Engagement Estimation for an E-Learning Environment Application

Win Shwe Sin Khine
School of Information Science
Japan Advanced Institute of
Science and Technology
Ishikawa, Japan
Email: winshwesinkhine@jaist.ac.jp

Shinobu Hasegawa
Research Center for
Advanced Computing Infrastructure
Japan Advanced Institute of
Science and Technology
Ishikawa, Japan
Email: hasegawa@jaist.ac.jp

Kazunori Kotani
School of Information Science
Japan Advanced Institute of
Science and Technology
Ishikawa, Japan
Email: ikko@jaist.ac.jp

***Abstract*—In this study, we conducted an estimation of engagement through virtual learning environment by using facial images. We aim to improve student learning rates and get a better understanding of them through facial expressions. Nowadays, computation power and memory capacity are available for analysis on large scale datasets. As a result, deep learning techniques can effectively extract useful features from the given dataset over traditional approaches. Unfortunately, deep learning-based methods require a massive amount of labeled data. Although there are many face datasets for face related problems, such as face detection and face recognition, it is still limited to facial expressions. To overcome this limitation, we use the advantages of the style transfer technique to obtain the basic features of the face and eliminate the features that are not useful for engagement estimation. In our experiment, we use the Visual Geometry Group-16 (VGG-16) face model to extract the prominent basic features of the face and eliminate the non-related features by differing peak and neutral frames. We demonstrated the practical use of our method through the efficiency of detecting student engagement. The results show that our proposed method provides 50% accuracy in engagement estimation.**

Keywords–E-learning; Engagement; Fine-tuning.

I. INTRODUCTION

Since Information Technology (IT) is incredibly improving in the 21st century, the usage of the virtual system is gradually increasing. Therefore, it is essential to maintain good experience and communication between users and the system. Human-Computer Interaction (HCI) becomes a significant concern in the IT field. From historical statistics, we know that, since the 1980s, there is a significant drop out rate, numerically between 25% and 60% shows by (see Larson and Richards [1]) for the participants in the learning system because students are extremely bored and not interested in their lectures. Maintaining the students' willingness, alternatively, is called engagement. Besides, interaction with a virtual system becomes important in HCI. As a result, it is a widely discussed topic in the educational area.

Learning methods of the adequate level of e-learning based on facial expressions can be divided into two categories. They are traditional hand-crafted based and deep-learning-based methods. The hand-crafted based methods typically consist of feature extraction and classification stages. In the feature extraction stage, appearance or geometric features are extracted by using traditional methods, such as Gabor filters [2], Local Binary Patterns (LBP) [3], Histograms of Oriented Gradients (HOG) [4]. Furthermore, the appearance features are depending on environmental settings, such as lighting, background, pose, and many other sensitive effects. On the other hand, geometric features are depending on prominent facial features and curvature of the face. The problem is that, when researchers use hand-crafted feature extraction methods, they need to set the constant environmental settings to get a stable result. However, facial expressions are depending on many variables and factors. Therefore, hand-crafted feature extraction is not feasible to extract facial features in the wild.

To overcome the difficulties of hand-crafted feature extraction methods, deep-learning-based feature extraction methods have been adopted. They provide impressive performance in face-related tasks, such as face detection, face recognition, facial expression recognition, and engagement estimation.

This paper is organized as follows; Section 2 focuses on the relevant study of facial expressions recognition related to deep-learning-based methods. Section 3 describes the dataset of this study. Section 4 presents the method used to conduct experiments in this study and discusses the obtained result. Section 5 summarizes our findings, draws conclusions based on our research objectives, and suggests potential improvements to this study.

II. LITERATURE REVIEWS

Mayya et al. [5] designed a deep neural network architecture called Deep Convolution Neural Network (DCNN) to learn features for recognizing six basic facial expressions from a single image. 96% recognition rate is achieved based on Extended Cohn-Kanade (CK++) and Japanese Female Facial Expressions (JAFFE) datasets, which are Action Unit (AU)-coded expression datasets. They discovered that more layers in the deep convolution neural network increases the ability to extract more features of an image when compared to few layers.

Jain et al. [6] designed a similar architecture with Mayya et al. [5]. However, the difference between these two models is that they use residual blocks after the convolution layer to prevent the degrading problem in which the gradient line cannot learn the data properly. It also saturates the accuracy at some point of time and finally degrades the model performance. Alternatively, it is a so-called gradient varnishing

 ISBN: 978-1-61208-761-0

problem. Therefore, they used two residual blocks, which can help prevent the degrading problem and improve the accuracy. After that, they trained the model with CK++ and JAFFE datasets and obtained 95% of recognition rates, a result that is close to Mayya's model.

