ACADEMIC FACIAL EXPRESSION DETECTION USING CNN-RNN(CLASSIFICATION) and CNN-LSTM(RECOGNITION) DEEP LEARNING

Shruti Sahare, Prof. Snehal Rathi
BRACT'S Vishwakarma Institute of Information Technology, Survey No. 3/4, Kondhwa (Budruk), Pune: 411048

ABSTRACT –

Online education systems have evolved during the past two years as physical teaching has been blocked due to the global pandemic, so this project plays an important role in understanding student's academic emotions such as Boredom, Confusion, Engagement, and Frustration during E-learning. We can track the progress of students using this technology for better understanding and communication between teachers and students. Our technology is basically focused on the deep learning approach, for classification we are using Convolutional Neural Network (CNN) along with Recurrent Neural Network (RNN,) and for recognition we are using Convolutional Neural Network (CNN) along with Long-Short Term Memory (LSTM).

In this project we have successfully created the real-time model with the accuracy ranging between 50-60 because the dataset isn't made for recognition purposes and to achieve a higher accuracy, we planned to create our own dataset with a group of students and plan to go further with this research.

Index Terms *- Long Short-Term Memory (LSTM), Convolutional Neural Network (CNN), Graphics Processing Unit (GPU), Recurrent Neural Network (RNN), Software Development Life Cycle (SDLC), Unified Modelling Language (UML), High Definition (HD), Red Green Blue (RGB)*

1. INTRODUCTION

As we know Facial Expression performs a key role in understanding human emotions (sadness, happiness, fear, anger, surprise, and disgust). Similarly, academic emotions are equally important because they can affect mental health. This state of feelings is associated with the thoughts and feelings of students which results in depression, anxiety, stress, etc. In past decades various methods have been used to detect facial expressions/emotions and also there are many algorithms already present which can detect basic human emotions very well. So, what is the need for developing a human emotion recognition system again? That's why we are aiming to develop an academic emotion recognition system to detect real-time emotions such as Boredom, Confusion, Frustration, and Engagement. We have created a framework for Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN,) and Long-Short Term Memory (LSTM).

In the past decades, various methods have been used to detect facial expressions/emotions and also there are many algorithms already present, which can detect basic human emotions very well. So, what is the need for developing a human emotion recognition system again?

That's why we are aiming to develop an academic emotion recognition system to detect real-time emotions such as Boredom, Confusion, Frustration, and Engagement. We have created a framework for Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN,) and Long-Short Term Memory (LSTM).

2. METHODOLOGY

This section gives us a detailed description of the methodology, features, parameters, and experimental setup we have used/done in this project. The project - Academic Emotion Detection is a Machine Learning based project in which a deep learning approach is followed. In this project, we are working to detect human emotions such as boredom, frustration, confusion, and engagement, naming them academic emotions.

We have used the DAISEE dataset to build the video classifier and real-time recognition(emotion) system. This dataset is commonly used to detect academic emotions itself which is divided into three different categories:

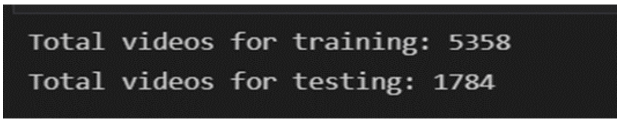
- Training data
- Testing data
- Validation data

So, the project can be broadly divided into two parts viz, Classification and Recognition. In the classification part, we have used a hybrid CNN-RNN architecture, In the real-time recognition part, we have used a CNN-LSTM architecture.

2.1 Video Classification with a CNN-RNN Architecture

Step 1: Start
The working of Academic Emotion Detection starts at this step.

Step 2: Data collection and preparation
The dataset consists of videos (test- 1784, train-5358) and is shown below.



```
Total videos for training: 5358
Total videos for testing: 1784
```

Step 3: Data Pre-processing and Feature Extraction:
As a part of the Pre-processing step, the videos were converted into frames (since videos are basically a collection of frames per second). The frames have been cropped to a specific size for efficient accuracy metrics, this conversion will help in optimum feature recognition.
We used a pre-trained network for efficient extraction of features from the extracted frames and also, we used the InceptionV3 model (Inception V3 is a CNN-based model (consisting of 42 layers) used for image classification.) which serves this purpose, it is a part of the Kera's Applications module pre-trained on the ImageNet-1k dataset.

