7-2020

Deep learning of facial embeddings and facial landmark points for the detection of academic emotions

Hua Leong FWA
Singapore Management University, hlfwa@smu.edu.sg

Deep learning of facial embeddings and facial landmark points for the detection of academic emotions

Fwa Hua Leong
keith_fwa@nyp.edu.sg
Nanyang Polytechnic, School of Information Technology
Singapore, Singapore

ABSTRACT

Automatic emotion recognition is an actively researched area as emotion plays a pivotal role in effective human communications. Equipping a computer to understand and respond to human emotions has potential applications in many fields including education, medicine, transport and hospitality. In a classroom or online learning context, the basic emotions do not occur frequently and do not influence the learning process itself. The academic emotions such as engagement, frustration, confusion and boredom are the ones which are pivotal to sustaining the motivation of learners. In this study, we evaluated the use of deep learning on FaceNet embeddings and facial landmark points for academic emotion detection on a publicly available dataset - DAiSEE that has been annotated with the emotional states of engagement, boredom, frustration and confusion. By modeling both the spatial and temporal dimensions, the results demonstrated that both models are able to detect incidences of boredom and frustration and can be used in the moment-by-moment monitoring of boredom and frustration of learners using a tutoring system either online or in a classroom.

CCS CONCEPTS

• **Computing Methodologies** → **Artificial Intelligence**; • **Applied Computing** → **Education**.

KEYWORDS

datasets, deep learning, emotions, facial emotion recognition

ACM Reference Format:
Fwa Hua Leong. 2020. Deep learning of facial embeddings and facial landmark points for the detection of academic emotions. In *2020 The 5th International Conference on Information and Education Innovations (ICIEI 2020), July 26–28, 2020, London, United Kingdom.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3411681.3411684

1 INTRODUCTION

Automatic emotion recognition is an actively researched area as emotion plays a pivotal role in effective human communications. Equipping a computer to understand and respond to human emotions or affective computing [21] has potential applications in many fields including education, medicine, transport and hospitality.

Among the techniques of automatic emotion recognition, the facial channel is universally recognized as the dominant channel of emotion expression in humans, resulting in facial expression recognition (FER) being the most researched among the various channels of emotion expression.

An early influential study in facial emotion by [6] concluded that there is a set of seven prototypical or basic emotions (happiness, sadness, fear, disgust, anger, contempt and surprise) that is recognizable even across different human cultures. He further postulated that the seven basic emotions can be described by combinations of 47 facial action units (AUs). The individual action units code the voluntary and involuntary movements of different facial muscles that occur during the expression of emotions. The labeling of the facial AUs, however is laborious and requires one to be trained for the accurate recognition of facial AUs [27]. As data labeling is an important prerequisite for the construction of machine learning model, the effort and expertise requirement constraints the use of facial AUs for emotion detection research. In addition, the link between facial AUs and the learning related affective states has also not been established [5].

Some researchers have argued that the set of seven basic emotions are not applicable for the different contexts in the real world. In a classroom or online learning context, the basic emotions do not occur frequently and even if they do, they have no little or no influence on the learning process itself. The academic emotions [19] e.g. frustration, confusion, engagement and boredom are the ones which influence the learning process. The ability of expert human tutors to achieve enhanced learning outcomes is widely attributed to their ability to sense the emotional states of the learners and to continually adapt their tutoring strategies in response to the dynamically changing emotional states throughout the tutoring session. By fostering the positive emotion e.g. engagement and suppressing negative emotions e.g. boredom, the tutor (which can be human or computer) can sustain the interest and motivation of the learner.

The traditional approaches in FER typically consists of 3 steps – face or facial component detection, feature extraction and emotion classification. After the detected facial components are formed into features, traditional machine learning techniques using handcrafted features are commonly employed to classify the various emotions. Many of these studies use machine learning techniques such as Hidden Markov Model (HMM) [15], Support Vector Machine (SVM) [22] and Random Forest (RF) [17]. In recent years however, deep

ICIEI 2020, July 26–28, 2020, London, United Kingdom

ACM ISBN 978-1-4503-7575-7/20/07…$15.00
https://doi.org/10.1145/3411681.3411684

learning techniques are showing much promise in FER and yielding state of the art results in many studies [10][26][1].

