cross-validates using a more accurate detector. Since the area to be scanned is smaller when compared to the original image; hence this verification phase executes faster.

Emotion Classification: After facial feature point detection by AAM, the desired measures are applied to the SVM classifier, which is initialized with already trained data. This stage classifies the detected face image into one of the seven emotions. The accuracy of classification can be improved by increasing the size of the training data set.

The previous approach is used to detect the face for a given frame of an image, but most of them fail to detect the faces, which are tilted, occluded, or with different illuminations. In this study, a novel face detection system in streaming data is proposed, which detects the faces that are tilted, occluded, or with different illuminations, any difficult pose. The proposed system is a desktop application with a user interface that not only collects the images from a web camera but also detects the faces in the image using a Haar cascaded classifier consisting of Modified Census Transform features. The problem with the cascaded classifier is that it does not recognize the tilted or occluded faces with different illuminations. Hence to overcome this problem, a system is proposed using Modified Affine Transformation with Viola-Jones.

3.1.1 Face Detection System in Streaming Data

Figure 3.1 shows the flow of how the face is detected, which takes input as the image then try to detect the face in that image via Viola-Jones Haar Cascade. If the face is not detected, then modified Affine Transformation is used to rotate the image to a degree and detect the face again, increment the degree by 3 for every step till face is detected in that image from -80° to +80°. If the loop completes with no face detected, then it can be concluded that there is no face in that image. For face detection in streaming data, frames in the video stream are used and detected independently. The following steps are performed to recognize and track the face in streaming data of an image: (i) Recognize face in the image, (ii) Follow the face in the subsequent frames.

Affine-Transformation: Linear 2-D geometric transformations consist of affine transformation which is one of the important class which maps variables or coordinates (e.g.,

in an input image values of the pixel intensity at position (x1, y1) into new variables or coordinates (e.g., (x2, y2) in the output image), this is done by applying a linear combination of scaling, translation, shearing, and rotation operations are applied to achieve this. The affine transformation is generally written in homogeneous coordinates as shown in the Equation 3.1.

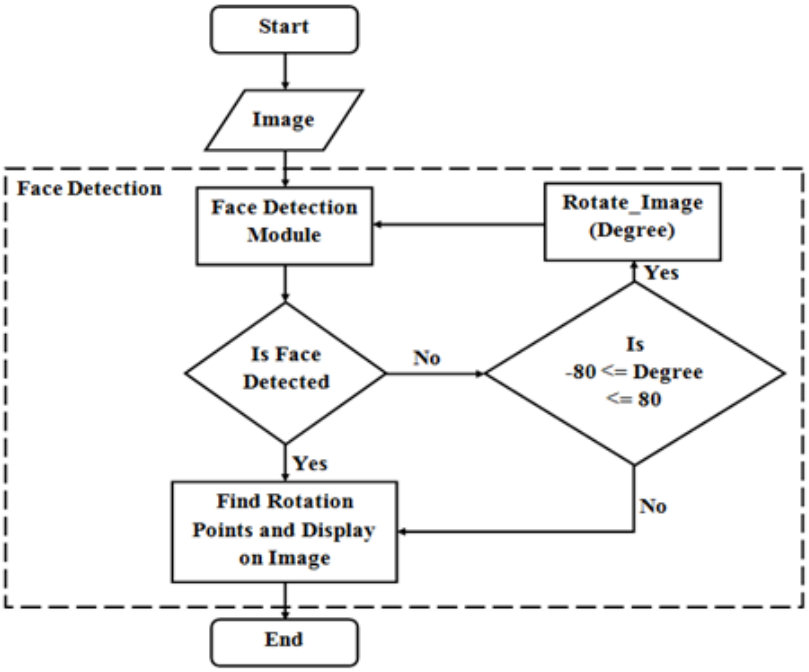


Fig. 3.1. Flowchart of proposed face detection method in streaming data.

$$\begin{bmatrix} X2 \\ X2 \end{bmatrix} = A * \begin{bmatrix} X1 \\ Y1 \end{bmatrix} + B \tag{3.1}$$

Matrix A is used for pure rotation and it is defined as (Equation 3.2):

$$A = \begin{bmatrix} cos\ \Theta & -\ sin\ \Theta \\ sin\ \Theta & cos\ \Theta \end{bmatrix}, B = \begin{bmatrix} 0 \\ 0 \end{bmatrix} (for\ clockwise\ rotations) \tag{3.2}$$

Here, image coordinates are considered, so the y axis goes downwards. The rotation formula is defined for the y axis goes upward. Now to get the new coordinates (x', y') for the image using the old coordinates (x, y), let us consider u = initial angle, i = angle of rotation.

$$x = r\ cos\ u \tag{3.3}$$

$$y = r\ sin\ u \tag{3.4}$$

$$x' = r\ cos(u+i)\ = r\ cosu\ *\ cosi\ -\ r\ sinu\ sini \tag{3.5}$$

$$y'\ =\ r\ sin(u+w)\ =\ r\ sinu\ *\ cosi\ +\ r\ cosu\ sini \tag{3.6}$$

hence:

$$x'\ =\ x\ cos(i)\ -\ ysin(i) \tag{3.7}$$

$$y'\ =\ y\ cos(i)\ +\ xsin(i) \tag{3.8}$$

Thus, when the original point (x, y) is rotated about another (origin) point (t, q) counterclockwise by an angle = θ, the new point's coordinates are calculated as follows (i) translating the entire plane so that (t, q) goes to the origin and (ii) perform the rotation and translate the entire plane back.

Therefore, the new point's coordinates are given below.

$$x' = (x-t)\ cos(\theta)\ -\ (y-q)\ sin(\theta)\ +\ t \tag{3.9}$$

$$y' = (x-t)\ sin(\theta)\ +\ (y-q)\ cos(\theta)\ +\ q \tag{3.10}$$

(Algorithm 3.3 uses these equations to calculate new point.)

Proposed Algorithm: The main goal of algorithm is to detect the face in the image obtained from streaming data with high accuracy. If viola Jones is not able to detect the tilted image face in image, then the proposed algorithm will rotate the image sequentially to detect the face, if image is rotated in all possible ways and still not able to detect then there is no face in that image, which is not even in the image train dataset. Affine-Transformation is used for rotating the image in a given plane.

Algorithm 3.1 is used to find the face in the image, if it is not able to detect any face then it calls Algorithm 3.2 to rotate the image and angle it in a sequential manner and sends the new image back to the Algorithm 3.1 and then the detected method again finds the face in that image. But the problem with image rotation is to obtain the coordinates of the face from where it will recognize the face using the rectangle to indicate that the face has been detected. Here, by using the above formula, the coordinates of the face

Algorithm 3.1. Image Rotation

```
Degrees = [-80 to 80]
For images in IMAGES {
For angles in Degrees {
Rotate_Image = rotate_image(image,angle)
DetectFace = face.detectMultiscale(rotateImage)
If (detectFace) {
//Face is detected at Angle angle
Break }
//Draw Rectangle over the Image
Rectangle(image,pt1,pt2,pt3,pt4)
Image.show() }}
```

are obtained, but these detected coordinates are for the face, which is detected in that rotated image and not the original image. Even if we have the coordinates for the face, and if we try to map these coordinates on the original image, we may get the wrong results. In some cases, we will get results i.e., if the face is tilted with a very small degree (around -150 to 150), but if the face is tilted to a greater degree, then we will miss the face coordinates in the original image.

Algorithm 3.2. Affine Transformation

```
//rotate image by degree angle
//(Using Affine Transformation)
rotate_image(image,angle) {
if(angle == 0 or -1 or 1){
return image
}
//get the height & width of image
Height,width = image.shape()
//image_matrix is rotated by using affine transformation
Rotate_matrix = getRotationMatrix (height, width, image, angle)
//warp the image
Result_Image = warpAffine(Image,Rotate_matrix)
Return Result_Image
}
```

To overcome this problem, a modified version of affine-transformation is used, which will rotate the image in its origin and then get the coordinates for the original image. As the image is made up of pixels, it can be represented in matrix form, and

Algorithm 3.3. Coordinate Detection

1: //find the Position or points (x,y) to draw rectangle
2: //get the points and rotate them in angle around the origin
3: Rotate_Point(Points[],image,angle){
4: x,y,h,w = Points[] //x,y coordinate and height and width from that x,y
5: P = image.shape(x)*0.4 //origin x coordinate
6: q = image.shape(y)*0.4 //origin y coordinate
7: new_point1_x = (x-p)cos(angle)-(y-q)sin(angle)+p
8: new_point1_y = (x-p)sin(angle)+(y-q)cos(angle)+q
9: new_point2_x = (x+h-p)cos(angle)-(y-q)sin(angle)+p
10: new_point2_y = (x+h-p)sin(angle)+(y- q)cos(angle)+q
11: new_point3_x = (x-p)cos(angle)-(y+wq) sin(angle)+p
12: new_point3_y = (x-p)sin(angle)+(y+wq) cos(angle) + q
13: new_point4_x = (x+h-p)cos(angle)-(y+wq) sin(angle)+p
14: new_point4_y = (x+h-p)sin(angle)+(y+wq) cos(angle)+q
15: return [(new_point1_x, new_point1_y), (new_point2_x, new_point2_y), (new_point3_x, new_point3_y), (new_point4_x, new_point4_y)] }

then affine-transformation is applied on the image matrix M given by Equation 3.11.

$$M = \begin{bmatrix} cos\ q & -sin\ q & 0 \\ sin\ q & cos\ q & 0 \\ 0 & 0 & 1 \end{bmatrix}, B = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \tag{3.11}$$

Till now, the proposed methods are able to recognize the emotions from facial expressions. In the next study, the recognition of emotions from both auditory and visual cues are explored and how best this can be used in the learning environments are analyzed.

3.1.2 Video Affective Content Analysis

Figure 3.2 shows the proposed affective computing module for video emotion recognition. This module consists of two major divisions, namely: Visual data affective content analysis part and Acoustic data affective content analysis part. The recognized Ekman's basic emotions from each analysis part is merged and validated using standard databases like HUMANE [1] and SAVEE [2]. Then the final step consists of recognized emotions.

[1] http://emotion-research.net/download/pilot-db/
[2] http://personal.ee.surrey.ac.uk/Personal/P.Jackson/SAVEE/

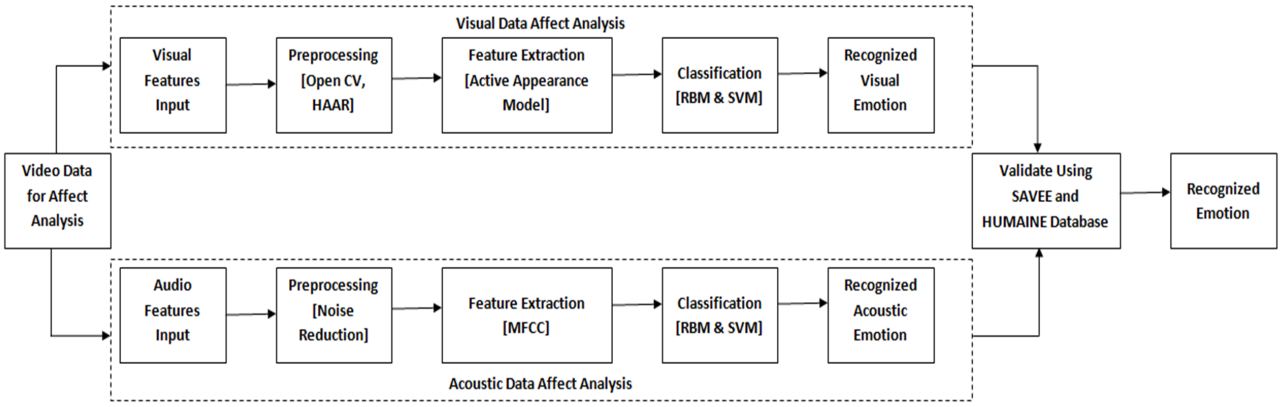


Fig. 3.2. Proposed framework for video affective content analysis.

The video affective content analysis module is divided into 2 phases: The training phase and the execution phase. The training phase is divided further divided into acoustic data training and visual data training phases.

(a) Training the Audio-Visual Data: SAVEE database is used for training the audio-visual data. The 60 percent of the Audio-Visual SAVEE database clips are considered for training, and the remaining 40 percent of the clips are used for testing. For every clip, the emotion from both Visual data and Audio data are obtained separately.

Acoustic Data Training Phase: Audio samples are trained using MFCC features of each audio sample that is extracted from video clips. The MFCC features are obtained as a 2-Dimensional vector of data with 13 features included for each of the audio samples, and the length of the vector depends on the size of the audio sample. Since the vectors consist of values of higher range, the MFCC data is normalized on a scale of 0 to 1 in order to reduce the computational cost. Hence the large floating-point values are normalized to values between 0 and 1 and thus significantly reduces the computational time of the training process without affecting the accuracy. Hence the data that needs to be trained in the normalized data of MFCC Features. As a part of the training process, SVM and RBM based classifiers are used.

The MFCC features are given as an input for the SVM. SVC (Support Vector Clustering) is used for the clustering of the data. If the data is not categorized initially, always a change in its parameters leads to an SVM, which gives different accuracy. So a perfect combination of the parameters is needed to obtain an SVM with high accuracy, and these are obtained by using Bayesian optimization. After obtaining the parameters, the SVM is used for testing the data.

Input for the visible layers of RBM is normalized MFCC features, and each node in the network works on a stochastic process for which each node needs to decide about to and from of data through the node. So it is a nondeterministic method of sending the data from the node. The weights of nodes are initially assigned by a random weight using a Gaussian distribution with mean 0 and standard deviation

0.1. Thus the MFCC is passed though the RBM network, which yields a trained net after a certain number of epochs.

Visual Data Training Phase: In the Visual Data Training Phase, initially, frame by frame images are obtained from videos. Then it is converted into a grayscale image. These grayscale images are cropped to increase the performance of feature selection. Detection of the frontal face is performed using Haar feature-based cascade classifiers. Then feature extraction is performed using AAM, which contains many distance points for eyelids width of mouth etc. Finally, the training phase is carried out by SVM. With the input as feature points, SVM uses a combination of Radial Basis Function and Polynomial Kernel to map each feature points to its respective emotion. This emotion mapping is performed for seven standard (basic) emotions such as anger, contempt, disgust, fear, happiness, sadness and surprise, and neutral emotions.

(b) Execution Phase: *Audio emotion recognition* is performed by extracting the MFCC features from the audio samples. The audio sample is extracted from the video on which emotion detection has to be done. Then the audio sample is divided into chunks so as to find the emotion based on audio at regular intervals of time similar to the way of detecting emotion from video frame by frame. But finding the emotion based on audio for each frame is not feasible because if the audio sample is extracted frame by frame, then the audio obtained always will be a just a small utterance of voice with different pitch and frequency. So the audio emotion is obtained for each second of the video. Thus the audio obtained from the video is divided into chunks of 1 second.

Visual emotion recognition starts with face detection using Local Binary Patterns (LBP). LBP is used for faster face detection and also minimizes the computational load of detecting the object for the feature extraction phase. Feature extraction is performed using Haar Cascades, but it is slow. Hence the regions which are detected as a face by the previous stage are applied to the execution stage where it cross-validates using a more accurate detector. Since the area to be scanned is smaller when compared to the original image; hence this verification phase executes faster. AAM is used for visual feature point detection, and these features are applied to already trained SVM classifiers. SVM classifies the detected face into its respective emotions.

3.1.3 Experimental Setup, Results, Analysis and Discussion

Multifacial Emotion Recognition

Experimental Setup:

- System Configuration:
  Processor: Intel Core I5
  RAM: 4GB
  OS: Windows 7
  GPU: Nvidia GeForce GT 830 M

- Datasets Used:
  Training Phase Dataset: Yale Face Database (YFD) (Georghiades et al., 1997) has been considered as a training dataset. Before using YFD as a training dataset, some pre-processing steps like cropping and thresholding are performed so as to increase the accuracy of feature extraction. But thresholding does not significantly increase the accuracy.

  Execution Phase Dataset: Once the training is done in Yale Face Database, the proposed method is tested for multiple face detection using Face Detection Dataset and Benchmark (FDDB) (Jain & Learned-Miller, 2010) and Labelled Faces in the Wild (LFW) (Zhang & Deng, 2016) face databases. The FDDB database is a dataset that consists of multiple faces in a single frame. Although the training is done using Yale Face Database, which trains for a single face, those features are used to detect multiple face emotions present in the FDDB and LFW face databases.

The FDDB dataset is designed for studying the problem of unconstrained face detection. This dataset is taken from the faces of wild dataset. The faces of wild dataset also known as labelled faces in the wild, consists of over 12000 images of faces which are collected from the web. These images are categorised into four different sets of images, namely:- i) original/normal images. ii) Aligned images, where the images are aligned in a proper manner such that the face detection can be done more accurately. iii) Funneled images, which are aligned images but each part or pose, are collected from different angles in different images. iv) Deep funnelled images are the aligned images that are developed using machine learning techniques where each feature can be tuned to represent the images that are at different resolutions. The results obtained are as shown in Table 3.1.

Table 3.1
Accuracy Analysis for the FDDB /LFW Dataset

Total Images	12620
Frontal Faces Detected	11232
Partial Detection	232
False Positive	0

Since GPU is tailor-made for image and video processing and image processing (Krizhevsky et al., 2012), GPU can significantly improve the speed of processing. The improvement in frame rate achieved with and without GPU is shown in Table 3.2. Once the face detection is done for multiple faces using both CPU and GPU, the corresponding emotion is detected.

Table 3.2
Performance of Proposed System

FPS with both CPU and GPU	17
FPS with only CPU	9
Speed Up Factor	≈ 2

FPS: Frames per second

Face Detection System on Streaming Data

Different datasets are used for the testing of our proposed algorithm. The following are the results with their runtime. It is evident that our proposed algorithm takes more time than Viola-Jones, but if it is compared to the number of faces detected, the proposed algorithm generates better results than Voila-Jones. To compare the algorithms with different approaches dataset like YALE (Figure 3.3 (a)) and for tilted, low resolution, occluded faces with difficult poses, and out of face images, FDDB dataset (Figure 3.3 (b)) and 'top 25 Google's searched "tilted face"' (Figure 3.3 (c)) images are used.

(a) (b) (c)

Fig. 3.3. Sample image frames from (a) Yale (b) FDDB and (c) Google top 25 tilted face

100% accuracy is obtained for YALE dataset using the proposed method. Also, 100% accuracy is obtained for those 25 images of 'top 25 Google's searched "tilted face"' whereas Viola Jones is able to detect only 2 images with an accuracy of 8%. For FDDB dataset, the proposed algorithms is tested on 3539 images and obtained an accuracy of 99.7% whereas Viola Jones is 93.5% accurate.

Results are shown in Table 3.3. The corresponding number of faces detected for each dataset is compared with the total number of faces present in that dataset and also the result of Voila Jones method. As it is observed from Table 3.3 that YALE database has same accuracy and runtime when compared to Voila Jones Method. This is because it contains only Frontal Face but there is significant difference in other datasets, since those datasets contain Tilted Face, Occlusion, different Pose and different Illuminations.

Out of 3539 images of FDDB dataset our proposed algorithm detected 3529 faces where as Voila Jones method detected only 3312. But it is also evident that our proposed algorithm took 55.88 sec runtime when compared to 52.12 sec of Voila Jones Algorithm.

Table 3.3
Results of Face Detection for Various Datasets

Technique	Dataset	Number of images	Number of faces detected	Runtime (sec)
Viola Jones	YALE	165	165	1.12
	Google top 25 "tilted face"	25	2	0.02
	FDDB	3539	3312	52.12
Proposed Method	YALE	165	165	1.12
	Google top 25 "tilted face"	25	25	1.28
	FDDB	3539	3529	55.88

For real time face detection webcam is used to capture the video stream data and then ran on the system. We are able to detect and track the face in streaming data as shown in Figures 3.4 and 3.5. These figures show the detected face when it is horizontally tilted and also occluded to some extent. Further, it shows the detected face when it is vertically tilted.

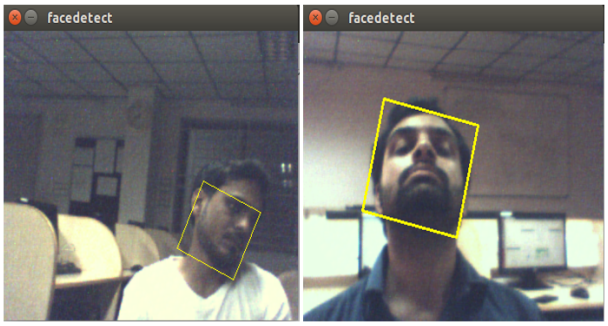


Fig. 3.4. Face detection results on streaming data.

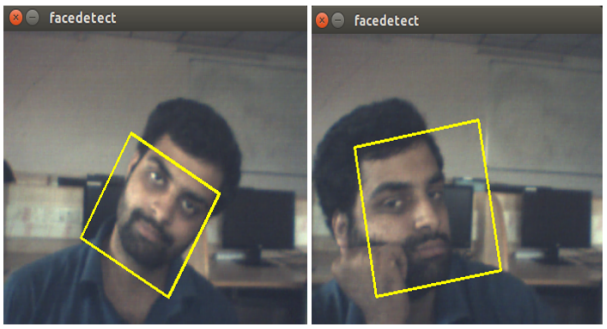


Fig. 3.5. Face detection results on streaming data.

Video Affective Content Analysis

In this subsection, details about datasets, streaming videos data using Web Camera, experimental results are discussed. The experiment is carried out on a system with the Intel core i7 processor, 3.40 Giga Hertz, with 8 Giga Bytes RAM running Windows 7 (64-bit). The proposed work is implemented in Python 2.7 programming language. Python is chosen for implementation because of its wide range of packages in Image Processing.

Datasets and Streaming Video Data

(a) Datasets:

HUMAINE: It is a database consisting of 46 videos, each video shows a person who is either performing or conversing. Snapshot of the HUMAINE database is given in Figure 3.6 (a).

(a) (b)

Fig. 3.6. Sample frame from (a) HUMAINE dataset and (b) SAVEE dataset.

SAVEE: SAVEE Audio-Visual data consists of 480 Audio-Visual clips of 4 different speakers namely DC, JK, KL and JE; each speaker has 120 clips of videos with 7 different emotions pertaining to the videos such as anger, disgust, fear, happy, neutral, surprise and sadness (Figure 3.6 (b)).

(b) Streaming Video Data

The method proposed here to recognize emotions is also put to test with the Streaming Video data obtained from the web cam (is a live streaming device which feeds or streams the image in real time) and Microsoft Kinect (is a motion sensing input used to recognize audio facial expression gesture and posture). The resulting recognised emotion is stored in the database. Figure 3.7 shows the snapshot of emotion recognition from live streaming video data (from Web Cam) for a given frame.

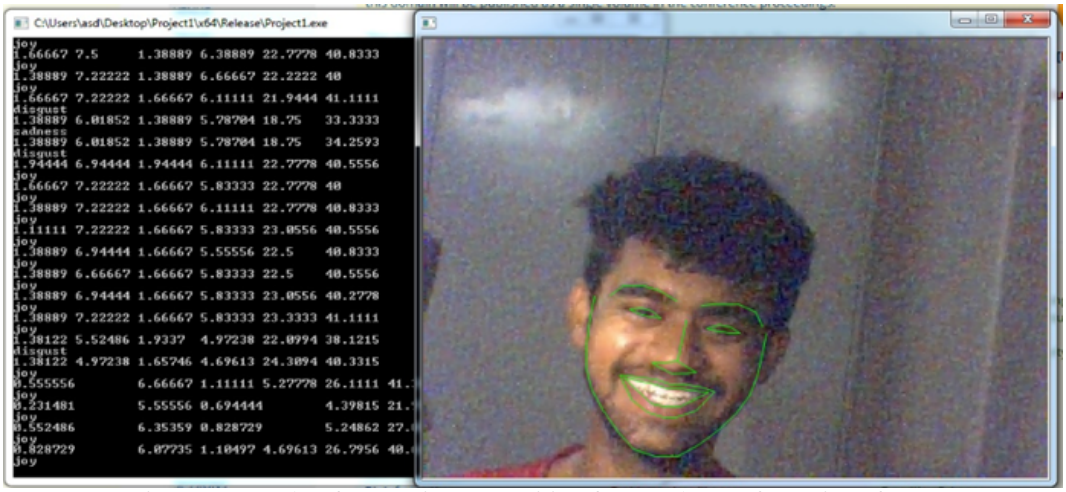


Fig. 3.7. Snapshot for emotion recognition from web-cam for a given frame

Experimental Setup, Results, Analysis and Discussion

Experiments are conducted on SAVEE and HUMAINE Datasets. The results of SVM classifier applied for both audio and visual data emotion recognition. After applying SVM on visual data using SAVEE dataset, got an accuracy of 70% for DC sample, 65% for JK sample, 68% for KL sample and 69% for JE sample. For audio, the accuracy is 75% for DC sample, 72% for JK sample, 70% for KL sample and 74% for JE sample using SVM classifier.

After applying RBM on Audio data, an accuracy of 79% is observed for DC sample, 78% for JK sample, 76% for KL sample and 80% for JE sample. The results obtained for emotion recognition using SVM classifier on visual data and RBM classifier on audio data are shown in Figure 3.8.

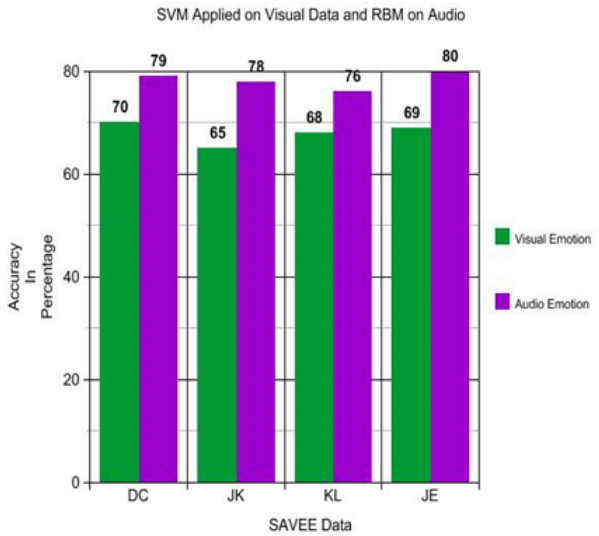


Fig. 3.8. SVM for visual and RBM for audio data

Further RBM on visual data is also experimented but results indicate that RBM underperforms (less than 50%) on visual data for both HUMAINE and SAVEE datasets. It is already shown that SVM outperforms RBM when applied on visual data (Wang & Ji, 2015). Hence the overall accuracy for SAVEE dataset is shown in Figure 3.9 with respect to visual data using SVM classifier and audio data with respect to both SVM and RBM classifiers.

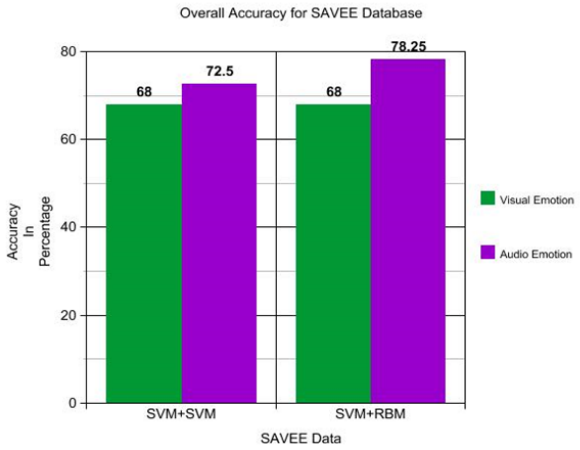


Fig. 3.9. Overall accuracy for SAVEE dataset

Similarly emotion recognition using both audio and visual feature are performed for HUMAINE Dataset. Video affective content analysis system is trained with SAVEE dataset for both audio and visual features. The accuracy after applying SVM on visual and audio data are 70% and 74% respectively. RBM applied on HUMAINE audio data has an accuracy of 79.4%.

The overall accuracy for both HUMAINE and SAVEE datasets is 70% (visual data classification using SVM) and for audio data classification, it is 74% using SVM Classifier and 79.4% with RBM Classifier as shown in Figure 3.10.

All datasets used for our experimentation are annotated and it is observed from Figure 3.10 that the combination of SVM classifier for visual data and RBM classifier for audio data is better than using only SVM or only RBM classifiers for video affective content analysis. 60% of SAVEE dataset is used for training, and tested on two different datasets, namely: SAVEE and HUMAINE for both audio and visual emotion recognition from the video. RBM is used for audio emotion recognition because, the accuracy remains almost same for both annotated and unannotated data, but it performs slightly better with unannotated data compared to annotated data for classification. Hence the

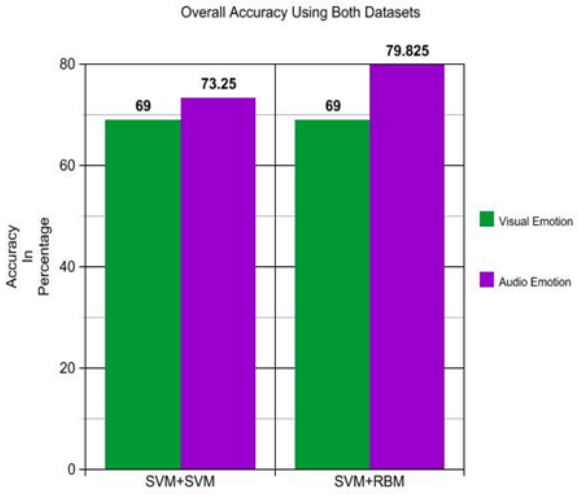


Fig. 3.10. Overall accuracy using both datasets

accuracy of our proposed SVM-RBM classifier remains almost the same for video affective content analysis for streaming data. On the other hand, the accuracy of SVM varies a lot with respect to accuracy when it comes to annotated and unannotated data.

The preliminary studies conducted to recognize the Ekman's basic emotions through audio and visual data performed better only for the data obtained from the webcam or any close range of facial expressions (visual data). And the action units used in these studies are not defined for learning-centered emotions. Hence, in the next section, convolutional neural network based architecture is proposed to address the above mentioned issue by predicting the learning-centered emotions of the students from both the e-learning and classroom environments.

3.2 Proposed Methodology for Recognizing Two different Affective States

The proposed methodology includes the database creation and the students' affective state recognition using convolutional neural networks. The database creation includes details such as affective state classification, affective state's labels & definitions and data annotation performed on both posed and spontaneous data. Further, the data augmentation to achieve the robustness of the proposed model is included and thus, two different datasets are created for the single and multiple students in a single image frame. Figure 3.11 shows the flow of the proposed architecture and the detailed explanation is provided in the following subsections.

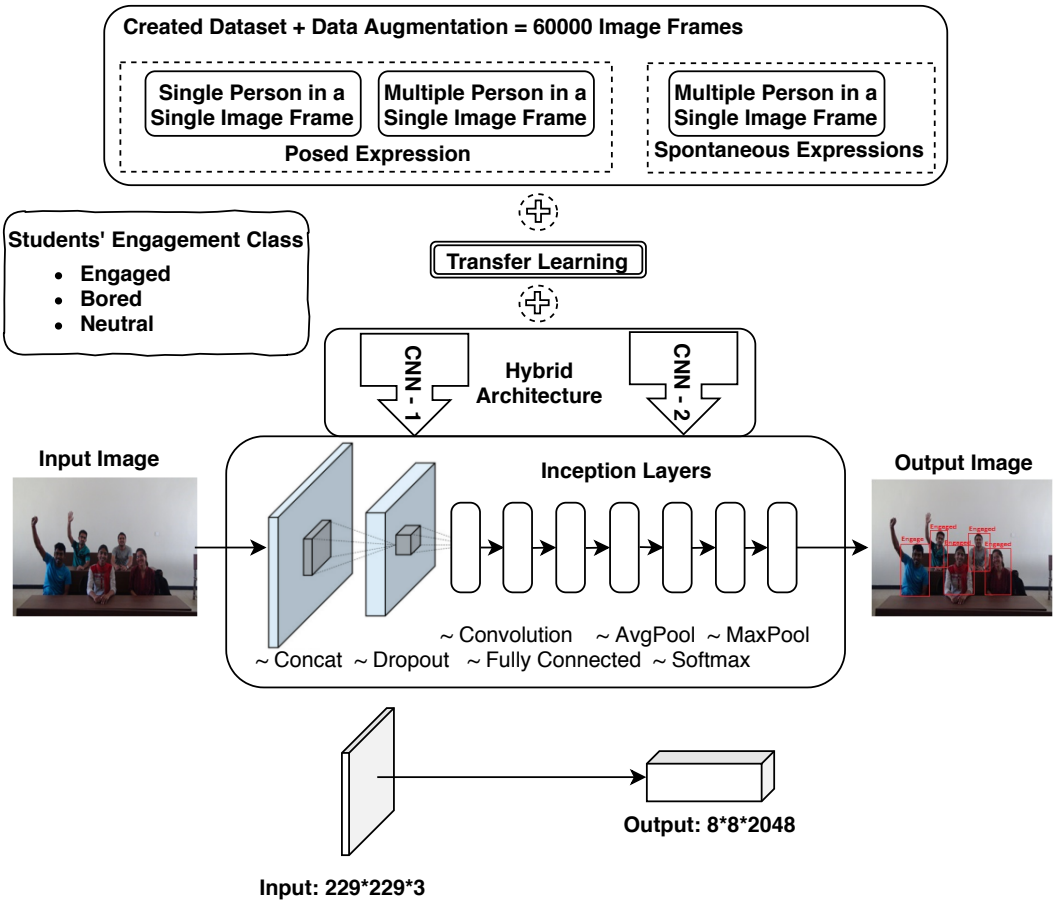


Fig. 3.11. Proposed students' affective state recognition architecture

3.2.1 Database Creation

Affective State Classification

Existing works include recognizing a student's facial expressions and classifying them into Ekman's basic emotions (Ekman, 1992). The study investigated by (D'Mello et al., 2010) showed that learning-centered emotions such as flow, boredom, and frustrated are more dominant and regularly observed in students than Ekman's basic emotions. (Whitehill et al., 2014) considered the body posture and eye movement along with facial expressions and classified the engagement into four types of engagement levels. The different types of engagement levels are monitored during the learning process, and corresponding affective trajectories are generated. The observed results show that flow or engaged students can finally end up getting bored if the confusion or frustration is not addressed, as shown in Figure 3.12: Russell's core affective framework (D'Mello, 2012). Hence, the engaged/flow, neutral, and boredom affective states are considered for this study.

Analyzing the students' emotions alone is not sufficient to classify among the above mentioned affective states. Behavioral patterns such as looking away from the task, eyes barely open, and complete lean on the desk are also beneficial to analyze the affective states. The facial expressions contain more information about the emotions, and hand gestures, and body postures contain more information about the behavioral patterns. Therefore, along with facial expressions for emotions, the hand gestures and body postures are also used to analyze the behavioral patterns to classify the affective states. So, the proposed classification includes both behavioral and emotional engagements of the students, along with some cognitive aspects involved in it. Based on this, the student's engagement is classified into three states, namely: engaged, boredom, and neutral.

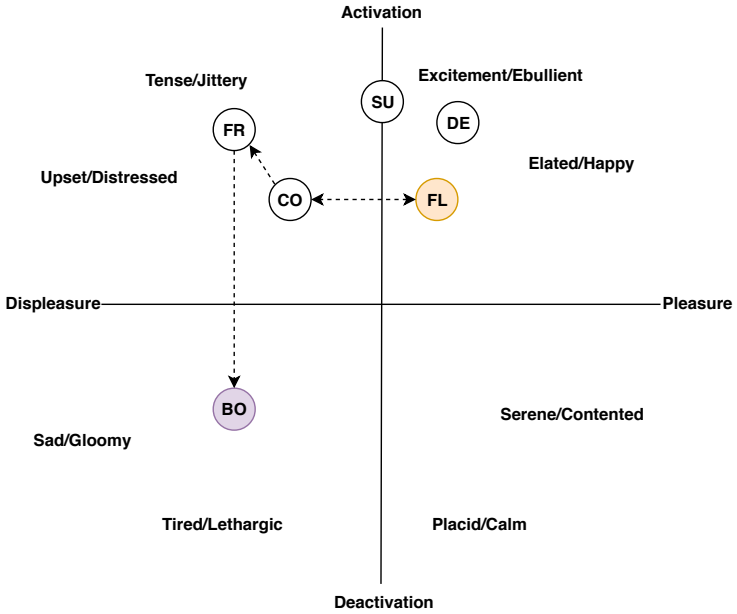


Fig. 3.12. Russell's core affective framework (FRustrated, COnfusion, FLow, BOredom, DElight, SUrprise) (D'Mello, 2012)

Labels and Definitions:

Learning-centered emotions and the behavioral patterns (head and eye movements) are considered as mentioned in (D'Mello et al., 2010) and (Whitehill et al., 2014), respectively. But, these works are on the single person in a single image frame. Also, the behavioral patterns mentioned in (Whitehill et al., 2014) did not include the body postures and hand gestures. Further, a few emotions are recognized accurately using both hand gestures and facial expressions instead of using only facial expressions. Hence, in this work, the existing classification standards are modified by adding the hand gesture and the body posture components. The label details are as follows.

Engaged: It includes both emotional and behavioral engagement aspects such as looking at the teacher/board, taking notes, listening, and discussions with the teacher and the standard facial expression action units of engaged/flow emotion.

Boredom: It includes looking away from the board/teacher, eyes barely open or completely closed and obviously not thinking about the task, leaning on the desk with heads down and the standard facial expression action units of boredom emotion.

Neutral: If there are no behavioral patterns and affective states recognized, with no facial expression, then it is labeled as neutral.

Data Annotation

Data annotation for posed expressions: To collect the posed expressions and behavioral patterns from the students, the situation and the required affective state expected from them are mentioned. We collected 2 minutes of video clips for all the variants of each affective state. We observed that the affective states are at its peak for about 2 to 8 seconds for each variant of an affective state in a 2 minutes video. We collected those peak frames and got manually verified by the student.

Annotations for classroom data in the wild: We also collected the students' 1-hour spontaneous (expressions and behaviors) classroom data with 24fps. The collected data is preprocessed by deleting the blurred and repeated frames. Here, repeated frames are those frames which had the same expressions and behaviors from all the students present in the single image frames. E.g., there are occasions where two successive image frames are obtained which are exactly the same. After pre-processing, 2400 image frames are obtained, and these image frames are manually annotated by three annotators (Sidney et al., 2005). One annotator is the student himself who did self-annotation, and the other two are expert annotators who followed the guidelines and definitions given for each label. We calculated Cohen's κ to check the reliability among annotators, and found that these annotators reliably agree when discriminating three different affective states with Cohen's $kappa$ = 0.59. Figure 3.13 shows the sample image snapshot of spontaneous classroom data.

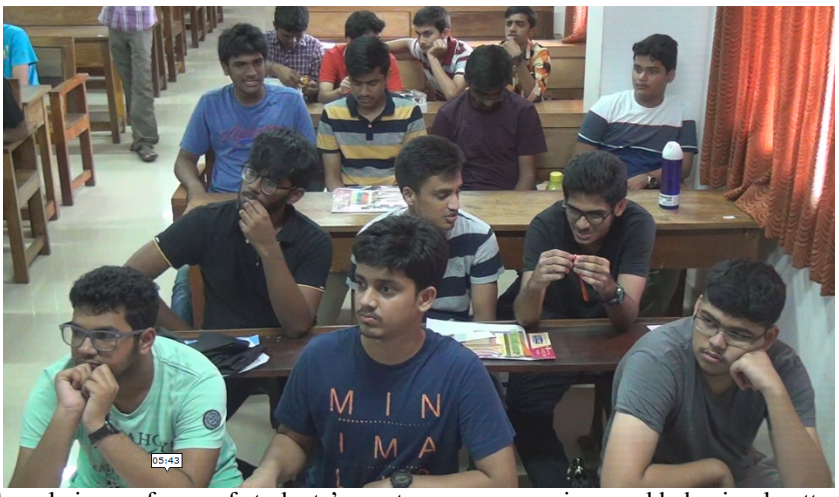


Fig. 3.13. Sample image frame of students' spontaneous expressions and behavioral patterns obtained from classroom data

3.2.2 Affective State Detection in a Classroom Environment

Here, two different methods are used for identifying the affective state of the students, which are as follows.

- Method1: - Detecting all the students in the image, classifying their affective states separately using each student's multi-modal data and then taking the collective average of all the students' affective states.
- Method2: - One complete image frame with all the multi-modal features of all students is considered to classify the group engagement score of that image frame.

Building Convolutional Neural Network Model

The functioning of convolutional neural networks is similar to that of a bunch of neurons collectively processing the input image and analyzing the data using axons, dendrites, and synapse. Similarly, CNN uses hidden layers, fully connected layers, and a classifier to classify the given input image frame data.