Step 4: The Sequence Model (pre-trained model):
The sequence model is considered to be suitable for a stack of layers where each layer has exactly one input tensor and one output tensor. Post the feature extraction, we pass the parameters through the Sequence Model consisting of recurrent GRU layers.

Step 5: Emotion Extraction:
After we are done with all the necessary steps, we run our main last function which gives us the resultant output, which is nothing but the emotion which got detected.

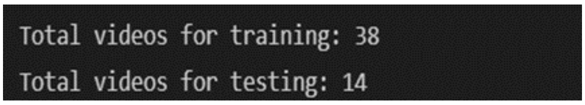
Step 6: Stop
Here the architecture stops as we get the desired output.

2.2 Real-Time Recognition with a CNN-LSTM

Architecture

Step 1: Start
The working of real-time Academic Emotion Detection starts at this step.

Step 2: Data connection and Preparation:
Here, the Daisee dataset has been trimmed according to our needs as it was sufficient for training and testing the model.
Now, the size of the dataset is given below.



```
Total videos for training: 38
Total videos for testing: 14
```

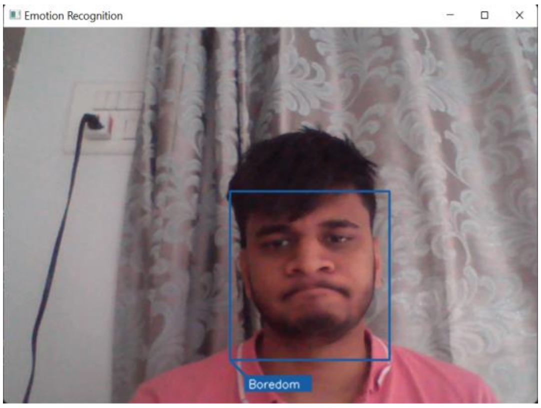
Step 3: Data Pre-processing and Feature Extraction:
As a part of the Pre-processing step, Videos were divided into frames 4 folders named after particular emotions created with each folder having the exact 1500 frames in it, Similar to the CNN-RNN algorithm, The frames have been cropped to a specific size for efficient accuracy metrics, this conversion will help in optimum feature recognition, But here we have used Haarcascade frontal face classifier so that the background gets excluded and the main focus is retained on the face.

Step 4: Creating a model:
Post the feature extraction procedure, we pass the results to the Sequential model which will pass a sequence of the frames as input to the LSTM layers in the form of a 5D tensor. Here we are using LSTM to construct our layers, which is a combination of CNN and LSTM.

Step 5: Exporting the Model
The model after being trained for a specific number of epochs, is then saved and it exports it as a h5 file which in turn is used by the driver code.

Step 6: Real-time recognition/Initializing webcam
The driver code basically consists of the code wherein the webcam gets initialized and real-time detection takes place. The model that was exported earlier in Step 4 will be used here for predictions.

Step 7: Result
After we are done with all the necessary steps, we run the driver code which gives us the resultant output, that is the emotion.
We can see the result in the following image



Step 8: Stop
Here the architecture stops as we get the desired output.

2.3 PROJECT SCOPE & LIMITATIONS

The scope of the project lies in various sectors like education, retail, marketing, and health care. It will help the education sector in various ways as the teaching faculty will be able to detect the emotions of the students and students in learning and understanding better way. This project will work in retail busy businesses that can detect the emotions of customers.

In the sector of marketing, this technology can be used to understand the demand and reaction of the customer towards a particular product, which can be helpful in detecting the purchase intent of the customer so that we can promote that product accordingly. It will really help us to transform the face of marketing/advertising and then adapt the customer views/experiences to these emotions in real-time. The sector of healthcare can help us to detect the emotion of the patient.

The system can help physicians determine whom to see first on the basis of the detected expression and engagement of that patient.

The limitations of the project come when there are a large number of faces to be recognized. To make it possible to detect many people we will have to use a larger GPU which will process many facial recognitions and detect their respective emotions. It has very limited use cases.