Deep learning based FER can be further categorized into static and dynamic FER. Static FER typically uses Convolutional Neural Network (CNNs) [13] to learn the weights of spatial filters from static input images while dynamic FER uses videos as input to capture not only the spatial but also the temporal dimensions. By including the modeling of temporal dimension, dynamic FER techniques are known to offer higher accuracies as compared to static FER techniques. Long Short Term Memory (LSTMs) [9] are commonly used to model the temporal dimensions in FER as they are fast and more importantly, they address the issue of vanishing gradients - the issue of gradients for deeper layers which is a product of gradients from earlier layers becoming zero.

FaceNet from Google [23] is a deep learning network which learns the Euclidean embedding of input face images. It is originally intended for use in facial recognition (FR) applications. When used in FR, the output facial embeddings from the FaceNet model for the reference facial image is compared using distance measures against the stored database of learned embeddings with the stored embedding with the closest match (or distance difference within a threshold) being the identified person. Analogous to Natural Language Processing (NLP) techniques where word embeddings [20] are learned for the inference of text sentiments, we hypothesize that the facial embeddings may similarly offer valuable information for the detection of emotions.

In facial detection, researchers have achieved high accuracy in the identification of facial landmark points or points which delineate the outline of various facial features e.g. eye-brow. In some studies, the facial landmark points are used for the automatic inference of human emotions. In the study by [25], features calculated from the facial landmark points are fed into two neural-network based architectures (one for the upper face and the other for the lower face) to identify the occurrences of facial AUs. An average accuracy rate of 93.3% for the recognition of 16 AUs was achieved on an independent sample of 122 subjects.

In this paper, we focus on the detection of academic emotions specifically frustration and boredom rather than the 7 basic emotions as the target emotions are more relevant to a classroom scenario.The detection of the academic emotions is challenging as they are more subtle as compared to the basic emotions and thus the significance of this research.

We also investigate the use of FaceNet embeddings and facial landmark points for academic emotion detection. Although, FaceNet was intended for use in facial recognition, the potential of its use in academic emotion detection has not been investigated.

One of the considerations in the use of deep learning algorithms is with their inference speed when deployed on mobile devices especially for models with deep number of layers and huge number of weights to be tuned. This has resulted in many researches being done recently to rectify the issue. The work by [3] in MobileFaceNets is able to achieve inference speed of 18 milliseconds on a mobile device for an input image resolution of 96 pixels by 96 pixels. We have developed a mobile application to test the speed of detection of facial landmark points using the dlib libary [11] on a mobile phone and we were able to achieve detection speed of 30 millisecond, thus demonstrating the viability of the proposed models for use in mobile devices. As such, in this study, we compare between the use of FaceNet embeddings and facial landmark points for use in emotion detection, specifically the academic emotions of frustration and boredom with consideration for eventual deployment on mobile devices.

2 RELATED STUDIES

Many studies have used CNNs for the inference of emotions. A study by [1] used CNNs across a few published Facial Emotion Recognition (FER) datasets for visualization of AUs activation and detection of 7 basic emotions. They also justified that their trained CNN model can generalize across the various FER datasets. In [16], CNNs were used for the detection of the 7 basic emotions. With the use of pre-processing techniques such as spatial and intensity normalization and generation of synthetic samples, the authors were able to achieve an accuracy of 96.76% accuracy on the CK+ database.

To model the temporal dependencies on top of the spatial features, [3] stacked LSTMs on top of CNNs to infer facial AUs which can in turn be correlated to the basic emotions. The authors postulated that temporal cues are pivotal to the accurate detection of the facial AUs. The results showed that the hybrid network architecture which addresses the spatial representation, temporal dependencies and AU correlation issues outperforms alternative models with an F1 score of 66.4.

In a recent work [18], the researchers trained the VGG-B [24] model on the Facial Emotion Recognition 2013 (FER-2013) [7] dataset before fine-tuning the model on their self-collected engagement dataset. The engagement dataset consists of videos of twenty students undertaking a learning scenario and were annotated by six trained psychologists. The results showed that their proposed model out-performed CNN based deep learning and traditional machine learning models. Deep convolutional architectures pre-trained on a public facial image dataset were employed by the authors for learning of the spatial dimensions of the facial images before fine-tuning it on their own dataset.