As shown in Figure 3.11, filters are convolved on the input image, and their dot product is calculated. These filters are used until they reach the first fully connected layer. Then ReLu is used as an activation function after every convolutional layer (CONV) or pooling layer (POOL). Finally, after the last fully connected (FC) layer, the softmax classifier is used to get the probability distribution for different classes which represents scores between 0 and 1. Figure 3.11 shows an overall view of the proposed CNN model where an image is given to the model, and then its pixels are used as an input

with their corresponding RGB (Red Green and Blue) values. This input feature (pixels) goes through a series of layers (CONV, Relu, Pool, and FC) and finally, it classifies the image as a class in terms of its score. Here the students' image is given as input and the output is the collective average of all of their affective states.

We used the inception v3 (Szegedy et al., 2016) model for affective state classification of students by providing a 299x299x3 RGB image to the model. To reduce the training time, only the final layer of inception v3 model is trained for affective state categories (engaged, boredom and neutral). We built two separate convolutional neural network models, CNN-1 for single student in a single image frame and CNN-2 for multiple students in a single image frame where CNN-2 consists of few extra layers. By combining both of these, emerged with the proposed hybrid architecture which yielded better accuracy.

The two CNN models are:

- CNN-1 - It is trained for classifying the affective state of a single student in a single image frame (Figure 3.14).
- CNN-2 - It is trained for classifying the affective state of multiple students in a single image frame (Figure 3.14). Further, the number of layers is increased by 20 from the base Inception-V3 model. The hyperparameters are fine-tuned separately.

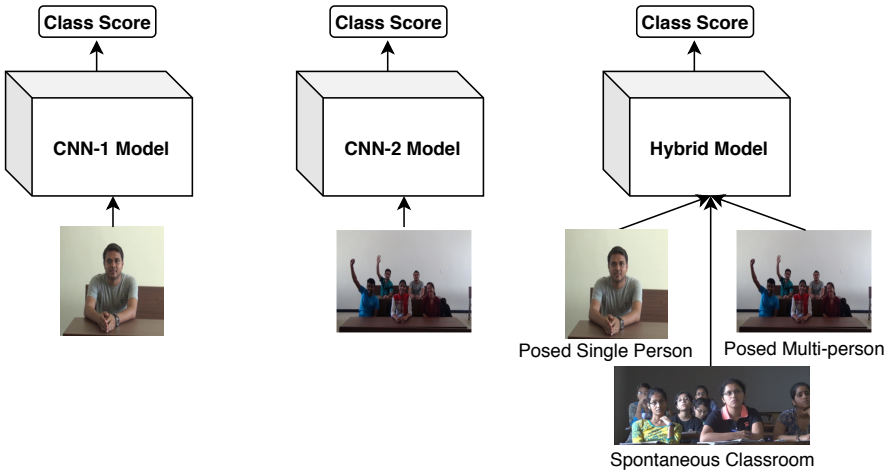


Fig. 3.14. Class score prediction using proposed architectures

Classification of the Affective State of Students for Entire Image using Convolutional Neural Network: The students' spontaneous data is analyzed using a hybrid model which is a combination of both CNN-1 and CNN-2 models. The number of layers in

the hybrid model is the same as that of CNN-2 model. The last two layers of CNN-1 model are also added in the hybrid model along with their weights. The model is trained for 299x299x3 image with RGB color values to downscale or upscale the image accordingly. An input image of size 299x299x3 is given to the hybrid model, and it generates scores for the three corresponding affective states. The class with the highest score is considered for the overall affective state of students in the classroom.

Collective Average Affective State Score: The students' posed data present in multiple people in a single frame image has one affective state for the entire image, but the students' spontaneous classroom data has different affective states in a single image frame. Hence feature fusion is used to calculate the same. The multi-modal (here, it is an intra-image multi-modality where the features of different students with their facial expressions, hand gestures and body postures present within that image frame are considered) feature fusion vector V_f for any pixel p_i and normalized prediction vector N_{P_i} uses normalized predicted probability distribution $N_{P_i,a}$ of class a using the softmax function (Equation 3.12).

$$N_{P_i,a} = \frac{e^{W_a^T V_f}}{\sum_{i \in classes} e^{W_i^T V_f}} \tag{3.12}$$

Where, W is the temporary weight matrix used to learn the features. The training generally converges in $T = 4000$ epochs. The final collective average affective state score A_{S_i} is given by Equation 3.13.

$$A_{S_i} = arg\,max\; N_{P_i,a}\; where\; a \in classes \tag{3.13}$$

The collective average affective state score is subjective because of which, we rendered a provision in research methodology where the instructors can just look at the individual scores of each student and analyse the group level affective state manually. In classes where large number of students are present or in webinars which are conducted at several places at a time, it may not be feasible to report the individual student's score and hence the instructors are provided with the students' collective average affective state score. This acts as a threshold and helps them to address those classes with lower scores.

3.2.3 Data Augmentation

The key to the robust deep learning model is the high quality data. But, it is a challenge to obtain such data. One better way to address this issue is the augmentation of datasets. Due to lack of data available for affective state analysis, data augmentation is used. Data augmentation has increased training data size by 10-fold. Following are the different data augmentation techniques used on the created datasets.

- channel_shift_range: Random channel shifts of the image.
- zca_whitening: Applies ZCA whitening to the image.
- rotation_range: Random rotation of image with a degree range.
- width_shift_range: Random horizontal shifts of the image with a fraction of total width.
- height_shift_range: Random vertical shifts of the image with a fraction of total height.
- shear_range: Shear intensity of the image where the shear angle is in the counter-clockwise direction as radian.
- zoom_range: Random zoom of the image where the lower value is 1-room_range and upper value is 1+zoom_range.
- fill_mode: If any of constant, nearest, reflect or wrap are filled according to the given mode, if any points outside the boundaries of the input.
- horizontal_flip: Randomly flip the inputs horizontally. Table 3.4 shows the details of different data augmentations performed on the created our dataset.

Table 3.4
Types of Data Augmentation Used

Type of Augmentation	Augmentation Value
channel_shift_range	20
zca_whitening	TRUE
rotation_range	40
width_shift_range	0.2
height_shift_range	0.2
shear_range	0.2
zoom_range	0.2
horizontal_flip	TRUE
fill_mode	Nearest

3.2.4 Creation of Datasets

We created a dataset considering two types of image inputs consisting of 50 Indian students each for both training and testing.

- Dataset-1: It contains 24000 posed images of 50 students with data augmentation for classification. Dataset consisting of images of a single student in a single

(a) (b) (c) (d)

Fig. 3.15. Sample images of dataset-1: single student in a single image frame

image frame with three different affective states, namely: engaged (Figure 3.15 (a)), boredom (Figure 3.15 (b) and (c)) and neutral (Figure 3.15 (d)) as shown in Figure 3.15. Every affective state has 8000 image frames each.

- Dataset-2: It contains 36000 images with data augmentation for classification of multiple students sitting in the classroom and their affective states are being classified into three different classes, namely: engaged (Figure 3.16 (d)), boredom (Figure 3.16 (a) and (b)) and neutral (Figure 3.16 (c)) with 12000 images each as shown in Figure 3.16. Dataset-2 also contains 2 hours of classroom video with 2400 students' images with spontaneous expressions.

Collection of posed data sets is necessary as it facilitated the training of the proposed architecture. It is also observed that the posed datasets for single and multiple students in a single frame image considered for training the proposed architecture increased the overall accuracy by 18% while testing the spontaneous classroom data (results are shown in Table 3.10).

These 24000 images of dataset-1 and 36000 images of dataset-2 contain the images with data augmentation as mentioned in Table 3.5.

(a) (b) (c) (d)

Fig. 3.16. Sample images of dataset-2: multiple students in a single image frame

Table 3.5

Details of Created Datasets for Posed Affective States

Affective States	Number of single students in a single image frames with data augmentation	Number of multiple students in a single image frames with data augmentation
Engaged	8000	12000
Board	8000	12000
Neutral	8000	12000

3.2.5 Experimental Setup, Results, Analysis and Discussion

Experimental Setup

For the current study, 8^{th} Generation Intel® Core™ $i7-4510U$ Processor, 8GB RAM, and 2GB NVIDIA® GeForce® $840M$ are used.

Table 3.6 shows the training setup attributes for both CNN-1 and CNN-2 models where each attribute has its corresponding values used for training both the models.

Retraining of the last pool layer (pooled three layers) took around 2 hours for three classes of 50 students with three affective states. Each training set consists of 800 images of students, with a different facial expressions, hand gestures and body postures.

Performance Evaluation of Posed Data

Convolutional neural networks work effectively for object classification in an image, even for detecting faces, gestures and the affective states of humans. The performance

Table 3.6
Training Setup for CNN-1 and CNN-2 Models

Attribute	Details	Attribute	Details
Validating and testing batch size	100 Images	Optimize	Adaptive Momentum
Epochs	4000	Loss	Categorical Entropy
Learning rate	0.1	Weight initialization	Pre Trained weights of Imagenet v3 model trained over imageNet. We fine tune that model by retraining bottleneck layer
Classifier at Bottleneck layer	Softmax classifier	Testing set	10% of training dataset
Number of classes	3	Validating set	10% of training dataset

evaluation of posed data is carried out using Method1 as mentioned in Section 3.2.2. Initially, similar results are obtained for both the methods (Method1 and Method2), but did not get better results when the number of students present in a single image frame is more than one. Table 3.7 shows the accuracy of different affective states for the proposed hybrid architectures.

The performance evaluations are carried out for both CNN-1 and CNN-2 models individually, but the proposed hybrid architecture performed better than the individual models. The results are projected in Table 3.10.

Table 3.7
Performance Evaluation of Posed Data

Model	Affective States	Test Accuracy (%)	Average Test Accuracy (%)
CNN-1	Engaged	95.2	94.0
	Boredom	93.1	
	Neutral	93.7	
CNN-2	Engaged	96.2	95.6
	Boredom	95.6	
	Neutral	95.0	
Hybrid	Engaged	96.2	95.8
	Boredom	96.1	
	Neutral	95.0	

The proposed CNN-1 model with the dataset of single students images got a final test accuracy of 94%.

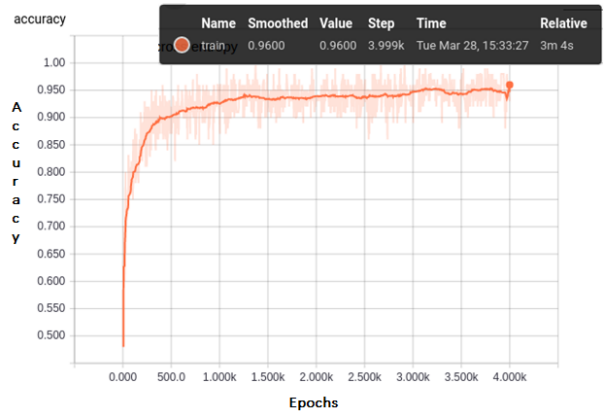


Fig. 3.17. Accuracy curve w.r.t epochs for training CNN-1 model

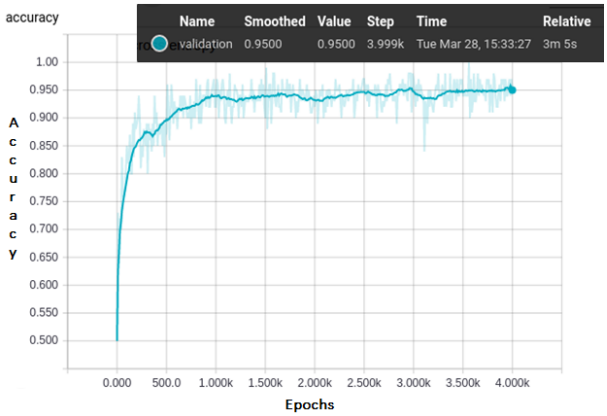


Fig. 3.18. Accuracy curve w.r.t epochs for validation CNN-1 model

The accuracy for three different affective states, i.e., engaged, boredom, and neutral are 95.2% , 93.1% and 93.7%, respectively. We also observed that training and validation accuracy is improving with each step or epoch, and reached the saturation after 1500 epochs as shown in Figure 3.17 and Figure 3.18. At the end of 4000 epochs, cross entropy of 0.1459 for training and 0.2045 for validation is obtained as shown in Figure 3.19 and Figure 3.20. Similarly, the experiment is performed on CNN-2 model and the results of test accuracy are shown in Table 3.7.

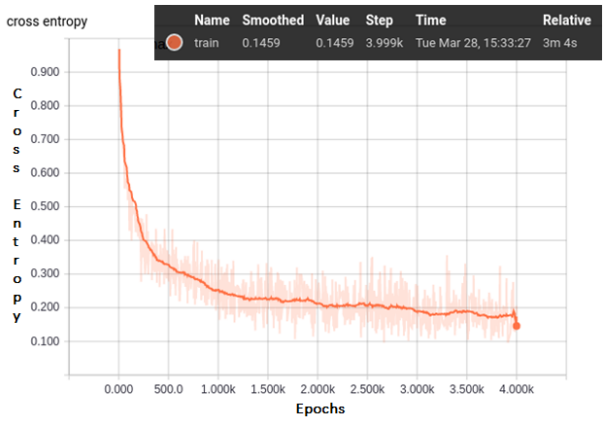


Fig. 3.19. Accuracy curve w.r.t cross entropy for training CNN-1 model

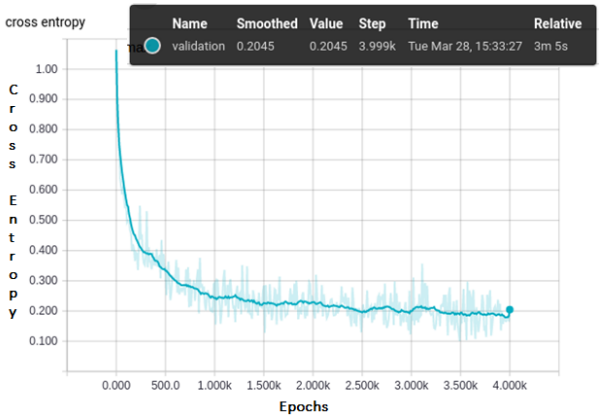


Fig. 3.20. Accuracy curve w.r.t cross entropy for validation CNN-1 model

An off-the-shelf CNN training method is used to overcome the problem of high computational training cost, and this process is referred to as transfer learning. Here, we used the Inception-V3 model which is trained over ImageNet dataset, weights to this model are used by training the last layer (pool 3 layers) of inception model as shown in Figure 3.11. Training took 1.5 hours for the CNN-1 model and 2 hours for the CNN-2 model as shown in Table 3.8. This helped us to reduce the training time and thus produce higher accuracy using Inception-V3 model for transfer learning.

We considered three different parameters for both the CNN-1 and CNN-2 models. Table 3.8 gives the details about the testing time for each model when given an image as input. Training time represents the time taken for training the model for a given set of images. Testing accuracy demonstrates the performance of the proposed model while classifying the given images. As the overall running time for CNN-1 is lesser than that of CNN-2, it can further be considered for use in e-learning and flipped classroom environments.

Table 3.8
Time and Accuracy Obtained for the Proposed Methods

Proposed Methods	Training time (Hrs)	Testing Accuracy (%)	Testing time (ms)
CNN-1	1.5	94.0	1200
CNN-2	2.0	95.6	1200

Performance Evaluation of Spontaneous Data

Testing on Spontaneous Classroom Data: The proposed techniques are tested on spontaneous classroom data, where the videos are recorded during the actual classroom lectures. The test data is student-independent and classroom-independent data (the students/classroom in test data are different from training data). Table 3.9 shows the

classification accuracies of affective states. The recognition of boredom affective state is better when CNN-1 model is used when compared to that of CNN-2, whereas CNN-2 performs better than CNN-1 for the engaged affective state. Hence, the hybrid of CNN-1 and CNN-2 models is used and obtained better accuracy as shown in Table 3.9.

Table 3.9
Comparison of Proposed Affective State Classification Techniques with Spontaneous Classroom Data

Model	Affective States	Accuracy (%)
CNN-1	Engaged	61.4
	Bored	66.3
	Neutral	70.0
CNN-2	Engaged	68.3
	Bored	63.1
	Neutral	71.1
Hybrid	Engaged	68.7
	Bored	66.8
	Neutral	73.7

Overall Results, Analysis and Discussion

The overall performance evaluation of the proposed method with the created database is shown in Table 3.10. The CNN-1, CNN-2, & hybrid models are tested on student-independent & classroom-independent 10-fold cross-validation, and the observed results are mentioned in Table 3.10. It is observed from Table 3.10 that both CNN-1 and CNN-2 are necessary for better classification of student's affective states. The hybrid model gave almost the same performance as that of CNN-1 and CNN-2 models for posed dataset but performed better than CNN-1 and CNN-2 models for the spontaneous dataset. F1-score is less for spontaneous data as the affective states are not equally distributed in the spontaneous classroom data.

Table 3.10
Overall Results of the Proposed Model

Performance Metrics	Affective State Classification					
	C-1	C-2	H	C-1	C-2	H
	Posed Dataset			Spontaneous Classroom Dataset		
Average Accuracy	0.84	0.85	0.86	0.65	0.67	0.70
Average Recall	0.87	0.86	0.89	0.66	0.71	0.72
Average Precision	0.91	0.88	0.91	0.69	0.78	0.77
Average F1-score	0.85	0.84	0.84	0.60	0.63	0.62
AUC	0.88	0.89	0.90	0.63	0.68	0.69

C-1: CNN-1 Model; C-2: CNN-2 Model; H: Hybrid Model

The created dataset also contains images from different camera positions (to make the recognition process more robust). Figure 3.21 is a sample image frame from the created dataset with different camera position/angle and is tested with the proposed method. It is observed from Figure 3.21 that the affective states are classified correctly for each student and also the overall class score for that frame is also calculated (top left corner in Figure 3.21). The proposed model is also tested on a few images of classroom subset data of Imagenet database (Deng et al., 2009) and a sample image snapshot is shown in Figure 3.22. Since the ImageNet database does not contain the annotations for students' affective states, the comparison of the results with the ground truth is not possible.

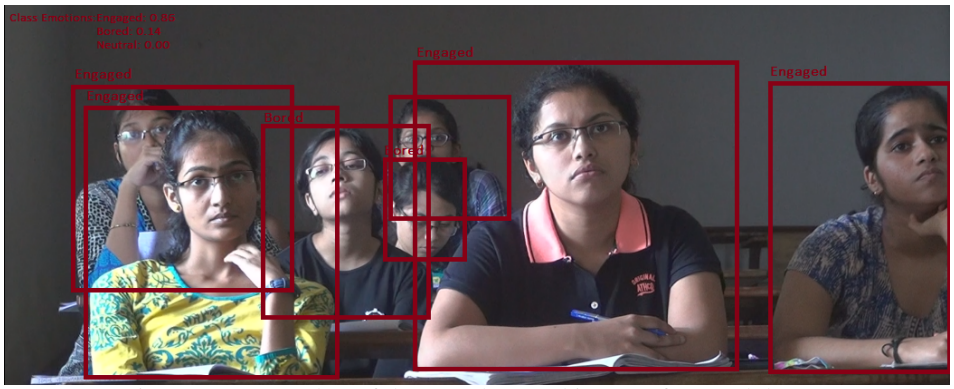


Fig. 3.21. Screenshot of the sample tested image of created dataset

Fig. 3.22. Screenshot of the sample tested image of ImageNet dataset

Comparison with state-of-the-art

State-of-the-art deep learning architectures: Deep learning methods are popular and give better results for multi-modal affective content classification. The existing deep learning techniques use the following architectures as backbones, which include, AlexNet, ResNet, VGGNet and Inception-V3. These architectures are tested on the created database and obtained an accuracy of 51% for AlexNet (Krizhevsky et al., 2012), 53% for ResNet (He et al., 2016), 61% for VGGNet (Simonyan & Zisserman, 2014). The results are encouraging for a single person in a single image frame but failed to perform better than the proposed model for images obtained from the classroom environment.

Students' affective state classification methods: There are few works on students' engagement. (Whitehill et al., 2014) used Gabor features with SVM and obtained an AUC of 0.729 to analyze the behavioral patterns. But this result is by testing on a single person in a single frame image. (Zaletelj & Košir, 2017) used the Kinect sensor & KNN and obtained an accuracy of 0.753 in the classroom environment. Though the Kinect considers multiple people in a single image frame, the range of capturing students is less. Hence, the students' detection accuracy decreases if the number of students are more than 10 in a single image frame. (Kahu, 2013) and (Bosch et al., 2016) used deep instance learning and WEKA (Waikato Environment for Knowledge Analysis) tools, but it is already evident from the literature that the handcrafted features are less efficient for faces in the wild. It is difficult to directly compare the proposed methodology with the existing works as the datasets, and the multimodalities are different, in spite of which the results are comparable and more robust in terms of AUC and accuracy.

3.3 Proposed Methodology for Students' Emotional Engagement Analysis

In the previous section, the proposed architecture classifies the students' emotions into two different affective states only, and this is not sufficient to analyze the students' emotional engagement (Sinatra et al., 2015). In this section, all the dominant learning-centered emotions are considered to classify the students' emotions by the proposed methodology. The proposed methodology is divided into two parts: first, the students' affective state analysis is performed for both single and multi-person in a single image frame scenario using facial expressions, hand gestures, and body postures. The second part involves object localization by recognizing the students' faces, hand gestures, and

body postures separately for every student using a bounding box.

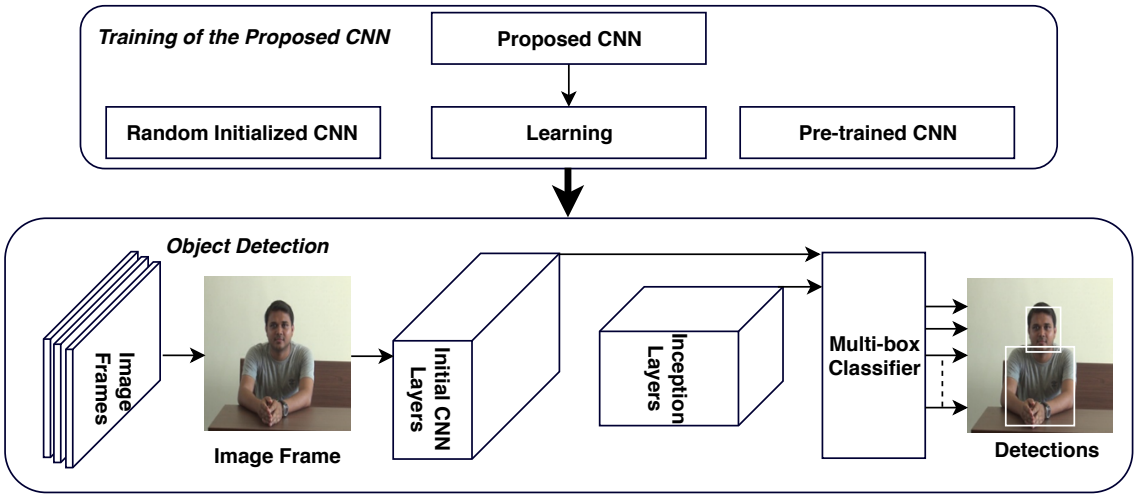


Fig. 3.23. The proposed architecture for affective state classification and object localization.

The entire architecture includes the students' affective state classification with object localization as shown in Figure 3.23. The input images are taken from all the four learning environments. These input image frames are used for training both the classification and localization models. The initial input is fed to the first convolutional layer and is then processed until the last convolutional feature map is obtained. For the classification of the students' affective states, a softmax classifier is utilized whereas for localization, another set of layers are deployed for the detection and plotting of the bounding box w.r.t. faces, hand gestures, and body postures. The details of affective state classification and object localization are mentioned in Section 3.3.2.

3.3.1 Data Collection and Annotation

Affective States: The entire student image frames data is classified into the learning-centered emotions as mentioned in D'Mello et al. (2010, 2007). The affective state definitions mentioned in D'Mello et al. (2010, 2007) are considered in this study. Though these definitions considered only the facial expressions to classify them into an affective state; hand gestures and body postures are also used in this proposed methodology along with the facial expressions to classify them into an affective state. Further, these definitions include not only the facial expressions of the emotional engagement, but also behavioural engagement along with some cognitive aspects (Whitehill et al., 2014; Picard, 1997; Grafsgaard et al., 2014, 2013; Bosch et al., 2016) to classify the recognized affective states using students' facial expressions, hand gestures and body postures. For

example, the definition of an engaged affective state is a state of interest that results from the involvement in an activity. This includes not only the facial expressions of students but also their behavioral aspects observed using hand gestures and body postures related to taking/writing notes, answering questions/asking the questions, paying attention towards the board or teacher etc.

Affective State Definitions (D'Mello et al., 2010; Whitehill et al., 2014; D'Mello et al., 2007):

- Boredom: uninterested in the current problem.
- Confusion: poor comprehension of material, attempts to resolve erroneous belief.
- Disgust: annoyance and irritation with the material and their abilities.
- Fear: feelings of panic and extreme feelings of worry.
- Sadness: feelings of melancholy, beyond negative self-efficacy.
- Frustration: difficulty with the material and an inability to fully grasp the material.
- Sleepy: extremely not interested and in a mental state of sleep.
- Happiness: satisfaction with performance, feelings of pleasure about the material.
- Neutral: displays no visible affect, at a state of homeostasis.
- Surprise: genuinely does not expect an outcome or feedback.
- Delight: a high degree of satisfaction.
- Engaged: a state of interest that results from involvement in an activity.

Participants: The proposed architecture is trained and tested on 350 undergraduate and postgraduate students of National Institute of Technology Karnataka (NITK) Surathkal, Mangalore, India. These student participants present in this database creation are in the age group of 20 to 26 years.

Posed and Spontaneous Expressions: We created the dataset consisting of students' face, hand gestures and body postures. Dataset contains both posed and spontaneous expressions which include single and multi-person in a single image frame scenario. Face of the students includes frontal, profile and tilted faces, hand gestures include raise of hands and body posture includes normal, half bent and full bent or completely lean on

the desk poses. The entire dataset includes variants such as occlusion, background clutter, pose, illumination, cultural & regional background, intra-class variations, cropped images, multipoint view and deformations.

The posed expression dataset is created to improve the accuracy of object localization and affective state classification by using it to train the proposed deep learning architecture. The posed or pre-informed expressions consist of both single and multi-person in a single image frame scenarios. This is performed with more than 30 students for 40 video clips of 2 minutes each. We performed State Trait Anxiety Inventory (STAI) and also asked them General Health Questionnaire (GHQ) to check their physical and mental distress before obtaining their posed expressions for the data collection. Further, the STAI and GHQ are performed to ensure that the students are in a normal state before the start of the posed emotion extraction process. These tests are not performed during spontaneous data collection process.

Spontaneous expressions are collected for more than 25 hours in a classroom environment. All spontaneous expressions are of multi-person in a single image frame scenario, but the number of persons in each frame vary depending on the subjects taught and the class strength.

Affective Annotation: We followed the Gold standard study, the standard annotation process used to label the learning-centred emotions (D'Mello et al., 2007) using multiple annotators. The detailed description on the participants, annotation, and all other related details regarding the created database is given in the separate Chapter, i.e., Chapter 6 of this thesis. The human annotators reliably agree when discriminating the recognized affective states with Cohen's $\kappa = 0.41$ for the entire database. Along with the annotation, object localization is also performed on each student by the annotator. The annotated image with object localization is stored in a JSON file.

3.3.2 Affective State Classification and Localization

Affective State Classification: CNN is a variant of multilayer perceptron and consists of several locally connected layers. CNNs exploit spatially-local correlation by enforcing a local connectivity pattern between neurons of adjacent layers. Figure 3.24 shows the proposed convolutional neural network architecture. The input image frame is directly fed to the first convolutional layer of dimension 1024*1024 with depth 3. The input

layer is convolved with the filter size of 5*5*3 (here 3 is the depth that extends upto the full depth of the input volume) with stride 2 and the depth increases according to the activation maps. If the filter size has a non-decimal value then zero padding is used to read the entire image. Leaky Relu is used as an activation function. The dimension of input image is reduced from the first convolutional layer to the next convolutional layer whereas the depth of the image is increased. The feature selection increases with increase in depth size.

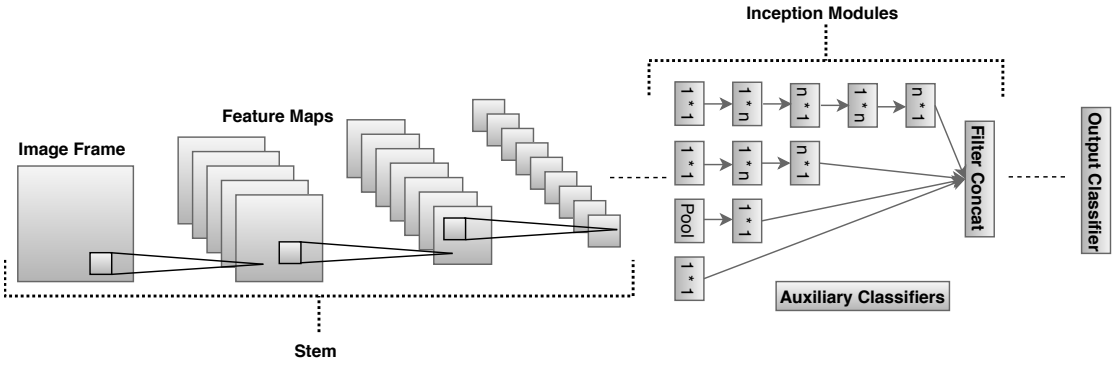


Fig. 3.24. Proposed CNN based architecture for affective state analysis of students.

Transfer learning helps in optimizing and fine-tuning the proposed architecture. Inception-v3 based FaceNet architecture recognizes the human faces in a single image frame with better accuracy than state-of-the-art methods (Schroff et al., 2015; Szegedy et al., 2016). The FaceNet is trained and tested on LFW and YouTube face database (Huang et al., 2007; Wolf et al., 2011). But, this architecture is not trained for gesture or posture recognition. In the proposed architecture, the convolution, activation and pooling operations are performed till a dimension less than 256*256 (3 extra layers are added to get to the dimension 256*256) is obtained, then the inception-v3 based FaceNet architecture is added which factorizes into smaller convolutions and uses auxiliary classifiers for more stable learning and better convergence.

Hence, the proposed architecture is divided into four major parts. The first part is stem, which contains the convolutional and pooling layers till the factorization of convolutional layers. The second part contains the factorization into smaller convolutions using filter concats. The third part contains the auxiliary classifiers and the last part contains fully connected layers and softmax classifier. The entire proposed architecture consists of 51 layers (3 initial layers and 48 Inception layers) with stride 2 and 6 pooling layers which are used for an independent evaluation over each activation maps and for

size reduction. The last pooling layer is connected to two fully connected layers. Those fully connected layers are connected to the softmax layer.

Training phase starts with the back-propagation using the recursive chain rule for computing the gradients for all inputs, parameters, and intermediates. The entire process of backpropagation is stored in a graph structure. The forward pass computes the results of an operation and stores the gradient value whereas the backward pass computes the gradient of loss function w.r.t. the inputs.

Equation 3.14 shows the loss function performed using multinomial logistic regression where y_i is the ith image label and x_i is ith image frame, using these, the loss for the ith image L_i is calculated. The overall mean of training loss of the entire training data L_t is defined as the total data loss given in Equation 3.15.

$$L_i = -log(\frac{e^{sy_i}}{\sum_j e^{s_j}}) \tag{3.14}$$

$$L_t = \frac{1}{N}\sum_{i=1}^{N} L_i \tag{3.15}$$

Since the data used for analyzing multiple class in a single image frame (dense labeling), the activation function is used which neither saturates for both positive and negative values nor dies for negative values. Hence Leaky Relu is used as an activation function (Equation 3.16) where x is an is an input column vector containing all pixel data of the image.

$$f\,(x) = max(\alpha x, x)\ where\ \alpha = 0.01 \tag{3.16}$$

Xavier initialization (Equation 3.17 where n_in & n_out are input neurons from current & next layer and r is to calculate zero mean) is used for weight initialization. Batch normalization is also used along with the hyperparameter optimization. Mini-batch stochastic gradient descent is used in a loop structure where it performs 4 major steps; Step 1: data is divided into N samples. Step 2: forward propagation to get the loss. Step 3: calculate the gradients using back propagation. Step 4: update the parameters using the calculated gradients.

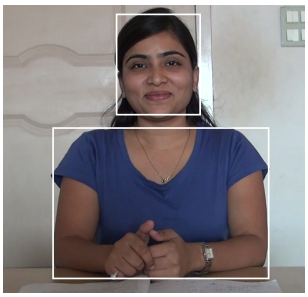
$$W = random.r(n\_in, n\_out)/sqrt(n\_in) \tag{3.17}$$

Parameters update is performed using RMSProp Tieleman & Hinton (2012). Learning rate (Equation 3.18 where α_0, and k are hyper parameters and t is the iteration number) is a hyperparameter for RMSProp. Regularization is performed using Dropout. Monte Carlo approximation is used in dropout where several forward passes with different dropout masks are performed and final prediction is average of all predictions. Numeric gradient is used for gradient check.

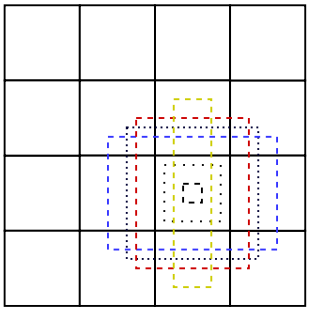
$$\alpha = \alpha_0 e^{-kt} \tag{3.18}$$

Object Localization: In object localization, the bounding box is generated around the students' faces, hand gestures and body postures separately for every student present in the image frame. Most of the images will have only two bounding boxes per student as the hand gestures will be in line with the body postures as shown in Figure 3.25 (a). In few other cases if the hand gestures are separated from the body postures like raising of hands, three bounding boxes will be used as shown in Figure 4.2c (If the background data is more in a bounding box, then the use of only one bounding box per person will lead to miss classifications (Figure 4.2a) (Yao & Fei-Fei, 2012). Using more bounding boxes leads to more computations as the number of default box increases (Liu et al., 2016). Hence, the optimal way to put the bounding box per person is the use of two bounding boxes and in special occasions where the intersection of hand gesture bounding box and the body posture bounding box is null, use three bounding boxes. Even during annotation, the bounding boxes are manually inserted by following the same). The proposed architecture for object localization contains six default boxes (Figure 3.25 (b)) which are mapped to the ground truth boxes. These default boxes are separately convolved for each pixel of every feature map. If the default boxes are above the defined threshold w.r.t. ground truth boxes then these default boxes are considered as positives and the rest as negatives.

Object localization is performed using the proposed CNN based localization model where the fully connected layers and the classification layer are removed and the last convolutional feature map is attached to another set of convolutional filters. This introduces the number of feature maps which allows to share the parameters for each object scales and smoothen the segmentation results.



(a) Sample image frame snapshot with bounding box.



(b) Six different default boxes with 4*4 feature maps.

Fig. 3.25. Sample bounding box image snapshot using default boxes.

The first layer of object localization contains the maximum number of feature maps and these feature maps are generated after performing both the pooling and deconvolution but, the remaining layers use only pooling. Each feature layer produces a fixed set of default boxes which are fed to fast non-max suppression for final detection. Training, matching, and hard negative mining is similar to that of SSD (Liu et al., 2016) where the overall objective loss function is the weighted sum of the localization loss (loc) and the confidence loss (conf) as shown in Equation 3.19.

$$L(x, c, l, g) = \frac{1}{N}(L_{conf}(x, c) + \alpha L_{loc}(x, l, g)) \tag{3.19}$$

$$L_{loc}(x, c) = -\sum_{i \in Pos}^{N} x_{ij}^{p} log(\hat{c}_i^0) - \sum_{i \in Neg} log(\hat{c}_i^0) \tag{3.20}$$

$$where\ \hat{c}_i^p = \frac{e^{c_i^p}}{\sum_p e^{c_i^p}} \tag{3.21}$$

Where, x_{ij}^{p} is a matching index for ith default box to jth ground truth box of category p. N is the number of matched default boxes. If $N = 0$, then set the loss to 0. The localization loss is a smooth $L1$ loss between the predicted box (l) and the ground truth box (g) parameters. We regress to offsets for the center (cx, cy) of the default bounding box (d) and for its width (w) and height (h).

The confidence loss is the softmax loss over multiple classes confidence (c) and the weight term α is set to 1 by cross-validation.

3.3.3 Experimental Setup, Results, Analysis and Discussion

Experimental Setup: Laptops with web-camera and Desktop C310 webcameras are used for data capturing. Two Tesla M40 GPUs with 128 GB RAM and 128 GB scrap space are used for deep learning computations.

Data Selection for Training: 3560 single person and 6650 multi-person in a single frame annotated images are obtained from the labellers. If the minimum and the maximum label given by the labeller differ by more than 1 then those images are discarded. If the labeller has marked the face as unclear then these images are also discarded. After discarding those images, 9423 image frames are finally obtained for the training purpose.

Affective State Classification and Localization Student independent (the students used in training are not used in testing) 10-fold cross-validation is performed on the dataset and the summary of the results is shown in Figure 3.26. From Figure 3.26, it is observed that the accuracy and precision values of a single person in a single image frame are very near for Resnet, Inception-v3 and the proposed architecture. But, in the multi-person in a single image frame scenario, the proposed model outperforms the other models. The improvements in accuracy and precision are observed in the classification of fear, sadness, engaged and neutral class of multi-person in single frame image. Other standard architectures such as VGG-16, VGG-19, ResNet-18, ResNet-34, ResNet-50, ResNet-101, ResNet-152, ResNet-200, BN Inception, Inception V2 are not tested as these architectures fail to perform better for dense labeling scenario such as classroom (Wang & Ji, 2015).

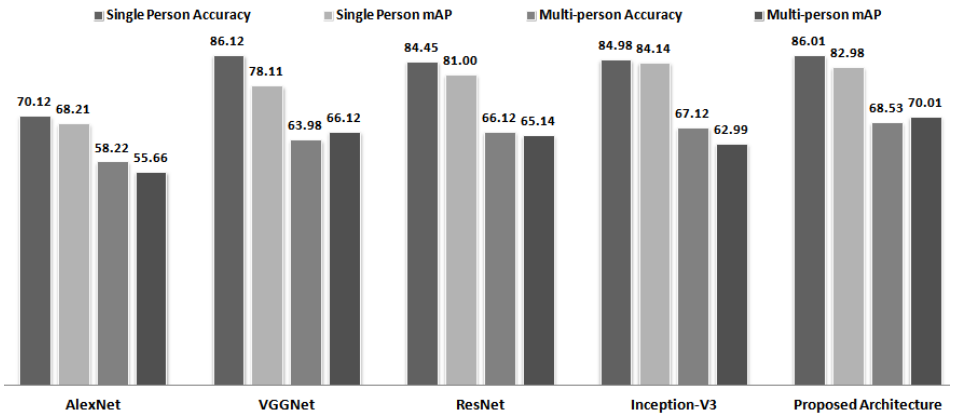


Fig. 3.26. Comparison with various affective state classification architectures.

Table 3.11
Confusion Matrix for Affective States of Single Person in Single Image Frame

In %	En	Ha	Su	De	Ne	Fe	Sa	Di	Bo	Co	Sl	Fr
En	81.00	7.50	1.20	2.10	2.70	0.80	0.70	0.90	0.50	1.10	0.70	0.80
Ha	3.17	83.12	1.11	0.78	7.80	0.68	1.10	0.62	0.64	0.11	0.17	0.70
Su	1.70	0.88	87.23	2.91	0.82	0.74	0.86	1.41	0.87	2.43	0.06	0.09
De	1.10	1.22	5.87	81.44	1.11	3.22	0.89	3.93	0.87	0.22	0.03	0.10
Ne	2.10	3.10	1.08	0.69	89.27	0.77	0.89	0.65	0.30	0.40	0.19	0.56
Fe	0.30	1.87	1.22	6.80	1.98	80.23	1.76	1.55	0.45	0.77	0.97	2.10
Sa	0.50	1.37	0.98	0.56	1.89	0.98	82.88	1.23	8.42	0.76	0.14	0.29
Di	0.70	0.83	1.23	1.45	0.88	1.87	0.98	85.00	0.98	5.75	0.03	0.30
Bo	0.90	0.98	0.92	0.87	1.23	0.89	1.98	1.11	82.44	0.98	2.78	4.92
Co	1.30	2.40	2.02	2.83	0.90	0.67	0.44	1.64	0.58	86.21	0.09	0.92
Sl	0.40	0.11	0.01	0.38	0.11	0.52	0.22	2.11	2.05	0.79	92.43	0.87
Fr	0.20	0.90	0.11	1.13	0.20	0.62	1.01	1.12	2.90	0.91	0.79	90.11

En: Engaged; Ha:Happiness; Su: Surprise; De: Delight; Ne: Neutral; Fe: Fear; Sa: Sadness; Di: Disgust; Bo: Boredom; Co: Confused; Sl: Sleepy; Fr: Fear

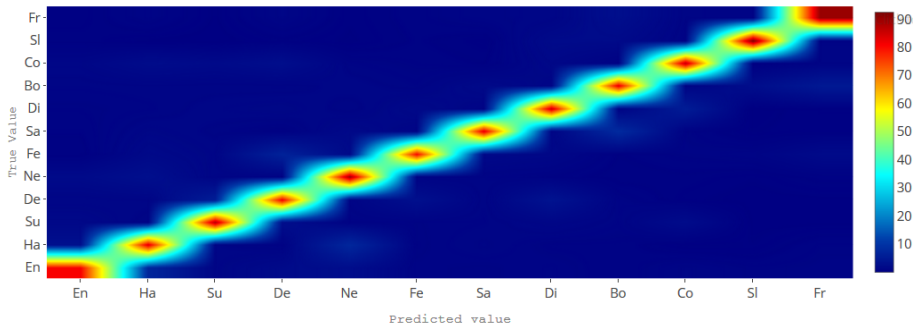


Fig. 3.27. Heatmap of confusion matrix for single person in single image frame.