3 REVIEW OF LITERATURE

The study of some recently published papers is as follows:

1. Detecting Boredom in Meeting

This paper was published in 2005. The author of this research paper is S. Korea.
The modality in this research is Face Expression, Object Tracking, and Fatigue Detection. The methodology used in this project is hand usage, eye movements, and mouth movements (closed, yawning, blinking). Though the insights and the conclusions we get from this research paper are very few we can conclude that boredom is a very complicated emotional state to be recognized or distinguished from other emotional states. To show boredom emotion, Humans exhibit various signs which tend to be misinterpreted sometimes. To achieve perfect accuracy and promising results this paper suggested a background removal process.

2. Detecting the Politeness and Frustration State of a Child in a conversational Computer Game

This paper has been written by Serdar Yildirim, Chul Min Lee, Sungbok Lee, Alexandros Potamianos, Shrikanth Narayanan. From this particular paper, we get the idea that it deals with two emotions specifically which are frustration and politeness in a child through a speech at the time of playing games to observe their mental state. Features here are extracted using acoustic and language information for different age groups using the speech database.

3. Unobtrusive Academic Emotion Recognition Based on Facial Expression

Using RGB-D Camera Using Adaptive-Network-Based Fuzzy Inference System (ANFIS). This paper was written by James Purnama and Riri Fitri Sari in January 2019. Here, The RGB-Depth Microsoft Kinect camera is used to detect emotion by placing video-capturing devices in front of the students, which are attached to the desks. Instead of attaching sensors to body parts because according to their research, it can distract the students.

4. Deep Learning-Based Emotion Recognition from Real-Time Videos

The research has been conducted by Wenbin Zhou, Justin Cheng, Xingyu Lei, Bedrich Benes(B), and Nicoletta Adamo. This study is based on deep neural networks VGG_S and Caffe, to detect human emotion in real-time (captured through a web camera), using a classifier. It is based on Russel's diagram which is divided into 4 quadrants [Activation, Deactivation, Pleasant, and Unpleasant], the research was small scale up to 10 people participated in this research to store the facial expressions as a dataset to perform training upon.

5. **Personalized Emotion Detection and Emotion Recognition**

The author of this project is Wenqiang Tian, Academic Editor - Sang-Bing Tsai. The paper was published on 27 May 2021. The Support Vector Machine algorithm is used for face detection, and it uses the temporal feature for emotion analysis and finally uses the classification method of machine learning to classify emotions into different categories.

4 PROBLEM DEFINITION AND OBJECTIVE

1. **Boredom**
   Boredom is a type of emotion that involves an emotional or psychological state sometimes, which is experienced when an individual is left with nothing to do, is not interested in the activity they are involved in or feels that the day is dull or tedious for them.

2. **Confusion**
   Confusion is the type of emotion that is associated with conflicting and contradictory information, such as when people appraise an event as unfamiliar and it is very hard to understand the topic.

3. **Frustration**
   This emotion is associated with feeling of being annoyed or less confident because you cannot achieve what you want, or something that makes you feel disappointed or discouraged. It is a very subjective feeling and depends from person to person.

4. **Engagement**
   Engagement is the amalgamation of focus, curiosity, and involvement. From a student's perspective engagement accounts for high levels of motivation and being in sync with their curriculum.

The objectives of our project is as follows:

- To get a deeper understanding of the different algorithms working together with a subjective dataset.
- To understand the various types of emotions and their implementation in the field of deep learning.

- To help the educational sector in understanding the emotions during the lectures which will help the students as well as the teachers.

5 HARDWARE, SOFTWARE & DATABASE REQUIREMENTS

5.1 Hardware Requirements

1. i7 configuration 8th Generation
2. Nvidia GeForce GTX GPU
3. Integrated Web-camera

5.2 Software Requirements

1. Visual Studio Code
2. Google Colab
3. Python 3.9.7 and its various libraries
4. Jupyter Notebook

5.3 Database Requirements

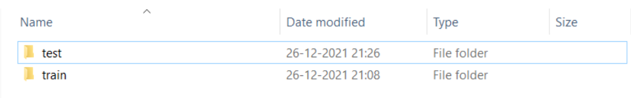
Our dataset is DAiSEE, the first multi-label video classification dataset consisting of 9068 video snippets captured from 112 users for recognizing the user affective states of boredom, confusion, engagement, and frustration.