Face detection is frequently performed as a prerequisite to facial landmark points detection. In his submission to the third Recognition in the Wild challenge in 2015 [4], the dlib library [11] is used to crop out the face and locate 68 facial landmark points. The distances between pairs of facial landmark points are then used as inputs into an SVM for classification of the seven basic emotions. The proposed model achieved an accuracy of 46.8% as compared to the baseline model's accuracy of 39.1%, demonstrating that the facial landmark points can be used to discriminate between the various emotions. Possible extensions to this study would be to use facial landmark points for detection of academic emotions such as boredom and frustration and to include the temporal dimensions into the modeling for enhancing the detection performance.

In this study, we propose to infer the emotions of frustration and boredom instead of AUs directly from either the facial embeddings or the facial landmark points. We hypothesize that we can derive valuable information for identification of frustration and boredom from the facial embeddings and the facial landmarks. As opposed to studies which model only the spatial dimensions of facial images,

we also include the temporal dimensions of the facial videos for the models used in this study.

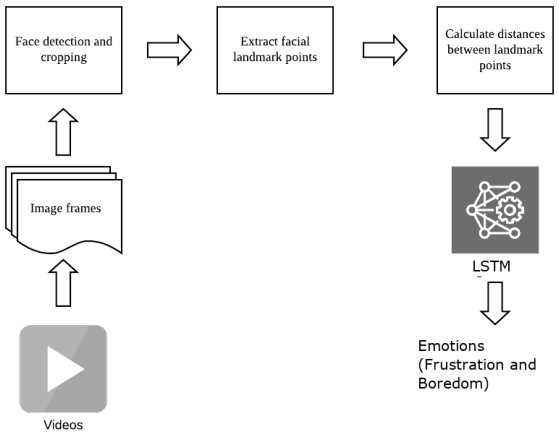

Figure 1: Processing workflow for Facial Landmark Point Model

3 METHODOLOGY

3.1 Dataset

We use DAiSEE (Dataset for Affective States in E-Environment) [8] for this study. DAiSEE consists of 9,068 video sequences collected from 112 Asian subjects. The provided dataset are also labeled through crowd annotation with the four learning related affective states of engagement, boredom, confusion and frustration. DAiSEE is used here as it is the only publicly available "in the wild" facial video dataset that is annotated with the states of engagement, confusion, frustration and boredom.

In this study, we only focus on the detection of boredom and frustration as these are the negative learning related affective states that are detrimental to the learning process. In a classroom or online learning context, with the successful detection of these negative states, the tutoring system can then enact appropriate pedagogical interventions to avert detrimental effects such as the learner giving up on the learning altogether.

3.2 Pre-processing

The videos with a resolution of 640 by 480 pixels and 10 seconds in duration each are first split into individual image frames, resulting in a total of 300 frame images per video. The videos are already divided into 60% for training, 20% for validation and 20% for test set in the original data set. In our case, we used the training set for the training of the models and the test set for evaluating the performance of the models. The videos are labeled with 4 levels of engagement, confusion, frustration and boredom but we used only frustration and boredom and re-labeled them with 2 levels (present or not present) instead. Pytorch is used for developing the codes for the models and for training and evaluating of the models described in the next section.

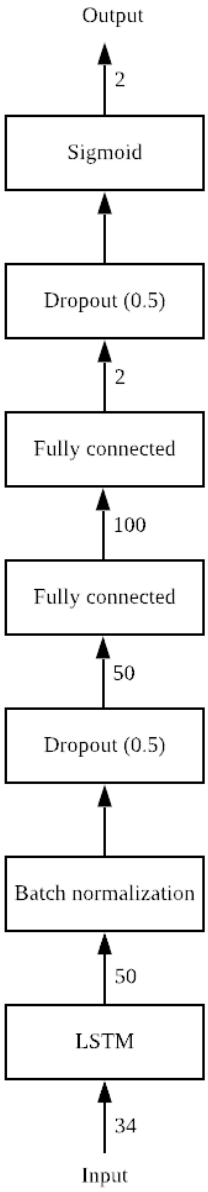

Figure 2: LSTM architecture for Facial Landmark Point Model

3.3 Facial Landmark Point Model (FacialLM)

The workflow for the facial landmark point model is shown in Fig 1. After the videos are split into frames images, the face would first have to be detected and cropped for the extraction of facial landmark points. The dlib library [11] is used to detect, crop the facial images from the frame images and to extract 68 facial landmark points from the cropped face. We then derive 34 distance features from