Proposed architecture is used to analyze the affective states of students separately for both single person (e-learning and flipped classroom) and multi-person (classroom and webinar) affective state classification. Tables 3.11 and 3.12 show the confusion matrix for different students' affective states for single and multi-person affective state classification, respectively. Figures 3.27 and 3.28 show the corresponding heatmaps generated from single and multi-person affective state classfication, respectively. It is observed from Table 3.11 that the affective state recognition is more accurate for surprise and less accurate for fear and sad affective states. Reason for the high accuracy for surprise affective state is that it has unique facial features which are recognized easily when compared to happiness and delight affective states where both have almost similar features of facial expression, hand gesture and body posture. Similarly, the sleepy affective state has high accuracy due to the unique features of facial expression, hand gesture and body posture. Whereas, sad affective state has similar features of

boredom, and hence this affective state has less accuracy. Engaged affective state gets misclassified with happy and neutral. Frustrated & boredom, surprise & delight are misclassified with each other. As already mentioned, CNN requires sufficient training data for better performance; further, delight, fear and sad affective states are less likely to appear in classroom scenario and hence less accuracy and recall is observed; only posed expressions contain delight, fear and sad affective states but students seldom express these affective states in the classroom environment and hence the accuracy is also less as shown in Table 3.12.

Table 3.12
Confusion Matrix for Affective States of Multi-Person in Single Image Frame

In %	En	Ha	Su	De	Ne	Fe	Sa	Di	Bo	Co	Sl	Fr
En	66.00	7.70	4.10	1.83	5.69	1.89	1.38	1.10	1.90	2.33	3.10	2.98
Ha	3.58	67.78	1.23	0.88	6.50	1.83	1.55	2.15	1.58	10.22	1.98	0.72
Su	4.11	2.99	72.11	6.33	1.23	1.33	0.89	2.49	2.33	3.33	1.33	1.53
De	4.23	7.32	9.54	62.53	1.82	0.79	0.88	2.27	2.40	3.22	1.89	3.11
Ne	3.26	10.34	1.45	1.77	69.22	1.22	3.11	2.13	1.92	0.98	3.10	1.50
Fe	2.00	4.26	2.20	5.40	5.22	60.01	10.45	2.82	1.89	1.77	2.10	1.88
Sa	3.82	3.22	2.87	3.10	4.10	13.42	59.23	2.23	1.44	2.12	2.11	2.34
Di	4.89	1.11	3.45	2.89	1.89	1.22	0.11	67.88	5.22	3.74	2.10	5.50
Bo	2.89	2.64	0.99	1.88	2.20	1.77	1.11	2.10	69.11	4.21	4.99	6.11
Co	3.12	8.11	5.23	2.33	4.24	1.10	1.40	2.10	2.40	65.98	1.88	2.11
Sl	3.33	2.70	1.80	2.80	3.10	0.78	0.98	1.66	3.55	2.20	74.00	3.10
Fr	4.10	2.55	0.98	2.12	2.89	1.22	2.89	3.11	4.88	2.87	1.17	71.22

En: Engaged; Ha:Happiness; Su: Surprise; De: Delight; Ne: Neutral; Fe: Fear; Sa: Sadness; Di: Disgust; Bo: Boredom; Co: Confused; Sl: Sleepy; Fr: Fear

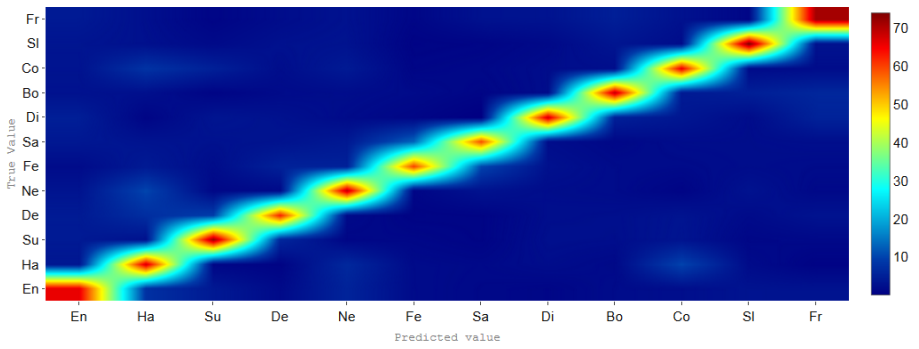


Fig. 3.28. Heatmap of confusion matrix for multi-person in single image frame.

Comparison among state-of-the-art architectures such as Alexnet, Resnet, VGGNet, and Google's Inception-v3 model for affective classification Krizhevsky et al. (2012); He et al. (2016); Simonyan & Zisserman (2014); Szegedy et al. (2016) w.r.t. mean average precision (mAP) is shown in Table 3.13 and Figure 3.29.

Table 3.13
Comparison of Different State-of-the-Art Architectures for Affective State Classification

Ar	mAP												
	OA	En	Ha	Su	De	Ne	Fe	Sa	Di	Bo	Co	Sl	Fr
Al	61.92	63.12	66.50	64.10	59.00	48.22	52.12	62.22	66.22	68.12	61.12	69.20	63.12
Re	73.03	75.50	76.12	72.30	71.11	72.13	71.22	71.34	71.11	73.32	74.17	73.90	74.12
VG	72.12	70.11	73.24	79.00	76.20	67.22	69.22	66.12	71.21	72.28	73.54	74.12	73.15
In	73.57	75.10	71.41	73.44	69.80	77.22	72.12	71.11	72.21	69.21	75.00	79.98	76.21
PA	76.10	73.50	75.45	79.67	71.98	79.24	70.12	71.05	76.44	75.77	76.09	83.21	80.66

Ar: Architecture; Al: Alex Net; Re: ResNet; VG: VGGNet; In: Inception V3; PA: Proposed Architecture; En: Engaged; Ha:Happiness; Su: Surprise; De: Delight; Ne: Neutral; Fe: Fear; Sa: Sadness; Di: Disgust; Bo: Boredom; Co: Confused; Sl: Sleepy; OA: Overall

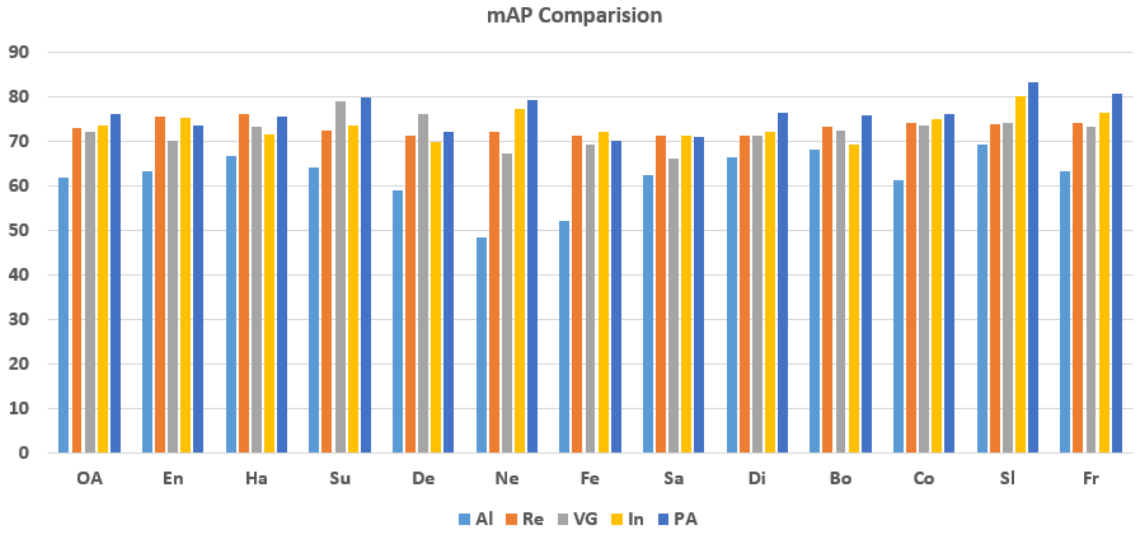


Fig. 3.29. Comparison with Different State-of-the-Art Architectures w.r.t. mAP.

Similarly, the object localization results of the proposed architecture and its comparison with other state-of-the-art techniques such as YOLO, YOLO-V2, SSD, SSD-300, SSD-500 Redmon et al. (2016); Liu et al. (2016); Ren et al. (2015); Girshick (2015); Girshick et al. (2014) are summarized in Figure 3.30. Architectures like YOLO has high fps but are less accurate. The proposed architecture outperforms other architectures as it accurately recognizes the students' faces even if they are sitting on the last bench or if their faces are partially occluded. The mAP of the face, hand gesture & body posture using different object localization architectures is shown in Table 3.14.

Table 3.14
Comparison of Different Object Localization Architectures

Object Localization Architectures	mAP			
	Overall	Face	Hand Gesture	Body Posture
YOLO	62.48	59.12	64.56	63.76
YOLOv2 544	70.76	67.45	73.71	71.12
SSD300	66.33	63.10	69.66	66.22
SSD512	69.96	67.30	71.22	71.35
Proposed Architecture 512	71.40	66.34	73.63	74.22
Proposed Architecture	74.49	71.20	77.11	75.17

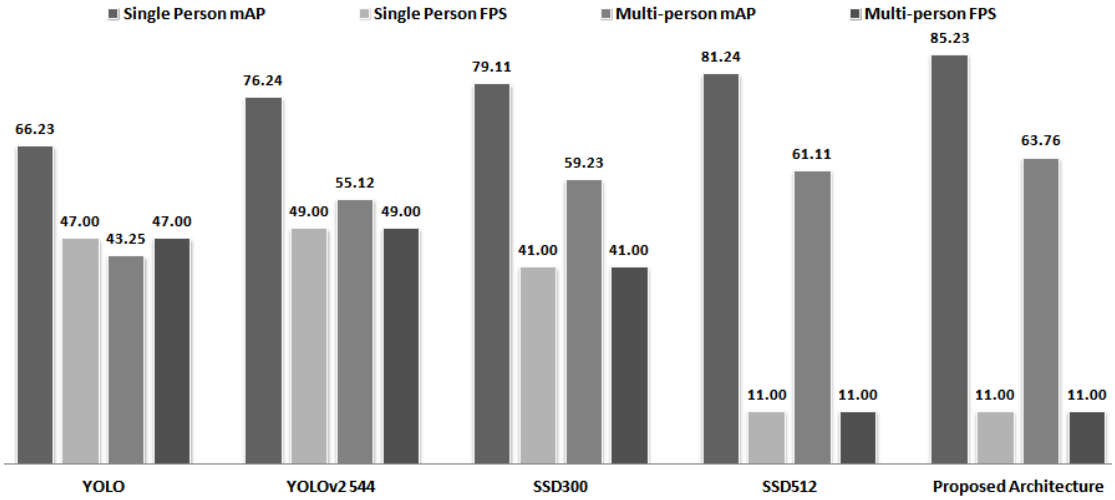


Fig. 3.30. Comparison of different architectures for object localization.

Overall Results, Analysis and Discussion: Table 3.15 projects the overall results of the proposed classification and localization architectures using the standard performance evaluation metrics. These are the aggregate results of the four different learning environments considered in this study.

Table 3.15
Overall Results of Detection and Classification

Performance Metrics	Detection	Classification
Average Accuracy	0.79	0.77
Average Recall	0.72	0.66
Average Precision	0.74	0.76
Average F-score	0.73	0.71
MCC	0.69	0.61
AUC	0.72	0.71

Statistical Analysis

We performed the following statistical analysis on the data obtained from the deep learning module.

The incidence of affective states recognized during the learning process: Repeated measure ANOVA test is conducted on the recognized affective states of students. Further, as mentioned in D'Mello et al. (2010), Bonferroni post-hoc test is conducted to analyze the pattern present in the recognized affective states, and tripartite classification of affective states are performed using one-sample t-test comparison, where chance (Equation 3.22 where, M is the mean of proportions and N is number of students that experienced the state at least once as shown in Table 3.16) is used to isolate the subset of affective states.

$$Chance = \frac{(1 - M_{neutral})}{N_{affect}} \tag{3.22}$$

Temporal dynamics of the affective states: Persistence, random and ephemeral are the three-way classification schemes in the temporal dynamics of affective states (D'Mello et al., 2010). Persistence is a property of the affective state observed at time t which is also observed at time $t + 1$. Random is the probability of affective state observed at time t but not related to the occurrence of affective state at $t + 1$. Ephemeral is the affective state at t that decreases its likelihood at $t + 1$. There exist no instances of ephemeral states in this study; hence the temporal data is classified into two-way classification using the likelihood and one-sample t-test analysis. The likelihood of affective states are calculated using the Equation 3.23, where the current affective state at time t is A_C and the next state at time $t + 1$ is A_N. If the current affective state and the next affective state are the same, then the persistence likelihood (L) of affective states can be calculated using Equation 3.24.

$$L(A_C \rightarrow A_N) = \frac{P(A_C|A_N) - P(A_N)}{1 - P(A_N)} \tag{3.23}$$

$$L(A_{C_t} \rightarrow A_{C_{t+1}}) = \frac{P(A_{C_{t+1}}|A_{C_t}) - P(A_{C_{t+1}})}{1 - P(A_{C_{t+1}})} \tag{3.24}$$

E-Learning Environment

We implemented the proposed methodology in every class ranging from 45 minutes to 1.5 hours in a classroom scenario whereas the e-learning duration ranges from 20 minutes to 1 hour. Initially, results and analysis of one class for a duration of 30 minutes is projected; subsequently the overall results are projected. Table 3.16 shows the distribution of affective states for all 30 students with $chance = (1 - 0.342)/11 = 0.059$ for the 30 minute e-learning class duration. Engaged affective state is significantly observed in the students. The number of students with sleepy and boredom affective states is also significant as they are observed under routine condition. Sporadic emotions includes confusion and frustration. Sadness, fear, surprise and disgust are observed under exceptional categories of emotions.

Table 3.16
Distribution of Affective States for 30 Minutes Learning Duration in E-Learning Environment

Affective States	Frequencies		Proportions		One-sample t-test		
	N	P	M	SD	$t(29)$	p	d
Routine							
Engaged	28	0.933	0.144	0.148	4.22	<0.010	0.66
Happiness	22	0.733	0.082	0.066	3.11	<0.001	0.43
Boredom	25	0.833	0.11	0.098	3.28	<0.001	0.47
Sleepy	27	0.900	0.123	0.127	4.09	<0.010	0.62
Sporadic							
Confused	18	0.600	0.047	0.049	-0.198	0.117	-0.29
Frustrated	16	0.533	0.041	0.048	-0.172	0.173	-0.27
Delight	17	0.567	0.051	0.060	0.311	0.298	0.05
Exceptional							
Sadness	4	0.133	0.02	0.031	-14.1	<0.001	-1.58
Fear	2	0.067	0.019	0.017	-17.3	<0.001	-2.8
Surprise	3	0.100	0.03	0.036	-6.3	<0.001	-0.79
Disgust	4	0.133	0.039	0.033	-5.8	<0.010	-0.89
Neutral	30	1.000	0.342	0.211			

N = number of students that experienced the state at least once
P = proportion of students that experienced the state at least once

Similarly, this experiment is conducted for all the students of e-learning courses taught during 2016 and 2017 academic years. After performing one-sample t-test with $chance = 0.61$, engaged, happiness, boredom, sleepy are observed as routine emotions as they occurred with a probability greater than 0.61. On an average these routine emotions comprised 60% of the observations and are experienced by 88% of the students. Confused, frustrated and delight are the three sporadic emotions comprising 17% of the

observations and occurred in 69% of students. Finally, the remaining four exceptional emotions comprising 23% of the observations and are observed in 39% of students.

Flipped Classroom Environment

In the flipped classroom, students learn using short e-learning modules, so there is not much difference when compared to e-learning. But from the observation, the boredom and sleepy routine emotions are seen more when compared to a regular e-learning environment. This is because of the higher course completion rate of the flipped classroom and the students have to complete the learning process before the start of the class (this is one of reasons we found from the students' oral survey). In contrast, the students can relax a bit in e-learning as their completion rate need not be so regular. Also, ANOVA and Bonferroni Post-hoc tests are performed and obtained the same routine, sporadic and exceptional affective state classification as of e-learning.

Classroom Environment

For classroom environment, the similar statistical analysis performed for e-learning is followed. A repeated measure ANOVA test indicated that there are significant differences in the proportion of affective states experienced by the students, [$F(11, 34840)$ = 131.33, $MSe = 0.008$, $p < 0.001$, $n2 = 0.499$]. Bonferroni post-hoc test and one sample t-test with $chance = 0.617$ revealed that the affective states engaged, boredom, sleepy and confused are observed as routine emotions. Frustrated and delight emotions are sporadic. Happiness is a routine emotion in an e-learning environment, whereas in a classroom environment it moved to exceptional, along with fear, sadness, surprise and disgust emotions.

Webinar Environment

The results obtained from the webinar data is similar to that of the classroom environment. Although the routine emotions are same as that of the classroom environment, for a few sessions the boredom affective state is observed with a P value of 0.998 for one of the webinar sessions. This may be true because all the faces are not visible to the speaker in a webinar scenario and the students can freely express their emotional states.

Temporal Dynamics of Affective States

The temporal dynamics of the affective states are observed for the entire data. We used a normalized base rate of 0 with a one-sample t-test. The observations are shown in Table 3.17. The persistence of engaged and boredom is high, and the primary emotions are random.

Table 3.17
Persistence of Affective States

Affective State	Descriptives (Likelihood)		One-sample t-test			
	M	SD	t	df	p	d
Persistence						
Engaged → Engaged	0.112	0.222	2.19	25	0.088	0.51
Frustration → Frustration	0.086	0.216	2.86	23	0.039	0.43
Confused → Confused	0.098	0.196	3.12	26	0.051	0.39
Boredom → Boredom	0.122	0.235	3.25	31	0.021	0.59
Sleepy → Sleepy	0.142	0.279	2.99	35	0.009	0.55
Random						
Happiness → Happiness	0.036	0.122	0.72	14	0.811	0.48
Surprise → Surprise	-0.005	0.098	-0.25	19	0.379	-0.09
Sadness → Sadness	0.002	0.153	1.8	32	0.584	0.14
Fear → Fear	-0.006	0.082	-0.39	13	0.662	-0.11
Disgust → Disgust	0.124	0.189	0.68	15	0.901	0.22
Delight → Delight	0.119	0.213	0.49	19	0.782	0.01

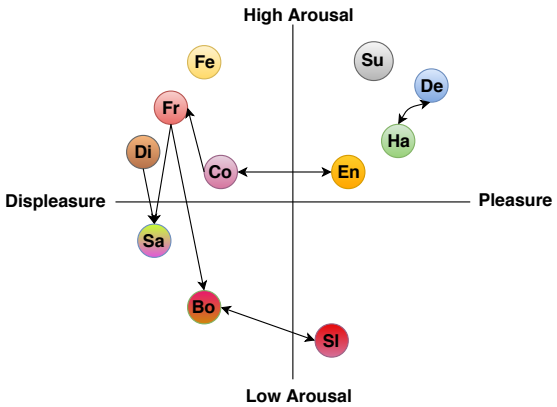


Fig. 3.31. Students' affective state transitions

We also performed the descriptive statistics for the likelihood of these recognized affective states. ANOVA test results demonstrated that there is a significant difference among all the affective states, [$F(10, 400) = 5.5$, $MSe = 0.05$ and $p < 0.5$]. Tukey HSD post-hoc test (D'Mello, 2012) is performed on the data, and it is summarized in Figure 3.31. The transitions from engaged to confusion are dominant. Happiness to

delight and delight to happiness transitions are observed. There are many instances where confusion lead to frustration, and the students get bored. Boredom to sleepy and vice versa is also observed. There are a few instances where transitions from frustrated to sadness and disgust to sadness are observed.

Initially, all learning-centered emotions are considered such as eureka, anxiety, curiosity and anger, but these emotions are observed rarely [$P < 0.01$ and $M < 0.001$ (Table 3.16)]. Hence, only dominant learning-centered emotions are considered for the analysis in this chapter.

3.4 Summary

The proposed method detects facial emotion recognition for multiple faces in a single frame. GPU along with the CPU helps in speeding up this process. The proposed algorithms based on Modified Affine Transformation and Viola Jones based Haar Cascades accurately detects the face in a given frame for real time face detection and tracking. Experimental results demonstrate that our proposed algorithm outperforms the existing Voila Jones algorithm by 6% for YALE, FDDB and 'top 25 Google's searched "tilted face"' datasets. These datasets consist of frontal, occluded and tilted faces with different illuminations. And also we tested on real time face detection and tracking using the web cam which gave better results with very good accuracy. Video affective content analysis is performed using both audio and visual features using SVM and RBM classifiers. From experimental results, it is observed that our proposed hybrid SVM-RBM classifier performs better than individual SVM and RBM classifiers for audio-visual emotion recognition with annotated data.

The current study explored the students' affective states in the classroom environment. Both emotional and behavioral engagements are considered to predict the students' affective states such as engaged and boredom along with the neutral. The multimodal analysis is performed using the students' facial expression, hand gesture, and body posture to increase the robustness of the method. Since the classroom image frame data contains multiple students in every image, a group engagement score is predicted for image frame data using the feature fusion technique. We proposed a deep learning-based hybrid CNN model to predict the students' affective states. A deep learning model should get trained well, and no standard datasets are available for analyzing stu-

dents affective states in the classroom environment. Hence, a dataset is created with two types of image input using the student's facial expression, hand gestures, and body postures. Dataset-1 consists of a single student in a single image frame, and dataset-2 consists of multiple students in a single image frame. For dataset-2, students' spontaneous expressions and their behavior in the classroom environment are also collected. Manual annotation is carried out by three annotators for annotating three different affective states, namely: engaged, boredom, and neutral. We obtained the reliability among annotators (Cohen's $\kappa = 0.59$) for spontaneous classroom data. We proposed both CNN-1 and CNN-2 models, for the affective state recognition of single and multiple students in a single image frame. We obtained an accuracy of 94% and 95.6% for the proposed CNN-1 and CNN-2 models, respectively, for posed classroom data. Further, 70% accuracy is obtained using the student & classroom independent 10-fold cross-validation for the proposed hybrid model, which is a combination of CNN-1 and CNN-2 for spontaneous classroom data. The proposed models outperformed the existing state-of-the-art techniques on both posed and spontaneous datasets.

The students' affective state classification with localization, using facial expressions, hand gestures and body postures for both single and multiple students in a single image frame is proposed. The proposed architecture is tested in e-learning, flipped classroom, classroom and webinar environments for 12 different class labels. After performing student-independent 10-fold cross-validation, we obtained an accuracy of 77% for the students' affective state classification and 79% for object localization. We performed statistical analysis and observed that the students' affective states such as engaged, boredom, sleepy and confused are categorized under routine emotions. Further, we observed some dominant transitions between the affective states such as engaged & confusion, confusion & frustration and frustration & boredom.

Similar to most of the research, the proposed methodology also has a few limitations. The current study focuses only on most frequently observed students' affective states but, a few less frequently observed learning-centered emotions like Eureka, and contempt are not considered to analyze its incidence and temporal dynamics during the teaching-learning process. The majority of students present in the created dataset are Indians. Hence, the working of the proposed model may differ when we test on other than Indian students. The proposed multimodal analysis is performed on sponta-

neous data obtained from a regular classroom environment. This will vary if we consider computer-enabled teaching laboratories or game-based learning classroom environments. This study considers only the emotional engagement as the engagement detection is performed by using the image-based affective content analysis. This limits us to analyze the cognitive engagement of students, for example, the student can be with the non-engaged affective states, but he(she) may be engaged. But, these aspects are not considered in this study.

In the next chapter, unobtrusive students' behavioral engagement is discussed where scale-invariant context-assisted single-shot CNN is proposed to classify the students' behavioral engagement into four different engagement levels.

Chapter 4

Behavioral Engagement Analysis

Pervasive intelligent learning environments can be made more personalized by adapting the teaching strategies according to the students' emotional and behavioral engagements. The students' engagement analysis helps to foster those emotions and behavioral patterns that are beneficial to learning, thus improving the effectiveness of the teaching-learning process. Unobtrusive students' engagement analysis is performed using the students' non-verbal cues such as facial expressions, hand gestures, and body postures. Though there exist several techniques for classifying the engagement of a single student present in a single image frame, there are limited works on the students' engagement analysis in a classroom environment. In the previous chapter, we discussed the unobtrusive emotional engagement analysis of the students, but not the students' behavioral engagement analysis. Students' behavioral engagement using their non-verbal cues is not explained much in both classrooms and computer-enabled teaching laboratories. Hence, in this chapter, we propose the CNN based architectures for unobtrusive students' engagement analysis using non-verbal cues in both classrooms and computer-enabled teaching laboratories.

The key contributions of this chapter are as follows:

- Proposed a novel scale-invariant context-assisted single-shot CNN architecture for the students' behavioral engagement analysis of multiple students in a single image frame in the classroom environment using their facial expressions, hand gestures, and body postures.
- Proposed an architecture for video surveillance camera-based students' behavioral engagement analysis in computer science and information technology teaching laboratories.
- Compared the students' engagement level score with their test performance for any possible correlations.

Since the features and classification patterns of students present in the classroom vary with the computer-enabled teaching laboratories, two different architectures are proposed for the behavioral engagement analysis. The first part of this chapter deals with the students' behavioral engagement in the classroom environment and the latter part discusses the computer-enabled teaching laboratories.

4.1 Proposed Methodology for Classroom Environments

Figure 4.1 shows the complete flow of the proposed methodology for the students' behavioral engagement analysis in the classroom environment, which includes the created dataset and the proposed engagement level classification method. The details are discussed in the following subsections.

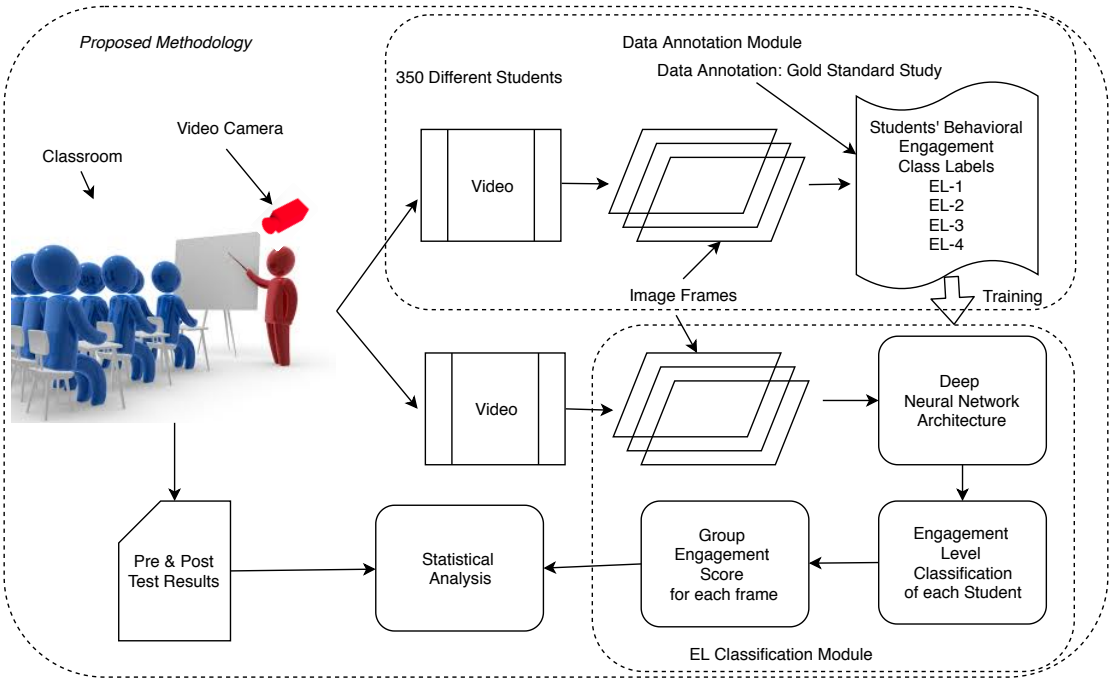


Fig. 4.1. The complete flow of the proposed methodology for students' behavioral engagement analysis.

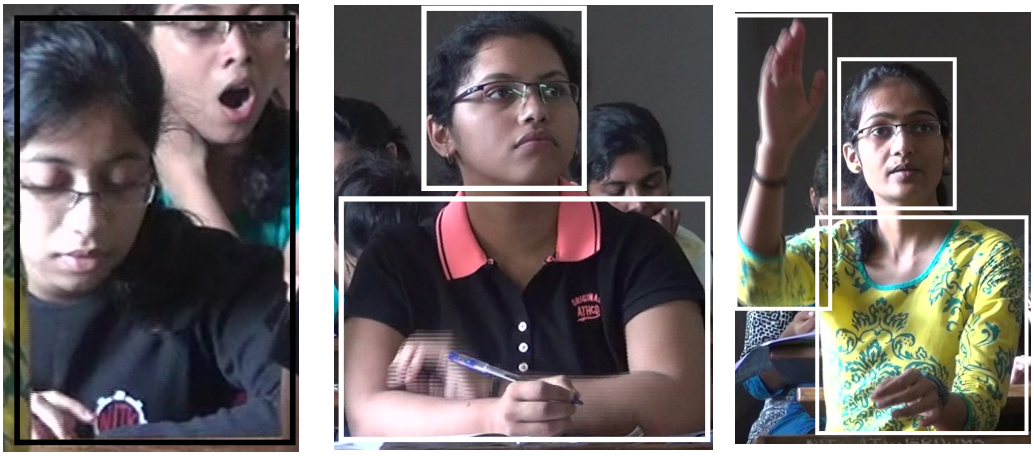
4.1.1 Students' Engagement Classification

Predicting the students' engagement is a very difficult task as there are challenges with both its conceptualization and measurement. Behavioral, emotional, cognitive, and agentic engagements are the four different types of engagement (Sinatra et al., 2015). Most popular works on student's engagement involve behavioral and emotional (learning-centered emotions) engagements with some cognitive aspects involved in it (Sidney et al., 2005; Whitehill et al., 2014). But these works contain a single person in a single image frame. Further, there exists no robust method which suits for a large classroom environment where all students are not clearly visible even after using high definition cameras. So, analyzing emotions using only the facial expressions in such a scenario is difficult. Hence, this study explored behavioral engagement (face, hand gestures, and body postures) involving some cognitive aspects.

The students' behavioral engagement is classified into four major engagement levels (ELs), as given in Whitehill et al. (2014). The guidelines designed for engagement levels classification shown in Whitehill et al. (2014) are modified by adding the features of the facial expression, hand gesture, and body posture for multiple students in a single image frame, but the ELs definitions remain the same.

- EL 1: Not engaged at all - e.g., looking away from the tutor or board and obviously not thinking about the task, eyes completely closed, etc.
- EL 2: Nominally engaged - e.g., eyes barely open, fully bent on the desk or the chair, no expression on the face, boredom, clearly not "into" the task.
- EL 3: Engaged in the task - a student requires no admonition to "stay on the task". Looking at the teacher/board, taking notes, listening, and discussions with the teacher, etc.
- EL 4: Very engaged - a student could be "commended" for his/her level of engagement in the task.
- X: The clip/frame is very unclear, or contains no person at all.

4.1.2 Participants and Engagement Level Annotation



(a) Single BBox for one student. (b) Multiple BBox for one student. (c) Multiple BBox for one student.

Fig. 4.2. Sample annotation of bounding boxes.

Subjects

The entire proposed architecture is trained and tested on 350 graduate and undergraduate students of National Institute of Technology Karnataka (NITK), Surathkal, Mangalore, India. These spontaneous expressions and body postures of students are collected for more than 10 hours from the classroom environment. All the classroom data has

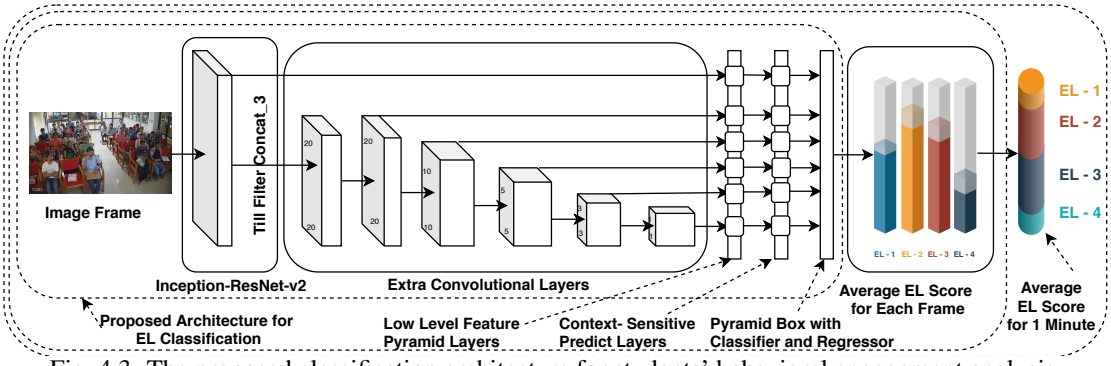


Fig. 4.3. The proposed classification architecture for students' behavioral engagement analysis.

multiple students in a single frame, but the number of people in each frame may vary depending on the subjects of discussion and the class strength.

The students present in this created database belong to the age group of 20 to 26 years. These students are undergraduate, postgraduate, and doctoral research students from India with different cultural and regional backgrounds.

Multiple labelers are used for the annotation process and more details on the camera setup, annotation process and the various image variants used in this study are given in Subsection 6.2.1 of Chapter 6. The reliability among the labelers are analyzed and found that they agree with Cohen's $\kappa = 0.43$.

Annotation of Bounding Box (BBox): The labelers also put the bounding boxes for each student present in the image. One bounding box for a student's face, hand gesture, and body posture will lead to more misclassifications as the background content will also have additional information (the deep learning method considers this as features) which are not required for the current EL class (Yao & Fei-Fei, 2012) (Figure 4.2a, where the bounding box contains another student's sleepy face, which alters the actual features of EL 3). To perform the optimal bounding box computations, one bounding box for the face, and one bounding box for both the hand gesture and the body posture (if both the hand gesture and body posture bounding boxes have an intersection of more than 70%) is used as shown in Figure 4.2b. Otherwise, each student will have three different bounding boxes (Figure 4.2c).

The annotated image with the class label and object localization is stored in the JSON file. Each recognized student will have an engagement level class label and corresponding bounding box coordinates (three sets of coordinates w.r.t. face, hand gesture, and body posture). If any of the coordinates are not recognized/required, then those are filled with null values. A few students are sitting in the last benches where only their faces are visible, in those cases, only face bounding box coordinates are stored, and remaining are filled with null values. To classify a given student in any engagement class, the face is must (even if the students are far from the camera and expressions are not clear, these are also considered). We did not classify the image into any class if only the hand gesture and the body posture of the student is present in that image frame.

4.1.3 Proposed Scale-Invariant Context-Assisted Single-Shot CNN

The proposed architecture for the students' engagement analysis in the classroom environment is based on the anchor-based detection framework (Zhang et al., 2017; Tang et al., 2018). Though the existing state-of-the-art techniques like SSD (Single-Shot MultiBox Detector) (Liu et al., 2016) provide better performance for object classification and localization, the performance of SSD drops for the smaller faces. Hence, to make the proposed architecture more robust for the classroom environment, low-level feature pyramid layers, context-sensitive predict layers and pyramid boxes are conglomerated with the anchor-based framework for both the students' bounding box detection and EL classification as shown in Figure 4.3.

The proposed framework is a scale-equitable anchor based framework. It consists of Inception-ResNet-V2 architecture till filter concat_3 (Szegedy et al., 2017) as the base convolutional layer. Then extra convolutional layers which decrease in size are progressively added, resulting in the multiscale feature maps. All high-level features are not helpful for detecting small, blurred, and occluded faces. Hence, the Low-Level Feature Pyramid Layers (LFPL) are used. It starts with a top-down structure from a middle layer with their receptive field close to half of the input size. The structure of each layer is the same as that of Tang et al. (2018), and L2 normalization is used to rescale these layers.

LFPLs are followed by context-sensitive predict module (CPM) (Tang et al., 2018). CPM uses pyramid anchors (PA) which contain contextual information regarding the

face, hand gesture, and body posture. The target students' face, hand gesture or body posture is localized at r_t $(r = region, t = target)$ at original image, the k^{th} pyramid anchor is defined as shown in Equation 4.1.

$$l_k(a_{i,j}) = \begin{cases} 1, if\ IOU\ (a_{i,j} \cdot s_i / {s_{pa}}^k, r_t) > t \\ 0,\ otherwise \end{cases} \tag{4.1}$$

Where, $a_{i,j}$ means the j^{th} anchor at the i^{th} feature map with stride s_i. for k = 0, 1, .. K, respectively, where s_{pa} is the stride of pyramid anchors. $a_{i,j} \cdot s_i$ denotes the corresponding region in the original image of $a_{i,j}$, and $a_{i,j} \cdot s_i / {s_{pa}}^k$ represents the corresponding down-sampled region by stride ${s_{pa}}^k$. The other anchor-based detector values are exactly the same for the threshold t. The hyperparameter is set as s_{pa} = 2 since the stride of the adjacent prediction modules is 2. Furthermore, the threshold is set to 0.35 and K to 2. Then l_0, l_1, and l_2 are labels of the face, hand gesture, and body posture, respectively. Here, it has three targets, namely: the face, the hand gesture, and the body posture associated with the face (occluded, background clutter, and other similar cases) in three continuous predictions.

Pyramid anchors perform both the classification and regression simultaneously. The loss function used here is PyramidBox Loss, as shown in Equation 4.2.

$$L(\{p_{k,i}\}, \{t_{k,1}\}) = \sum_k \lambda_k L_k(\{p_{k,i}\}, \{t_{k,i}\}) \tag{4.2}$$

where, the k^{th} pyramid-anchor loss is given by Equation 4.3.

$$\begin{aligned} L_k(\{p_{k,i}\}, \{t_{k,1}\}) = \frac{\lambda}{N_{k,cls}} \sum_{i_k} L_{k,cls}(p_{k,i}, p^*_{k,i}) + \\ \frac{1}{N_{k,reg}} \sum_{i_k} p^*_{k,i} L_{k,reg}(t_{k,i}, t^*_{k,i}) \end{aligned} \tag{4.3}$$

Here, k is the index of pyramid-anchors (k = 0, 1, and 2 represents for face, hand gesture, and body posture, respectively), and i is the index of an anchor and $p_{k,i}$ is the predicted probability of anchor i being the k^{th} object (face, hand gesture or body

posture). The ground-truth label is defined by Equation 4.4.

$$p^*_{k,i} = \begin{cases} 1, if\ the\ anchor\ down\_sampled\ by\ stride\ s^*_{pa} \\ is\ positive \\ 0,\ otherwise \end{cases} \tag{4.4}$$

$t_{k,i}$ is a vector representing the four parameterized coordinates of the predicted bounding box, and $t^*_{k,i}$ is that of ground-truth box associated with a positive anchor, defined by Equation 4.5. Most of the images will have only two bounding boxes per student as the hand gestures will be in line with the body postures, as shown in Figure 4.5. In a few other cases, if the hand gestures are separated from the body postures like raining of hands, three bounding boxes will be used, as shown in Figure 4.2c.

$$\begin{aligned} t^*_{k,i} = (t^*_x + \frac{1 - s^k_{pa}}{2} t^*_w s_{w,k} + \Delta_{x,k}, t^*_y + \frac{1 - s^k_{pa}}{2} t^*_h s_{h,k} \\ + \Delta_{y,k}, s^k_{pa} t^*_w s_{w,k} - 2\Delta_{x,k}, s^k_{pa} t^*_h s_{h,k} - 2\Delta_{y,k}) \end{aligned} \tag{4.5}$$

where, $\Delta_{x,k}$ and $\Delta_{y,k}$ denote offset of shifts, $s_{w,k}$ and sh, k are scale factors with respect to (w.r.t.) width and height, respectively. In our experiments, the set values are $\Delta_{x,k} = \Delta_{y,k} = 0$; $s_{w,k} = sh, k = 1$ for $k < 2$ and $\Delta_{x,2} = 0$; $\Delta_{y,k} = t^*_h$; $s_{w,2} = 7/8$; $s_{h,2} = 1$ for $k = 2$. The classification loss $L_{k,cls}$ is softmax loss, and the regression loss $L_{k,reg}$ is the smooth L1 loss [13]. The regression loss is activated only for positive anchors and disabled for others as indicated by the term $p^*_{k,i} L_{k,reg}$. The balancing weights λ, and λ_k for $k = 0, 1, 2$ and the two terms are normalized using $N_{k,cls}$, $N_{k,reg}$.