The dataset has four levels of labels namely - very low, low, high, and very high for each of the affective states, which are crowd annotated and correlated with a gold standard annotation created using a team of expert psychologists.

Even though it is a very subjective dataset. We believe that DAiSEE provides the research community with challenges in feature extraction, context-based inference, and the development of suitable machine learning methods for related tasks, thus providing a springboard for further research.
Link to dataset: https://iith.ac.in/~daisee-dataset/

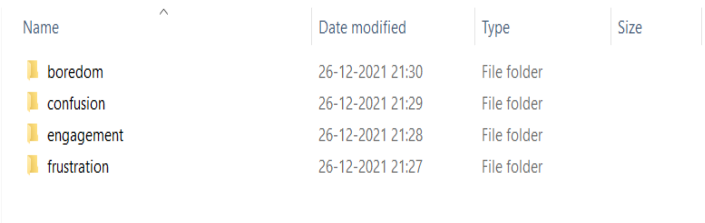
Preparation of Dataset:

We got to know that the DAiSEE dataset has around 9000 videos in different levels, but this wasn't our requirement so we had to cut down the number of videos to 48 accordingly into two different folders: Train and Test.



Name	Date modified	Type	Size
test	26-12-2021 21:26	File folder	
train	26-12-2021 21:08	File folder	

In these folders (train and test), four sub-folders have been created named as per the academic emotions (boredom, confusion, frustration and engagement). So now our dataset is no longer 'level based' but instead 'class based'.

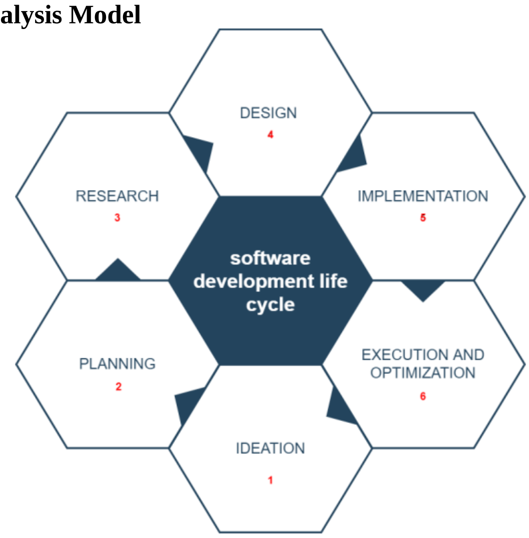


Name	Date modified	Type	Size
boredom	26-12-2021 21:30	File folder	
confusion	26-12-2021 21:29	File folder	
engagement	26-12-2021 21:28	File folder	
frustration	26-12-2021 21:27	File folder	

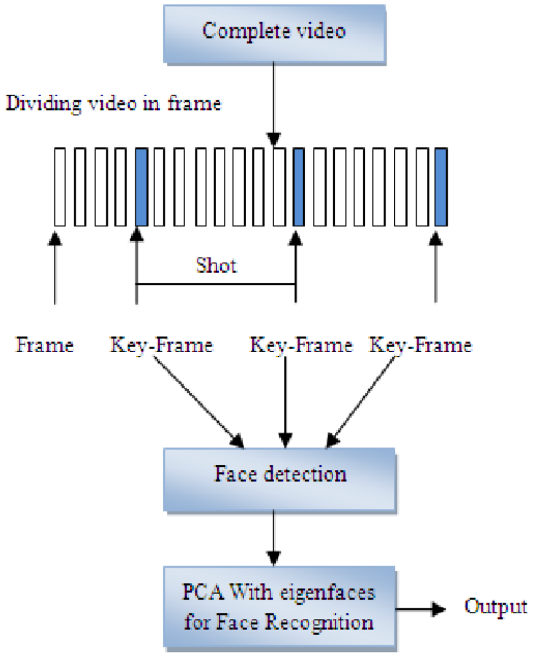
Now that our video dataset is ready, we further proceed to extract the frames from all the videos and store them in a new folder named frames. Each video is extracted into 300 frames so 1500 per emotion.
Now our images dataset of a total of 6000 images is ready to use.