He and Zhang [7] designed a model consisting of two networks. The first network is a binary positive or negative classification model which is used to disintegrate the emotions into a binary class, and serves as an extra input to the second one for specific emotion recognition. When the input patches are fed into the network, the first model gives a positive or negative result to the second network and serves as prior knowledge. The second model extracts the features from input patches and gives the classification result based on the learned features and the prior binary result. From the study, they achieved up to 64% of overall accuracy using the Image Emotion Dataset, which contains downloaded images from Flickr and Instagram with searching eight emotions as keywords.

Chen et al. [8] developed a model for recognition of basic emotions with a limited amount of training images. It consists of three parts. The first part is the extraction of face-related features with deep face model VGG-16 [9]. The results of VGG-16 show that the adopted model obtained high performance in feature extraction. However, some features are not necessary for the recognition of facial expressions. Therefore, in the second part, they used k-means clustering to cluster all frames into two groups, such as peak-like frames and neutral-like frames. After clustering, they used the semi-classification method like Support Vector Machine (SVM++) to determine the clustered groups and retrieved the critical peak and neutral frames which are closer to the centers. After getting keyframes, they calculated differences between the key peak frame and the key neutral frame in order to eliminate the face identity information. Finally, in the last part, they perform multi-classification for basic classified emotions. Their experimental results show that their model achieved 78.4% of recognition rate using the Binghamton University 3D Facial Expression (BU3DFE) dataset.

Sabri and Kurita [10] show that the performance of the Convolution Neural Network (CNN) is dependent on the labeled data. With a limitation of labeled data, it is not feasible for CNN learn and extract information from the data. To get the general property of CNN, Koch et al. [11] proposed a Siamese network that can learn unfamiliar features by utilizing extra information about the relationship between input pairs. However, compared to the features representations that are explicitly learned by the model, the learned features by the Siamese network produces results lower than the average. Therefore, a triplet network model is introduced by Wang and Gupta [12] to overcome this problem. It consists of three input vectors instead of two in the Siamese network. Their results are comparable to the ones that are explicitly learned by the model. Moreover, the triplet network does not require the class labels of the processing input. Inspired by the benefits of the Siamese and Triple network models, Sabri and Kurita [10] use three data frames in their approach. The data frames consist of the onset, neutral, and apex of spatio-temporal data. They are used for the estimation of facial expression intensity against the basic emotions. Their study shows that it obtained the accuracy of up to 86% on the Cohn-Kanade (CK) facial expression dataset.

III. DAiSEE: Dataset for Affective States in E-Environment

To train our model, we utilize DAiSEE, which is a multi-labeled classification dataset. The number of interactions with computers has greatly increased in recent years, and the interaction between users and computers has become more significant than in the past. The interaction of the user with the system can change the response that the system returns, depending on the level of engagement of each user. Based on this effect, Gupta et al. [13] simulated the environment and constructed DAISEE dataset for testing the developed model. The dataset is used to recognize human affective states, which are based on how much users of an online system are engaged or satisfied while using the system. The examples of the application are online shopping, health care system, e-advertisement, online learning system, and others.

Figure 1. Examples images from DAiSEE dataset with engagement levels 0 (leftmost), 1, 2 and 3 (rightmost) respectively

DAiSEE consists of 9,068 videos with 10 seconds duration that is captured from 112 users for recognizing user affective states, such as boredom, confusion, engagement, and frustration. Gupta et al. [13] defined that levels of labels have four states ranging the values from 0 to 3. The engagement values '0', '1', '2' and '3' namely refer to 'very low level', 'low', 'high' and 'very high' level of engagement respectively. The example images with different engagement levels from the DAiSEE dataset are shown in Figure 1.

IV. The Applied Method

In this section, we describe the applied methods in the implementation stage. Figure 2 shows our proposed model.

Training a deep convolution neural network requires a massive amount of labeled data. If the labels are not enough for training, the model cannot learn properly on unknown data and lead to an over-fitting problem. Since research topics related to face, such as face detection and recognition, have been developed from a few decades ago, the techniques and learning methods to solve these types of problems are nearly optimum. Besides, there are different types of large scale face datasets for evaluation of these methods. Unfortunately, these datasets are useful for face detection and recognition but not for engagement. To overcome this issue, we use the transfer learning technique, which can transfer the learned features from one specific problem to another if both problems are similar. For example, face recognition and engagement recognition are quite similar to each other, because both are recognition based on the face. In face recognition that is based