Group Engagement Level Classification

The classroom image frame data contains multiple students with different engagement levels in a single image frame. Hence, feature fusion is used to calculate the same. The multimodal feature fusion vector V_f for any pixel p_i and normalized prediction vector N_{P_i} use normalized predicted probability distribution $N_{P_i,a}$ of class a using the softmax function (Equation 4.6).

$$N_{P_i,a} = \frac{e^{W_a^T V_f}}{\sum_{i \in classes} e^{W_i^T V_f}} \tag{4.6}$$

Where, W is the temporary weight matrix used to learn the features. The training generally converges in $T = 4000$ epochs. The final collective average engagement level score A_{S_i} is given by Equation 4.7.

$$A_{S_i} = arg\,max\, N_{P_i,a}\ where\ a \in classes \tag{4.7}$$

After obtaining the students' engagement level classification results for each frame, the average engagement level value for each minute is calculated, and the variation in the students' engagement level for every minute is stored.

4.1.4 Experimental Setup, Results, Analysis and Discussion

Experimental Setup

For the current study, 8^{th} Generation Intel® Core™ $i7 - 4510U$ Processor, 8GB RAM, and 2GB NVIDIA® GeForce® $840M$ are used for engagement level classification and localization.

Data Selection for Training

4560 multiperson in a single frame annotated images are obtained from the labelers. If the minimum and the maximum labels given by the labelers differ by more than one, then these images are discarded. Even if one labeler has marked that the face/faces are unclear, then these images are also discarded. After discarding these images, finally a total of 4423 image frames are obtained for the training purpose.

Cohen's κ is computed to compare the accuracy of deep learning classification technique with human annotations (Whitehill et al., 2014). We obtained an average κ value, which varies between 0.36 and 0.78 for the classroom environment where multiple students are present in a single image frame.

Data Augmentation

From the previous step, 4423 image frames are obtained, which contains 23%, 33%, 30%, and 14% for EL 1, EL 2, El 3, and EL 4, respectively. These images contain small, burred, occluded students (faces or postures). To make the proposed architecture more robust, data augmentation is used to increase the size of training data. Data anchor sampling (Tang et al., 2018) is used to increase the diversity of face samples by increasing the proportion of small faces to larger ones and vice versa. A few other data augmentation techniques are performed on our dataset and the details are given below. After augmentation, a minimum of 20000 instances of each EL class label are obtained, as shown in Table 4.1.

Table 4.1
EL Class Label Instances Used for Training

Class Label	No of students in each class label
EL 1	21000
EL 2	24000
EL 3	23000
EL 4	20800

- channel_shift_range: Random channel shifts of the image.
- zca_whitening: Applies ZCA whitening to the image.
- rotation_range: Random rotation of the image with a degree range.
- width_shift_range: Random horizontal shifts of the image with a fraction of total width.
- height_shift_range: Random vertical shifts of the image with a fraction of total height.
- shear_range: Shear intensity of the image where the shear angle is in the counter-clockwise direction as radian.
- zoom_range: Random zoom of the image where the lower value is 1-room_range and upper value is 1+zoom_range.
- fill_mode: If any of constant, nearest, reflect or wrap are filled according to the given mode, if any points outside the boundaries of the input.
- horizontal_flip: Randomly flip the inputs horizontally. Table 4.2 shows the details of different data augmentations performed on our dataset.

Table 4.2
Types of Data Augmentation Used

Type of Augmentation	Augmentation Value
channel_shift_range	20
zca_whitening	TRUE
rotation_range	40
width_shift_range	0.2
height_shift_range	0.2
shear_range	0.2
zoom_range	0.2
horizontal_flip	TRUE
fill_mode	Nearest

Performance Evaluation of Proposed Architecture

Figure 4.4 shows the comparison of the proposed architecture for the students' EL classification with other state-of-the-art architectures such as Inception-V3, Hyperface, S^3FD (Single-Shot Scale-invariant Face Detector), MCCN (Multitask Cascaded Convolutional Networks) (Szegedy et al., 2017; Zhang et al., 2017; Ranjan et al., 2019; Zhang et al., 2016). It is observed that the proposed architecture obtained a better accuracy of 71% as it is able to detect most of the scale-variant faces, hand gestures, and body postures. One of the major contributors to this better accuracy is context-sensitive predict layers and intra-image multimodality, where the other methods missed some of the features of face and hand gestures of the students sitting in the last benches. A few other major contributors include anchor-based frameworks, low-level feature pyramid layers, pyramid boxes, and base architecture.

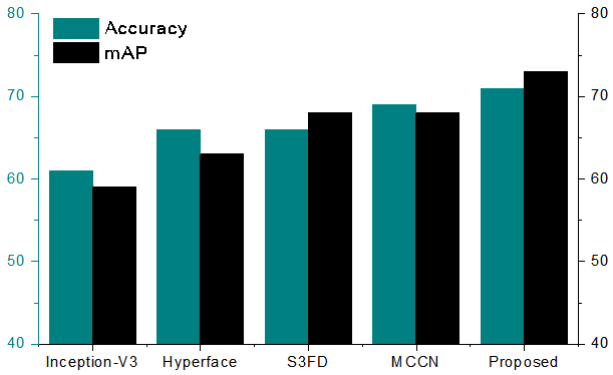


Fig. 4.4. Comparison with various EL classification architectures.

Since the created dataset is collected from the classrooms, it contains an unequal proportion of engagement level class labels. Hence, the MCC (Matthews correlation

coefficient) is performed and obtained a value of $+0.638$. The AUC value is 0.701. Further, a mAP (mean Average Precision) of $0.735, 0.741$ and 0.755 is obtained for IOUs (Intersection over Union) $>= 0.9, 0.8$ and 0.7, respectively. A sample snapshot of engagement level classified data using the proposed method is shown in Figure 4.5.

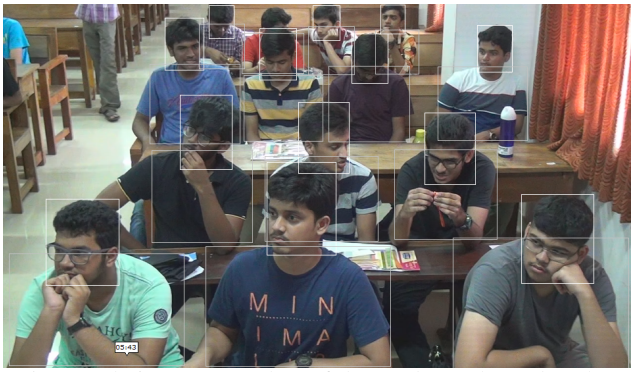


Fig. 4.5. Sample image snapshot of the students' boundary box plot.

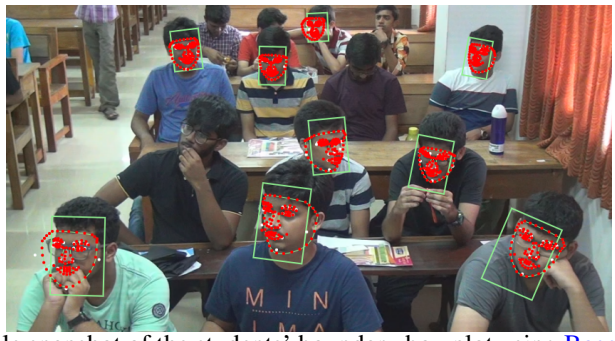


Fig. 4.6. Sample snapshot of the students' boundary box plot using Bosch et al. (2016).

Figure 4.6 shows the results of existing method (Bosch et al., 2016). It is evident from Figure 4.6 that all the students present in the classroom are not detected and hence, the proposed method outperforms the existing method. We used student-independent 10-fold cross-validation (the students present in the training set are not present in the test data) for the entire dataset (Bosch et al., 2016). We also performed cross-day, cross-gender and cross-period generalization for student independent 10-fold cross-validation and obtained +1.89%, -1.33% and +2.4% increase from the overall accuracy (Figure 4.4). For all the three generalizations, 67% student independent random data is used for training Bosch et al. (2016) and the remaining data for testing. The data is run for over 150 iterations to obtain these results.

We obtained an mAP of 73.22% using the proposed method on the test set, whereas the standard SSD and S^3FD gave an mAP of 59.11% and 66.73%, respectively. Nvidia GeForce 840 M is used during the test phase. 600ms is the average prediction time

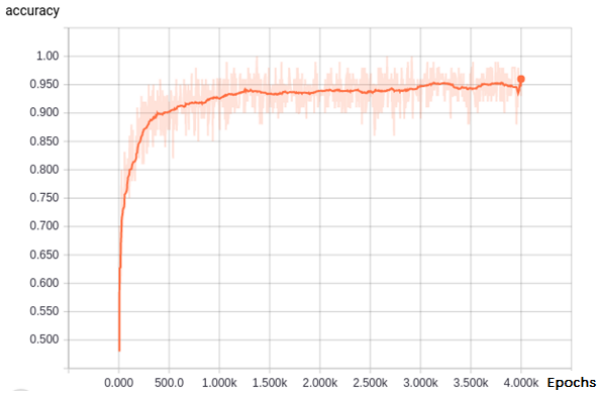


Fig. 4.7. Accuracy curve w.r.t epochs for training the proposed model.

for each image frame using SSD (Liu et al., 2016). The proposed method contains context-sensitive layers & feature pyramid layers, and the average prediction time for each image frame is 2153ms.

We observed that training and validation accuracy improved with each step or epoch and reached saturation after 1500 epochs. Figure 4.7 shows the training accuracy obtained for the created dataset without considering student-independent validation (but, the overall results discussed in 4.1.6 considered the student-independent 10-fold cross-validation). Similarly, at the end of 4000 epochs, we got a cross-entropy of 0.1459 for training and 0.2045 for validation.

Dissection of Multimodal Analysis

We analyzed the impact of each multimodal data for every student. Using localization, the multimodal data is divided into a face, hand gesture, and body posture to analyze the engagement levels. The proposed combined model performs better when compared to the use of a single mode to analyze the students' engagement, as shown in Figure 4.8. The facial expressions gave better classification accuracy but failed in few instances like, if the students' face is downwards, then it recognizes them under EL 2 or EL 1 whereas, from hand and body postures, it can be easily classified them under EL 3 as the student is taking down the notes. For the same scenario of taking the notes, if the body posture is bent backward and the facial expression is neutral, then hand gesture plays a major role in the classification. Similarly, there are various instances where each multimodal component contributed to better classification engagement levels.

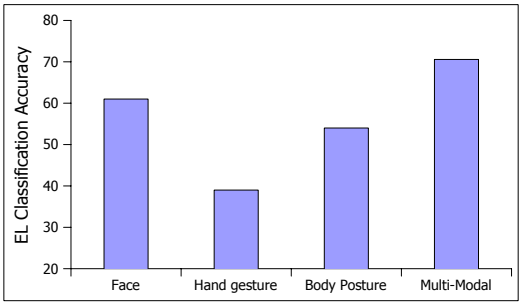


Fig. 4.8. Accuracy comparision among different multimodalities.

Group Engagement Level

Figure 4.9 shows the two samples of predicted average engagement levels for two classroom videos of 40 minutes each. For Class_1 most of the predicted values of engagement range between 0.5 and 1.5, whereas for Class_2 the average engagement level is between 2 and 3. In general, the engagement level value ranges between 0 and 3 (EL 1 to EL 4, respectively) in one single class, but clustered engagement level value variations are observed, as shown in Figure 4.9.

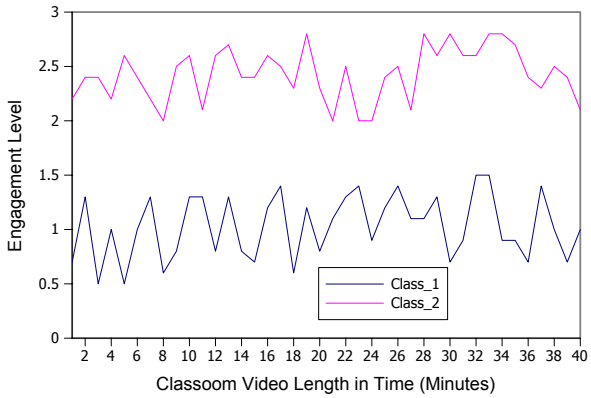


Fig. 4.9. Group engagement score in two different classroom videos.

4.1.5 Comparision of Proposed Method

Comparision with Popular Survey based Methods

We considered the most popular survey based students' engagement analysis methods such as NSSE (Kuh, 2003) and AUSSE (Australasian Survey of Student Engagement) (Kuh et al., 2008). After the completion of each class, the data related to the students' engagement is collected, and the marks obtained by the students in the post-test questionnaires are correlated using the Pearson correlation function. Pre-test analysis is also conducted by us to ensure that the students of all the sessions of the class are not familiar with the concepts (Rajendran et al., 2018).

We considered ten classes of around 40 minutes each in the classroom environment. Table 4.3 shows the results using the Pearson correlation coefficient. It is observed from Table 4.3 that our proposed students' engagement analysis has a high positive correlation with their test performance when compared to NSSE and AUSSE survey-based methods. Further, Table 4.3 shows a very less positive correlation between AUSSE and students' marks, which could be caused due to biased self-reports.

Table 4.3
Compassion of Proposed Methods with the Most Popular Survey based Methods for Students' Engagement

Correlation Metric	Proposed Model	NSSE	AUSSE
Pearson Correlation Coefficient Value	0.51	0.33	0.11

Comparison with State-of-the-art Methods

A few recent studies are available for students' behavioral engagement analysis in the learning environment. Whitehill et al. (2014) used Gabor features with SVM (Support Vector Machine) for a single face in a single image frame and obtained an AUC of 0.729. But this result is tested on a single person in a single frame image. Zaletelj & Košir (2017) used the Kinect sensor and KNN; thus obtained an accuracy of 0.753. Though the Kinect considers multiple people in a single image frame, the range of capturing students is less. Hence, the students' detection accuracy decreases if the number of students is more than 10 in a single frame image. Kahu (2013) and Bosch et al. (2016) used deep instance learning and WEKA tools, but it is already evident from the literature that the handcrafted features are less efficient for faces in the wild. It is difficult to directly compare the proposed methodology with the existing works as the datasets, and the multimodalities are different, in spite of which our results are comparable and more robust in terms of AUC and accuracy (Table 4.4).

4.1.6 Overall Results, Analysis and Discussion

The dataset is created with more than 4000 image frames with multiple students in a single image frame obtained from the classroom. The created dataset is classified into four different engagement levels. In order to increase the robustness of the training data, data anchor sampling, and data augmentations are used; thus, the size of the dataset is increased by five folds. We obtained an accuracy of 71%, MCC and AUC of 0.638 and

0.701, respectively. We performed student-independent 10-fold cross-validation, where the students present in training data are not present in the test image frames and thus obtained a mAP of 73.22%. The use of feature pyramid and context-sensitive layers in proposed architecture enhanced its performance leading to outperform the existing state-of-the-art architectures such as Inception and Hyperface. Even after the addition of feature pyramid and context-sensitive layers, the proposed method is able to classify the engagement levels with a predict time of 2153ms per frame.

Table 4.4
Comparison of Proposed Method with State-of-the-art Student Engagement Analysis Methods

Literature	DSIF		Data (S)	RT	Technique	PM	GEA
	SF	MF					
Whitehill *et al.* 2014	✓	×	34	iPad	SVM (Gabor)	AUC: 0.729	×
Zaletej *et al.* 2017	✓	✓	18	K	KNN, and other classifiers	Acc: 0.753	×
Thomas *et al.* 2017	✓	×	10	C	SVM, and LR (Logistic Regression)	AUC: 0.708	×
Bosch *et al.* 2016	✓	×	30	WC	14 different classifiers, including Bayesian classifiers, and LR	AUC: 0.790	×
Psaltis *et al.* 2018	✓	×	72	K	Kinect SDK	AUC: 0.850	×
Yun *et al.* 2018	✓	×	18	WC and K	VGG Face Network	AUC: 0.814	×
Henderson *et al.* 2019	✓	×	119	K	Deep neural network	AUC: 0.708	×
Tiam-Lee *et al.* 2019	✓	×	73	WC	Classifier model using Weka and OpenFace	AUC: 0.752	×
Proposed Method	✓	✓	350	C	Scale-invariant CNN architecture	AUC: 0.701	✓

DSIF: Detection within a Single Image Frame; SF: Single Face; MF: Multiple Face; RT: Recording Tool S: Students K: Kinect; C: Camera; WC: WebCamera; Acc: Accuracy

The existing systems used facial expressions significantly for the prediction of the students' engagement, but the use of multimodality in the proposed method improved the classification accuracy by 10% when compared to the use of facial features alone. As shown in Table 4.4, most of the existing literature is tested on e-learning or online learning environments with a single student in a single image frame. A few current works are tested on the classroom environment using a Kinect sensor. The range of Kinect devices is limited and can classify a maximum of 6 students in a single image frame. Further, these studies did not perform any group engagement analysis.

The proposed method is the first of its kind, which introduced a group engagement score for multiple students in a single image frame using the feature fusion technique. Most of the existing literature considered learning-centered emotions, but the proposed work explored behavioral patterns of students in the classroom environment. The proposed system outperformed even the popular and widely used survey-based methods such as NSSE and AUSSE for the students' engagement analysis. From Table 4.4, it is observed that the proposed method performs better than the existing techniques for the students' engagement analysis in the classroom environment with the use of contextual features, behavioral patterns, multimodality, and group engagement analysis.

4.1.7 Further Analysis

The standard statistical analysis performed in D'Mello et al. (2010) is used for the created dataset. The analyzed results are mentioned in the following subsections. The proposed method is also tested on classroom subset of ImageNet dataset (Deng et al., 2009) and the details are mentioned in the following subsections.

The Frequency of Engagement Levels

Table 4.5 shows the analysis for a sample of a 20 minutes classroom video with 16 post-graduate students of the NITK Surathkal, Mangalore, India. The engagement level frequency analysis is performed on these students. The predicted engagement levels are statistically analyzed using repeated measure ANOVA test, and it is observed that there is a significant difference in the proportion of engagement levels experienced by the students $F(4, 81300) = 421.83$, $MSE = 0.022$, $n^2 = 0.211$. The Bonferroni post-hoc

test revealed the following pattern $((EL\ 4 = EL\ 3) > EL\ 1)$ with $(p < 0.05)$ and tried to isolate these engagement levels using base as neutral (which is present in $EL\ 2$) using the chance ($Chance = (1 - M_{EL\ 2})/N_{EL} = (1 - .359)/4 = 0.16$) and performed t-test analysis on the data with chance level as 0.16. It is observed that there are only routine and sporadic engagement levels for the proposed four engagement level classification of students, as shown in Table 4.5. Similar results are observed when the same test is conducted for the entire dataset collected from students present in the classroom.

Table 4.5
Distribution of Engagement Levels

ELs	Frequencies		Proportions		One-sample t-test		
	N	P	M	SD	t(15)	p	d
Routine							
EL 4	11	0.638	0.121	0.077	3.715	<0.010	0.310
EL 3	13	0.761	0.111	0.117	3.212	<0.001	0.390
Sporadic							
EL 1	6	0.662	0.048	0.055	0.211	0.021	0.041
EL 2	15	1.000	0.359	0.313			

N = number of students that experienced the EL at least once
P = proportion of students that experienced the EL at least once
M = median and SD = standard deviation

Students' Engagement-Test Performance Relationship: Table 4.6 shows a sample analysis of the EL-test performance relationship for one class using Pearson's coefficient r. We repeated the same for the entire data. Better students' performance is more positively correlated with student's ELs 3 and 4 ($r = 0.568, p < 0.05$).

Table 4.6
Engagement Levels and Test Performance Relationship

Routine	*r*	**Sporadic**	*r*
EL 4	0.561	EL 1	-0.254
EL 3	0.246		
EL 2	-0.139		

Temporal Dynamics of Engagement Levels

Results are also analyzed for the persistence of the engagement levels. *Persistence* refers to a property in which the engagement level (S_t) at time t is also observed at time $t + 1$ (S_{t+1}). An engagement level (S_{t+1}) can be considered to be persistent if its experience at one time interval increases the likelihood of experiencing the engagement

level at the subsequent time interval i.e. $(S_t \rightarrow S_{t+1})$. Similarly, an engagement level is *ephemeral* if its experience at one time interval decreases the likelihood that will be observed at $t+1$. Finally, for a *random* engagement level, if an engagement is observed at time t then it is not related to the probability of its occurrence at $t+1$.

The likelihood metric (Equation 4.8) is used in an attempt to characterize the engagement levels along with this tripartite classification scheme. The metric quantifies the likelihood that the current state (S_t) influences the next state (X) after correcting the base rate of X. According to this metric, if $L(S_t \rightarrow X) \approx 1$, then the state X reliably follows state (S_t) above and beyond the prior probability of state X. If $L(S_t \rightarrow X) \approx 0$, then X follows (S_t) at the chance level. Furthermore, if $L(S_t \rightarrow X) < 0$, then the likelihood of state X following state (S_t) is much lower than the base rate of X.

$$L(S_t \rightarrow X) = \frac{P(X|S_t) - P(X)}{1 - P(X)} \tag{4.8}$$

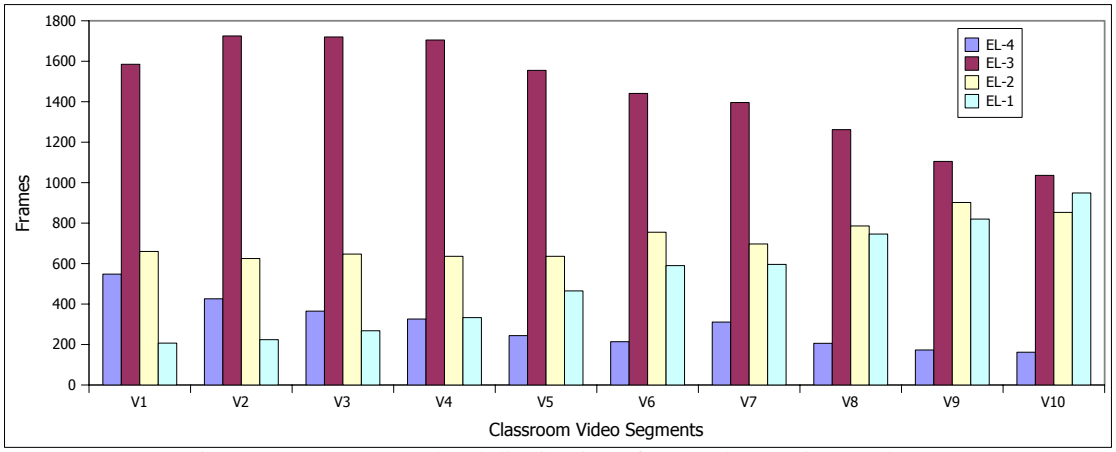


Fig. 4.10. Engagement level distribution of a sample 20 minutes class.

The main goal is to assess the likelihood that engagement level (S_t) observed at time t is also observed at time $t+1$ (S_{t+1}). This can be easily accomplished by modifying the metric such that the current engagement level (S_t) and the next engagement level (X) are the same (Equation 4.9).

$$L(S_t \rightarrow S_{t+1}) = \frac{P(S_{t+1}|S_t) - P(S_{t+1})}{1 - P(S_{t+1})} \tag{4.9}$$

In order to detect the significant engagement level persistence, the likelihood of each

engagement level repeating itself and it is hypothesized mean of 0 (normalized base rate) wihch is compared using a one-sample t-test. The results of the tests are presented in Table 4.7 where it appears that the data supports a one-way classification scheme (persistent) instead of a three-way classification scheme, as there are no instances of random and ephemeral states.

Table 4.7
Persistence of Engagement Levels

Enagagement Levels	**Descriptive Measurement (Likelihood)**		**One-sample t-test**			
	M	**SD**	**t**	**df**	**p**	**d**
Persistent						
EL 4 ->EL 4	0.151	0.249	3.210	11	0.008	0.390
EL 3 ->EL 3	0.401	0.232	3.888	15	0.005	0.560
EL 2 ->EL 2	0.230	0.121	1.678	09	0.053	0.320
EL 1 ->EL 1	0.122	0.168	2.220	11	0.038	0.390

It is observed from Table 4.7 that there are no random and ephemeral engagement levels in the proposed engagement level classification. This infers that the four different engagement levels have a significant impact on the students' behavioral engagement analysis. Its prediction is sufficient to analyze the overall classroom engagement.

Figure 4.10 contains a sample image from a classroom video clip of 20 minutes long, where the duration of every segmented video is 2 minutes, and 300 frames from each video segment are extracted at the rate of 5 frames/second. It is observed that the first segment video engagement level has 2732 judgments, the subsequent video engagement levels have judgments of 2880, 2882, 2901, 2800, 2880, 2830, 2820, 2810, 2753. The distribution of engagement levels for a particular student may be different, but when the entire class is considered, there exist enough instances of engagement levels for possible likelihood in the temporal dynamics of engagement levels. It is also observed that similar results are obtained for the entire collected data.

EL Transitions

To check for any possible pattern in EL transitions for the created dataset, Tukey HSD post-hoc test (D'Mello, 2012) is used on the data, and it is summarized in Figure 4.11. The transitions between EL 4 to EL 3 are dominant. EL 3 to EL 2 and EL 2 to EL 1 transitions are observed. There are many instances where EL3 lead to EL 2, and then

the students moved to EL 1. EL 3 to EL 4 and vice versa is also observed. There are a few instances where transitions from EL 1 to EL 3 or EL 4 are observed, but the frequency of that is less.

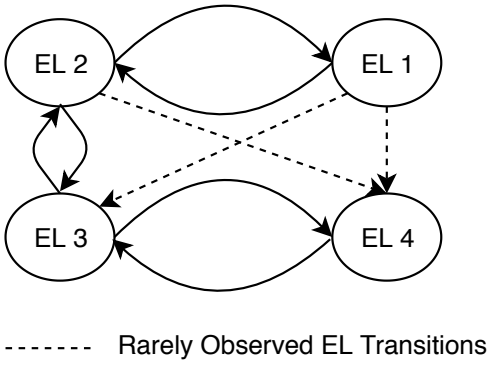


Fig. 4.11. Students' EL transitions.

Testing on ImageNet Dataset

The proposed model is tested on the images obtained from classroom subset of ImageNet database. Though the ImageNet database contains images with students' present in the classroom, they are not annotated for EL classification and student identification. Hence, ground truth is not present for the calculation of the performance evaluation metrics. But, the proposed model is able to recognize the students' engagement level, and few snapshots of the same are shown in Figures 4.12 and 4.13. From these figures, it is observed that the engagement levels are classified for the students' data with different angles and blurred images.

Fig. 4.12. Snapshot of proposed methodology tested on ImageNet.

Testing of Various Image Variants in the Classroom Environment

The proposed method is tested with the image frames obtained from the CCTV cameras, and it is able to detect the bounding box as well as the EL class labels, as shown

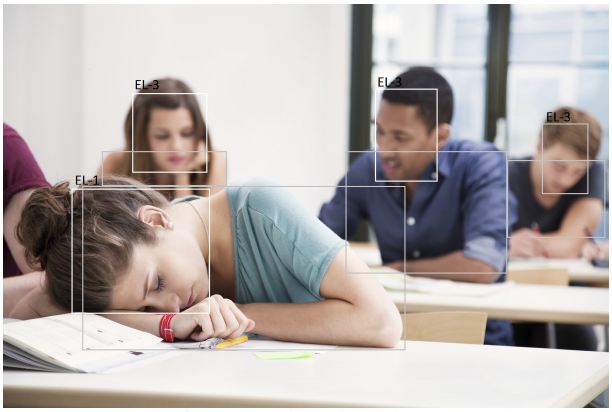
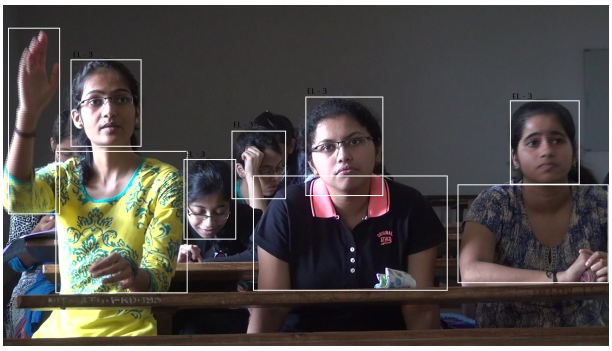
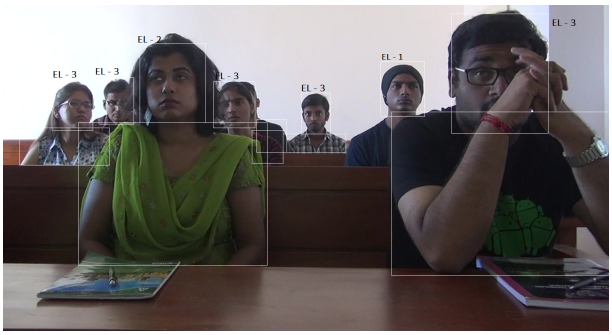


Fig. 4.13. Snapshot of proposed methodology tested on ImageNet.

(a) Sample snapshot obtained from the CCTV camera.



(b) Classroom data obtained using different camera angle.



(c) Classroom data obtained using different camera angle.

Fig. 4.14. Sample snapshots tested using the proposed methodology.

in Figure 4.14a. We also tested it on image frames obtained from the classroom using different camera angles, as shown in Figures 4.14b and 4.14c. From Figure 4.14b, it is

observed that the left-most student (student with the hands raised) has three different bounding boxes corresponding to face, hand gesture, and body posture. Here the intersection of hand gesture and body posture bounding boxes is null (less than 70%). Hence three different bounding boxes are used. Further, Figures 4.14a, 4.14b, and 4.14c are analyzed for group engagement analysis, and all the three image frames are classified under EL 3 using the proposed method.

4.2 Proposed Methodology for Computer-Enabled Teaching Laboratories

In the previous section, we discussed the first part of this chapter where the students' behavioral engagement analysis is performed in the classroom environment. In this section, unobtrusive students' behavioral engagement analysis in computer-enabled teaching laboratories (lab) using their non-verbal cues are discussed. The previous proposed deep learning architecture works better only when there are students in the classroom environment or something similar, here we have data which not only includes students but also includes computers and other lab-related accessories. Also, there are several other variants such as the position of sitting, the seating arrangement, occlusions from a computer cabin and so on makes it difficult to recognize and classify the students' behavioral patterns. Hence, a separate methodology is proposed to classify the students' behavioral patterns where the input data is obtained from the video surveillance cameras.

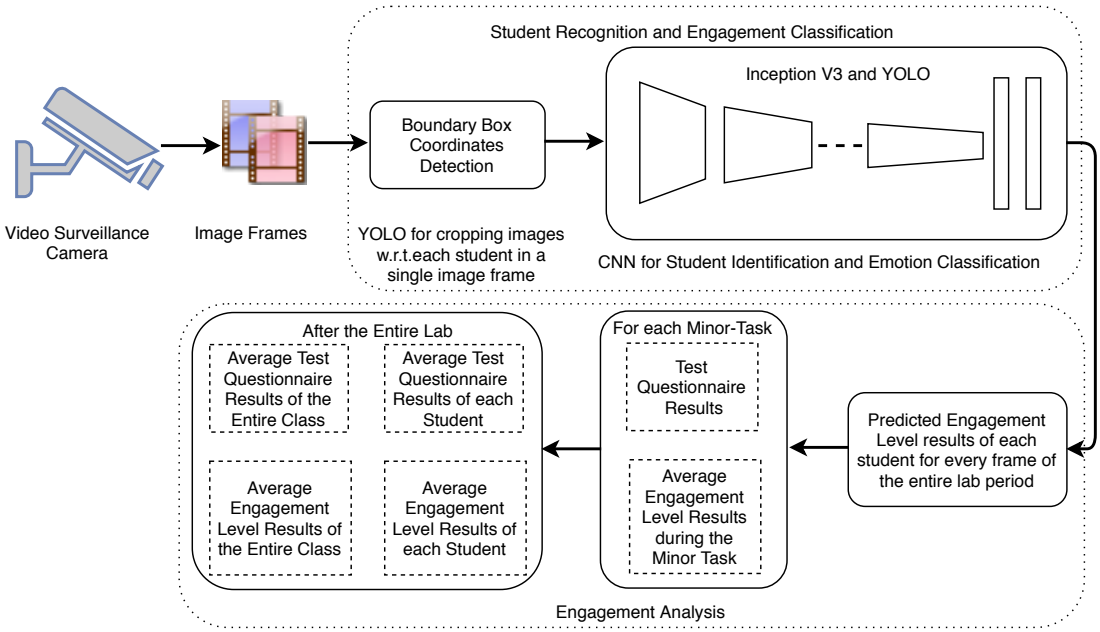


Fig. 4.15. Proposed architecture for students' engagement analysis

Figure 4.15 shows the proposed methodology for the student's engagement analysis. The still images are obtained from video surveillance cameras. These images are taken as an input by You Only Look Once (YOLO) architecture (Redmon et al., 2016). YOLO performs the cropping of images concerning each student and those cropped images are sent to the next layer, i.e. Convolutional Neural Networks (CNN) (Szegedy et al., 2016). CNN uses the trained data to classify the emotions/moods into the student's engagement level. These are calculated for each frame, for every student and the set of images for the entire teaching lab.

Entire lab assignments are divided into minor-tasks and the time duration taken to complete the minor-task for each student is noted. After each minor-task and also at the end of lab session, test questionnaires are given and these are evaluated. Students' engagement level is mapped with the learning of the students for each minor-task. Also an average of total marks obtained and the overall engagement score for the entire lab session are also calculated. Further, overall engagement score of the entire class is calculated by taking an average of engagement score of all the students and the marks obtained to analyse the students' engagement.

4.2.1 Students' Engagement Classification

The student's behavioral engagement class label remains the same. As discussed in the beginning of this chapter, the class labels are classified into four major engagement levels and the definition of each class labels remains the same (Whitehill et al., 2014).

4.2.2 Detection and Classification

The proposed architecture for detection, classification uses YOLO (Redmon et al., 2016) and Inception V3 Model (Szegedy et al., 2016), respectively. This model is optimized for faster and better processing and classification as shown in Figure 4.15. Instead of using the entire input image for engagement classification of each student using Inception V3 and then performing object location using YOLO, the YOLO architecture is used initially to obtain the boundary box coordinates of each student present in one image frame. Then these image frames are cropped according to the boundary box coordinates and then sent to Inception V3 model for the students' engagement classification. Further, the cropped images are used by CNN architecture for individual

student identification using the extracted local feature from the train data to calculate the similarity measurement.

Each lab is given a set of specific tasks (minor-tasks) to the students to implement (they can also learn if the concepts are not clear). After completion of each task the time taken for the completion is noted and test questionnaires are provided to evaluate their understanding for each minor-task and accordingly marks are awarded. The engagement classification results obtained from CNN architecture are used to calculate the engagement of each student in the entire teaching lab for every minor-task.

4.3 Experimental Setup, Results, Analysis and Discussion

4.3.1 Experimental Setup

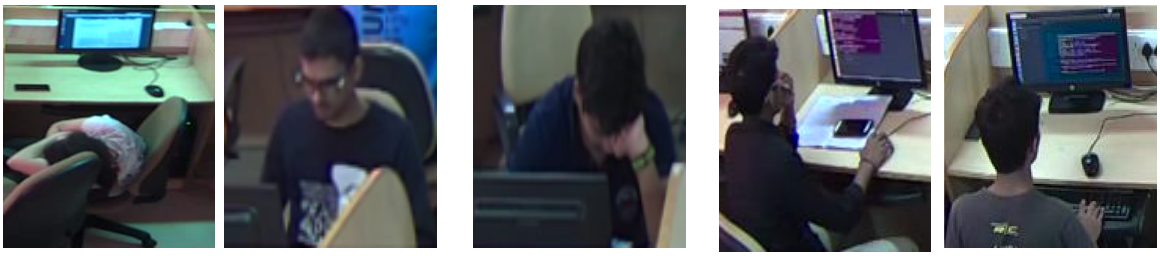
For the current study, 8^{th} Generation Intel® Core™ $i7-4510U$ Processor, 8GB RAM, and 2GB NVIDIA® GeForce® $840M$ are used for engagement level classification and localization.

4.3.2 Dataset

Data from different laboratory courses of computer science and information technology for B.Tech and M.Tech students is collected from the video surveillance cameras. Two cameras are present for each lab and if the students are not visible in one camera then they are visible in the other (Figures 4.17 (b), 4.17 (d) and 4.17 (e)). 5 hours 43 minutes of data is collected and manually annotated for training purpose. To annotate the database, proper guidelines are given to the annotator and minimum three different annotators perform the annotation. More details on the annotation process for the data used in this study is given in Section 6.2.1 of Chapter 6.

Fig. 4.16. Sample lab image frame with boundary box coordinates

A sample lab image frame retrieved with boundary box coordinates are shown in Figure 4.16 (the head region is detected using black boundary box and the entire body is detected using white boundary box). This image is not cropped concerning individual student for CNN identification and the classification phase. As mentioned earlier, from this camera angle only a few students can be observed with frontal pose whereas for a few students the system is visible with their back turned towards the camera. Figure 4.16 also shows some of the students relaxing whereas other students are engaged.



(a) (b) (c) (d) (e)

Fig. 4.17. Sample cropped image frames obtained for video surveillance camera during training phase

4.3.3 Detection and Classification Accuracy

The database is tested for teaching lab data of five different courses with a total of 243 students. The students present in the test data is same as the train data since the students identification can be performed only if they are already present in the training data. The detection and classification accuracy of the test dataset are mentioned in Table 4.8. 60 fps still image frames are collected from the surveillance video server and tested with the CNN model. The detection accuracy is high since both YOLO and CNN models use pretrained weights (these pretrained weights are trained with Labeled Faces in the Wild (LFW) dataset [1]) along with the manually annotated training data.

Table 4.8
Overall Results of Detection and Classification

Performance Metrics	Detection	Classification
Average Accuracy	0.94	0.89
Average Recall	0.94	0.91
Average Precision	0.93	0.88
Average F-score	0.92	0.87
AUC	0.89	0.86

The classification model also performed better than other standard deep learning

[1] http://vis-www.cs.umass.edu/lfw/

architectures like AlexNet (Krizhevsky et al., 2012) (71% accurate), ResNet (He et al., 2016) (73% accurate), VGGNet (Simonyan & Zisserman, 2014) (81% accurate). If the images are not clear, then they are discarded. But for those detected images the classification performance metrics are better than any state-of-the-art techniques.

The initial accuracy of the students' identification is 61%. Since the students sit in one position for the entire teaching lab session, the recognition code is optimized using the following steps. If the student's face detected image frame has an accuracy of 0.95 concerning similarity measure of train data for more than 30 image frames, then all remaining frames are discarded for that student's identification. This helped us to increase the identification accuracy to 97%.

4.3.4 Engagement Analysis

Engagement analysis is performed for each lab. For each image frame of the student, the predicted engagement level is used for the calculation of engagement analysis. For every student, engagement analysis is performed for each minor-task. The predicted engagement levels are given a numeric value ranging from 1 to 4 where four being the highly engaged. If the minor-task duration is 35 minutes, then an average score is calculated from engagement scores predicted from all the image frames during that period. Here, the minor-task duration is fixed for each student, if they complete faster then they can go to the next minor-task immediately. Even though the minor-task completion time is different for each student, the time accordingly for each student during the engagement analysis is considered.

Table 4.9
Overall Students' Engagement Analysis Results

Lab Course No.	Student Engagement Score	NSSE Score	AUSSE Score	Marks Obtained
Lab Course 1	3.1	2.3	3	839.66
Lab Course 2	3.3	3.4	2.7	823.66
Lab Course 3	3.5	3	3.1	964.33
Lab Course 4	2.8	2.4	2.9	762.66
Lab Course 5	3.6	3.3	2.3	922.33

Students are also told to complete the implementation and learning process for the entire lab without any deadline. We found many students relaxing, getting engaged in group discussions, sleeping etc. A sample image snapshot is shown in Figures 4.16, 4.17 (a) and 4.17 (c) where the student is not engaged during lab hours. But, Surprise

tests are given at the end of these labs to analyze and map the engagement level with their learning. With five different lab courses, a total of 243 students with 16 minor-tasks are evaluated, the obtained engagement score of each student is correlated with their marks obtained using Pearson correlation function.

Table 4.9 shows the correlation among the marks obtained in each lab course and the corresponding engagement score. Marks score is defined as the average of marks score obtained by all the students for that particular lab course. Similarly, the engagement score is also the average engagement score of all the students for that particular lab course. After applying the Pearson correlation function (Sedgwicket al. , 2012), a positive correlation of +0.8976 is obtained.

Further, existing and the most popular survey based student engagement analysis methods, i.e., NSSE and AUSSE data are also collected related to the students' engagement after the completion of each lab and the results are correlated with marks obtained by the students using the Pearson correlation function. We also conducted pre-test analysis to confirm that the concepts are not already known to the students for "learn and implement sessions" of the lab. Also, there is no deadline oriented questions, so the slow learners can take sufficient time to learn a single minor-task comfortably.