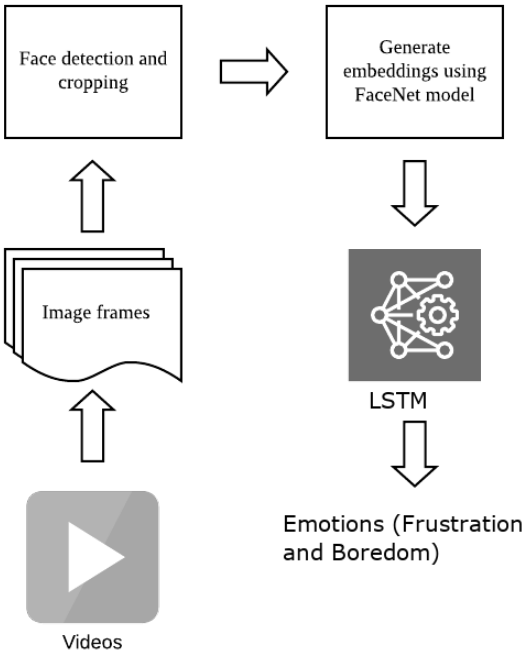

Figure 3: Processing workflow for FaceNet Embedding Model

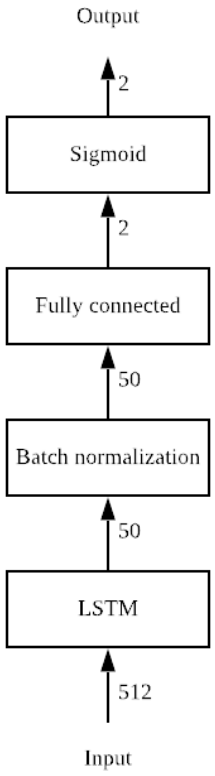

Figure 4: LSTM architecture for FaceNet Embedding Model

the facial landmark points by calculating the Euclidean distances between the points similar to the approach by [14]. The distance features are then passed into a LSTM.

The architecture for the LSTM is shown in Fig 2. Trading off between accuracy and computational speed, we opted for a single layer of LSTM. For each batch of 64, 30 sequences (only 1 out of every 10 frames are used) of facial distances with a dimensional size of 34 are passed into the LSTM.

For the LSTM, after batch normalization and drop out, the output is passed through 2 fully connected layers (with 50 and 100 neurons respectively). The final prediction is derived from a sigmoid layer with 2 prediction scores – one for boredom and the other for frustration. A sigmoid is used here as we model this as a multi-label classification problem where both boredom and frustration may be occurring for the same instance. To train the model, the Stochastic Gradient Descent (SGD) is used with Adam technique [12] for optimization and with a binary cross-entropy loss function. An initial learning rate 0.001 is exponentially decayed with a gamma of 0.95 by multiplying the learning rate by 0.95 every epoch. We then trained the model for a total of 50 epochs with a batch size of 64 before passing in the test data to test the generalizability of the model.

3.4 FaceNet Embedding Model (FaceNetEmbed)

The processing workflow for the FaceNet embedding model as shown in Fig 3 is similar to the facial landmark point model except that MTCNN library [26] is used for face detection and cropping (instead of dlib) and the extraction of facial landmark points is replaced with the extraction of facial embeddings using FaceNet model. The base model used in FaceNet is InceptionResnet pretrained with VGGFace2 dataset [2].

The LSTM architecture as shown in Fig 4 is used to model the temporal relations of the facial embeddings similar to the facial landmark model. For each batch of 64, 30 sequences (only 1 out of every 10 frames are used) of facial embedding with a dimensional size of 512 are passed into the LSTM model. The architecture of the LSTM model is shown in Fig 3. To train the model, Stochastic Gradient Descent (SGD) was used with Adam technique for optimization and with a binary cross-entropy loss function. An initial learning rate 0.001 is exponentially decayed with a gamma of 0.95 by multiplying the learning rate by 0.95 every epoch. We then trained the model for a total of 20 epochs with a batch size of 32 before passing in the test data to test the generalizability of the model.