6 ANALYSIS MODEL & SYSTEM DESIGN

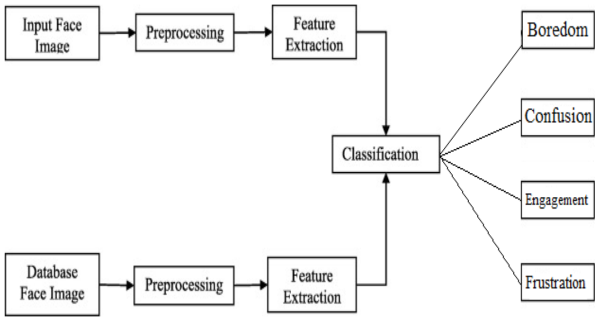
6 .1 Analysis Model



6.2 System Design



7 SYSTEM ARCHITECTURE

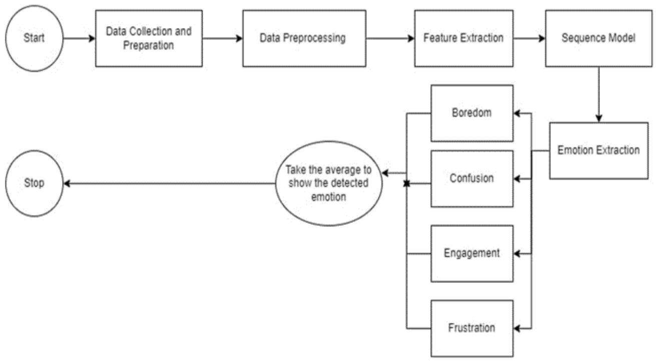


Algorithm 1: CNN-RNN:

The CNN-RNN algorithms are used together in a way that they provide optimum results. CNN extracts the features from the dataset.

The results are then fed to the RNN and the RNN is trained on the features from the earlier step. Following is the model structure:

Real-Time Recognition with CNN-RNN Architecture:



Algorithm 2: CNN-LSTM:

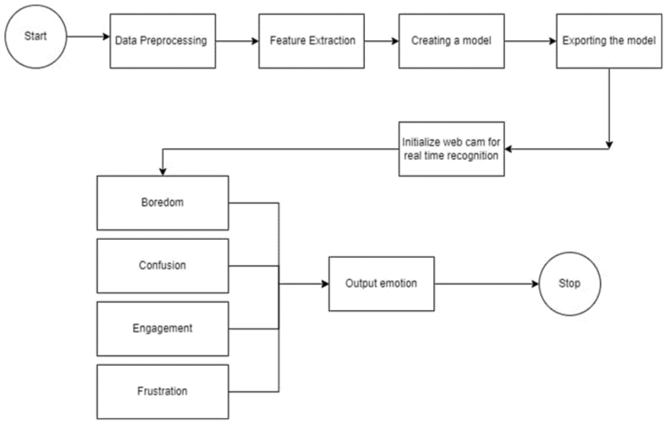
The CNN algorithm extracts features from the dataset. The data that is being fed to LSTM is of the features type. In order to input this data some sort of pre-processing is performed.

A 3D input is passed to LSTM which consists of three constraints. viz. Batch size, timesteps, and features. 30 timesteps along with features are fed to the LSTM model. Activation functions used: reLU, SoftMax.

The model after training and evaluation predicts emotions.

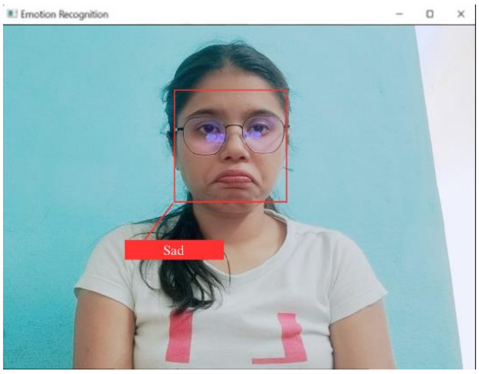
Following is the model structure:

Real-Time Recognition with CNN-LSTM Architecture:

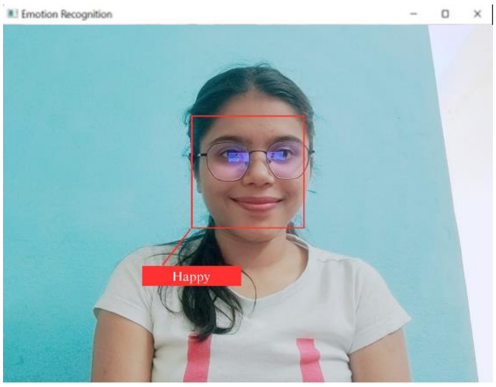


8 RESULTS

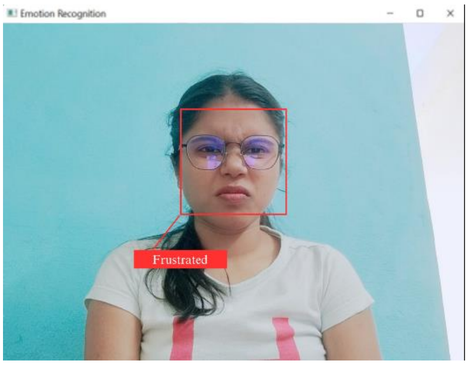
8.1 – Academic Emotion - Sad



8.2 – Academic Emotion - Happy



8.3 – Academic Emotion - Frustrated



9 FUTURE SCOPE & APPLICATIONS

9.1 Future Scope

Our project is a small-scale project for now as only limited to a certain extent but it has a very good future scope even though now it's only based on four academic emotions that is boredom, engagement, frustration, and confusion there are many other emotions a student or a person goes through while in an interview, an exam or in a meeting which is yet to be explored. And we look forward to going more in-depth with it. In this research, we even included a part of a questionnaire to compare our project with actual personal feelings to see if they match.

9.2 Applications

1. **Education -** This project will be very useful to the educational field especially. We can detect the interest of the students by observing and analyzing the results obtained by this emotion detection system. It will be a great help in improving the teaching strategy on e-learning platforms in this present situation of a pandemic.

2. **Retail -** This technology can help in the field of retail too. For a personalized experience, emotion detection plays an important role to identify the needs of the customers. After facial expression detection of customers, this deep technology can be useful to present various purchasing recommendations.

3. **Marketing/advertising -** Emotion detection is clearly bringing a revolution in the field of marketing /advertising. This technology can be used to understand the demand and reaction of the customer towards a particular product, which can be helpful in detecting the purchase intent of the customer so that we can promote that product accordingly. It will really help us to transform the face of marketing/advertising and then adapt the customer views/experiences to these emotions in real-time.

4. We can use this project in the field of healthcare as it can help us to detect the emotion of the patient. The system can help physicians determine whom to see first on the basis of the detected expression and engagement of that patient.

10 CONCLUSION

As we know Facial Expression performs a key role in understanding human emotions (Anger, surprise, sadness, happiness, fear, and disgust). In a similar manner, Academic emotions are equally important because they can affect mental health. This state of feelings is associated with the thoughts and feelings of students which results in depression, anxiety, stress, etc. In this project, we have successfully created the real-time model with an accuracy ranging between 50-60 because the dataset isn't made for recognition purposes and to achieve a higher accuracy, we planned to create our own dataset with a group of students and plan to go further with this research. In terms of technical conclusion and as a comparative study we would like to conclude by saying CNN is best suited for image classification whereas both RNN and LSTM are not as suited as CNN in terms of both accuracy and structure, even though all these models are part of deep learning but still they serve different purposes.

In terms of non-technical roles, our project plays an important role in the academic field to help both teachers and students. Understanding the emotion of a particular student or what they are going through is really important cause if not it could lead to depression and anxiety but our project serves its purpose aptly which is to help with academic emotion.