From Table 4.10, it is observed that our proposed engagement analysis system has a high positive correlation with the students' learning when compared to the survey based engagement analysis of NSSE and AUSSE. Further, it is observed from Table 4.10 that there exists less positive correlation between AUSSE and students marks, and this could be because of biased student's self-reports, being one of the reasons.

Table 4.10
Performance Evaluation of Proposed Model

Correlation Metric	**Proposed Model**	**NSSE**	**AUSSE**
Pearson Correlation Coefficient Value	0.8976	0.4796	0.2985

Table 4.11 shows the correlation among the marks obtained from 7 minor-tasks of Lab Course 1 and the corresponding engagement scores. Each minor-task score is an average of marks obtained by all the students in that particular minor-task. Similarly, the engagement score is also the average engagement score of all the students for that particular minor-task. After applying the Pearson correlation function, a positive correlation of +0.9227 is observed. Similarly, all the remaining minor-tasks of all the courses

Table 4.11
Pearson Correlation for Students' Engagement vs Each Minor-Tasks

Minor-Task No	Marks Obtained	Engagement score
Minor-Task 1	16	2.7
Minor-Task 2	17	2.6
Minor-Task 3	17	3.4
Minor-Task 4	12	2.6
Minor-Task 5	20	3.1
Minor-Task 6	15	2.5
Minor-Task 7	22	3.8

Table 4.12
Pearson Correlation for Student's Engagement

Student No	Total Marks Obtained	Engagement Score
Student 1	88	3.1
Student 2	85	2.2
Student 3	89	3.5
Student 4	86	2.1
Student 5	88	2.1
Student 6	93	3.8
Student 7	90	3
Student 8	88	4
Student 9	95	4
Student 10	90	2.9
Student 11	95	3.4
Student 12	95	3.8
Student 13	88	2.4
Student 14	93	3.7
Student 15	88	3
Student 16	91	4

are performed and the range of correlation coefficient value is in between 0.86 and 0.93.

Table 4.12 shows the correlation among the marks obtained for one minor-task of Lab Course 2 by a student and the corresponding engagement score. This correlation is tested for 16 students. After applying the Pearson correlation function, a positive correlation of +0.7146 is obtained. Similarly, all the remaining minor-tasks of all the courses for all the students are considered and found that the range of correlation coefficient value is in between 0.7 and 0.88.

4.4 Summary

The students' behavioral engagement analysis is proposed and implemented in the classroom environment using their facial expressions, hand gestures, and body postures.

The proposed scale-invariant context assisted single-shot CNN architecture performed well for multiperson in a single image frame. It is also observed that the results are better for multimodality than single modality. We could recognize most of the students in the wild and predict four different behavioral engagement levels. A single group engagement level score for each frame is obtained using the proposed feature fusion technique. The students' engagement analysis is performed for more than ten classes of 40 minutes each. Manual annotations are carried out for ground truth validation. Pre-Post test analysis is performed to check the correlation between the students' behavioral engagement and their test performance. The proposed multimodal analysis outperformed the popular survey-based methods (NSSE, and AUSSE) for student engagement analysis by showing a positive correlation between behavioral engagement and test performance. Further, frequency, temporal dynamics, and distribution of the engagement levels are analyzed. The proposed method is also tested on the classroom subset of the ImageNet database.

Video surveillance based students' behavioral engagement analysis is also proposed and implemented in computer science and information technology teaching laboratories. We obtained a good accuracy rate using Convolutional Neural Network for the students' identification and engagement classification. Students' Engagement analysis is performed for each minor-task and also for the entire lab session. Overall the students' engagement analysis for the entire lab is also performed. There is a positive correlation between students' engagement and learning. Also, this engagement analysis system outperformed the existing survey-based engagement analysis systems.

The current study focuses only on four different engagement levels based on the students' behavioral patterns, but the combination of emotional and behavioral engagement using the non-verbal cues is one of limitations of this study. The next chapter describes the use of recognized students' affective states as feedback to the teacher in real-time to enhance the teaching-learning process. Here, the students' affective states will be used for developing an automatic inquiry intervention system to enhance the teaching-learning process.

Chapter 5

Automatic Inquiry Intervention

Effective teaching strategies improve the students' learning rate within academic learning time. Inquiry-based instruction is one of the effective teaching strategies used in the classrooms. But these teaching strategies are not adapted in other learning environments like intelligent tutoring systems, including auto-tutors. In the previous chapters, we did not use the recognized students' affective states in real-time as feedback to the teacher to improve the teaching-learning process. There are several existing works that use self-reports, agents, and text-based analysis as feedback to improve the students' engagement but, the unobtrusive students' engagement analysis in real-time as feedback to enhance the teaching-learning process is not explored in the literature. Hence in this chapter, an architecture is proposed for classifying the students' affective states into teacher-centric attentive and in-attentive states using learning-centered and Ekman's basic emotions. Affective state recognition of each student is performed in real-time using his(her) facial expressions, hand gestures and body postures. The proposed architecture includes student's identification, affective state classification, student's affective state transition and automated inquiry-based instruction teaching strategy using inquiry interventions and it is tested in e-learning, flipped classroom, classroom and webinar environments for both the individual and the group of students.

The proposed architecture is divided into two modules: the first module includes the identification & localization of students and recognition of their affective states using deep learning techniques; the second module uses Inquiry Interventions (InIvs) which are used to optimize the students' performance within the academic learning time so that the student becomes more active and attentive during the learning process. For e-learning and flipped classroom environments, a separate tool is designed to support the real-time affective feedback and the students' performance analysis. Students' performance analysis is performed using the marks obtained by them in the Test Questionnaires (TQ) before and after the introduction of InIvs with the standard pre-test and post-test analysis.

The key contributions of this chapter are as follows:

- A novel deep learning architecture for unobtrusive affective state recognition with localization using students' facial expressions, hand gestures and body postures for: (i) e-learning & flipped classroom environments (single person in a single image frame) and (ii) classroom & webinar environments (multiperson in a single image frame).
- A novel real-time students' engagement analysis using the proposed teacher-centric attentive and in-attentive states for both the individual (e-learning and flipped classrooms) and the group of students (classrooms and webinars),
- Use of the proposed affective states as immediate feedback to the teacher to introduce the InIvs and thus analyze the impact of InIvs on the students.

The details of the proposed real-time students' affective state recognition, incorporating the above contributions in using the recognized affective states as feedback to the teacher is given in the following sections.

5.1 Data Collection and Participants

Participants: The proposed architecture is trained and tested for 350 undergraduate, postgraduate and doctoral students of National Institute of Technology Karnataka (NITK) Surathkal, Mangalore, India with different cultural and regional backgrounds. The participant details are shown in Table 5.1. The students who participated in this study are in the age group of 20 to 26 years. For training, both the posed and spontaneous expressions are collected from the students. The test dataset is collected on a real-time basis.

The created dataset consists of students' faces, hand gestures, and body postures. The dataset contains both posed and spontaneous expressions which include single and multiple persons in a single image frame. The faces of the students include frontal, profile and tilted faces; hand gestures include hands raise and body posture includes normal, half bent and full bent or completely lean on the desk pose. The entire dataset includes variants such as occlusion, background clutter, pose, illumination, cultural & regional background, intra-class variations, cropped images, multipoint view, and deformations. Images are captured from different camera positions to ensure the presence of such variants.

Table 5.1
Student Participant Details

Participant Details		E-L	C	W	F
Even Semester 2016	Classes	10	10	10	10
	Students	30	130	45	30
	Hours	7	9	5	4
	Course	2	4	4	1
Odd Semester 2016	Classes	3	4	4	3
	Students	40	125	50	10
	Hours	5	4	4	3
	Course	1	2	2	1
Even Semester 2017	Classes	10	10	10	10
	Students	40	111	50	30
	Hours	7	10	5	4
	Course	2	4	4	1
Odd Semester 2017	Classes	3	4	4	3
	Students	40	125	50	30
	Hours	3	3	3	2
	Course	1	2	2	1

E-L: E-Learning; C: Classroom; W: Webinar; F: Flipped Classroom

Affect Annotation Manual annotation, verification, and validation is performed using the gold standard study as mentioned in Sidney et al. (2005) where participants, novice judges, and expert judges are the three gold standards used for annotation. After annotation, even object localization is performed on each student. The annotated image with object localization is stored in the JSON file.

Affective States The image frame data obtained from learning environments are classified into twelve different classes, namely: happiness, sadness, delight, fear, disgust, surprise, sleepy, boredom, frustrated, confused, engaged and neutral. This classification includes Ekman's basic emotions and learning-centered emotions of students. These affective states are obtained from the literature and uses the standard definitions mentioned in D'Mello et al. (2010); Whitehill et al. (2014); D'Mello et al. (2007). The definitions of each affective state are mentioned below.

1. Boredom: uninterested in the current problem.
2. Confusion: poor comprehension of material, attempts to resolve erroneous belief.
3. Disgust: annoyance and/or irritation with the material and/or their abilities.
4. Fear: feelings of panic and/or extreme feelings of worry.
5. Frustration: difficulty with the material and an inability to fully grasp the material.
6. Happiness: satisfaction with performance, feelings of pleasure about the material.

7. Neutral: displays no visible affect, at a state of homeostasis.
8. Sadness: feelings of melancholy, beyond negative self-efficacy.
9. Surprise: genuinely does not expect an outcome or feedback.
10. Delight: a high degree of satisfaction.
11. Sleepy: extremely not interested and in a mental state of sleep.
12. Engaged: a state of interest that results from involvement in an activity.

More details on Database creation such as participants, data collection, affect annotation, data used for training and testing and inter-rater agreement among the annotators are mentioned in Section 6.1 of Chapter 6.

5.2 Proposed Methodology for Automatic Inquiry Intervention

Figure 5.1 shows the proposed methodology. It consists of two modules: the first module uses deep learning to recognize the students and their corresponding affective states using facial expression, hand gesture, and body posture. The second module uses the recognized affective states of the student and classifies them into two affective states, namely: teacher-centric attentive and in-attentive affective states. Group engagement analysis is also performed using the classified affective states. Further, the second module uses the data related to recognized moods of each student and accordingly the inquiry intervention (InIv) and test questionnaires (TQ) are asked. Further, the output of InIv and TQ is used to calculate and analyze the students' engagement and their corresponding learning rate. This entire process is trained and tested in both e-learning and classroom environments. Further, the teacher[1] can get the real-time feedback for affective states of the individual student in a classroom environment.

5.2.1 Proposed Affective State Classification

Predicting students' engagement is an arduous task as there are challenges with both the conceptualization and measurement of students' engagement. Behavioral, emotional, cognitive, and agentic engagements are the four different types of engagement (Sinatra et al., 2015). Most popular works on unobtrusive student engagement involve behavioral and emotional engagement with some cognitive aspects involved in it (as they are analyzed using non-verbal cues), but these works are performed on a single person in a single image frame. There exists no robust method which suits for either e-learning and classroom environments, hence a deep learning architecture is proposed which works

[1]The word "teacher" used in this chapter includes the faculty member or instructor

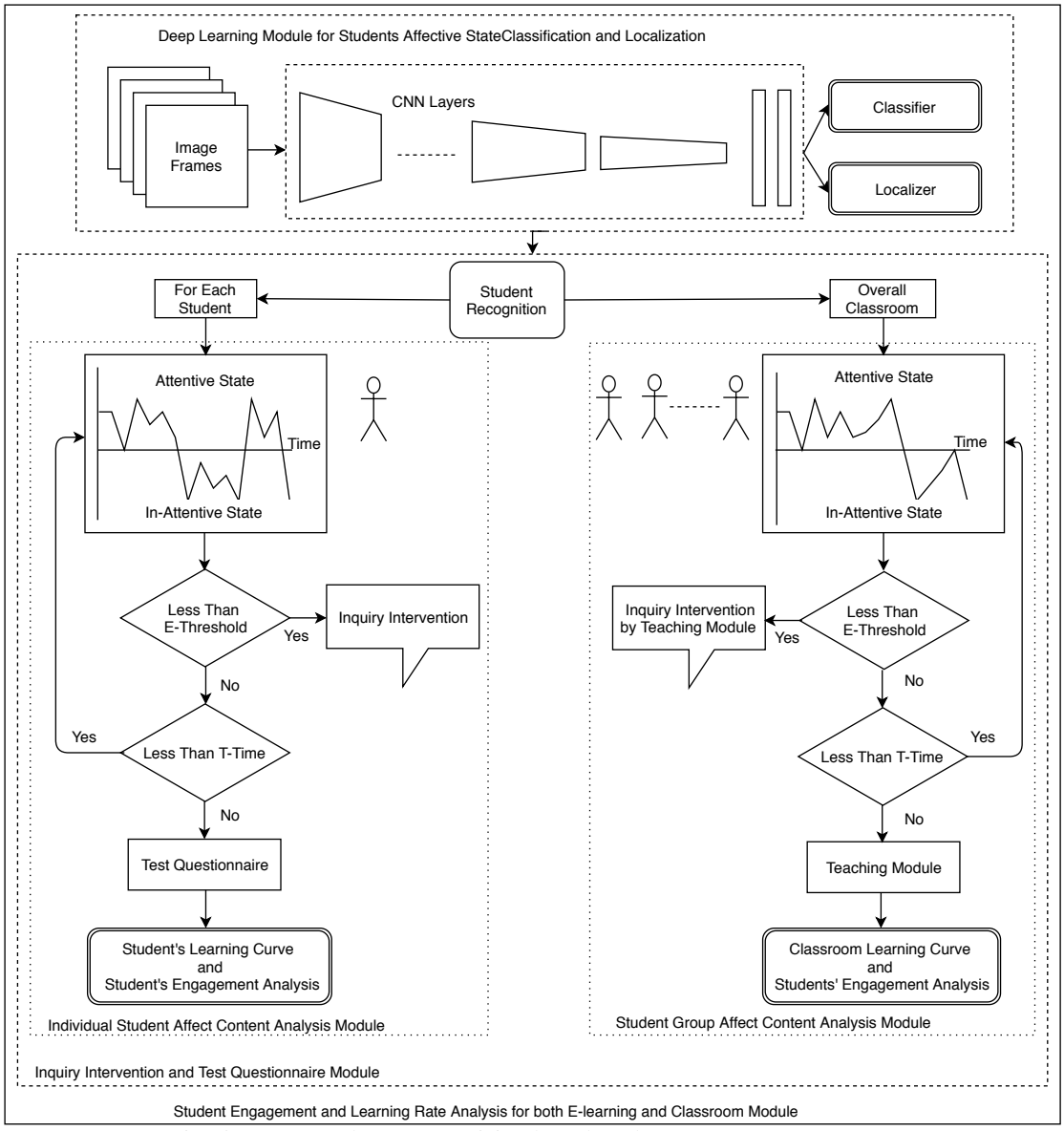


Fig. 5.1. Proposed computer vision based students' engagement system

in all the four learning environments to recognize the students' emotional engagement (facial expressions), and behavioral engagement (hand gestures and body postures) with some cognitive aspects involved in it. In a dense labeling scenario such as classrooms, there exist a few works which explore students' engagement analysis for either behavioral (Ashwin & Guddeti, 2019b; Zaletelj & Košir, 2017) or emotional engagement (Yu, 2017; Bosch et al., 2016). The prediction of both the emotional and behavioral engagement for multiple persons in a single frame using machine learning techniques is a difficult task. Hence, in this proposed work, both Ekman's basic emotions, as well as learning-centered emotions (both the students' emotional and behavioral aspects), are considered for analyzing the students' engagement.

The image frame data obtained from learning environments are classified into twelve different classes, namely: happiness, sadness, delight, fear, disgust, surprise, sleepy, boredom, frustrated, confused, engaged and neutral. This classification includes Ekman's basic emotions and the learning-centered emotions of students. Further, two affective states are proposed, namely: teacher-centric attentive and teacher-centric in-attentive for categorizing the recognized eleven different classes (excluding neutral) for the entire student database. Works on affective state recognition of students are not only considered (Ekman, 1992; Calvo & D'Mello, 2010), but also the works on behavior engagement, students' psychology, and their cognitive aspects (Whitehill et al., 2014) are considered to classify the recognized affective states into the two different affective states. For example, The definition of the engaged affective state is a state of interest that results from involvement in an activity. This includes not only the facial expressions related to engagement but also the behavioral aspects observed using hand gestures and body postures related to taking/writing notes, answering questions/asking the questions, paying attention toward the board or teacher and so on.

Table 5.2
Proposed Affective State Classification

TC Attentive Affective States				**TC In-Attentive Affective States**						
EB Emotions			LC Emotions	EB Emotions			LC Emotions			
Ha	Su	De	En	Sa	Fe	Di	Fr	Bo	Sl	Co

Ha: Happiness, Su: Surprise, De: Delight, En: Engaged, Sa: Sadness, Fe: Fear
Di: Disgust, Fr: Frustrated, Bo: Boredom, Sl: Sleepy, Co: Confused
TC: Teacher-Centric; EB: Ekman's Basic; LC: Learning-Centered

The proposed affective state classification (attentive and in-attentive) is based on the students' affective state analysis from the teacher's perspective. If the teacher can continue with the lecture without changing the teaching strategy, then all those students' affective states are considered as attentive and vice versa. For instance, from a students' perspective, sadness emotion can be considered as an attentive affective state where students are trying hard to understand the topic. However, from the teacher's perspective, sadness emotion can be considered as an in-attentive state where the students are not entirely engaged as the learning curve is not progressing at a regular rate. Hence the teacher changes the teaching strategy in such a manner that the students understand those topics easily. Thus, the recognized students' affective states are classified into

teacher-centric attentive affective state and teacher-centric in-attentive affective state for both the e-learning and classroom environments and the mapping of the same is shown in Table 5.2.

Module 1: Deep Learning for Affective State Classification and Localization

Student's Affective State Classification: The affective state classification is performed using the proposed Convolutional Neural Networks (CNN) architecture. The camera is used to obtain the input video stream. Then the video data is converted into image frames (60 frames per second (fps)). These image frames are preprocessed and normalized to 512*512 image size. Each frame contains face, gestures, and postures of the student. Further, there exist frames that contain multiple students in a single image frame. These image frames are considered as input for the convolutional layer. The proposed CNN architecture consists of 80 convolutional layers in which each layer extracts the features by convolving around the entire image frame. The final convolutional layer is attached to two fully connected layers which are connected to a softmax classifier which predicts the student affective states[1].

Students' Identification and Localization: The proposed architecture for the student's identification is based on YOLOv2 architecture (Redmon et al., 2016) where the YOLOv2 version is modified by adding a deconvolutional layer which is attached at the end of YOLOv2 architecture. By using deconvolution, large-Kernel Convolutions are made more accessible to approximate and thus perform better for blurred and noisy image frames.

Module 2: Inquiry Intervention (InIv) and Test Questionnaire (TQ)

The affective state classification results obtained from the previous module (Section 5.2.1) are used for the classification of teacher-centric attentive and in-attentive affective states. Here two components are considered for the analysis of engagement and learning curve, namely: individual student's affective content analysis and classroom (group) affective content analysis.

The individual affective content analysis module receives the affective state of the student for each frame (60fps) and generates an attentive/in-attentive affective transition

[1]More details on Affective state classification and localization are mentioned in the Supplementary Details 5.4 of Chapter 5

graph over time. The transition graphs are generated using a script. It uses the students' affective state for each frame and calculates the mode average for 60 frames to get the affective state for each second and the total time duration for which the student is present in a particular affective state is stored. These data are used for the engagement analysis of the student for the entire lecture. Generally, in the e-learning environment, there will be a smooth transition from one state to another.

For a real-time classroom engagement analysis, the students' affective state classification data obtained from the deep learning module are used to calculate the group engagement score. As per the definition of engagement, it is highly related to participation. For group engagement score calculation, two major factors considered are students' affective states and the balance/similarity in the participation (which determines whether the group members have the same engagement level or not). For example, a group could have an engaged affective state, but results are calculated by considering just a few students. The average affective state score can be computed using fusion techniques. There are a few effective fusion techniques, namely: feature fusion and decision fusion. Decision fusion is not considered in this study as it classifies the affective states even if there is any marginal difference. For example, a student's affective state is classified as engaged if the prediction probability value is 0.51, even if the 0.49 belongs to boredom it classifies as engaged. To reduce this carryover effect, feature fusion is used where all the features of attentive and in-attentive are considered before the affective state prediction. The balance in participation is the second factor, and the existing works use dispersion indicators such as Shannon entropy, the Gini coefficient, the Ricci-Schutz coefficient (also named Pietra's measure) and the Atkinson's index to measure this aspect. These metrics are normalized between zero and one, they compare the proportion of participation of each student rather than the absolute contribution, and standard implementations are facilitating their adoption. There exists literature on students' group engagement prediction using text-based analysis which demonstrates that Atkinson's index performs with least error rate (Castellanos et al., 2017). The text-based analysis and image-based predictions have similar features for the level of participation, and hence, Atkinson's index is used and the feature fusion to predict the students' group aggregate score.

The created training data set with annotated images (both posed and spontaneous) are considered. Every student's affective state is given a unique value from 1 to 12, including the neutral (as mentioned in Section 5.2.1). The balance in participation is addressed using Atkinson's index, and the decision trees are used to model high and low engagements, and the ties among them are resolved by choosing the high level (as mentioned in Castellanos et al. (2017)). For the proposed method, the use of Atkinson's index is modified in the following way. All the students in a large classroom cannot be in one affective state (for example, all 35 students present in the single image frame are frustrated), this is practically impossible unless it is a posed image frame. And, the inquiry intervention uses attentive and in-attentive affective states. Hence, it is sufficient to check the level of participation of students' w.r.t. attentive and in-attentive states.

The created dataset with multiple students in a single image frame is considered. Every student's affective state value present in that image frame is collected and classified into two numerical values corresponding to attentive and inattentive states and 0 for neutral. The number of times attentive, in-attentive, and neutral affective states observed in each image frame is stored. This data is iterated for more than 1500 epochs using decision trees to compute the model. This model ranges from the sum of one single affective state (attentive, in-attentive or neutral) present in the image frame is equal to the sum of all the students present in that image (i.e., all the students have same affective state (this data is collected from posed expressions)) and vice versa. Atkinson's index is performed for decision trees with a mean error rate of 4%. The images used for train, test, and validation of decision trees are also verified by the annotators, and they agreed with the classification of high and low levels ($\kappa = 0.873$).

Group Engagement: Students' affective state obtained from the feature fusion method is classified into attentive, inattentive, or neutral. If the predicted affective state of the image frame obtained from the feature fusion method and the corresponding similarity in participation is high then, that particular affective state is considered as a group's affective state for that frame. If the similarity in participation is low, then one cannot specifically mention any affective state for that image (this means 50% of the students are attentive and the other 50% are inattentive). Less than 1% of the time these results are observed where the similarity of participation is low, and the feature fusion showed a specific affective state.

Each student's attentive state ranges from 0 to 1, and in-attentive state ranges from -1 to 0. The surprise and delight affective states are given weights between 0.5 and 1. Similarly, the weights for happiness & engaged; confused, bored & sleepy; and fear, disgust & frustrated affective states are 0 to 0.5, -0.5 to 0 and -1 to -0.5, respectively. This weight classification is performed based on the "two-factor structure of affect" (basic two-dimensional affective state) (Watson & Tellegen, 1985) and "Russell's Circumplex model" (Russell, 1980) where the affective states are distributed within four quadrants of the 2D space. For the group or class score, the feature fusion method provides a probability value between 0 and 1 for the students' affective state (for example, from feature fusion the probability score of a spontaneous image frame for that particular frame is engaged = 0.63, means the feature fusion technique has classified the image frame as engaged). Both the probability value and the affective state are considered for mapping to the attentive (0 to 1) and in-attentive (-1 to 0) scores.

Mapping of deep learning results to attentive and in-attentive scores: The proposed deep learning module provides the probability score of the affective state ranging from 0 to 1. This value is normalized between 0.5 and 1 for both surprise and delight; 0 to 0.5 for both happiness and engaged. For inattentive states, though the predicted probability values are positive, these are normalized to negative values ranging from -0.5 to 0 and -1 to -0.5 for bored & sleepy; and fear, disgust & frustrated affective states.

Every concept is subdivided into mini-concepts which can be explained in 5 to 12 minutes duration. If the student stays in the in-attentive affective state for more than 30% duration of the mini-concept, then a question is posed to make the student active, interactive and engaged. This is called Inquiry Intervention (InIv).

The active learning teaching strategy is popular and widely used by teachers to enhance learning (Bonwell & Eison, 1991). Active learning has intellectually, socially, and physically active learning strategies (Edwards, 2015). Inquiry activities are one of the instructional strategies related to intellectually active learning. The Inquiry activity uses purposeful questions to enhance students' engagement. Purposeful questions are broadly classified into Creative, Critical, and Curiosity thinking questions (Eison, 2010). In this study, the proposed method automatically intervenes and poses the purposeful questions to the students based on their engagement.

The purposeful questions used in the InIvs are prepared manually by the teachers and are stored in the database. The prepared questions are either descriptive or fill in the blanks or multiple-choice questions. A threshold score is set for each mini concept and in case of lower threshold rate, InIvs pop up on the screen in case of e-learning and flipped classroom environment and the same is directly posed by the teachers to the students in a classroom or webinar environments. Since these questions are used to make the students attentive, three categories of questions are asked depending on the concept taught and its context, like:

1. Questions which stimulate the *critical thinking* about the observation,
2. Questions which stimulate the *creative thinking* about the observation,
3. Questions which stimulate the *curiosity thinking* about the observation[1].

At the end of every mini-concept, a set of questions are posed to the students. These questions are used as a continuous assessment component of students in the learning process. These set of questions are called Test Questionnaires (TQs). This is defined manually by an expert teacher according to the course content. Without answering the TQs, students cannot proceed to the next mini-concept. These manually stored TQs popup on the screen after every mini-concept video in case of e-learning & flipped classroom environment. The same is collected in the written form in the classroom & webinar environment.

5.2.2 Proposed Affective State Transition Diagram

From the feedback of faculties, it is observed that visualization and remembering the flow of affective states for each student and the entire class with just the numerical data provided by the proposed method is difficult. The existing competency maps, dashboards, and others did not suit for the visualization of affective states. Hence, an affective state transition diagram is proposed based on the conceptual framework of finite automata.

The proposed affective state diagram helps in visualizing the transition between attentive and in-attentive affective states of the student. Figure 5.2 shows the sample affective state transition diagram of a student for a 20 minutes video lecture. Once the student starts the learning process (video lectures in e-learning and teacher's audio in the classroom), the start state of the transition diagram gets triggered, and it stops with the

[1]Sample questions used in the inquiry intervention module is given in the Supplementary Details 5.4 at the end of this Chapter

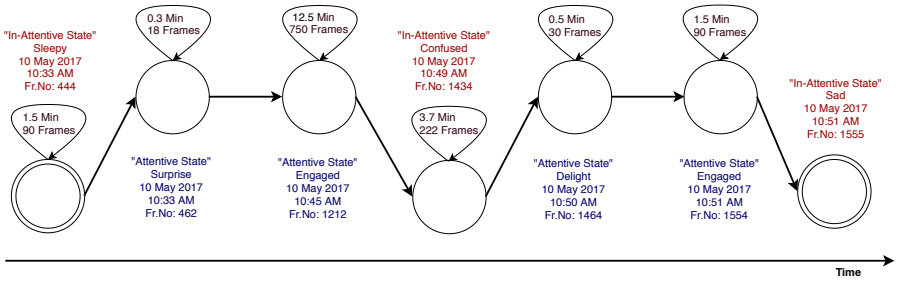


Fig. 5.2. Sample affective state transition diagram of a student

end of the video or the classroom lecture with the final state of the transition diagram. The start and end states are represented as concentric circles shown in Figure 5.2. Each transition state contains the information about the affective state and its corresponding affective state, date & time of the image frame captured and the frame number of last accessed image frame as shown in Figure 5.2. Students tend to stay in the same affective state for a fixed period, and this is shown using a self-loop in the transition diagram which contains the details such as the time duration of stay and the number of frames captured during that time. If there is an immediate transition of affective state, then that state does not contain any self-loop as shown in the final state of Figure 5.2. The proposed affective state transition diagram is also used to visualize the affective state transition between attentive and in-attentive affective states for the entire class (all the students present in one classroom are considered as one group).

When a student or the group is found in the in-attentive affective state for more than 30% of the duration of the lecture, InIvs are introduced. A sample affective state transition diagram with InIv for a 7 minutes mini-concept video is shown in Figure 5.3.

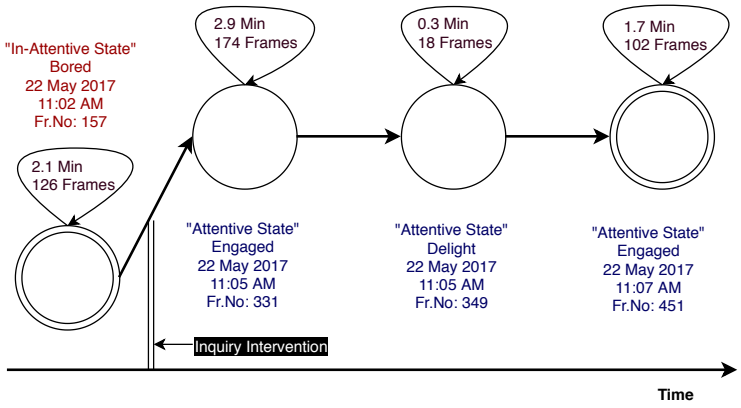


Fig. 5.3. Sample affective state transition diagram of a student with InIv

5.2.3 Model Implementation

The proposed model is implemented in four different learning environments. The working of the same is discussed as under.

E-Learning and Flipped Classroom Environments

To integrate the real-time engagement analysis module in the e-learning environment, an e-learning web portal is designed. All the standard e-learning websites' components are considered during the website design. Some functionalities are further added such as affective state transition visualization, mapping of their understanding and the attentiveness shown by the students. Further, appropriate components for the proposed InIv and TQ module is incorporated. Figure 5.4 shows a sample e-learning UI components which contain a lecture video along with its completion and buffer length. The toggle button placed next to date and time indicates if the student is online or not. Notes are also provided along with the video clips and the button for the same is placed below the chapter heading, i.e., top left. If the web-page remains inactive for more than 20 minutes while reading the notes, and for more than 5 minutes if the student is not visible within the frame then it automatically stops the video streaming of the students' face and body postures to the server. Students' details along with the date and time appear at the bottom right. In the case of group activity, more than one student can use the UI (bottom right). In this case, the video stream data contains multiple persons in a single frame image, and the affective state classification is performed individually.

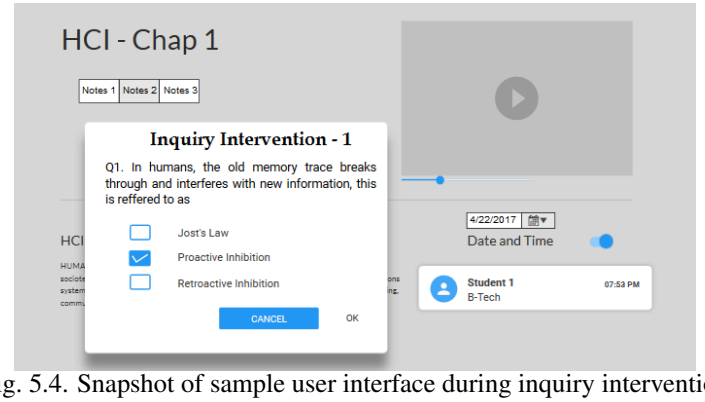


Fig. 5.4. Snapshot of sample user interface during inquiry intervention

Students of two different semesters are considered for four different courses over a period of two years for this study. On an average one video lecture consists of 1 hour 15 mins which are divided into 6 to 8 mini-concept video clips, each video clip

consisting of 10 to 15 minutes explaining a part of the content present in the entire one hour lecture. On average one complete chapter takes 2 to 3 hours of a video lecture.

InIvs: When the student is found in-attentive for the said threshold limit, InIv pops-up on the screen and is evaluated, marks distribution varies according to the questions. These marks are stored in the database.

Fig. 5.5. Spontaneous affective state recognized in e-learning environment

TQs: A Mini-concept can be taught using any source such as video lectures, reading material, animations, and code implementation. If the student understands the concept in lesser time, i.e., before the entire video is played or the study material gets completed, then the student can manually click on TQs and proceed. The marks scored by the students on answering the TQs are also stored in the database. Marks obtained from the TQs for each mini-concept, average student engagement score during the mini-concept duration and number of times InIv triggered are stored for further analysis. Figure 5.5 shows a snapshot of a sample image frame from an e-learning environment.

Classroom and Webinar Environments

In classroom and webinar environments, the cameras mentioned in camera setup are used to capture the students. These images are sent to the server, and the analyzed affective state for the entire class is shown in a tab that is placed next to the faculty member who is teaching. The course content of an hour class is divided into 2 to 3 mini-concepts. The tab displays the average engagement score of the entire class. If more than 10% of the class is observed in the in-attentive affective state, then a red flash will blink in the tab. If more than 30% of the students are present in the in-attentive affective state, then the vibrator attention is given using the smart-watch, so that they can pay attention towards the tab once (smart-watch vibrates for every 1 minute time interval if more than 30% of students stay in the in-attentive affective state). Once the teacher clicks the InIv button, even the affective state transition diagram for the

entire class is shared with the teacher so that he(she) can ask the InIvs and modify the teaching strategy accordingly. A snapshot of each image along with the location details are provided in the tab so that the faculty can locate the student easily even in the webinar environment. A sample image snapshot of the classroom environment is shown in Figure 5.6.

Fig. 5.6. Spontaneous affective states recognized in classroom environment

InIvs: If more than 30% of the students are in in-attentive affective state and then the teacher can teach up to 50% time duration of the mini-concept, introduce the InIvs and then proceed with that mini concept. If the faculty member wants to ask the InIvs, then he(she) has to press the InIv button on the tab and then proceed so that the proposed method can save the time stamp and the number of instances InIvs are asked. The faculty member can also give them a small task as a part of InIvs.

TQs: As discussed earlier, after every mini concept, the teachers collect the written form of the TQs and evaluate them accordingly. Marks obtained from the TQs for each mini-concept, average student engagement score during the mini-concept duration and number of times InIvs triggered are stored for further analysis.

In addition to the study made for the entire classroom as a group and aggregated group engagement score, it is also possible to analyze the affective state of each student and display the same in the tab. Whenever the teacher finds the time, he(she) can see the individual student's engagement score and the corresponding affective states. Figure 5.7 shows a snapshot of the sample image frame from the webinar environment.

Fig. 5.7. Spontaneous affective states recognized in webinar environment

5.3 Experimental Setup, Results, Analysis and Discussion

5.3.1 Experimental Setup

Two Tesla M40 GPUs with 256 GB RAM and 256 GB scrap space are used for deep learning computations.

5.3.2 Baseline for Affective State Classification and Localization

Twelve different students' affective states classification (including neutral) is performed in real-time using various state-of-the-art techniques, namely: AlexNet, ResNet-500, Inception-V3/V4 and compared with the proposed Modified-Inception-V3 (M-I-V3). The recommended metrics for affective state classification are meanAveragePrecision (mAP) and accuracy. Since it is performed in real-time, the classification time in seconds is also considered.

Object localization is performed using various techniques such as Regions with Convolutional Neural Networks (RCNN), Fast-RCNN, Faster RCNN, You Only Look Once (YOLO), YOLO-V2, Single-Shot Multibox Detector (SSD), SSD-300, SSD-500 and compared with the proposed Modified-YOLO. The standard metrics used for the object localization are accuracy, mAP, and frames per second (fps).

There are many students' affective state classification and localization methods but, the results are not used in real-time applications. Hence, there is no baseline for comparing the InIv and TQ module.

5.3.3 Students' Affective State Classification and Localization

Initially, the model is tested with state-of-the-art recognition and object localization techniques. Inception v3 performed better for student's identification and localization. The Inception v3 model is further optimized by adding depthwise separable convolution, Leaky ReLU (Rectified Linear Unit) as an activation function, RMSprop with root mean square, minibatch stochastic gradient descent and the number of layers increased to 80. Many hyper-parameters are also fine-tuned according to the input data.

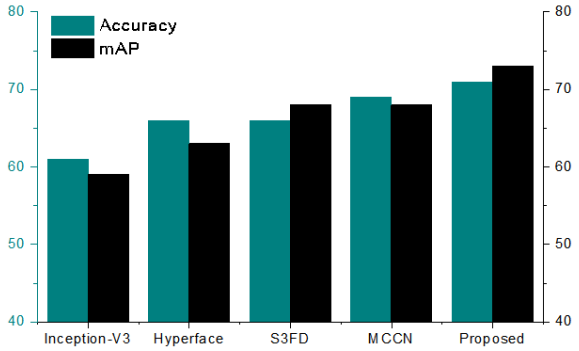


Fig. 5.8. Comparison with various affective state classification architectures

Figure 5.8 shows the summary of the comparison of the proposed architecture for affective state classification with other state-of-the-art architectures. The data is trained and tested with different architectures such as AlexNet, ResNet-500, Inception-V3/V4 and Modified-Inception-V3 (M-I-V3) (Krizhevsky et al., 2012; He et al., 2016; Simonyan & Zisserman, 2014; Szegedy et al., 2016, 2017) as mentioned above. It is observed that a few architectures like Inception-V3 are most accurate but took more time to execute and vice versa, whereas the proposed architecture gives better accuracy with minimum classification time such that the analysis can be performed in real-time. Other standard architectures such as VGG-16, VGG-19, ResNet-18, ResNet-34, ResNet-50, ResNet-101, ResNet-152, ResNet-200, BN Inception, Inception-V2 are not tested as these architectures fail to perform better for dense labeling scenario such as classroom.

Similarly, the object localization results of proposed modified-YOLO (M-YOLO) and its comparison with other state-of-the-art techniques such as Regions with Convolutional Neural Networks (RCNN), Fast-RCNN, Faster RCNN, You Only Look Once (YOLO), YOLO-V2, Single-Shot Multibox Detector (SSD), SSD-300, SSD-500 (Redmon et al., 2016; Liu et al., 2016; Ren et al., 2015; Girshick, 2015; Girshick et al., 2014)

are summarized in Figure 5.9. Faster-RCNN (FRCNN) and RCNN have better precision than M-YOLO, but they fail to perform in real-time. YOLO and SSD have high fps, but the precision is less than M-YOLO. Hence, the proposed M-YOLO gives better performance in real-time for student's localization when compared to state-of-the-art methods.

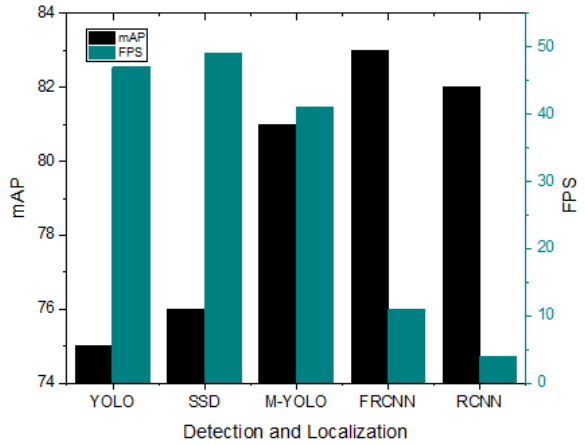


Fig. 5.9. Comparison with various object localization architectures

Student-independent 10-fold cross-validation is used for the entire dataset. Also performed cross-day, cross-gender and cross-period generalization for student independent 10-fold cross-validation and obtained +1.89%, -1.33% and +2.4% increase from the overall baseline accuracy. For all the three generalizations, 67% student independent random data are used for training (Bosch et al., 2016) and the remaining data for testing. More than 150 iterations are run to obtain these results.

There are several existing tools for real-time emotion recognition with object localization. Tools such as Affectiva[2], Imotions[3], Emotient[4], IBM Watson[5], and Microsoft Asure[6] are also tested on the created dataset, and obtained an accuracy of less than 9%, 11%, 17%, 6%, and 14% respectively. These tools failed to recognize even the basic emotions (for which these tools are trained for) for the images obtained from the classroom environment. The less accuracy in the detection and classification of the majority of these existing tools are due to the following: (i) Lack of multi-face detection, (ii) Inadequate Learning centered emotions consideration, (iii) Absence of multimodality as

[2]https://www.affectiva.com
[3]https://imotions.com
[4]http://emotient.com
[5]https://developer.ibm.com/exchanges/models/all/max-facial-emotion-classifier
[6]https://www.programmableweb.com/api/microsoft-project-oxford-face

only facial expressions are considered, (iv) Aggregation of emotions or group emotion is not present, and so on. Apart from these tools, there are state-of-the-art architectures proposed for emotion recognition (Gupta et al., 2016). The created database is tested on those networks. For deep learning architectures, student independent 10 fold cross-validation is used and the obtained results (accuracy in %) are: (i) FaceNet: 41%, (ii) EmotiNet: 43%, (iii) LBP with SVM: 37%, and (iv) Viola-Jones Haar Cascades: 33%.

Accuracy Comparision: Cohen's κ is computed in order to compare the accuracy of deep learning classification technique with human annotations (Whitehill et al., 2014). The results are mentioned in Table 5.3 where it is observed that the average κ value varies between 0.89 and 0.97 for both human annotation and machine classification in e-learning and flipped classroom environments. This is due to the fact that the image frame contains a single person in a single image frame whereas the average κ value varies between 0.56 and 0.81 for classroom and webinar environments where multiple students are present in a single image frame, and the system fails to perform better.