Table 1: Accuracy of models

Models	Boredom	Frustration
EmotioNet	35.89%	73.09%
FacialLM	58.78%	60.94%
FaceNetEmbed	52.15%	70.67%

Table 2: Precision, Recall and F1 score of models

Models	Boredom			Frustration		
	Precision	Recall	F1 score	Precision	Recall	F1 score
FacialLM	0.593	0.842	0.696	0.29	0.75	0.419
FaceNetEmbed	0.56	0.56	0.56	0.28	0.489	0.356

4 EXPERIMENT

4.1 Evaluation metrics

In this paper, the performance of the models are evaluated on the test data set. We report the accuracy, precision, recall and F1 score for both models. The computations for precision (P), recall (R) and F1 score is given below.

$$P = \frac{TP}{TP + FP}$$

$$R = \frac{TP}{TP + FN}$$

$$F1 = \frac{2 * (P * R)}{P + R}$$

where TP denotes True Positive, TN denotes True Negative, FP denotes False Positive, FN denotes False Negative.

4.2 Results and Discussion

The accuracy of the models are shown in Table 1. In [8], the authors used EmotioNet, a CNN used for emotion recognition in photographs of human faces as a benchmarking model. As we are using DAiSEE as the dataset for training and evaluating our models as well, we thus quote the evaluation performance (only accuracy is reported in [8]) of EmotioNet on DAiSEE as reported by the authors.

As can be seen from the table, FaceNetEmbed model achieved higher accuracies for detection of both boredom and frustration when compared against the FacialLM model. As compared to the EmotioNet model, accuracy for detection of boredom is higher by 16.26% while accuracy for frustration is lower by 2.42%.

The precision, recall and F1 score performance of the models are shown in Table 2. The F1 score for the FacialLM model for the detection of boredom is higher than that of FaceNetEmbed model by 0.136 and F1 score for detection of frustration for FacialLM model is also higher than that of FaceNetEmbed model by 0.063.

In terms of the recall figures, the FacialLM model outperforms the FaceNetEmbed model by a huge margin. We attempted to improve the recall for FaceNetEmbed model by changing the class weight (i.e. penalizing wrong classification of incidences of frustration more) but the recall figure did not change much. The recall for FacialLM model on the other hand, in the detection of boredom is higher than that for FaceNetEmbed by 0.282. Similarly, the recall for FacialLM model in the detection of frustration is higher than that for FaceNetEmbed by 0.261.

In terms of accuracy, the performance of the proposed FaceNetEmbed model is comparable to that of the benchmark EmotioNet model (with better accuracy for detection of boredom). The FacialLM model has a lower accuracy though for detection of frustration. However, the FacialLM model achieves higher recall and F1 score than FaceNetEmbed model.

We intend to use the model in a tutoring system deployed on a mobile device which will detect the boredom and frustration of learners on a moment-to-moment basis. The high recall in the detection of both frustration and boredom would mean that most instances of frustration and boredom can be detected albeit at the expense of lower precision or a higher false positive rate. Depending on the design of the tutoring system, a higher number of false positives in detecting frustration of learners might just result in dispensing of pedagogical responses (with the intention of lowering frustration levels) when learners don't really need them which should not really impact on the learning itself. Similarly, autonomous intervention by the tutoring system to mitigate boredom when the learners are not bored would not have much impact on learning motivation as well.

5 CONCLUSION

A tutoring system that can autonomously detect incidences of academic emotions such as frustration and boredom in a learner would result in enhanced tutoring outcomes by sustaining the motivation of the learners. The detection of these academic emotions is however challenging as they are more subtle as compared to the basic emotions.

In this study, we evaluated the use of deep learning on FaceNet embeddings and facial landmark points for academic emotion detection on a publicly available dataset - DAiSEE that has been annotated with the emotional states of engagement, boredom, frustration and confusion. Other than the spatial dimensions, we also modeled the temporal dimensions of the facial videos to enhance the accuracy of detection.

The results demonstrated that though the accuracy of detection is higher for the FaceNet embeddings model, the facial landmark

points model is better at distinguishing between incidences of occurrence and non-occurrence for both boredom and frustration. This would be pivotal to the moment-by-moment monitoring of boredom and frustration of learners using a tutoring system either online or in a classroom. As a further extension to this study, we intend to deploy the model in a tutoring software hosted on mobile phones and tablets. Trials would be conducted to capture contextual logs of learners e.g. difficulty level of question, the duration of time spent on each question e.t.c. The contextual logs would serve as additional features for input to the model to enhance the accuracy of detection of the academic emotions, bringing us a step closer to autonomous personalized tutoring for every learner.

6 ACKNOWLEDGMENTS

This work was supported by the Singapore Ministry of Education Translational R&D and Innovation Fund (MOE-TIF) 11th Grant Call.