11 REFERENCES

1. A Blaszczynski, N McConaghy, A Frankova. Boredom Proneness in Pathological Gambling, Psychological Reports, Vol 67, 1990
2. W. Mikulas and S. Vodanovich. The Essence of Boredom, Psychological Record, Vol. 43 Issue 1, 1993
3. Becker, S. Kopp and I. Wachsmuth. Why should conversational agents be able to cry? Conversational Informatics, John Wiley & Sons, 2001
4. H. Wallbott. Bodily Expression of Emotion. European Journal of Social Psychology, Vol 28, 1998
5. Andersen, M. R., Jensen, T., Lisouski, P., Mortensen, A. K., Hansen, M. K., Gregersen, T., & Ahrendt, P. (2012). Kinect depth sensor evaluation for computer vision applications. Århus Universitet.
6. Arroyo, I., Cooper, D. G., Burleson, W., Woolf, B. P., Muldner, K., & Christopherson, R. (2009). Emotion Sensors Go To School. In AIED (pp. 17--24).
7. Azcarraga, J., Ibañez, J. F., Lim, I. R., Lumanas, N. J., Trogo, R., & Suarez, M. T. (2011). Predicting academic emotion based on brainwaves signals and mouse click behavior. Chiang Mai, Thailand: Asia-Pacific Society for Computers in Education.
8. Azcarraga, J., Suarez, M. T., & Inventado, P. S. (2010). Predicting the Difficulty Level Faced by Academic Achievers based on Brainwave Analysis. learning, 1, 6.
9. J. Ang, R. Dhillon, A. Krupski, E. Shriberg, and A. Stol-cke, "Prosody-based automatic detection of annoyance and frustration in human-computer dialog," in Proc. of IC-SLP, Denver, CO, 2001.
10. S. Arunachalam, D. Gould, E. Andersen, D. Byrd, and S. Narayanan, "Politeness and frustration language in child-machine interactions," in Proc. Eurospeech, 2001, pp. 2675–2678.
11. A. Batliner, K. Fischer, R. Huber, J. Spiker, and E. Noth, "Desperately seeking emotions: Actors, wizards, and hu-man beings," in Proc. ISCA Workshop on Speech and Emotion, Belfast, 2000, pp. 195–200.
12. F. Dellaert, T. Polzin, and A. Waibel, "Recognizing emo-tion in speech," in ICSLP '96, Philadelphia, PA, 1996.
13. Aifanti, N., Papachristou, C., Delopoulos, A.: The MUG facial expression database. In: 11th International Workshop on Image Analysis for Multimedia Interactive Services, WIAMIS 2010, pp. 1–4. IEEE (2010)
14. Allen, I.E., Seaman, J.: Staying the Course: Online Education in the United States. ERIC, Newburyport (2008)
15. Alsop, S., Watts, M.: Science education and affect. Int. J. Sci. Educ. 25(9), 1043–1047 (2003)
16. Ark, W.S., Dryer, D.C., Lu, D.J.: The emotion mouse. In: HCI (1), pp. 818–823 (1999)
17. Bakhtiyari, K., Taghavi, M., Husain, H., 2015. Hybrid affective computing—keyboard, mouse and touch screen: from review to experiment. Neural Comput. Appl. 26 (6), 1277–1296.
18. Bernardo, A.B., Ouano, J.A., Salanga, M.G.C., 2009. What is an academic emotion? Insights from Filipino bilingual students' emotion words associated with learning. Psychol. Stud. 54 (1), 28–37.
19. D'Errico, F., Paciello, M., Cerniglia, L., 2016. When emotions enhance students' engagement in e-learning processes. J. e-Learn. Knowl. Soc. 12 (4).
20. D'Mello, S., Graesser, A., 2012. Dynamics of affective states during complex learning. Learn. Inst. 22 (2), 145–157.
21. Mao, X., Li, Z.: 'Agent-based affective tutoring systems: a pilot study', Comput. Educ., 2010, 55, (1), pp. 202–208
22. Boekaerts, M.: 'Understanding students' affective processes in the classroom', in Schutz, Paul A., Reinhard, Pekrun (Eds.): 'Emotion in education' (Elsevier, USA, 2007), pp. 37–56
23. Kim, C.M., Hodges, C.B.: 'Effects of an emotion control treatment on academic emotions, motivation and achievement in an online mathematics course', Instr. Sci., 2012, 40, (1), pp. 173–192
24. Pekrun, R., Goetz, T., Frenzel, A.C., et al.: 'Measuring emotions in students' learning and performance: the achievement emotions questionnaire (AEQ)', Contemp. Educ. Psychol., 2011, 36, (1), pp. 36–48
25. Hussain, M.S., AlZoubi, O., Calvo, R.A., et al.: 'Affect detection from multichannel physiology during learning sessions with AutoTutor'. Int. Conf. on Artificial Intelligence in Education, Auckland, New Zealand, 2011, pp. 131–138