Table 5.3
Cohen's κ for Student-Independent Train-Test Results

Cohen's κ (and std. dev.)

Environment	DL Technique	Human Annotation
E-Learning	0.94 (.83)	0.97 (0.91)
Flipped Classroom	0.93 (0.86)	0.97 (0.94)
Classroom	0.61 (0.83)	0.69 (0.81)
Webinar	0.73 (0.88)	0.77 (0.91)

5.3.4 Impact of Inquiry Intervention on Individual Students

The Mann-Whitney test (Rajendran et al., 2018) is used to analyze the impact of InIv during the learning process for students of all the four learning environments. This section is divided into two subsections corresponding to the single person in a single image frame (e-learning and flipped classroom) and multiperson in a single image frame (classroom and webinar environments). Further, the Pearson correlation function is used to analyze the correlation between students' attentive and in-attentive affective states and the marks obtained by students in TQs with and without InIv module.

E-Learning Environment

The proposed methodology is implemented in an e-learning environment, duration ranges from 20 minutes to 1 hour. Initially, results and analysis of one class for a duration 30 minutes is projected; Subsequently, the overall results are projected.

30 students completed the Human-Computer Interaction (HCI) course which spanned five sessions over a week. These sessions did not include the proposed InIv module. TQs are considered to map their performance with the attentiveness in the class. A total of 17 TQs for each student are conducted, corresponding to 17 concepts taught in five sessions and the results are stored for the comparison. A student may have a bad day and may perform poorly. Hence generalized least-squares estimates are considered during the calculation.

The same 30 students are given similar complex course content for Human-Computer Interaction to learn 17 concepts. But this time they used our proposed InIv module. Their corresponding recognized affective states and test questionnaires marks are collected. It is observed that the sum of in-attentive state instances reduced to 21% using the InIvs module.

For every mini-concept duration, an average of attentive and in-attentive affective state scores [(-1 to +1) Engagement Score] for each student is calculated, and corresponding marks obtained in TQ is correlated using the Pearson's correlation function. After the completion of the entire course, it is observed that the student's engagement score and the marks obtained by the student positively correlates with a value of +0.675.

Similarly, this entire process is performed for all the students of even and odd semester batches of 2016 and 2017 academic years as shown in Table 5.1. The average engagement score of all the students and class average (in marks) obtained for both InIv modules (with and without) are analyzed. It is observed that the student is more engaged using the InIv and hence understanding the concepts is better when compared to regular e-learning class. The Mann-Whitney statistical test is performed for analyzing the impact of InIv on student affective states. Table 5.4 shows that the sum of in-attentive affective state instances is reduced using InIv. The Pearson correlation test is also used on our entire e-learning data for the student's engagement score and marks obtained. An overall positive correlation value of +0.622 is observed.

Table 5.4
Impact of Inquiry Intervention on In-Attentive Affective State in E-Learning Environment

Students	NMC	Without IIM		With IIM		Mann-Whitney's Significant Test
		SISI	Median	SISI	Median	
ES 2016	52	2216	42	714	14	$p<0.05$
OS 2016	40	1211	30	800	20	$p<0.05$
ES 2017	55	1913	35	491	9	$p<0.05$
OS 2017	21	1619	77	384	18	$p<0.05$

NMC: Number of Mini-Concepts, IIM: Inquiry Intervention Module,
ES: Even Semester, OS: Odd Semester, SISI: Sum of In-Attentive State Instances

Flipped Classroom Environment

The proposed real-time engagement analysis is tested for four semester's flipped classroom data and observed that there is a 43% reduction in in-attentive affective state instances by using InIv module. Pearson correlation coefficient is 0.89 for the students' engagement score and marks obtained.

Classroom Environment

The proposed methodology is implemented in every class ranging from 45 minutes to 1.5 hours in the classroom scenario. The same procedure (as in e-learning) of implementing the InIvs module (i.e., with and without InIvs) is followed. The average engagement scores of all the students and class average in marks are obtained. Pearson correlation coefficient is +0.742 for student's engagement score, and the marks obtained. Overall, for all the 180 mini-concepts, and observed a 43% reduction of in-attentive instances by using InIv module in the classroom environment (Table 5.5).

Table 5.5
Impact of Inquiry Intervention on In-Attentive Affective State in Classroom Environment

Students	NMC	Without IIM		With IIM		Mann-Whitney's Significant Test
		SISI	Median	SISI	Median	
ES 2016	82	11280	141	5988	71	$p<0.05$
OS 2016	35	5323	147	2902	80	$p<0.05$
ES 2017	90	9351	101	6393	73	$p<0.05$
OS 2017	28	4111	153	1826	66	$p<0.05$

NMC: Number of Mini-Concepts, IIM: Inquiry Intervention Module,
ES: Even Semester, OS: Odd Semester, SISI: Sum of In-Attentive State Instances

Webinar Environment

Here the number of InIvs asked is more frequent than any other learning environment as discussed. Pearson correlation for engagement score and marks obtained is +0.833, and also there is 53% decrease in in-attentive affective state instances using InIv module.

Overall Results on Inquiry Intervention Module

The number of attentive and in-attentive affective state instances for all the students in all four learning environments is summarized in Table 5.6. It is observed that the in-attentive affective state instances are less in courses that involved the inquiry intervention module.

Table 5.6
Impact of Inquiry Intervention on In-Attentive Affective State Instances in Learning Environments

Students	NMC	Without IIM		With IIM		Mann-Whitney's Significant Test
		SISI	Median	SISI	Median	
ES 2016	194	39312	204	17921	94	p<0.05
OS 2016	141	36456	250	16946	117	p<0.05
ES 2017	190	38824	196	17989	96	p<0.05
OS 2017	88	13096	148	8883	98	p<0.05

NMC: Number of Mini-Concepts, IIM: Inquiry Intervention Module,
ES: Even Semester, OS: Odd Semester, SISI: Sum of In-Attentive State Instances

The affective states which have a significant impact ($p < 0.05$) on student's performance are shown in Table 5.7. Bonferroni post-hoc test is used to analyze the impact of affective states on student's performance, and the mean and standard deviation of affective states for both the correct and incorrect answers are shown in Table 5.7.

Table 5.7
Affective States and Performance Relationship

Affective State	Correct		Incorrect		Student's Performance
	M	SD	M	SD	
Engaged	0.67	0.71	0.37	0.55	Beneficial
Bored	0.37	0.43	0.83	0.88	Harmful
Sleepy	0.20	0.21	0.92	0.97	Harmful
Frustrated	0.33	0.39	0.77	0.81	Harmful
Confused	0.33	0.37	0.70	0.79	Harmful

5.3.5 Performance Evaluation

Table 5.8 shows the overall comparison of existing works with the proposed method. It is observed that the existing works lack the use of unobtrusive students' affective video content data in real-time for either feedback or interventions. Even group engagement score and multi-modality are not explored in most of the existing works. The proposed method with multi-person detection, multi-modality, group engagement score, inquiry intervention, and with an accuracy of 0.77 for a test data of 350 students, outperforms the existing methods.

Table 5.8
Comparison of Proposed Method with State-of-the-art Student Engagement Analysis Methods

Authors	DSF		MM	GEA	F/I	DS	RT	Technique	PM
	SF	MF							
Whitehill et al. (2014)	Yes	No	No	No	No	34	iPad	SVM (Gabor)	AUC: 0.729
Zaletelj & Košir (2017)	Yes	Yes	No	No	No	18	K	KNN, and other classifiers	Ac: 0.753
Thomas & Jayagopi (2017)	Yes	No	No	No	No	10	C	SVM, and LR (Logistic Regression)	AUC: 0.708
Bosch et al. (2016)	Yes	No	No	No	No	30	W	14 different classifiers, including Bayesian classifiers, and LR	AUC: 0.790
Tiam-Lee & Sumi (2019)	Yes	Yes	No	No	No	73	W	OpenFace	Ac: 0.750
Yun et al. (2018)	Yes	No	Yes	No	No	18	W\K	VGG	Ac: 0.866
Psaltis et al. (2017)	Yes	No	Yes	No	No	72	K	ANN	Ac: 0.850
Proposed Method	Yes	Yes	Yes	Yes	Yes	350	C\W	CNN	Ac: 0.770

DSF: Detection within a Single Image Frame; SF: Detects Single Face
MF: Detects Multiple Face; DS: Number of Students used in the study
GEA: Group Engagement Analysis; F/I: Feedback/ Intervention
MM: Uses Multi-modality for affective state prediction; RT: Recording Tool
C: Camera; W: Webcamera; K: Kinect; Ac: Accuracy; PM: Performance Metric

Adaptive feedback and interventions are used in literature in the form of hints and motivational messages but they are performed using text-based analysis. They collect the log file of students while interacting with the social media platforms, and those data

are used for text data analysis, and accordingly, feedback and interventions are used (Rajendran et al., 2018; Silva et al., 2019; Psaltis et al., 2017; Moore & Stamper, 2019).

5.3.6 Further Analysis

This subsection consists of three parts: the first part deals with the impact of using similar complexity content and further discusses whether there is a reduction in in-attentive affective state instances on the introduction of the same. The second part deals with the student's affect-performance relationships. ANOVA (D'Mello et al., 2010) test is performed to analyze the relationship between students' affective states and their performance using the overall marks obtained by the students from test questionnaires with and without InIvs. The Bonferroni post-hoc test (D'Mello et al., 2010) is used to check which affective state is more significantly related to students' performance. The correlation between students' learning rate and their performance is also discussed using Pearson correlation function (D'Mello et al., 2010). The third part deals with the random InIvs and InIvs to high attentive engagement level students.

Impact of using Similar Complex Content: The course content with similar complexity is used to compare the impact of InIvs on in-attentive affective states. There may be a possibility that the reduction in in-attentive affective state instances maybe because of the course content with similar complexity. Considering this as the null hypothesis and to address this carryover effect, the same experiment is conducted for two different concepts of the same complexity and performed Mann Whitney test to check whether any significant difference is present in the number of in-attentive affective state instances. In this experiment, the InIv module is not used and obtained $r = 0.45$ from the Mann Whitney test for $p > 0.05$, thus fail to reject the null hypothesis. Hence it is clear that the reduction in in-attentive affective state instances is due to the InIv module. Further, it is also experimented by giving the InIv first and then without InIv. Even then there is a significant impact of InIvs on reducing the in-attentive affective state instances.

Affective-Performance Relationship: The ANOVA test is conducted for e-learning, flipped classroom, classroom and webinar environments and confirmed that there is a significant relationship between student's affective states and marks obtained. After Bonferroni post-hoc test, the following are observed: Experiences of engagement

are positively correlated to the positive outcome, [Engagement$_{\text{Correct}}(M = 0.67, SD = 0.71)$ > Engagement$_{\text{Incorrect}}(M = 0.37, SD = 0.43)$] and vice versa for frustration, [Frustration$_{\text{Correct}}(M = 0.33, SD = 0.39)$ < Frustration$_{\text{Incorrect}}(M = 0.77, SD = 0.81)$]. Similarly, all the experienced emotions with the corresponding marks obtained are considered. It is observed that the emotions of attentive affective state are more positively correlated to the positive outcome than in-attentive affective states, [Attentive$_{\text{Correct}}(M = 0.80, SD = 0.88)$ > Attentive$_{\text{Incorrect}}(M = 0.20, SD = 0.27)$], [In-Attentive$_{\text{Correct}}(M = 0.32, SD = 0.37)$ < In-Attentive$_{\text{Incorrect}}(M = 0.71, SD = 0.81)$].

Overall students' learning curve is calculated using the students' marks obtained. The standard academic learning time based learning curve equation (Brown & Saks, Brown & Saks) is adopted and accordingly modified this learning curve equation by adding engagement analysis factor, also adjusted the variables based on mini-concepts. In the standard equation mentioned in Equation 5.1, X_i variable is considered as a constant value 1 assuming that the students are completely engaged during the learning process. But, for the proposed work, the equation is modified by introducing a variable value corresponding to students attentive and inattentive states. The proposed learning curve equation is shown in Equation 5.1. Where, MC_A, MC_B, and MC_C are the marks of students during three different intervals of a course completion time (the number of intervals increases with an increase in the number of mini-concepts taught in the class). $TIME_{AB}$ and $TIME_{BC}$ are timings allotted between A & B and B & C respectively. X_i is the proportion of time being attentive in the learning process (X value varies from -1 to +1, in-attentive to attentive). e_B and e_C are the error terms that are mapped to other influencing factors. a_1, a_0, b and c are constants.

$$
\begin{aligned}
log\,(MC_C) - log\,(MC_B) &= (a_1 - a_0) \\
&+ b\,[log\,(MC_B) - log\,(MC_A)] \\
&+ c\,[log\,(TIME_{BC}) - log\,(TIME_{AB})] \\
&+ \sum_i (X_{iBC} - X_{iAB}) + (e_C - e_B)
\end{aligned} \tag{5.1}
$$

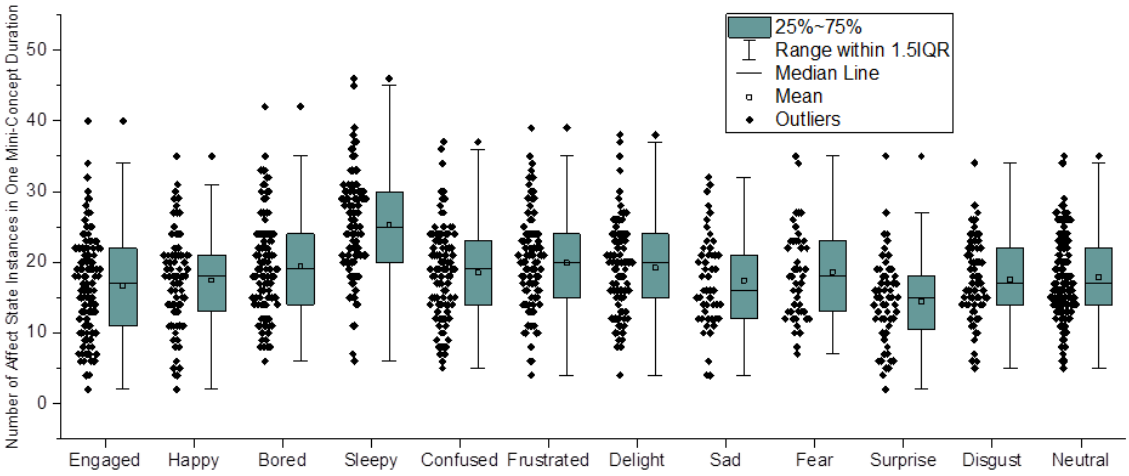


Fig. 5.10. Student affective state transitions

Students' learning rate (Equation 5.1) is calculated for each student in all the four learning environments. Each student's attentive state ranges from 0 to 1, and in-attentive state ranges from -1 to 0. The surprise and delight affective states are given weights between 0.5 and 1. Similarly, the weights for happiness & engaged, confused, bored & sleepy, and fear, disgust & frustrated affective states are 0 to 0.5, -0.5 to 0 and -1 to -0.5, respectively. For the group or class, the overall probability value obtained for each frame is considered. Table 5.9 shows a sample analysis of the affective state-learning rate relationship for one student in the e-learning environment using Pearson's coefficient r. The same is repeated for all the students. The higher learning rate is more positively correlated with the student's attentive affective state ($r = 0.598, p < 0.05$).

Table 5.9
Affective States and Performance Relationship

Routine	*r*	**Sporadic**	*r*	**Exceptional**	*r*
Engaged	0.581	Confusion	-0.354	Sadness	0.005
Happiness	0.216	Frustrated	-0.291	Fear	0.008
Boredom	-0.212	Delight	0.012	Surprise	0.219
Sleepy	-0.301	**Neutral**	-0.039	Disgust	-0.116

Figure 5.10 shows the number of affective state instances experienced by the students in one hour class. A few outliers are observed in the classroom environment where the students are completely in-attentive throughout the class and still managed to obtain above-average scores. This may be because of those topics which are trivial to the student or already known concepts. These outliers are not observed in e-learning, flipped classroom and webinar environments.

Random InIvs and InIvs to High Attentive Level: In order to test the impact of InIvs, they are posed randomly during the mini concepts irrespective of the students' affective states. There may be a possibility that only InIvs would lead to the improvement of the performance irrespective of their affective states. Considering this as a null hypothesis, the Mann Whitney test is performed, and a value of 0.00134 for $(p > 0.05)$ is obtained. There is no significant improvement in the performance, hence rejecting the null hypothesis.

The feedback is obtained from the students of all the four environments that random InIvs, at times, interrupted the flow of conceptual understanding as in case of e-learning the InIvs on the subtopics yet to be covered also popped up. On the other hand, if the random InIvs are posed at low engagement students, it is reasonably effective whereas if it is posed at high engagement level students, it is similar to that discussed below.

Further, InIvs are also tested for students with high engagement level. Pearson's correlation between students' engagement score and performance is 0.128 showing no significant difference as compared to the previous test with no InIvs.

The feedback from the students of all the learning environments when InIvs are posed at them is obtained, few of them got frustrated as though, the mini concepts are quite clear and additional learning is not happening, they had to answer the InIvs. It is also reported that the learning rate is hindered within the academic learning time due to unnecessary InIvs.

5.4 Supplementary Details

1. *Inquiry Intervention Module*

   These are the sample questions asked in the inquiry intervention module.

   (a) Examples of questions to stimulate critical thinking.

   Q: What point you think the Schachter and Singer is making in their proposed third interpretation [312a]?

   This illustrates a question that asks students for clarification.

   Q: What are the notable similarities and/or differences between the items on your list and those provided by D A Norman's Emotion and Design?

   This question challenges the students to compare and contrast ideas provided by two sources.

   Q: What are the possible underlying assumptions behind Shneiderman's criteria?

   This question challenges the students to demonstrate the critical skill of identifying assumptions made by an author.

   Q: What alternative assumptions might an HCI Analyst make?

   This question further probes students' understanding of assumptions by asking to identify alternative assumptions that might be made.

   Q: Do you personally agree with Shneiderman's criteria and what types of evidence can you offer to support your position?

   These two critical thinking questions ask students to both take a personal position and to provide relevant information or evidence to support their decision.

   (b) Examples of questions to stimulate creative thinking.

   Q: How many different heuristics can you list that are the most important and will receive attention in heuristic evaluation stated in Jakob Nielsen's ten heuristics?

   This question challenging students to come up with as many possible different ideas as they can and elicits the type of creative thinking process that has been described in the literature as fluency.

   Q: If Shneiderman wanted to identify Eight Golden Rules of Interface Design instead of only five, what three additional possibilities would you suggest to him to add to his list?

   This question, requiring students to embellish or expand upon ideas, elicits a type of creative thinking process that has been described in the literature as elaboration.

   Q: What novel, unique, or unusual types of course activities and assignments do you think they are most likely to inspire today's college and university students?

   This type of question illustrates one approach to stimulate the creative thinking skill of originality.

(c) Examples of questions to stimulate curiosity.

Q: Who is Don Norman and why might people be interested in his perspective on this topic?

Q: What was Jakob Nielsen like as a student?

The questions asked are not having a similar pattern mentioned in the examples for all the subjects, but the underlying concepts for framing the questions remain the same.

2. *Affective State Classification and Localization Technique*

Inception-V3 (Szegedy et al., 2016) and YOLOv2 (Redmon et al., 2016) are the base architectures for the affective state classification and localization, respectively. The existing architecture is modified to perform better in real-time for students' affective state classification and localization. The details are mentioned as follows. The proposed architecture is divided into four major parts. The first part is a stem, which contains the convolutional and pooling layers until the factorization of convolutional layers. The second part contains the factorization into smaller convolutions using filter concats. The third part contains the auxiliary classifiers and the last part contains fully connected layers and softmax classifiers. The entire proposed architecture consists of 80 layers with stride 2 and 6 pooling layers which are used for an independent evaluation over each activation map and for size reduction. The last pooling layer is connected to two fully connected layers. Those fully connected layers are connected to the softmax layer.

The training phase starts with the back-propagation using the recursive chain rule for computing the gradients for all inputs, parameters, and intermediates. The entire process of backpropagation is stored in a graph structure. The forward pass computes the results of an operation and stores the gradient value whereas the backward pass computes the gradient of loss function w.r.t. the inputs.

Equation 2 shows the loss function performed using multinomial logistic regression where y_i is the ith image label and x_i is ith image frame, using these, the loss for the ith image L_i is calculated. The overall mean of training loss of the entire training data L_t is defined as the total data loss given in Equation 5.2.

$$L_t = \frac{1}{N}\sum_{i=1}^{N} L_i \tag{5.2}$$

$$where,\ L_i = -log(\frac{e^{sy_i}}{\sum_j e^{s_j}})$$

Since the data is used for analyzing multiple class in a single image frame (dense labeling), The activation function which neither saturates for both positive and negative values nor dies for negative values is used. Hence Leaky ReLU (Rectified Linear Unit) is used as an activation function (Equation 5.3) where x is an

input column vector containing all pixel data of the image.

$$f(x) = max(\alpha x, x) \; where \; \alpha = 0.01 \tag{5.3}$$

Xavier initialization (Equation 5.4 where n_in & n_out are input neurons from current & next layer and r is to calculate zero mean) is used for weight initialization. Batch normalization is also used along with the hyperparameter optimization. Minibatch stochastic gradient descent is used in a loop structure where it performs 4 major steps; Step 1: data is divided into N samples. Step 2: forward propagate to get the loss. Step 3: calculate the gradients using backpropagation. Step 4: from the calculated gradients, update the parameters.

$$W = random.r(n\_in, n\_out)/sqrt(n\_in) \tag{5.4}$$

Parameters update is performed using RMSProp. The learning rate is a hyperparameter for RMSProp. Regularization is performed using Dropout. Monte Carlo approximation is used in dropout where several forward passes with different dropout masks are performed and the final prediction is average of all predictions. The numeric gradient is used for gradient check.

Group Engagement Score: Generally, the data collected from the e-learning environment contains single student in a single image frame, but the students' spontaneous classroom data can have different affective states for each student present in a single image frame. Hence feature fusion is used to calculate the same. The multi-modal (here, it is an intra-image multi-modality where the features of different students with their facial expressions, hand gestures and body postures present within that image frame are considered) feature fusion vector V_f for any pixel p_i and normalized prediction vector N_{P_i} uses normalized predicted probability distribution $N_{P_i,a}$ of class a using the softmax function (Equation 5.5).

$$N_{P_i,a} = \frac{e^{W_a^T V_f}}{\sum_{i \in classes} e^{W_i^T V_f}} \tag{5.5}$$

Where, W is the temporary weight matrix used to learn the features. The training generally converges in $T = 6000$ epochs. The final collective average affective state score A_{S_i} is given by Equation 5.6.

$$A_{S_i} = arg\,max\; N_{P_i,a} \; where \; a \in classes \tag{5.6}$$

Data Augmentation: Data augmentation has increased training data size by 10-fold. More details on the different data augmentation techniques that are performed on our datasets are given in 3.2.3.

3. *Student Engagement Level Analysis using its Graphical Representation*

   A sample snapshot of a student's engagement level for a mini-concept of 15 minutes duration is shown in Figure 5.11. There is a drop in the engagement score

(Attentive State to In-Attentive state) of the student from 10 to 12.5 minutes range. Initially, a similar drop is observed after 5 to 6 minutes of mini-concept video completion for more than 30% of students in the e-learning environment. After introducing InIvs, the engagement drop is reduced from 30% to 7%.

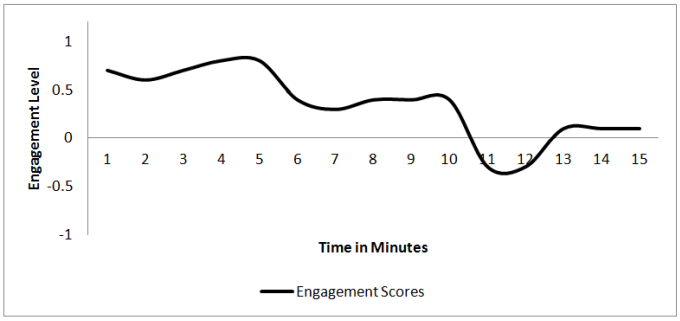


Fig. 5.11. A sample snapshot of student's engagement level for a mini-concept

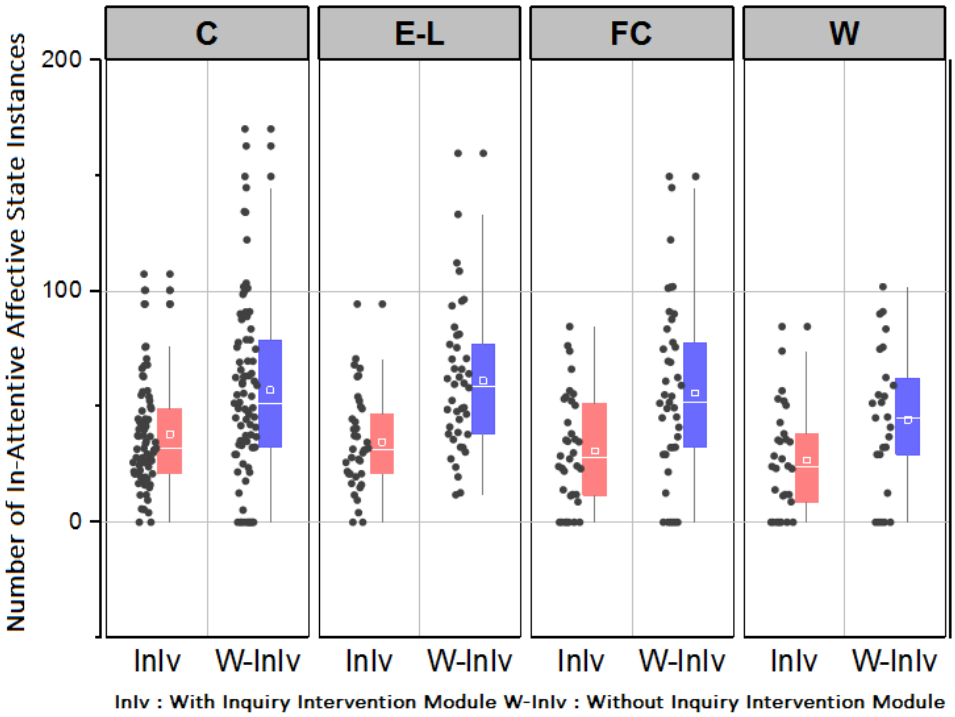


Fig. 5.12. The distribution of attentive and in-attentive state instances for the sample mini-concept duration in all four learning environments.

The distribution of attentive state and in-attentive state instances for a sample 5 minutes duration of a mini-concept (the same concept is taught in all four learning environments) is shown in Figure 5.12. Though the number of students is significantly more in the classroom environment, it is observed that the median is almost the same.

5.5 Summary

In this chapter, a real-time students engagement analysis is proposed for both the individual and group of students based on their facial expressions, hand gestures and body posture for e-learning, flipped classroom, classroom, and webinar environments. The proposed model effectively classifies the students' affective states into teacher-centric attentive and in-attentive affective states. To address the negative impact of in-attentive affective states on the performance of students, InIvs are proposed. Experimental results demonstrated that there is a positive correlation between the students learning rate and their attentive affective state engagement score for both individual and group of students. The proposed affective state transition diagram and appropriate visualizations helped students and the faculty members to improve the teaching-learning process.

There are a few limitations of this study. The students' affective states distribution varies according to the learning environments and teaching strategies. The current study focuses only on the above-discussed environments and inquiry-based instruction strategy in such environments, but the results could vary in other cases. For instance, in case of game-based learning, the engaged affective state may be higher as compared to the results of this study. In a dynamic learning environment, there will be various aspects to study in order to implement the proposed methodology. Even if a few such aspects are considered for the study, it is hard to control all of them. The InIvs are not tested on such uncontrollable aspects.

In the next chapter, the creation of students' affective state database is described. The created database contains data from various learning environments such as e-learning, classrooms, webinars, and flipped classrooms. The created database is also benchmarked with various existing algorithms.

Chapter 6

Database Creation

Unobtrusive automatic recognition of the students' engagement is a challenging task. Students' engagement is recognized either using their emotional or behavioral patterns. The emotional and behavioral patterns are recognized using the students' facial expressions, hand gestures, and body postures. From the existing literature, it is observed that the affective states are used to analyze the students' emotional engagement and engagement levels are used to analyze students' behavioral engagement. Intelligent tutoring system and smart classroom environments can be made more personalized using students' affective states and engagement levels prediction and analysis. This can be performed using machine or deep learning techniques. Effective recognition of affective states or engagement levels is mainly dependent on the quality of the database used. From the literature, there exists no standard database for students' affective state or engagement level recognition and its analysis in both e-learning and classroom environments.

The different learning environments considered in this study include e-learning & flipped classroom which contains a single person in a single image frame; classroom or webinar environments which contain image frames with multi-persons in a single image frame; and computer-enabled teaching laboratories which contains different types of occlusions (though it contains multi-persons in a single image frame). In this study, different types of students' engagement analysis are explored in different learning environments such as e-learning, flipped classroom, classroom, webinar, and computer-enabled teaching laboratories to analyze the best suited unobtrusive method for each environment. Hence, in the current study, a new database is created for affective state or engagement level recognition of students in different learning environments by considering:

- both single and multi-persons in a single image frame,
- both Ekman's basic emotions and learning-centered emotions for students' affective state analysis,
- behavioral patterns for students' engagement analysis,

- classification of students' affective states or engagement levels with object localization, and
- multimodality: students' facial expressions, hand gestures, and body postures are used for classification and localization.

The key contributions of this chapter are as follows:

- Created and analyzed a database for students' emotion engagement analysis using their affective states in different learning environments such as e-learning, classroom, webinar, and flipped classroom.
- Created and analyzed a database for students' behavioral engagement analysis using their engagement levels in classroom and computer-enabled teaching laboratories.

The rest of this chapter is organized as follows: Section 6.1 describes the database of students' affective states. Section 6.2 explains the created database for students' behavioral engagement. The performance evaluation of the created database is mentioned in Section 6.3. Finally, Section 6.4 summarizes the entire chapter.

6.1 Students' Affective States Database

In order to create an affective database for both e-learning and classroom environments, data which contains both single and multi-person in a single image frame is collected with their corresponding affective states. We also collected posed as well as spontaneous expressions to train the deep learning architecture, which facilitates better accuracy in classification and localization. The pre-informed or posed expressions are initially collected from 700 single-person image frames with Ekman's basic emotions, namely: happiness, surprise, delight, sadness, fear, and disgust, along with the neutral. Further, 50 video clips of 2 minutes each is collected for five learning-centered emotions of students, such as engaged, boredom, sleepy, frustrated, and confused. The basic emotions could be captured as images when they are posed, but learning-centered emotions are continual expressions which required the change in the medium to a video clip. Then, 1450 multi-person single image frames are collected for posed expressions where all the students are instructed to express a said emotion. Thus a total of 2900 image frames are collected with posed expressions; all these image frames with frontal posed expressions are annotated w.r.t. eleven affective states, along with the neutral.

For the students' spontaneous affective states analysis, we collected the classroom data with 350 students for more than 20 hours of classes where different subjects of computer science/information technology are taught. The total number of image frames obtained is more than 72000. For multi-person in a single image frame, we eliminated the duplicates, where duplicates are the frames with 80% or more similarity in facial expressions, hand gestures, and body postures from its previous image frame. Initially, annotations are performed manually. Once sufficient annotated image frames are obtained for training, we used the semi-automatic annotation process (Sekachev et al., 2019) to make the annotation process faster. Here, we used deep learning architectures for object localization and classification, where the deep learning architecture predicts the bounding box and the affective state. The annotator manually checks and corrects if the bounding boxes and classified affective state is not correct. For example, if the bounding box predicted by the deep learning architecture is not a tight bounding box, then the annotator adjusts it to make it close instead of drawing the bounding box from scratch for all the multitude of modalities present in that image frame.

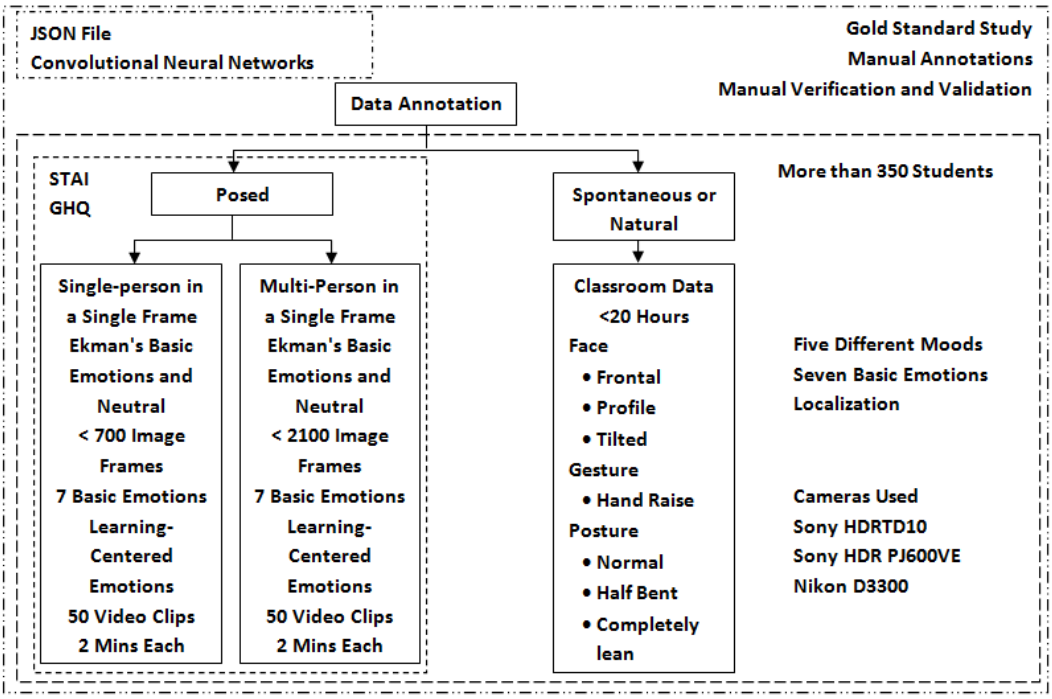


Fig. 6.1. Flow of database creation and its performance evaluation.

A pre-trained Convolutional Neural Network (CNN) based on GoogleNet architecture Krizhevsky et al. (2012) is used, and the results are manually verified for validation of detection and classification of affective states. Initially, 500 spontaneous random images from one hour class are taken and annotated manually using the standards referred

to under the gold standard study Sidney et al. (2005). Then, these annotated images are used to train the CNN model, thus detecting the multimodal features of remaining frames of the same one hour lecture. These image frames include faces, hand gestures, and body postures. For each of them, a bounding box is used for object localization, and the corresponding affective states are annotated. We also cropped the image frames whenever the students' visibility is too low. We collected one image frame for each basic emotion from the students' posed data. From two minutes video clip of learning-centered emotions, annotators manually considered the average among peak affective states of all the students present in that image frame, and accordingly select one image frame among them (in a few cases, the annotators selected 2 to 3 image frames as peak expression or patterns (all three frames have different expressions or patterns for the same affective state) of that affective state. But, this happened only in 7.27% of the entire database). For students' spontaneous affective states analysis, we manually annotated 1000 image frames and also manually verified 2000 CNN classified image frames. Also, a few basic emotions like fear, sadness, and disgust are less observed in the classroom environment; hence, we induced these less observed affective states in the classroom using the method mentioned in D'Mello et al. (2010). These induced affective states are present only in the first two hours of the collected classroom data.

Figure 6.1 shows the complete flow of database creation and its performance evaluation. It is observed from Figure 6.1 that the classification of eleven affective states (happiness, sadness, surprise, fear, disgust, anger, engaged, sleepy, boredom, frustrated and confused) including neutral for both posed and spontaneous data using face, hand gestures and body postures is considered. The detection and classification of faces include frontal, profile, and tilted faces; hand gestures include gestures associated with affective states and raised hands; body postures include normal or straight, bent over backward/forward, and lean completely on the desk. Further, we used cameras to capture the visible behavior. We manually annotated using the gold standard study and stored the same in a JSON file. We also used the proposed CNN architecture for database creation, verification, and validation. Figure 6.1 is explained in detail in the following subsections.

6.1.1 Camera Setup

Three different cameras are used for the collection of videos to generate the database, namely:

- *Sony HDR - TD10 Handycam*,
- *Sony HDR - PJ600VE Handycam* and
- *NIKON D-3300 DSLR Camera*.

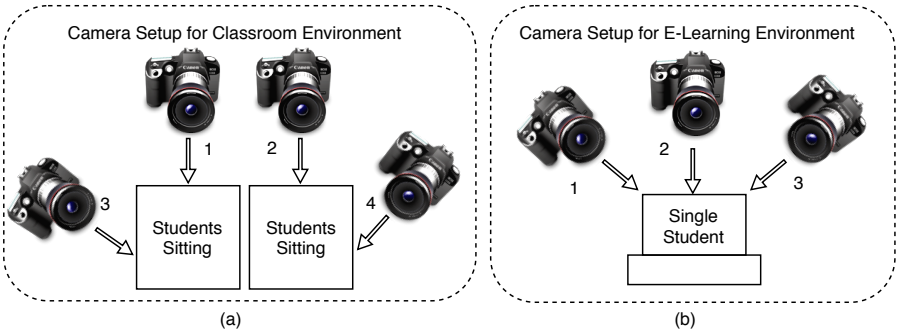


Fig. 6.2. Camera setup

The top view of camera setup for the classroom data collection is shown in Figure 6.2 (a). Most of the classroom data are collected using the Cameras 1 and 2. The Cameras 1 and 3 are used if there are less than 50 number of students in the classroom. The Cameras 1, 2, and 3 or 1, 2, and 4 are used if the number of students present in the classroom is below 100. All the four cameras shown in Figure 6.2 (a) are used if the number of students present in the classroom is more than 100. Depending on the lighting conditions and the visibility of students, these cameras are accordingly placed in the concerned classrooms.

For the posed expressions, students are made to sit in front of the camera, and they are asked to express the mentioned expressions, and the camera setup for the same is shown in Figure 6.2 (b). Camera 2 is used for most of the posed expressions; other cameras are used only for the data collection w.r.t. to multi-point view data.

For real-time data capturing in classroom and webinar environments, we used Sony HDR - TD10 Handycam and Sony HDR - PJ600VE Handycam cameras. In a few classroom and webinar environments, surveillance cameras are already installed, and the data is collected directly from it. The camera setup is as same as shown in Figure 6.2 (a). For real-time data capturing of e-learning and flipped classroom environments, we used laptops with web-camera and desktop C310 webcameras. The camera setup is the same as shown in Figure 6.2 (b) with only Camera 2.

6.1.2 Affective State Classification

The existing work includes recognizing the students' facial expressions and classifying them into Ekman's basic emotions (Ekman, 1992). The study investigated by Dhall et al. (2015) showed that learning-centered emotions such as flow, boredom, frustrated are more dominant and regularly observed in students. Whitehill et al. (2014) considered body postures and eye movements along with facial expressions, and classified the engagement into four different types (including both emotional and behavioral engagements). Analyzing only the students' emotion/affective state is not sufficient. Behavioral patterns such as looking away from the task, eyes barely open, complete lean on the desk are also important. Hence, in this study, we considered facial expressions as well as hand gestures and body postures for both basic and learning-centered emotions. Though not all the emotions are significantly observed in the teaching-learning process (emotions like fear, frustrated and disgust), those who are seldom observed are also considered under this study.

Labels and Definitions: Learning-centered emotions and the behavioral patterns of head and eye movements are considered as mentioned in Dhall et al. (2015) and Whitehill et al. (2014), respectively. But, these works are on a single person in a single image frame. Also, the behavioral patterns mentioned in Whitehill et al. (2014) did not include the body postures and hand gestures of multiple students in a single image frame. Further, a few emotions are recognized accurately using both hand gestures and facial expressions instead of using only facial expressions. Hence, in the proposed classification, we modified the existing classification standards by adding the hand gesture and the body posture components, but the standard label definitions remain the same, and the details are mentioned below.

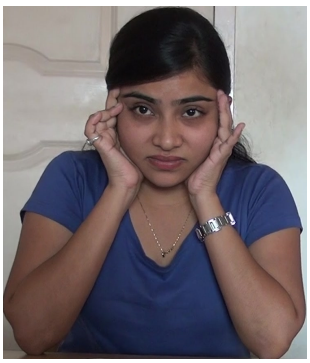
Ekman's basic emotions: Facial Action Coding System (FACS) is widely used to recognize the facial expressions, proposed by Ekman & Rosenberg (1997). Automatic Face Analysis (AFA) (Tian et al., 2001) further optimized the FACS to recognize the fine-grained changes in facial expressions. While creating the database, we used both AFA and FACS to classify the facial expressions into Ekman's basic emotions. For body postures and hand gestures, we followed the standards mentioned in Bimodal Face and Body Gesture (FABO) (Gunes & Piccardi, 2006).