REFERENCES

[1] Ran Breuer and Ron Kimmel. 2017. A deep learning perspective on the origin of facial expressions. *arXiv preprint arXiv:1705.01842* (2017).
[2] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. 2018. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE, 67–74.
[3] Wen-Sheng Chu, Fernando De la Torre, and Jeffrey F Cohn. 2017. Learning spatial and temporal cues for multi-label facial action unit detection. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*. IEEE, 25–32.
[4] Matthew Day. 2016. Exploiting facial landmarks for emotion recognition in the wild. *arXiv preprint arXiv:1603.09129* (2016).
[5] M Ali Akber Dewan, Mahbub Murshed, and Fuhua Lin. 2019. Engagement detection in online learning: a review. *Smart Learning Environments* 6, 1 (2019), 1.
[6] Paul Ekman. 1992. An argument for basic emotions. *Cognition & emotion* 6, 3-4 (1992), 169–200.
[7] Ian J Goodfellow, Dumitru Erhan, Pierre Luc Carrier, Aaron Courville, Mehdi Mirza, Ben Hamner, Will Cukierski, Yichuan Tang, David Thaler, and Dong-Hyun Lee. 2013. Challenges in representation learning: A report on three machine learning contests. In *International Conference on Neural Information Processing*. Springer, 117–124.
[8] Abhay Gupta, Arjun D'Cunha, Kamal Awasthi, and Vineeth Balasubramanian. 2016. Daisee: Towards user engagement recognition in the wild. *arXiv preprint arXiv:1609.01885* (2016).
[9] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.
[10] Heechul Jung, Sihaeng Lee, Junho Yim, Sunjeong Park, and Junmo Kim. 2015. Joint fine-tuning in deep neural networks for facial expression recognition. In *Proceedings of the IEEE international conference on computer vision*. 2983–2991.
[11] Davis E King. 2009. Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research* 10, Jul (2009), 1755–1758.
[12] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).
[13] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. 1998. Gradient-based learning applied to document recognition. *Proc. IEEE* 86, 11 (1998), 2278–2324.
[14] Chuanhe Liu, Tianhao Tang, Kui Lv, and Minghao Wang. 2018. Multi-feature based emotion recognition for video clips. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*. 630–634.
[15] Zhilei Liu and Shangfei Wang. 2011. Emotion recognition using hidden Markov models from facial temperature sequence. In *International Conference on Affective Computing and Intelligent Interaction*. Springer, 240–247.
[16] André Teixeira Lopes, Edilson de Aguiar, Alberto F De Souza, and Thiago Oliveira-Santos. 2017. Facial expression recognition with convolutional neural networks: coping with few data and the training sample order. *Pattern Recognition* 61 (2017), 610–628.
[17] MINP Munasinghe. 2018. Facial expression recognition using facial landmarks and random forest classifier. In *IEEE/ACIS 17th International Conference on Computer and Information Science (ICIS)*. IEEE, 423–427.
[18] Omid Mohamad Nezami, Mark Dras, Len Hamey, Deborah Richards, Stephen Wan, and Cécile Paris. 2018. Automatic Recognition of Student Engagement using Deep Learning and Facial Expression. *arXiv preprint arXiv:1808.02324* (2018).
[19] Reinhard Pekrun, Thomas Goetz, Wolfram Titz, and Raymond P Perry. 2002. Academic emotions in students' self-regulated learning and achievement: A program of qualitative and quantitative research. *Educational psychologist* 37, 2 (2002), 91–105.
[20] Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 1532–1543.
[21] Rosalind Wright Picard. 1995. Affective computing. (1995).
[22] KM Rajesh and M Naveenkumar. 2016. A robust method for face recognition and face emotion detection system using support vector machines. In *2016 International Conference on Electrical, Electronics, Communication, Computer and Optimization Techniques (ICEECCOT)*. IEEE, 1–5.
[23] Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 815–823.
[24] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).
[25] Y-I Tian, Takeo Kanade, and Jeffrey F Cohn. 2001. Recognizing action units for facial expression analysis. *IEEE Transactions on pattern analysis and machine intelligence* 23, 2 (2001), 97–115.
[26] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. 2016. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters* 23, 10 (2016), 1499–1503.
[27] Lei Zhang, Yan Tong, and Qiang Ji. 2008. Interactive labeling of facial action units. In *19th International Conference on Pattern Recognition*. IEEE, 1–4.