Learning-Centered Emotions: We used the definitions mentioned in D'Mello et al. (2010) for learning-centered emotions, namely: Frustration- dissatisfaction or annoyance; Confusion- noticeable lack of understanding; Engaged/Flow- interest in the activity; Boredom- being weary or restless through lack of interest; Sleepy- extremely not interested and in a mental state of sleep; Neutral- no apparent emotion or feeling. The facial expressions, hand gestures, and body postures are classified as per the standards mentioned in D'Mello et al. (2007, 2010); Mota & Picard (2003). Since there are no standards for multimodal affective state prediction of students in the classroom environment, we followed the express, observe, and validate method where the students' express a particular emotion in their way and self annotate them. The expert annotators further affirmed the affective state and corrected if needed (Picard, 1997). A few sample image frames for the students' affective state are shown in Figure 6.3. The hand gestures such as hands near the head, neck resting on the hand, body shift, change orientation are observed in the students' affective states. The engaged affective state is predominantly observed in the classroom environment, and the impact of hand gestures and body postures is very high for the engaged affective state. From Figure 6.4a, it looks like the student is sleepy, when we consider the hand gesture and body posture (Figure 6.4b), it becomes evident that it belongs to the engaged affective state. Similarly, the facial features in Figure 6.4c look like happy, but the student is in the confused affective state (Figure 6.4d).

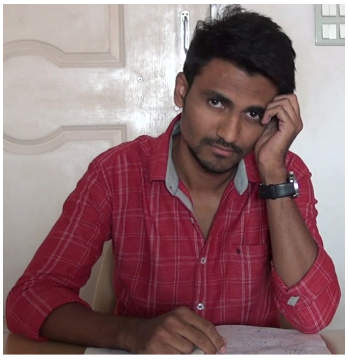
6.1.3 Annotation using Gold Standard Study

The affective states are classified using all the three components, namely: facial expressions, hand gestures, and body postures of the students. The entire process of affective state elicitation is performed using the gold standard study. The database is manually annotated separately for each person, and the boundary box plot is also manually performed separately for face, hand gestures, and body postures. The three gold standards considered for data annotation, are: participants, novice judges, and expert judges.

Participant annotators are the student participants who have self-annotated their affective states. On average, this self-annotation is performed after one week of data selection. Enough time gap is ensured for self-annotation to reduce biased and inaccurate annotations.



(a) Frustrated affective state.



(b) Bored affective state.

(c) Confused affective state..

(d) Surprised affective state..

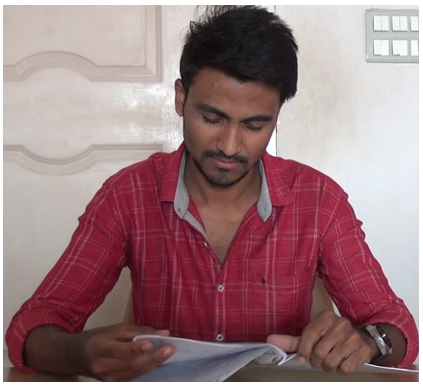
Fig. 6.3. Image snapshots of student's affective state with multimodality.

Novice judges are the students who are not trained for image data annotation but have annotated the students' affective states for the database.

Expert judges are the faculty members who annotate the affective states of the students' database using the standard guidelines mentioned in Bosch et al. (2016). Standard guidelines include FACS, AFA, BROMP, and FABO. The experts in the emotion recognition domain (includes psychologists) design these guidelines.

Labeling Process: The definitions of each affective state are provided to the labelers. Though all image frames are not annotated by all the labelers (more than 30 annotators), we made sure that each image frames are annotated by multiple labelers. The labelers are instructed to annotate images for "which affective state the student appears to be in" and not try to infer what is really going on in their mind. We used the standard annotation approach, where static images are viewed, and a single number is used to rate each image (1 to 11 corresponding to 11 affective states mentioned in Table 5.2). This method has the consequence that when the student blinks his(her) eyes, then that particular image frame is annotated as the sleepy affective state. But, we observed that these momentary affective states would not have a significant impact

(a) Student's facial expression with sleepy affective state.



(b) Student's multimodal data with engaged affective state.

(c) Student's facial expression with happy affective state.

(d) Student's multimodal data with confused affective state.

Fig. 6.4. Importance of hand gestures and body postures in affective state recognition.

when we average across all image frames over a period. Also, these static images are not in the streaming order as the labeler could get influenced by previous image frames and guess the students' affective states. Since the number of annotators is more than two, we used quadratic-weighted Cohen's κ, and the leave-one-labeler-out agreement as shown in Whitehill et al. (2014). The annotators reliably agree when discriminating against the recognized affective states with Cohen's $\kappa = 0.48$. In case of semi-automatic annotation process, the expert judges randomly picked the deep learning technique annotated documents from JSON file and cross-verified it using the standard guidelines. In this database annotation process, the self-annotation and the verification of expert judge annotations are not performed on the entire dataset of spontaneous expressions.

6.1.4 Storing Annotated Data

Object localization helps to focus on the candidate regions such as the student's face, hand gesture, and body postures for identifying the correct location of the target object. The object localization helps in student identification and better classification of their affective states (especially in spontaneous expressions data). We manually detected the students from the image. Object localization is performed by putting the bounding box on the face, body posture, and hand gesture separately for each student, as shown in Figure 6.5. Once the detection part is done, then the annotation/labeling is performed, and all the attributes shown in Figure 6.5 are stored in a JSON file. The purpose of using a JSON file over text or .csv file is that the JSON file format is lightweight, self-describing, and easy to understand data-interchange format. Further, as per the standard format of the annotation file, we also converted the JSON file into CVAT XML 1.1 format for storing the annotated data (Sekachev et al., 2019). Instead of storing the attributed of annotated data in the hierarchical format, the attributes such as affective states, bounding boxes are stored separately for each frame. The use of storing the data in the standard format is that the architectures used for the classification can use any data fusion techniques for the affective state classification such as decision-level, feature-level, score-level, image-level, confidence-level fusion and so on.

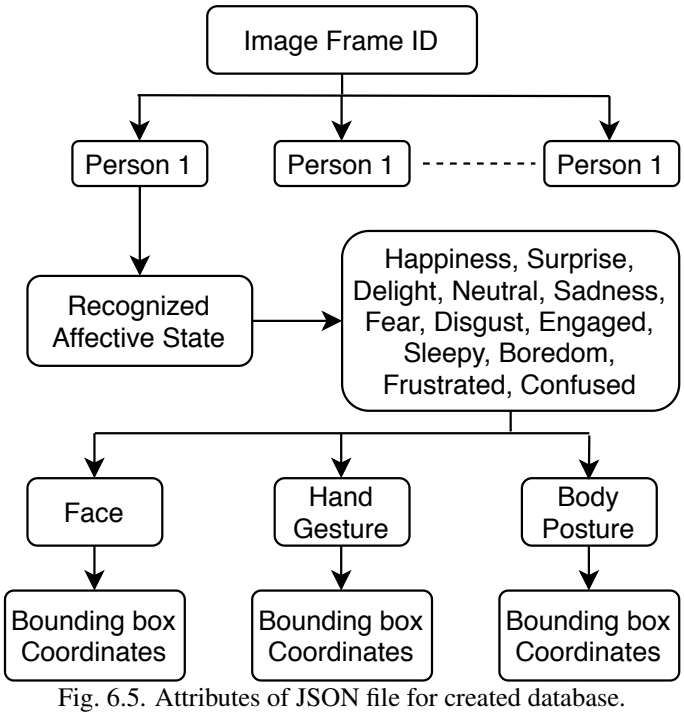


Fig. 6.5. Attributes of JSON file for created database.

Figure 6.5 shows the attributes present in the JSON file. A unique image frame id is generated for all the image frames present in the database. Every image frame contains the details of each person present in that image frame. A person's details include the person id, followed by the affective state and three different boundary box coordinates (four coordinate points for each boundary box) for a detected face, gesture, and body posture. If the posture and gesture of a person are not visible in that image frame, then only the face coordinates are stored, and the remaining will be filled with null values and vice versa.

(a) Single person with disgust posed expression.

(b) Multi-person with happiness posed expression.

(c) Uncropped image frame with spontaneous expressions.

(d) Cropped image frame with spontaneous expressions.

Fig. 6.6. Sample single frame images of created students' affective database.

6.1.5 Database Content

For the posed expressions, a video of one minute per person is taken with a posed affective state. From this video clip, we manually considered one image frame when the expression is at its peak. If there are multi-person in a single image frame, then we manually considered the average among the entire persons' peak affective state, and the

annotators accordingly select one image frame among them.

The created database with posed expressions contains the frontal face, profile face, and tilted face (upwards/downwards). Body posture is divided into a normal, half bent, and resting the upper body on the desk or complete lean pose. Hand gestures are significantly observed during the question answering sessions (when a student either wants to clarify a doubt or to answer a question).

A sample image of a single-person with posed expression using face, hand gesture, and body posture is shown in Figure 6.6 (a). Figure 6.6 (b) shows the sample image frame of multi-person in a single image frame with posed expressions corresponding to the happiness. In the creation of the database with posed expressions for multi-person in a single image frame, we avoided overlapping of persons to a maximum extent by adjusting the camera position. We also used spontaneous multi-person image data for training, even in case of overlapping or occlusions.

Spontaneous expressions are observed from classroom data where multi-person are present in a single image frame. Here, we collected more than 20 hours of data with 60 frames per second. This data includes different poses, gestures, postures, tilted, occluded, and posed faces corresponding to affective states (happiness, sadness, surprise, neutral, anger, fear, disgust, engaged, boredom, frustrated, sleepy and confused). An uncropped sample image is shown in Figure 6.6 (c), where we can observe an occluded face of the student on the first bench. Further, we observe that the image contains tilted faces, bent posture, etc. Figure 6.6 (d) shows the cropped image, which contains students for the major part of the image. We also performed manual cropping for those images where fewer students are present, and more of the classroom structures like walls, windows, and benches are detected to ensure better clarity of the students' features.

Further, we created a database with only faces for posed expressions by cropping the facial part of the image. Thus, we eliminated the hand gestures and body postures from the image frame. This face database is created with better clarity while comparing it with existing state-of-the-art face emotion detection techniques in the learning environment.

6.1.6 Variants of the Database

There are several variants considered for making the database more robust. Some of the major variants are as follows:

Occlusion: Our database consists of several images with occlusion for all the three components (face, hand gesture, and body posture). Some sample occluded images are shown in Figures 6.6 (a) and 6.6 (c). These occluded images are present in both single and multi-person in a single image frame. Further, these occluded images are present in this database with both posed and spontaneous expressions.

Background Clutter: In order to increase the detection accuracy during the feature selection process of posed images, we removed the background clutter for a few images where background clutter has occurred. Background cluttering is not performed on the image with spontaneous expressions.

Illumination: Proper lighting is a basic necessity while recording videos. We provided extra lighting/illumination for proper recognition of the face, for the posed expressions. But for the classroom scenario, no such extra lighting/illumination arrangements are used during the collection of data to make a real-time classroom scenario and also to make the data more robust for real-time affective state recognition environment. Since these video image frames of real-time classrooms are used for training and testing, the classroom database contains image frames with high, medium, and low light intensities.

Cultural and Regional Variations: Our database consists of students from different states with different regional and cultural backgrounds. Though out of the annotated students, 39% are from South India, 26% are from North India, 20% are from West India, and the remaining 15% are from East India, there is no significant difference in the way they express their facial expressions and behavioral patterns.

Intra-Class Variation: The number of participants for the classroom environment is 350, and these participants are from different regions and cultural backgrounds. Therefore, there exist several intra-class variations in the annotated image frames.

Image Cropping: We performed image cropping on various occasions mainly to increase the visibility of students and also to avoid the background (windows, fans, etc.). Further, we manually cropped only the face part for posed expressions and created

a separate face dataset, so that it can be utilized for verification and validation of any face recognition system.

Deformation: Our database contains deformed images w.r.t. body postures and gestures. These images are considered to reduce the data augmentation task required during the training phase of the classroom response system.

Pose: The single-person posed images are considered with different body postures or angles. Poses can be images with the face tilted to some degree, bent upwards and downwards, or can be profile photos. These variants in pose increase the robustness in detection and classification machine and deep learning architectures.

Multipoint View: Posed images are taken from all the three camera positions to facilitate better accuracy levels due to varied images used in the training phase. Further, a multipoint view image reduces the need for data augmentation during the training phase. Also, we obtained the data from different camera angles and image resolutions to make the database more robust.

6.1.7 Duplications

The image frames for classroom data with spontaneous expressions are collected at the rate of 60 fps. There is a chance that two successive frames may have more than 80% similarity, as shown in Figure 6.7. In order to avoid this, we considered the image frames with a gap of 5 frames. For multi-persons' posed expressions, we selected 2 to 5 image frames with peak affective state intensity where all the students are present in that single image frame. If the expressions or poses in those image frames are similar (more than 80%), then the redundant image frames are discarded. Thus, we avoided duplications.

6.2 Students' Engagement Level Database

The second database is created to analyze the students' behavioral patterns using their face, hand gestures, and body postures unobtrusively for computer vision techniques. Students' engagement is classified into four major engagement levels (ELs) as mentioned in Whitehill et al. (2014). The guidelines designed for engagement levels mentioned in Whitehill et al. (2014) are modified by adding the features of the facial expression, hand gesture, and body posture for multiple students in a single image frame.

Fig. 6.7. Two image frames with a gap of 60 frames with more than 80% similarity of face, hand gestures and body postures.

- EL 1: Not engaged at all - e.g., looking away from the tutor or board and obviously not thinking about the task, eyes completely closed etc.
- EL 2: Nominally engaged - e.g., eyes barely open, fully bent on the desk or the chair, no expression on the face, boredom, clearly not "into" the task.
- EL 3: Engaged in the task - a student requires no admonition to "stay on the task". Looking at the teacher/board, taking notes, listening, and discussions with the teacher etc.
- EL 4: Very engaged - a student could be "commended" for his/her level of engagement in the task.
- X: The clip/frame is very unclear, or contains no person at all.

6.2.1 Participants and Engagement Level Annotation

The entire proposed architecture is trained and tested on 350 undergraduate, postgraduate, and doctoral research students of National Institute of Technology Karnataka (NITK), Surathkal, Mangalore, India. These naturalistic expressions and body postures of students are collected for more than 10 hours from the classroom environment. All the classroom data has multiple students in a single frame, but the number of people in each frame may vary depending on the subjects of discussion and the class strength.

For the given engagement level definitions, manual annotation, verification, and validation are performed using the Gold standard study (Sidney et al., 2005). The same

steps are followed even for the engagement analysis, the camera setup, labeling process, annotation, storing annotated data, variants present in the image data and removal of duplicants remains the same. Instead of 12 different class labels used in the previous study, only four different class labels are used. The analyzed reliability among the labelers is found that they reliably agree with Cohen's $\kappa = 0.43$. Further, we also created a dataset with similar steps with three different class labels, namely: engaged, bored and neutral and the details including the performance evaluation is mentioned in Chapter 3 (4.1.4).

Subjects

The students present in this created database belong to the age group of 20 to 26 years. These students are undergraduate, postgraduate and doctoral research students from India with different cultural and regional backgrounds.

6.3 Performance Evaluation

We evaluated the created database with the state-of-the-art techniques for the detection, feature extraction, dimensionality reduction and classification of the face, hand gesture and body posture. The performance evaluation of students' emotional engagement database with 12 different class labels is given below. The performance evaluation related to behavioral engagement with four different class labels and engaged, bored and neutral class labels are explained in Chapter 4 (4.1.2,4.1.4) and Chapter 3 (3.2.1,3.2.5), respectively.

6.3.1 Facial Feature Extraction

Initially, the created database with single-face in a single image frame that contains tilted and occluded images is tested with Haar cascades (Viola & Jones, Viola & Jones) for face detection and got an accuracy of 73%, and an accuracy of 60% is obtained for multi-face in a single image frame. The snapshot of a sample image frame on Haar is applied for multi-face in a single image frame as shown in Figure 6.8. It is observed from Figure 6.8 that the detection accuracy is less. Even dark regions, objects on the clothes etc. are also recognized as the face. The detection accuracy of Haar cascades is poor due to the fact that it fails to recognize occluded and tilted faces.

Fig. 6.8. Face detection using Haar cascades.

The accuracy of expression recognition highly depends on the selection of appropriate features to represent the expressive face. We experimented with different types of feature extraction techniques as mentioned below.

Local Binary Pattern (LBP) (Ojala et al., 2002) recognizes the pattern from the given image by comparing the neighboring pixels and represents it in a binary format. The performance of LBP is better even for illumination variant images. We used LBP histograms, uniform LBP histograms, and rotational invariant uniform LBP patterns in this experiments.

Gabor Filters (Jain & Farrokhnia, 1991) are linear filters with frequency and representation similar to that of the human visual system. As different frequencies and orientations are used to extract features, we used 2D Gabor filters.

Local Gabor Binary Pattern (LGBP) (Zhang, Shan, Gao, Chen & Zhang, Zhang et al.) uses both LBP and Gabor filters. One of the major advantages of LGBP is its robustness towards both illumination variants and misalignments.

PHOG Descriptor (Bosch et al., 2007): Position of orientations are added to Histogram of Gradients (HOG) to get Pyramid Histogram of Gradients (PHOG), which is robust w.r.t. shape change, rotation and illumination variations.

We performed student-independent 10-fold cross validation (Bosch et al., 2016) on the entire database with cropped images of facial features and tested them among the techniques such as LBP, Gabor filters, PHOG & LGBP and the corresponding results are shown in Table 6.1.

Table 6.1
Accuracy of Different Feature Extraction Techniques

Feature	Average Accuracy in % for the Entire Dataset	Average Accuracy in % for the Single Face in a Frame Dataset
LBP	58.73	82.00
Gabor Filters	55.84	81.78
PHOG	47.15	68.00
LGBP	61.11	85.43

We also tested the entire database using machine learning algorithms which use LBP and Haar cascades for feature extraction, Principle Component Analysis (PCA) for dimensionality reduction and Support Vector Machine (SVM) for classification (Ashwin et al., 2015). Though, these systems perform better for single-person in a single image frame for frontal face data (as compared to the results mentioned in Table 6.1), they failed to perform better for the created database. We obtained an accuracy of less than 50% for the entire created database.

6.3.2 Classification of Expressions

The classification of expressions is performed using state-of-the-art deep learning technique i.e, convolutional neural networks (CNN) (Krizhevsky et al., 2012). The main benefit of using deep learning over other machine learning or computer vision techniques is that deep learning uses a cascade of nonlinear processing units for feature extraction & transformation and thus uses the raw features for detection and classification. It automatically generates the higher level features from lower level features to recognize the pattern. Further, it also generates a sparse connectivity which reduces the cost of matrix computation whereas, other techniques use hand-crafted features with no automatic generation of higher-level features from lower level features.

Hence, we used inception v3 model of CNN based on GoogleNet architecture Szegedy et al. (2016) for the recognition of affective states for the entire database. Inception v3 model consists of convolutional layers, average and max-pooling layers, dropout, fully connected layers and softmax classifier. RGB image data is the input for the training phase. As mentioned in (Szegedy et al., 2016), the process starts with the convolutional layer, where 3*3 convolutions are performed with 3 traditional inception modules at 1080*1080 with 288 filters each. Then, the activation function is used before pooling.

These three layers loop around for many iterations, occasionally a dropout layer follows to avoid overfitting. Finally, it is connected to the fully connected layer which classifies the given affective state using the final layer with the softmax classifier.

Data augmentation[1]: Since the data augmentation increases the accuracy, hence we included augmented images. A total of 24000 image frames are fed to training and testing phases of affective state classification. We used student-independent 10-fold cross validation as mentioned in Bosch et al. (2016) for the results obtained by the CNN using GoogleNet architecture.

Detection of Face, Hand Gesture and Body Posture for Affective State Classification: The data obtained from the fully connected layer is used to plot the boundary box coordinates (Redmon et al., 2016). The obtained results after student-independent 10-fold cross validation for detection of the face, hand gesture and body posture are shown in Table 6.2.

Table 6.2
Detection Results w.r.t. Face, Gesture & Posture

Performance Metrics	Face	Gesture	Posture
Accuracy	77.00	91.00	86.00
Recall	00.67	00.86	00.79
Precision	00.72	00.89	00.81
F1-score	00.74	00.91	00.83
MCC	00.70	00.83	00.80
AUC	00.71	00.82	00.79

It is observed from Table 6.2 that the detection accuracy of hand gesture is high. This is due to the fact that the hand raise is one of the hand gestures and it becomes easy for the CNN architecture to extract features even for the last bench students. We also obtained the results w.r.t. the detection of single and multi-person in a single image frame as shown in Table 6.3. The clear distinction of all the features makes single-person in a single image frame data more accurate than multi-person in a single image frame data.

Our database contains classes of different sizes. For example, in spontaneous expressions, the fear affective state is not frequently observed when compared to happiness or engaged class. Hence, we used the Matthews Correlation Coefficient (MCC) for

[1]Same data augmentation parameters used in Section 3.2.3 Chapter 3 are used in this study.

Table 6.3
Detection Results w.r.t. Single and Multi-Person

Performance Metrics	Single Person	Multi-Person
Accuracy	87.80	79.11
Recall	00.78	00.68
Precision	00.79	00.71
F1-score	00.83	00.74
MCC	00.71	00.63
AUC	00.77	00.67

analyzing the results for the database. MCC values range from -1 to +1 (worst-case to best-case) where the prediction rate versus the observed rate becomes the same. From Tables 6.2 and 6.3, it is observed that the MCC value is above 0.6. This shows that the prediction rate is high even with classes of different sizes. Further, a better prediction rate is ascertained from the area under the curve (AUC) from Tables 6.2 and 6.3.

Table 6.4
Classification Results for Different Affective States

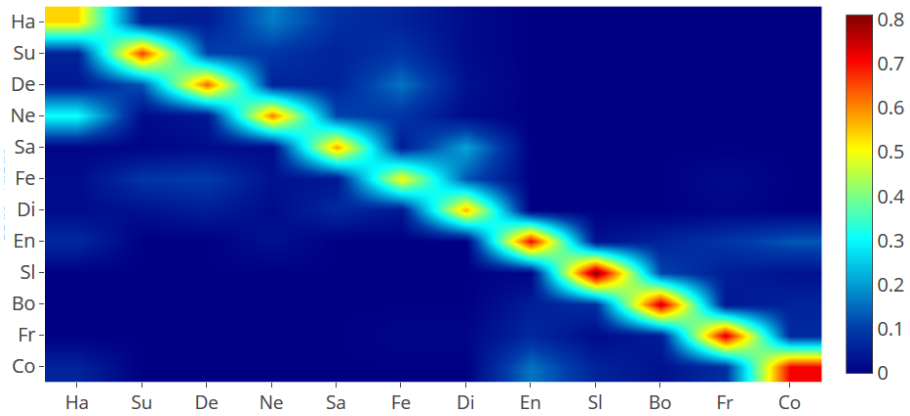
PM		**Ha**	**Su**	**De**	**Ne**	**Sa**	**Fe**	**Di**	**En**	**Sl**	**Bo**	**Fr**	**Co**
PD	**Accuracy**	0.82	0.84	0.82	0.80	0.84	0.82	0.86	0.85	0.84	0.83	0.79	0.77
	Recall	0.76	0.79	0.77	0.69	0.71	0.69	0.72	0.70	0.80	0.75	0.71	0.73
	Precision	0.78	0.81	0.77	0.71	0.73	0.74	0.73	0.71	0.81	0.79	0.74	0.76
	F1-Score	0.81	0.83	0.80	0.72	0.77	0.76	0.75	0.74	0.85	0.83	0.77	0.80
	MCC	0.69	0.75	0.71	0.69	0.65	0.68	0.67	0.64	0.71	0.74	0.68	0.71
	AUC	0.74	0.76	0.72	0.71	0.67	0.71	0.70	0.68	0.76	0.75	0.69	0.73
SD	**Accuracy**	0.71	0.63	0.65	0.70	0.57	0.59	0.58	0.61	0.71	0.67	0.63	0.55
	Recall	0.69	0.61	0.59	0.68	0.53	0.52	0.55	0.60	0.69	0.63	0.61	0.51
	Precision	0.75	0.66	0.61	0.72	0.55	0.55	0.57	0.60	0.69	0.66	0.65	0.54
	F1-Score	0.71	0.62	0.62	0.67	0.51	0.51	0.56	0.61	0.68	0.67	0.61	0.53
	MCC	0.69	0.53	0.63	0.70	0.43	0.53	0.51	0.63	0.69	0.66	0.60	0.54
	AUC	0.71	0.69	0.60	0.70	0.52	0.50	0.55	0.61	0.67	0.62	0.61	0.56

A: Accuracy; P: Precision; R: Recall; M: MCC; A:AUC
PM: Performance Metrics; PD: Posed Dataset (single and multi-person); SD: Spontaneous Dataset (multi-person); Ha: Happiness; Su: Surprise; De: Delight; Ne: Neutral; Sa: Sadness; Fe: Fear; Di: Disgust; En: Engaged; Sl: Sleepy; Bo: Boredom; Fr: Frustrated; Co: Confused.

Classification of Affective States: The classification results of 12 affective states (including neutral) for posed and spontaneous data are shown in Table 6.4. It is observed from the results that the accuracy, precision, recall, and F1-score values are better for posed data when compared to spontaneous data. This is because the posed dataset contains an almost equal number of image frames for each affective states. Whereas the spontaneous data contains different number of image frames for each affective state.

For e.g., engaged affective state is observed more compared to fear affective state and hence, the number of image frames for each affective state is different. Hence, we performed MCC on spontaneous data and the results suggest that the performance of the proposed model is better even for datasets of different number of image frames.

Figure 6.9 shows the heatmap generated from the confusion matrix. It is observed from heatmap that the affective state recognition is more accurate for surprise and less accurate for fear and sadness affective states. Reason for the high accuracy for surprise affective state is that it has unique facial features which are recognized easily when compared to happiness and neutral affective states with almost similar features of the face, hand gestures and body postures.



Ha: Happiness; Su: Surprise; De: Delight; Ne: Neutral; Sa: Sadness;
Fe: Fear; Di: Disgust; En: Engaged; Sl: Sleepy; Bo: Boredom;
Fr: Frustrated; Co: Confused.

Fig. 6.9. Heatmap for confusion matrix.

Similarly, the sleepy affective state has high accuracy due to unique features of the face, hand gesture and body posture. Whereas sadness affective state has similar features of boredom and hence this affective state has less accuracy. The engaged affective state gets misclassified with happiness and neutral. Frustrated & boredom, surprise & delight are misclassified with each other. As already mentioned, CNN requires sufficient training data for better performance, frustrated, fear and sadness affective states are less likely to appear in classroom scenario and hence we have less accuracy/recall; only posed expression contains frustrated, fear and sadness affective states but students seldom express these affective states in the classroom environment.

The results of overall detection and classification accuracy for affective state recognition using the CNN model based on the GoogleNet architecture are shown in Table 6.5. The detection and classification accuracies mentioned in Table 6.5 are computed

based on the average of single person posed data, multi-person posed data and spontaneous expressions data. After student-independent 10-fold cross validation, we obtained an accuracy of 88% for single person posed data, 79% for multi-person posed data and 61% for spontaneous expressions data.

Table 6.5
Overall Results of Detection and Classification

Performance Metrics	Detection	Classification
Average Accuracy	0.83	0.76
Average Recall	0.76	0.73
Average Precision	0.80	0.76
Average F1-score	0.82	0.79
MCC	0.75	0.65
AUC	0.77	0.72

Affective State Recognition of Single and Multi-person: We experimented the created database separately for the single and multi-person in a single image frame and the results of detection and classification accuracy w.r.t. each fold are shown in Figure 6.10. Similarly, Figure 6.11 shows the results of detection and classification accuracy w.r.t. face, hand gesture and body posture.

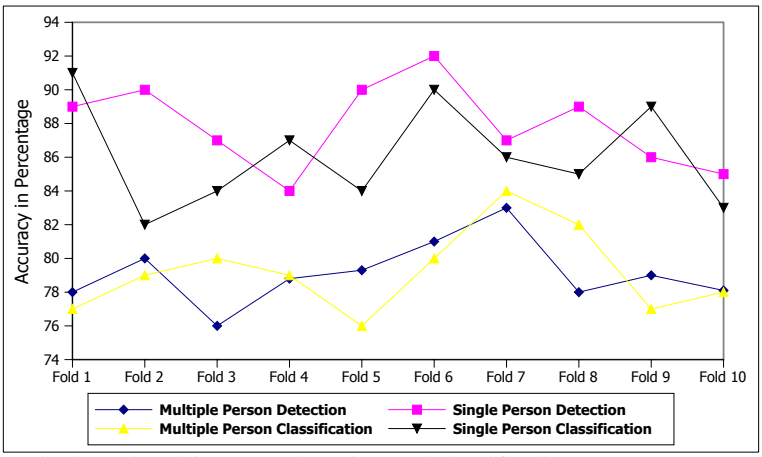


Fig. 6.10. Single and multi-person detection and classification accuracy w.r.t. each fold.

It is observed from Figures 6.10 and 6.11 that detection and classification accuracy of single-person (e-learning scenario) is high since those image frames are completely visible with the face, hand gesture, and body posture. Whereas in the multi-person (classroom environment), the face is the only visible part for all the students in the entire class. This is due to various variants (mentioned in Section 6.1.6) which may not be properly trained for that particular affective state and also all the faces, hand gestures and body postures are not visible for all the students in a classroom scenario. It is also

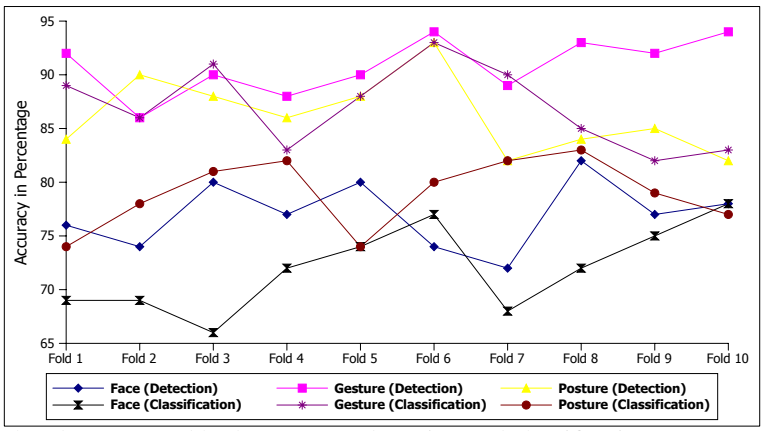


Fig. 6.11. Face, hand gesture and body posture's detection and classification accuracy w.r.t. each fold.

observed from Figures 6.10, 6.11 and Table 6.1 that the CNN performs better than any other feature extraction technique interms of detection and classification accuracy for single-person (e-learning scenario).

The overall classification accuracy of affective states w.r.t. face, hand gesture and body posture is shown in Figure 6.12.

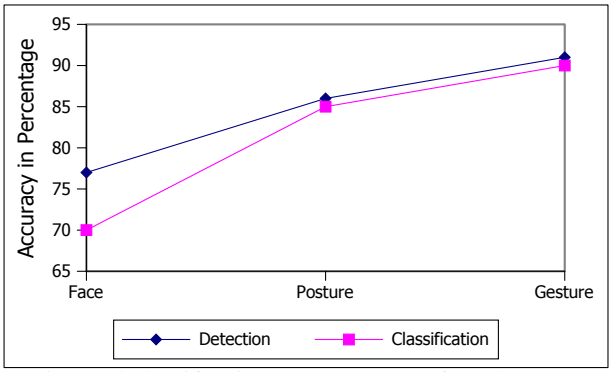


Fig. 6.12. Overall detection and classification accuracy w.r.t. face, hand gesture and body posture.

Dataset Dissemination

This created database can be accessed currently by approaching the authors. Once the journal papers gets accepted, it will be disseminated to the users by signing in the EULA (End-User License Agreement) through the link hccg.nitk.ac.in.

Ethics Statement and Acknowledgment

The experimental procedure, participants, and the course contents used for the all the experiments conducted and mentioned in this thesis are approved by the Institutional Ethics Committee (IEC) of NITK Surathkal, Mangalore, India. The participants were also informed that they had the right to quit the experiment at any time. The video

recordings of the subjects were included in the experiment only after they gave written consent for the use of their videos for this research experiment. All the subjects were also agreed to use their facial expressions, hand gestures, and body postures for all the process involved in the completion of the entire project.

The authors wish to thank undergraduates (B.Tech.), postgraduates (M.Tech.), and doctoral research (Ph.D.) students of Department of Information Technology, National Institute of Technology Karnataka Surathkal, Mangalore, India for their voluntary help for creating the student engagement database in learning environments.

6.4 Summary

The multimodal affective database for both e-learning (single student in a single image frame) and classroom environments (multiple students in a single image frame) is successfully created using the students' facial expressions, hand gesture, and body posture. Both posed and spontaneous expressions are collected to make the training set more robust. Also, various image variants are considered during the dataset creation. Annotations are performed using the gold standard study for 11 different affective states, including neutral. The annotators reliably agree when discriminating against the recognized affective states with Cohen's κ = 0.48. Further, we created a behavioral engagement level dataset and a dataset with students' engaged, bored and neutral affective states. Object localization is performed on each modality of every student, and the bounding box coordinates are stored along with the affective state. We obtained an accuracy of 83% and 76% for the detection and classification of affective state classification. Further, we cropped only the face part of the image and thus created the dataset with posed facial expressions for both Ekman's basic emotions and learning-centered emotions for better comparison with other datasets.

One of the major limitations of this study is that the majority of students present in the created dataset are Indians. Hence, the working of the deep learning architectures for emotion classification and localization may differ when we test on other than Indian students. The next chapter, concludes the research work presented in this thesis with future directions.

Chapter 7

Conclusions and Future Directions

7.1 Conclusions

Pervasive intelligent learning environments can be made more personalized by adapting the teaching strategies according to the students' emotional and behavioral engagements. The students' engagement analysis helps to foster those emotions and behavioral patterns that are beneficial to learning, thus improving the effectiveness of the teaching-learning process. Unobtrusive students' engagement analysis is performed using the students' non-verbal cues such as facial expressions, hand gestures, and body postures. Though there exist several techniques for classifying the engagement of a single student present in a single image frame, there are limited works on the students' engagement analysis in different learning environments such as the classroom, webinar, flipped classroom environments, and so on. Further, there exists no standard database with students' affective states in different learning environments. Classification of students' affective states in real-time and use it as feedback to enhance the teaching-learning process is also not explored in the literature. Hence, in this work, we addressed these issues, and the details of the contributions are given below.

The first set of contributions address the issues related to students' emotions engagement analysis, where the multi-facial emotion recognition from facial expressions are addressed using the proposed algorithms based on Modified Affine Transformation and Viola Jones based Haar Cascades. Experimental results demonstrate that our proposed algorithm outperforms the existing Viola-Jones algorithm by 6% for YALE, FDDB, and top 25 Google's searched "tilted face"' datasets. Video affective content analysis is performed using both audio and visual features using SVM and RBM classifiers. From experimental results, it is observed that our proposed hybrid SVM-RBM classifier performs better than individual SVM and RBM classifiers for audio-visual emotion recognition with annotated data. Also, a deep learning-based hybrid CNN model is proposed to predict the students' affective states, such as bored and engaged in the classroom environment. The multimodal analysis is performed using the students' facial expressions, hand gestures, and body posture to increase the robustness of the method. Since

the classroom image frame data contains multiple students in every image, a group engagement score is predicted for image frame data using the feature fusion technique. Further, a CNN based architecture is proposed to test in e-learning, flipped classroom, classroom, and webinar environments for 12 different students' affective states. After performing student-independent 10-fold cross-validation, we obtained an accuracy of 77% for the students' affective state classification and 79% for object localization. As an enhancement to the proposed work, a few less frequently observed learning-centered emotions like Eureka, and contempt can also be considered to analyze its incidence and temporal dynamics during the teaching-learning process. One of the limitations of this study is that the combination of the emotion is not tracked as only the affective state prima facie will be judged, and the cognitive aspects are not completely considered. For example, a student may be sad but engaged. This will be classified under sadness and not engaged.

The second set of contributions address the issues related to students' behavioral engagement. The students' behavioral engagement analysis method is proposed and implemented in the classroom environment using their facial expressions, hand gestures, and body postures. The proposed scale-invariant context assisted single-shot CNN architecture performed well for multiperson in a single image frame. It is also observed that the results are better for multimodality than single modality. We could recognize most of the students in the wild and predict four different behavioral engagement levels. A single group engagement level score for each frame is obtained using the proposed feature fusion technique. We performed student-independent 10-fold cross-validation, where the students present in training data are not present in the test image frames and thus obtained a mAP of 73.22%. The use of feature pyramid and context-sensitive layers in proposed architecture enhanced its performance leading to outperform the existing state-of-the-art architectures such as Inception and Hyperface. Even after the addition of feature pyramid and context-sensitive layers, the proposed method is able to classify the engagement levels with a predict time of 2153ms per frame. The proposed multimodal analysis outperformed the popular survey-based methods (NSSE, and AUSSE) for student engagement analysis by showing a positive correlation between behavioral engagement and test performance. Further, frequency, temporal dynamics, and distribution of the engagement levels are analyzed. The proposed method is also tested on

the classroom subset of the ImageNet database. Video surveillance-based students' behavioral engagement analysis is also proposed and implemented in computer science and information technology teaching laboratories. There is a positive correlation between students' engagement and learning. Also, this engagement analysis system outperformed the existing survey-based engagement analysis systems. This study limits only to the classrooms and computer-enabled teaching laboratories, but various other asynchronous learning environments such as m-learning, collaborative e-learning, social e-learning are not explored.

The third set of contributions address the issues related to the use of recognized students' affective state as feedback to enhance the teaching-learning process in real-time. Here a real-time students' engagement analysis is proposed for both the individual and group of students based on their facial expressions, hand gestures and body posture for e-learning, flipped classroom, classroom, and webinar environments. The proposed model effectively classifies the students' affective states into teacher-centric attentive and in-attentive affective states. To address the negative impact of in-attentive affective states on the performance of students, InIvs are proposed. The proposed method with multi-person detection, multi-modality, group engagement score, inquiry intervention, and with an classification accuracy of 0.77 for a test data of 350 students, outperforms the existing methods. Experimental results demonstrated that there is a positive correlation between the students learning rate and their attentive affective state engagement score for both individual and group of students. The proposed affective state transition diagram and appropriate visualizations helped students and the faculty members to improve the teaching-learning process.

Though there are various teaching strategies adapted, it is not possible to put an end to distraction in the learning environments which include students bending down to use cellular phones, turning to pass notes or any such activities where the face may not be detected clearly. This limits the accuracy of affective state classification. In a dynamic learning environment, there will be various aspects to study in order to implement the proposed methodology. Even if a few such aspects are considered for the study, it is hard to control all of them. The InIvs are not tested on such uncontrollable aspects.

The fourth set of contributions is related to the creation of the affective database with students' facial expressions, hand gestures, and body postures in different learning environments. The multimodal affective database for both e-learning (single student in a single image frame) and classroom environments (multiple students in a single image frame) is successfully created using the students' facial expressions, hand gestures, and body postures. Both posed and spontaneous expressions are collected to make the training set more robust. Also, various image variants are considered during the dataset creation. Annotations are performed using the gold standard study for 11 different affective states, including neutral. The annotators reliably agree when discriminating against the recognized affective states with Cohen's κ = 0.48. Further, we created a behavioral engagement level dataset and a dataset with students' engaged, bored, and neutral affective states. Object localization is performed on each modality of every student, and the bounding box coordinates are stored along with the affective state. We obtained an accuracy of 83% and 76% for the detection and classification of affective state classification. Further, we cropped only the face part of the image. We thus created the dataset with posed facial expressions for both Ekman's basic emotions and learning-centered emotions for better comparison with other datasets. One of the limitations of this study is that the majority of students present in the created dataset are Indians. Hence, the working of the deep learning architectures for emotion classification and localization may differ when we test on other than Indian students.

In summary, we proposed different architectures to classify the students' affective states in different learning environments for both emotional and behavioral engagement analysis. The multitude of modalities, group engagement analysis, and object localization for each student is considered in the study. Different databases are created and benchmarked, such as (i) students' affective state database for 11 different learning-centered emotions and the neutral; (ii) students' behavioral engagement dataset with four different engagement levels, and (iii) students' affective state database in the classroom environment with bored and engaged affective states. Real-time students' affective state classification is used as feedback to the teacher to perform automatic inquiry interventions and enhance the teaching-learning process. Various transition diagrams and visualizations are proposed to better understand the students' engagement in different learning environments such as e-learning, classroom, webinar, and flipped class-

room environments. Experimental results demonstrate that our proposed algorithms outperform the state-of-the-art techniques.

The gaps mentioned in Chapter 2.6 are addressed through the following contributions as summarized in Table 7.1.

Table 7.1
Overall Results of Detection and Classification

Gaps	Contributions
Data from large classrooms, webinars, and flipped classroom environments	The created students' database considering all these environments
Multimodal emotion recognition for multiple students in a single image frame	Automatic detection of the students' emotion in image frames obtained from the classroom environments
Image/video frame based group engagement analysis or group level score prediction using multiple students in a single image frame	Group engagement analysis using feature fusion and Atkinson's index
Behavioral engagement of the students in the classroom environment.	Students' behavioural engagement in classroom environment using four different engagement levels
The students' engagement analysis in computer-enabled teaching laboratories	Students' behavioural engagement using four different engagement levels in computer-enabled teaching laboratories
Adapting the teaching strategy based on the students' affective states	Automated InIvs using students affective states
Multitude of modalities for each student present in the image frame along with the object localization	Intra-image multimodality with object localization for the entire created database

7.2 Future Directions

The research work conducted in this study has several future directions to improve the current work and the details are mentioned below.

Additional factors such as time and context can be explored for better engagement analysis of students. Existing time-aware and context-aware recommendation systems architectures can be explored and finetuned for its applicability in the students' affective state prediction with additional parameters such as time and context. The temporal dynamics of the students' affective states can be explored using the framework which considers the previous image frame data as a memory (architectures such as RNN, LSTM, and so on).

Mapping of a teacher's affective states & behavior with the students' affective states & behavior and analyzing the impact of the teacher's verbal and non-verbal cues to improve the effectiveness of the teaching-learning process can be explored. Here, both audio and visual data need to be processed. And the issue of synchronizing the cameras (both capturing the student and the teacher) along with the voice of both the students' and the teacher is challenging.

Object localization and student identification can be introduced to enhance the auto-tutors' performance by making the teaching-learning process more personalized. The auto-tutors are currently confined only to the traditional e-learning environment. Social and collaborative e-learning contains image frame data with more than one student in a single image frame. In this context, it is necessary to differentiate each student and classify their affective states using their multitude of modalities, object localization, and student identification.

The data obtained from the proposed method can also be used to map various student assessments such as diagnostic, formative, and summative assessments and to check for possible correlation among them. The predicted students' affective states can be mapped to the students' understanding using the various student assessments to get an automated assessment of each course w.r.t. course plan, course objective, programming objective, and so on. This can also be used to effectively analyze the students' understanding within the academic learning time.

The proposed method can also be explored for different users to know the user experience in various other domains such as entertainment, marketing, healthcare, and so on. A few affective states considered in this study, such as engaged, frustrated, confusion, boredom, and so on, are also observed in other domains. Finetuning of the proposed architecture is required as different features are observed at various other domains (For example, user affective state prediction in a shopping mall should consider the person who is standing or walking (which cannot be trained using the created database)). Still, the affective state features from facial expression remain informative for such predictions as well.

Testing the proposed models for the engagement analysis of students in a smart campus where collaborative and social learning platforms are adopted, and an unobtrusive

students' engagement analysis can be used to make intelligent tutoring systems more personalized in these platforms. For such platforms, the deep learning architecture may run on the local machine and hence, should be a lightweight application that consumes less bandwidth and installation space.

More annotations can be made to the created database to identify the students with gender, demography, and so on. The annotations can also be performed for context, time, course, instructor, year of study, and so on, such that more precise and possible correlations can be explored.

Different feedback mechanisms can be tested based on the real-time students' affective state classification. Active learning has intellectually, socially, and physically active learning strategies. Inquiry activities are one of the instructional strategies related to intellectually active learning. Different active learning stratergies other than intellectually active learning can be automated and explored for enhancing the teaching-learning process.

The accuracy of the students' affective state classification and localization can be improved using various low-level, mid-level, and high-level feature construction methods. The image frame-based data can be used to analyze the temporal and spacial features to enhance the affective state prediction accuracy using various state-of-the-art deep learning techniques.

References

Ahlfeldt, S., Mehta, S., & Sellnow, T. (2005). Measurement and analysis of student engagement in university classes where varying levels of pbl methods of instruction are in use. *Higher Education Research & Development*, *24*(1), 5–20.

Alameda-Pineda, X., Staiano, J., Subramanian, R., Batrinca, L., Ricci, E., Lepri, B., Lanz, O., & Sebe, N. (2016). Salsa: A novel dataset for multimodal group behavior analysis. *IEEE transactions on pattern analysis and machine intelligence*, *38*(8), 1707–1720.

Alizadeh, S. & Fazel, A. (2017). Convolutional neural networks for facial expression recognition. *arXiv preprint arXiv:1704.06756*.

Andriluka, M., Pishchulin, L., Gehler, P., & Schiele, B. (2014). 2d human pose estimation: New benchmark and state of the art analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3686–3693.

Arroyo, I., Cooper, D. G., Burleson, W., Woolf, B. P., Muldner, K., & Christopherson, R. (2009). Emotion sensors go to school. In *AIED*, volume 200, 17–24.

Arroyo, I., Woolf, B. P., Burelson, W., Muldner, K., Rai, D., & Tai, M. (2014). A multimedia adaptive tutoring system for mathematics that addresses cognition, metacognition and affect. *International Journal of Artificial Intelligence in Education*, *24*(4), 387–426.

Ashwin, T. & Guddeti, R. M. R. (2018). Unobtrusive students' engagement analysis in computer science laboratory using deep learning techniques. In *2018 IEEE 18th International Conference on Advanced Learning Technologies (ICALT)*, 436–440. IEEE.

Ashwin, T. & Guddeti, R. M. R. (2019a). Automatic detection of students' affective states in classroom environment using hybrid convolutional neural networks. *Education and Information Technologies*, 1–29.

Ashwin, T. & Guddeti, R. M. R. (2019b). Unobtrusive behavioral analysis of students in classroom environment using non-verbal cues. *IEEE Access*, *7*, 150693–150709.

Ashwin, T., Jose, J., Raghu, G., & Reddy, G. R. M. (2015). An e-learning system with multifacial emotion recognition using supervised machine learning. In *2015 IEEE Seventh International Conference on Technology for Education (T4E)*, 23–26. IEEE.

Balaam, M., Fitzpatrick, G., Good, J., & Luckin, R. (2010). Exploring affective technologies for the classroom with the subtle stone. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1623–1632. ACM.

Baveye, Y., Chamaret, C., Dellandréa, E., & Chen, L. (2017). Affective video content analysis: A multidisciplinary insight. *IEEE Transactions on Affective Computing*, *9*(4), 396–409.

Bian, C., Zhang, Y., Yang, F., Bi, W., & Lu, W. (2018). Spontaneous facial expression database for academic emotion inference in online learning. *IET Computer Vision*, *13*(3), 329–337.

Bodily, R. & Verbert, K. (2017). Review of research on student-facing learning analytics dashboards and educational recommender systems. *IEEE Transactions on Learning Technologies*, *10*(4), 405–418.

Bonwell, C. C. & Eison, J. A. (1991). Active learning: Creating excitement in the classroom. 1991 ashe-eric higher education reports. ERIC.

Booth, B. M., Ali, A. M., Narayanan, S. S., Bennett, I., & Farag, A. A. (2017). Toward active and unobtrusive engagement assessment of distance learners. In *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*, 470–476. IEEE.

Bosch, A., Zisserman, A., & Munoz, X. (2007). Representing shape with a spatial pyramid kernel. In *Proceedings of the 6th ACM international conference on Image and video retrieval*, 401–408. ACM.

Bosch, N., D'mello, S. K., Ocumpaugh, J., Baker, R. S., & Shute, V. (2016). Using video to automatically detect learner affect in computer-enabled classrooms. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, *6*(2), 17.

Brown, B. W. & Saks, D. H. Measuring the effects of instructional time on student learning: Evidence from the beginning teacher evaluation study. *American Journal of Education*, *94*(4), 480–500.

Burkert, P., Trier, F., Afzal, M. Z., Dengel, A., & Liwicki, M. (2015). Dexpression: Deep convolutional neural network for expression recognition. *arXiv preprint arXiv:1509.05371*.

Burnik, U., Zaletelj, J., & Košir, A. (2017). Video-based learners' observed attention estimates for lecture learning gain evaluation. *Multimedia Tools and Applications*, 1–24.

Calvo, R. A. & D'Mello, S. (2010). Affect detection: An interdisciplinary review of models, methods, and their applications. *IEEE Transactions on affective computing*, *1*(1), 18–37.

Castellano, G., Kessous, L., & Caridakis, G. (2008). Emotion recognition through multiple modalities: face, body gesture, speech. In *Affect and emotion in human-computer interaction* 92–103. Springer.

Castellanos, J., Haya, P. A., & Urquiza-Fuentes, J. (2017). A novel group engagement score for virtual learning environments. *IEEE Transactions on Learning Technologies*, *10*(3), 306–317.

Chi, M., VanLehn, K., Litman, D., & Jordan, P. (2011). Empirically evaluating the application of reinforcement learning to the induction of effective and adaptive pedagogical strategies. *User Modeling and User-Adapted Interaction*, *21*(1-2), 137–180.

Coffrin, C., Corrin, L., de Barba, P., & Kennedy, G. (2014). Visualizing patterns of student engagement and performance in moocs. In *Proceedings of the fourth international conference on learning analytics and knowledge*, 83–92. ACM.

Conati, C. (2002). Probabilistic assessment of user's emotions in educational games. *Applied artificial intelligence*, *16*(7-8), 555–575.

Craig, S., Graesser, A., Sullins, J., & Gholson, B. (2004). Affect and learning: an exploratory look into the role of affect in learning with autotutor. *Journal of educational media*, *29*.

DeFalco, J. A., Rowe, J. P., Paquette, L., Georgoulas-Sherry, V., Brawner, K., Mott, B. W., Baker, R. S., & Lester, J. C. (2018). Detecting and addressing frustration in a serious game for military training. *International Journal of Artificial Intelligence in Education*, *28*(2), 152–193.

Dehghan, A., Ortiz, E. G., Shu, G., & Masood, S. Z. (2017). Dager: Deep age, gender and emotion recognition using convolutional neural network. *arXiv preprint arXiv:1702.04280*.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.

Dhall, A., Goecke, R., & Gedeon, T. (2015). Automatic group happiness intensity analysis. *IEEE Transactions on Affective Computing*, *6*(1), 13–26.

Dhall, A., Goecke, R., Lucey, S., & Gedeon, T. (2011). Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark. In *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, 2106–2112. IEEE.

Dhall, A., Joshi, J., Sikka, K., Goecke, R., & Sebe, N. (2015). The more the merrier: Analysing the affect of a group of people in images. In *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, volume 1, 1–8. IEEE.

Dhamija, S. (2017). Learning based visual engagement and self-efficacy. In *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*, 581–585. IEEE.

Dhamija, S. & Boult, T. E. (2017). Automated mood-aware engagement prediction. In *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*, 1–8. IEEE.

Ding, C. & Tao, D. (2015). Robust face recognition via multimodal deep face representation. *IEEE Transactions on Multimedia*, *17*(11), 2049–2058.

Ding, C. & Tao, D. (2018). Trunk-branch ensemble convolutional neural networks for video-based face recognition. *IEEE transactions on pattern analysis and machine intelligence*, *40*(4), 1002–1014.

D'Mello, S. (2012). Monitoring affective trajectories during complex learning. In *Encyclopedia of the Sciences of Learning* 2325–2328. Springer.

D'mello, S. & Graesser, A. (2012). Autotutor and affective autotutor: Learning by talking with cognitively and emotionally intelligent computers that talk back. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, *2*(4), 23.

D'Mello, S., Picard, R. W., & Graesser, A. (2007). Toward an affect-sensitive autotutor. *IEEE Intelligent Systems*, *22*(4).

D'Mello, S. K., Lehman, B., & Person, N. (2010). Monitoring affect states during effortful problem solving activities. *International Journal of Artificial Intelligence in Education*, *20*(4), 361–389.

Duncan, D., Shine, G., & English, C. (2016). Facial emotion recognition in real time.

Edwards, S. (2015). Active learning in the middle grades. *Middle School Journal*, *46*(5), 26–32.

Eison, J. (2010), Using active learning instructional strategies to create excitement and enhance learning.

Ekman, P. (1992). An argument for basic emotions. *Cognition & emotion*, *6*(3-4), 169–200.

Ekman, P. & Rosenberg, E. L. (1997). What the face reveals: Basic and applied studies of spontaneous expression using the facial action coding system (facs). Oxford University Press, USA.

Ezen-Can, A., Grafsgaard, J. F., Lester, J. C., & Boyer, K. E. (2015). Classifying student dialogue acts with multimodal learning analytics. In *Proceedings of the Fifth International Conference on Learning Analytics And Knowledge*, 280–289. ACM.

Fan, Y., Lu, X., Li, D., & Liu, Y. (2016). Video-based emotion recognition using cnn-rnn and c3d hybrid networks. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, 445–450. ACM.

Filntisis, P. P., Efthymiou, N., Koutras, P., Potamianos, G., & Maragos, P. (2019). Fusing body posture with facial expressions for joint recognition of affect in child-robot interaction. *arXiv preprint arXiv:1901.01805*.

Garber-Barron, M. & Si, M. (2012). Using body movement and posture for emotion detection in non-acted scenarios. In *2012 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, 1–8. IEEE.

Gavrilescu, M. (2015). Recognizing emotions from videos by studying facial expressions, body postures and hand gestures. In *Telecommunications Forum Telfor (TELFOR), 2015 23rd*, 720–723. IEEE.

Georghiades, A., Belhumeur, P., & Kriegman, D. (1997). Yale face database. *Center for computational Vision and Control at Yale University, http://cvc. yale. edu/projects/yalefaces/yalefa*, *2*.

Girshick, R. (2015). Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, 1440–1448.

Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 580–587.

Glowinski, D., Dael, N., Camurri, A., Volpe, G., Mortillaro, M., & Scherer, K. (2011). Toward a minimal representation of affective gestures. *IEEE Transactions on Affective Computing*, *2*(2), 106–118.

Grafsgaard, J. F., Wiggins, J. B., Boyer, K. E., Wiebe, E. N., & Lester, J. C. (2013). Embodied affect in tutorial dialogue: student gesture and posture. In *International Conference on Artificial Intelligence in Education*, 1–10. Springer.

Grafsgaard, J. F., Wiggins, J. B., Vail, A. K., Boyer, K. E., Wiebe, E. N., & Lester, J. C. (2014). The additive value of multimodal features for predicting engagement, frustration, and learning during tutoring. In *Proceedings of the 16th International Conference on Multimodal Interaction*, 42–49. ACM.

Grann, J. & Bushway, D. (2014). Competency map: Visualizing student learning to promote student success. In *Proceedings of the fourth international conference on learning analytics and knowledge*, 168–172. ACM.

Gunes, H. & Piccardi, M. (2006). A bimodal face and body gesture database for automatic analysis of human nonverbal affective behavior. In *18th International Conference on Pattern Recognition (ICPR'06)*, volume 1, 1148–1153. IEEE.

Guo, X., Zhu, B., Polanía, L. F., Boncelet, C., & Barner, K. E. (2018). Group-level emotion recognition using hybrid deep models based on faces, scenes, skeletons and visual attentions. In *Proceedings of the 2018 on International Conference on Multimodal Interaction*, 635–639. ACM.

Gupta, A., D'Cunha, A., Awasthi, K., & Balasubramanian, V. (2016). Daisee: Towards user engagement recognition in the wild. *arXiv preprint arXiv:1609.01885*.

Gupta, S. K., Ashwin, T. S., & Guddeti, R. M. R. (2019). Students' affective content analysis in smart classroom environment using deep learning techniques. *Multimedia Tools and Applications*, *78*(18), 25321–25348.

Happy, S., Patnaik, P., Routray, A., & Guha, R. (2015). The indian spontaneous expression database for emotion recognition. *IEEE Transactions on Affective Computing*, *8*(1), 131–142.

Hayashi, Y. (2019). Detecting collaborative learning through emotions: An investigation using facial expression recognition. In *International Conference on Intelligent Tutoring Systems*, 89–98. Springer.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Holmes, M., Latham, A., Crockett, K., & O'Shea, J. D. (2018). Near real-time comprehension classification with artificial neural networks: Decoding e-learner non-verbal behavior. *IEEE Transactions on Learning Technologies*, *11*(1), 5–12.

Hrastinski, S. (2008). Asynchronous and synchronous e-learning. *Educause quarterly*, *31*(4), 51–55.

Hu, M. & Li, H. (2017). Student engagement in online learning: A review. In *Educational Technology (ISET), 2017 International Symposium on*, 39–43. IEEE.

Huang, G. B., Ramesh, M., Berg, T., & Learned-Miller, E. (2007). Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, Technical Report 07-49, University of Massachusetts, Amherst.

Huang, T., Mei, Y., Zhang, H., Liu, S., & Yang, H. (2019). Fine-grained engagement recognition in online learning environment. In *2019 IEEE 9th International Conference on Electronics Information and Emergency Communication (ICEIEC)*, 338–341. IEEE.

Huang, X., Dhall, A., Goecke, R., Pietikäinen, M., & Zhao, G. (2018). Multimodal framework for analyzing the affect of a group of people. *IEEE Transactions on Multimedia*, *20*(10), 2706–2721.

Jain, A. K. & Farrokhnia, F. (1991). Unsupervised texture segmentation using gabor filters. *Pattern recognition*, *24*(12), 1167–1186.

Jain, D. K., Zhang, Z., & Huang, K. (2017). Multi angle optimal pattern-based deep learning for automatic facial expression recognition. *Pattern Recognition Letters*.

Jain, V. & Learned-Miller, E. (2010). Fddb: A benchmark for face detection in unconstrained settings. *University of Massachusetts, Amherst, Tech. Rep. UM-CS-2010-009*, *2*(7), 8.

Kahou, S. E., Bouthillier, X., Lamblin, P., Gulcehre, C., Michalski, V., Konda, K., Jean, S., Froumenty, P., Dauphin, Y., Boulanger-Lewandowski, N., et al. (2016). Emonets: Multimodal deep learning approaches for emotion recognition in video. *Journal on Multimodal User Interfaces*, *10*(2), 99–111.

Kahu, E. R. (2013). Framing student engagement in higher education. *Studies in higher education*, *38*(5), 758–773.

Kaur, A., Mustafa, A., Mehta, L., & Dhall, A. (2018). Prediction and localization of student engagement in the wild. In *2018 Digital Image Computing: Techniques and Applications (DICTA)*, 1–8. IEEE.

Kim, Y., Jeong, S., Ji, Y., Lee, S., Kwon, K. H., & Jeon, J. W. (2015). Smartphone response system using twitter to enable effective interaction and improve engagement in large classrooms. *IEEE Transactions on Education*, *58*(2), 98–103.

Klein, R. & Celik, T. (2017). The wits intelligent teaching system: Detecting student engagement during lectures using convolutional neural networks. In *Image Processing (ICIP), 2017 IEEE International Conference on*, 2856–2860. IEEE.

Kleinsmith, A. & Bianchi-Berthouze, N. (2013). Affective body expression perception and recognition: A survey. *IEEE Transactions on Affective Computing*, *4*(1), 15–33.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, 1097–1105.

Ku, K. Y., Ho, I. T., Hau, K. T., & Lai, E. C. (2014). Integrating direct and inquiry-based instruction in the teaching of critical thinking: an intervention study. *Instructional Science*, *42*(2), 251–269.

Kuh, G. D. (2003). What we're learning about student engagement from nsse: Benchmarks for effective educational practices. *Change: The Magazine of Higher Learning*, *35*(2), 24–32.

Kuh, G. D., Cruce, T. M., Shoup, R., Kinzie, J., & Gonyea, R. M. (2008). Unmasking the effects of student engagement on first-year college grades and persistence. *The journal of higher education*, *79*(5), 540–563.

Kulik, J. A. & Fletcher, J. (2016). Effectiveness of intelligent tutoring systems: a meta-analytic review. *Review of Educational Research*, *86*(1), 42–78.

Li, H., Sun, J., Xu, Z., & Chen, L. (2017). Multimodal 2d+ 3d facial expression recognition with deep fusion convolutional neural network. *IEEE Transactions on Multimedia*, *19*(12), 2816–2831.

Li, J., Liu, L., Li, J., Feng, J., Yan, S., & Sim, T. (2017). Towards a comprehensive face detector in the wild. *IEEE Transactions on Circuits and Systems for Video Technology*.

Lin, T. Y., Dollár, P., Girshick, R., He, K., Hariharan, B., & Belongie, S. (2017). Feature pyramid networks for object detection. In *CVPR*, volume 1, $\tilde{4}$.

Liu, J., Wang, G., Duan, L.-Y., Abdiyeva, K., & Kot, A. C. (2018). Skeleton-based human action recognition with global context-aware attention lstm networks. *IEEE Transactions on Image Processing*, *27*(4), 1586–1599.

Liu, M., Calvo, R. A., Pardo, A., & Martin, A. (2015). Measuring and visualizing students' behavioral engagement in writing activities. *IEEE Transactions on learning technologies*, *8*, 215–224.

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C. Y., & Berg, A. C. (2016). Ssd: Single shot multibox detector. In *European conference on computer vision*, 21–37. Springer.

Liu, Y. & Jiang, C. (2019). Recognition of shooter's emotions under stress based on affective computing. *IEEE Access*, *7*, 62338–62343.

Lopes, A. T., de Aguiar, E., De Souza, A. F., & Oliveira-Santos, T. (2017). Facial expression recognition with convolutional neural networks: coping with few data and the training sample order. *Pattern Recognition*, *61*, 610–628.

Lyons, M. J., Akamatsu, S., Kamachi, M., Gyoba, J., & Budynek, J. (1998). The japanese female facial expression (jaffe) database. In *Proceedings of Third International Conference on Automatic Face and Gesture Recognition*, 14–16.

Maneeratana, K., Tiamsa-Ad, U., Ruengsomboon, T., Chawalitrujiwong, A., Aksomsiri, P., & Asawapithulsert, K. (2017). Class-wide course feedback methods by student engagement program. In *Teaching, Assessment, and Learning for Engineering (TALE), 2017 IEEE 6th International Conference on*, 393–398. IEEE.

Marin, G., Dominio, F., & Zanuttigh, P. (2014). Hand gesture recognition with leap motion and kinect devices. In *2014 IEEE International Conference on Image Processing (ICIP)*, 1565–1569. IEEE.

Martinez, A. M. (1998). The ar face database. *CVC Technical Report*, *24*.

Matar, G., Lina, J.-M., Carrier, J., Riley, A., & Kaddoum, G. (2016). Internet of things in sleep monitoring: An application for posture recognition using supervised learning. In *2016 IEEE 18th International Conference on e-Health Networking, Applications and Services (Healthcom)*, 1–6. IEEE.

Mollahosseini, A., Chan, D., & Mahoor, M. H. (2016). Going deeper in facial expression recognition using deep neural networks. In *Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on*, 1–10. IEEE.

Monkaresi, H., Bosch, N., Calvo, R. A., & D'Mello, S. K. (2017). Automated detection of engagement using video-based estimation of facial expressions and heart rate. *IEEE Transactions on Affective Computing*, *8*(1), 15–28.

Moore, S. & Stamper, J. (2019). Decision support for an adversarial game environment using automatic hint generation. In *International Conference on Intelligent Tutoring Systems*, 82–88. Springer.

Mota, S. & Picard, R. W. (2003). Automated posture analysis for detecting learner's interest level. In *2003 Conference on Computer Vision and Pattern Recognition Workshop*, volume 5, 49–49. IEEE.

Muhammad, K., Hussain, T., & Baik, S. W. (2018). Efficient cnn based summarization of surveillance videos for resource-constrained devices. *Pattern Recognition Letters*.

Ng, H. W., Nguyen, V. D., Vonikakis, V., & Winkler, S. (2015). Deep learning for emotion recognition on small datasets using transfer learning. In *Proceedings of the 2015 ACM on international conference on multimodal interaction*, 443–449. ACM.

Noroozi, F., Kaminska, D., Corneanu, C., Sapinski, T., Escalera, S., & Anbarjafari, G. (2018). Survey on emotional body gesture recognition. *IEEE transactions on affective computing*.

Ojala, T., Pietikainen, M., & Maenpaa, T. (2002). Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on pattern analysis and machine intelligence*, *24*(7), 971–987.

Patwardhan, A. S. & Knapp, G. M. (2014). Affect intensity estimation using multiple modalities. In *The Twenty-Seventh International Flairs Conference*.

Phillips, P. (2004), The facial recognition technology (feret) database.

Picard, R. W. (1997). Affective computing. *The MIT Press, Cambridge (MA)*, *167*, 170.

Poria, S., Hazarika, D., Majumder, N., Naik, G., Cambria, E., & Mihalcea, R. (2018). Meld: A multimodal multi-party dataset for emotion recognition in conversations. *arXiv preprint arXiv:1810.02508*.

Psaltis, A., Apostolakis, K. C., Dimitropoulos, K., & Daras, P. (2017). Multimodal student engagement recognition in prosocial games. *IEEE Transactions on Games*, *10*(3), 292–303.

Radeta, M. & Maiocchi, M. (2013). Towards automatic and unobtrusive recognition of primary-process emotions in body postures. In *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction (ACII)*, 695–700. IEEE.

Rahmani, H., Mian, A., & Shah, M. (2018). Learning a deep model for human action recognition from novel viewpoints. *IEEE transactions on pattern analysis and machine intelligence*, *40*(3), 667–681.

Rajendran, R., Iyer, S., & Murthy, S. (2018). Personalized affective feedback to address students' frustration in its. *IEEE Transactions on Learning Technologies*, *12*(1), 87–97.

Ramirez L, Yao W, C. E. . S. B. (2019). Toward instrumenting makerspaces: Using motion sensors to capture students' affective states and social interactions in open-ended learning environments., pp. 639–642.

Ranganathan, H., Chakraborty, S., & Panchanathan, S. (2016). Multimodal emotion recognition using deep learning architectures. In *Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on*, 1–9. IEEE.

Ranjan, R., Patel, V. M., & Chellappa, R. (2019). Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *41*(1), 121–135.

Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 779–788.

Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, 91–99.

Riaz, S. & Mushtaq, A. (2016). Emerging themes analysis of learner's aesthetic-emotions in e-learning environments. In *2016 Sixth International Conference on Information Science and Technology (ICIST)*, 399–405. IEEE.

Rowe, J., Mott, B., McQuiggan, S., Robison, J., Lee, S., & Lester, J. (2009). Crystal island: A narrative-centered learning environment for eighth grade microbiology. In *workshop on intelligent educational games at the 14th international conference on artificial intelligence in education, Brighton, UK*, 11–20.

Russell, J. A. (1980). A circumplex model of affect. *Journal of personality and social psychology*, *39*(6), 1161.

Sapiński, T., Kamińska, D., Pelikant, A., Ozcinar, C., Avots, E., & Anbarjafari, G. (2018). Multimodal database of emotional speech, video and gestures. In *International Conference on Pattern Recognition*, 153–163. Springer.

Schroff, F., Kalenichenko, D., & Philbin, J. (2015). Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 815–823.

Sedgwick, P. et al. (2012). Pearson's correlation coefficient. *Bmj*, *345*(7).

Sekachev, Boris, Nikita, M., & Andrey, Z. (2019), Computer vision annotation tool: A universal approach to data annotation.

Setty, S., Husain, M., Beham, P., Gudavalli, J., Kandasamy, M., Vaddi, R., Hemadri, V., Karure, J., Raju, R., Rajan, B., et al. (2013). Indian movie face database: a benchmark for face recognition under wide variations. In *2013 Fourth National Conference on Computer Vision, Pattern Recognition, Image Processing and Graphics (NCVPRIPG)*, 1–5. IEEE.

Sidney, K. D., Craig, S. D., Gholson, B., Franklin, S., Picard, R., & Graesser, A. C. (2005). Integrating affect sensors in an intelligent tutoring system. In *Affective Interactions: The Computer in the Affective Loop Workshop at*, 7–13.

Silfver, E., Jacobsson, M., Arnell, L., Bertilsdotter-Rosqvist, H., Härgestam, M., Sjöberg, M., & Widding, U. (2018). Classroom bodies: affect, body language, and discourse when schoolchildren encounter national tests in mathematics. *Gender and Education*, 1–15.

Silva, P., Costa, E., & de Araújo, J. R. (2019). An adaptive approach to provide feedback for students in programming problem solving. In *International Conference on Intelligent Tutoring Systems*, 14–23. Springer.

Simonyan, K. & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Sinatra, G. M., Heddy, B. C., & Lombardi, D. (2015), The challenges of defining and measuring student engagement in science.

Sindagi, V. A. & Patel, V. M. (2018). A survey of recent advances in cnn-based single image crowd counting and density estimation. *Pattern Recognition Letters*, *107*, 3–16.

Singh, A., Karanam, S., & Kumar, D. (2013). Constructive learning for human-robot interaction. *IEEE Potentials*, *32*, 13–19.

Subramanian, R., Wache, J., Abadi, M. K., Vieriu, R. L., Winkler, S., & Sebe, N. (2018). Ascertain: Emotion and personality recognition using commercial sensors. *IEEE Transactions on Affective Computing*, *9*(2), 147–160.

Sun, B., Wei, Q., Li, L., Xu, Q., He, J., & Yu, L. (2016). Lstm for dynamic emotion and group emotion recognition in the wild. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, 451–457. ACM.

Sun, M. C., Hsu, S. H., Yang, M. C., & Chien, J. H. (2018). Context-aware cascade attention-based rnn for video emotion recognition. In *2018 First Asian Conference on Affective Computing and Intelligent Interaction (ACII Asia)*, 1–6. IEEE.

Szegedy, C., Ioffe, S., Vanhoucke, V., & Alemi, A. A. (2017). Inception-v4, inception-resnet and the impact of residual connections on learning. In *AAAI*, 4278–4284.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2818–2826.

Tang, X., Du, D. K., He, Z., & Liu, J. (2018). Pyramidbox: A context-assisted single shot face detector. *arXiv preprint arXiv:1803.07737*.

Tarrés, F. & Rama, A. (2012). Gtav face database. *GVAP, UPC*.

Thomas, C. & Jayagopi, D. B. (2017). Predicting student engagement in classrooms using facial behavioral cues. In *Proceedings of the 1st ACM SIGCHI International Workshop on Multimodal Interaction for Education*, 33–40. ACM.

Tiam-Lee, T. J. & Sumi, K. (2019). Analysis and prediction of student emotions while doing programming exercises. In *International Conference on Intelligent Tutoring Systems*, 24–33. Springer.

Tian, Y. I., Kanade, T., & Cohn, J. F. (2001). Recognizing action units for facial expression analysis. *IEEE Transactions on pattern analysis and machine intelligence*, *23*(2), 97–115.

Tieleman, T. & Hinton, G. (2012). Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning*, *4*(2), 26–31.

Tompson, J., Stein, M., Lecun, Y., & Perlin, K. (2014). Real-time continuous pose recovery of human hands using convolutional networks. *ACM Transactions on Graphics (ToG)*, *33*(5), 169.

Tucker, B. (2012). The flipped classroom. *Education next*, *12*(1), 82–83.

Tzirakis, P., Trigeorgis, G., Nicolaou, M. A., Schuller, B., & Zafeiriou, S. (2017). End-to-end multimodal emotion recognition using deep neural networks. *arXiv preprint arXiv:1704.08619*.

Valstar, M. & Pantic, M. (2010). Induced disgust, happiness and surprise: an addition to the mmi facial expression database. In *Proc. 3rd Intern. Workshop on EMOTION (satellite of LREC): Corpora for Research on Emotion and Affect*, 65.

Van der Sluis, F., Ginn, J., & Van der Zee, T. (2016). Explaining student behavior at scale: The influence of video complexity on student dwelling time. In *Proceedings of the Third (2016) ACM Conference on Learning@ Scale*, 51–60. ACM.

Varol, G., Laptev, I., & Schmid, C. (2018). Long-term temporal convolutions for action recognition. *IEEE transactions on pattern analysis and machine intelligence*, *40*(6), 1510–1517.

Viola, P. & Jones, M. Rapid object detection using a boosted cascade of simple features. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001.*, volume 1, I–I. IEEE.

Walker, E., Ogan, A., Aleven, V., & Jones, C. (2008). Two approaches for providing adaptive support for discussion in an ill-defined domain. *Intelligent Tutoring Systems for Ill-Defined Domains: Assessment and Feedback in Ill-Defined Domains*, *1*.

Wang, S. & Ji, Q. (2015). Video affective content analysis: a survey of state-of-the-art methods. *IEEE Transactions on Affective Computing*, *6*(4), 410–430.

Watson, D. & Tellegen, A. (1985). Toward a consensual structure of mood. *Psychological bulletin*, *98*(2), 219.

Whitehill, J., Serpell, Z., Lin, Y.-C., Foster, A., & Movellan, J. R. (2014). The faces of engagement: Automatic recognition of student engagementfrom facial expressions. *IEEE Transactions on Affective Computing*, *5*(1), 86–98.

Wolf, L., Hassner, T., & Maoz, I. (2011). Face recognition in unconstrained videos with matched background similarity. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, 529–534. IEEE.

Wu, H., Zhang, K., & Tian, G. (2018). Simultaneous face detection and pose estimation using convolutional neural network cascade. *IEEE Access*, *6*, 49563–49575.

Xia, X., Liu, J., Yang, T., Jiang, D., Han, W., & Sahli, H. (2018). Video emotion recognition using hand-crafted and deep learning features. In *2018 First Asian Conference on Affective Computing and Intelligent Interaction (ACII Asia)*, 1–6. IEEE.

Xie, S. & Hu, H. (2019). Facial expression recognition using hierarchical features with deep comprehensive multipatches aggregation convolutional neural networks. *IEEE Transactions on Multimedia*, *21*(1), 211–220.

Yang, B., Cao, J., Ni, R., & Zhang, Y. (2018). Facial expression recognition using weighted mixture deep neural network based on double-channel facial images. *IEEE Access*, *6*, 4630–4640.

Yao, B. & Fei-Fei, L. (2012). Recognizing human-object interactions in still images by modeling the mutual context of objects and human poses. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *34*(9), 1691–1703.

Yin, X. & Liu, X. (2018). Multi-task convolutional neural network for pose-invariant face recognition. *IEEE Transactions on Image Processing*, *27*(2), 964–975.

Yousuf, B. & Conlan, O. (2017). Supporting student engagement through explorable visual narratives. *IEEE Transactions on Learning Technologies*, *11*(3), 307–320.

Yu, Y. C. (2017). Teaching with a dual-channel classroom feedback system in the digital classroom environment. *IEEE Transactions on Learning Technologies*, *10*(3), 391–402.

Yun, W. H., Lee, D., Park, C., Kim, J., & Kim, J. (2018). Automatic recognition of children engagement from facial video using convolutional neural networks. *IEEE Transactions on Affective Computing*.

Zaletelj, J. & Košir, A. (2017). Predicting students' attention in the classroom from kinect facial and body features. *EURASIP Journal on Image and Video Processing*, *2017*(1), 80.

Zhang, K., Zhang, Z., Li, Z., & Qiao, Y. (2016). Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, *23*(10), 1499–1503.

Zhang, N. & Deng, W. (2016). Fine-grained lfw database. In *2016 International Conference on Biometrics (ICB)*, 1–6.

Zhang, S., Pan, X., Cui, Y., Zhao, X., & Liu, L. (2019). Learning affective video features for facial expression recognition via hybrid deep learning. *IEEE Access*.

Zhang, S., Zhu, X., Lei, Z., Shi, H., Wang, X., & Li, S. Z. (2017). Sˆ 3fd: Single shot scale-invariant face detector. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, 192–201. IEEE.

Zhang, W., Shan, S., Gao, W., Chen, X., & Zhang, H. Local gabor binary pattern histogram sequence (lgbphs): A novel non-statistical model for face representation and recognition. In *Tenth IEEE International Conference on Computer Vision, 2005. ICCV 2005.*, volume 1, 786–791. IEEE.

Zhang, X., Yin, L., Cohn, J. F., Canavan, S., Reale, M., Horowitz, A., Liu, P., & Girard, J. M. (2014). Bp4d-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database. *Image and Vision Computing*, *32*(10), 692–706.

Zhao, J. & Itti, L. (2017). Improved deep learning of object category using pose information. In *Applications of Computer Vision (WACV), 2017 IEEE Winter Conference on*, 550–559. IEEE.

Zilvinskis, J., Masseria, A. A., & Pike, G. R. (2017). Student engagement and student learning: Examining the convergent and discriminant validity of the revised national survey of student engagement. *Research in Higher Education*, 1–24.

Publications

Journal Papers

1. **Ashwin T S**, and Ram Mohana Reddy Guddeti. (2020). Affective database for e-learning and classroom environments using Indian students' faces, hand gestures and body postures. *Future Generation Computer Systems*, Volume 108, July 2020, Pages 334-348. DOI: 10.1016/j.future.2020.02.075.

2. **Ashwin T S** and Ram Mohana Reddy G. (2020). Impact of Inquiry Interventions on Students in E-Learning and Classroom Environments using Affective Computing Framework. *User Modeling and User-Adapted Interaction* 1-43, Springer. DOI:10.1007/s11257-019-09254-3.

3. **Ashwin T S**, and Ram Mohana Reddy Guddeti. (2019). Unobtrusive Behavioral Analysis of Students in Classroom Environment Using Non-Verbal Cues. *IEEE Access* , Vol. 7, pp. 150693-150709. DOI: 10.1109/ACCESS.2019.2947519.

4. **Ashwin T S**, and Ram Mohana Reddy Guddeti. (2019). Automatic detection of students' affective states in classroom environment using hybrid convolutional neural networks. *Educational and Information Technologies* 1-39, Springer. DOI: 10.1007/s10639-019-10004-6.

5. Sujit G, **Ashwin T S**, and Ram Mohana Reddy Guddeti (2019) Students' affective content analysis in smart classroom environment using deep learning techniques. *Multimedia Tools and Applications*, 78(18), 25321-25348, Springer. DOI: 10.1007/s11042-019-7651-z.

6. **Ashwin T S**, and Ram Mohana Reddy Guddeti. Unobtrusive Students' Affective Engagement Analysis in E-Learning and Classroom Environments Using Deep Learning. *Journal of Ambient Intelligence & Humanized Computing*, Springer (Revision).

Conference Papers and Book Chapters

1. **Ashwin T S**, and Ram Mohana Reddy Guddeti. (2018). Unobtrusive Students' Engagement Analysis in Computer Science Laboratory using Deep Learning Techniques. *18th IEEE International Conference on Advanced Learning Technologies (ICALT 2018)*, July 9-13, 2018, IIT Bombay, India. DOI: 10.1109/ICALT.2018.00110 *(Core Conference)*

2. **Ashwin T S**, Sai Saran, and Ram Mohana Reddy Guddeti. (2016) Video Affective Content Analysis Based on Multimodal Features Using a Novel Hybrid SVM-RBM Classifier. *2016 IEEE Uttar Pradesh Section International Conference on Electrical, Computer and Electronics Engineering (UPCON), Varanasi, 2016, pp. 416-421*. DOI: 10.1109/UPCON.2016.7894690. *(Best Paper Award)*

3. **Ashwin T S**, Kartik S, Mohammad A, Anmol V and Ram Mohana Reddy Guddeti. (2016) Virtual Slate: Microsoft Kinect Based Text Input Tool to Improve Hand-writing of People. *Proceedings of the 24th International Conference on Computers in Education. IIT Bombay, 2016, pp. 12-19*. India: Asia-Pacific Society for Computers in Education ISBN: 9789869401210 *(Core Conference)*

4. **Ashwin T S**, Jijo J, Raghu G, and Ram Mohana Reddy Guddeti. (2015). An E-Learning System with Multifacial Emotion Recognition Using Supervised Machine Learning. *2015 IEEE Seventh International Conference on Technology for Education (T4E), Warangal, 2015, pp. 23-26.* DOI: 10.1109/T4E.2015.21

5. **Ashwin T S**, Vishal P and Ram Mohana Reddy Guddet. (2016). Unobtrusive Intelligent Smartphone Systems for Enhancing the User Experience Using Emotion Recognition. *The 3rd IEEE Annual Conference on Cognitive Science (ACCS 2016)*, 2016, Indian Institute of Technology Gandhinagar (IITGN), India. *(Poster Presentation)*

6. Sharma R, **Ashwin T S**, and Ram Mohana Reddy Guddeti. (2019) A Novel Real-Time Face Detection System Using Modified Affine Transformation and Haar Cascades. *5th International Conference on Advanced Computing, Networking, and Informatics, Jun 1 - 3 Goa, India.* (proceedings by Springer) *In Recent Findings in Intelligent Computing Techniques*, pp. 193-204. Springer, Singapore, DOI: 2019.10.1007/978-981-10-8639-7_20 *(Book Chapter)*

Brief Bio-Data

Mr. Ashwin T S
Full-Time Research Scholar
Department of Information Technology
National Institute of Technology Karnataka
P.O. Srinivasanagar, Surathkal
Mangalore-575 025

Permanent Address

Ashwin T S
s/o Dr. T M Sadashiva
Sri Balaji, Malligenahalli - 577205
Shimoga District, Karnataka, India.
Email: ashwindixit9@gmail.com
Mobile: +919164617445.

Academic Records

1. M.Tech. in Computer Science and Engineering from Manipal Institute of Technology, Manipal, Karnataka, India, 2013.
2. B.E. in Computer Science and Engineering from VTU University, Belgaum, Karnataka, India, 2011.

Research Interests

Affective Computing
Human Computer Interaction
Computer Vision