 ISBN: 978-1-61208-761-0

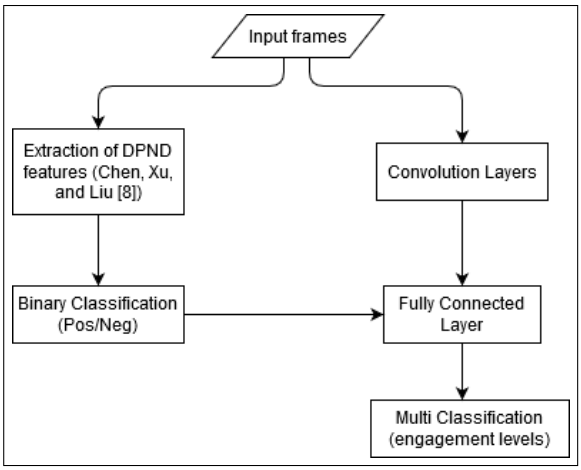

Figure 2. The proposed system design

on the deep learning method, it needs to extract prominent face features for classification. The outcome is similar to engagement recognition, that classifies the level of engagement based on facial expressions; it also requires prominent facial features that are changing in image sequences. As a result, we can transfer learned features from face problems to engagement estimation problems.

For a base model, we used the VGG-16 face model, which is pre-trained on large-scaled face datasets, comprising 2.6 million images. Parkhi et al. [9] collected the face images and classified them into 2,622 classes of celebrities. The based-model achieved over 98% accuracy in face recognition. Recognizable results of the VGG-16 model proved that it could learn face features for face recognition by all means. Inspired by the advantage of the VGG-16 model, we adopted the VGG-16 model for the extraction of deep representations of input frames.

Unfortunately, features that are extracted by the model for face recognition may contain face identity information, which is useful for face recognition but not for engagement estimation, because VGG-16 is explicitly trained for recognizing faces. Therefore, it is necessary to eliminate face-related information. We used the technique from Chen et al. [8] to eliminate face identity information. In their method, they used Deep Peak Neutral Difference (DPND) features for training the model. In DPND features, they removed face identity information by finding the differences between key peak and key neutral frames. The reason is that appearances are changing only in the face in different frames. As a result, by finding differences among these keyframes, we can obtain the features for recognition of engagement only.

After we obtained the extracted features, the features are classified into binary class, namely positive and negative classes. In parallel, we fed the same input to convolution layers to extract features and used the binary classification result to assist the final multi-classification. The result is for recognizing levels of engagement, ranging values from 0 to 3.

A. Preprocessing

DAiSEE dataset provides the videos with 10 seconds duration in which students' facial expressions are recorded while playing lectures. To feed these videos to the model, they are converted into frames sequences by using Fast Forward Motion Picture Experts Group (FFmpeg) [14] video converter. It converts the videos into frame sequences with 30 frames per second (fps) and gives 300 frames per video. In Gupta et al. [13], they defined each video with engagement labels. Therefore, each frame sequence that is extracted from videos is also labeled the same as videos.

B. Extraction of Deep Representations of Frames

Inspired by the advantages of the VGG-16 face model, which is developed by Parkhi et al. [9], we used their model as a based development of our model network to extract prominent features of the face. In the architecture of the VGG-16 model, it uses five blocks of convolution layers, followed by max-pooling layers.

After constructing the model, the input frames are fed into the network; it converts the images into matrix representation and passes it to convolution layers. At the convolution layers, it does pairwise multiplication with kernels throughout the whole matrix to get the general features of the entire image and gives the feature maps as an outcome. This result is passed to max-pooling layers to reduce the dimension of feature maps, followed by non-linear activation, so-called Rectified Linear Unit (relu), to return the result. The activation unit activates the neuron if the inputs are greater than zero; otherwise, it does not activate and returns zero as a result. In the configuration of the VGG-16 model, the input image is limited to 224 by 224 dimensions. The 3 by 3 as kernels are used and defined as the first block of the convolution layer. Similar to the first block, the next four convolution blocks are designed and followed by fully connected layers to perform multi-classification.

1) Extraction Features by VGG-16 Model with Engagement Labels: In the first step, before inputting the frames into the VGG-16 model, we used the same processing steps of the VGG-16 model by Parkhi et al. [9]. In their study, they fed the model with cropped 224 by 224 patches of input images from four corners and centers. They also performed data-augmentation of horizontal flipping with a 50% probability during training. For our study, we performed the preprocessing steps similar to Parkhi's study and visualized the result of preprocessing, as shown in Figure 3.

Figure 3. Visualization Preprocessing Result of VGG-16

 ISBN: 978-1-61208-761-0

In order to get a better understanding of extracted features from convolution layers, the feature maps are visualized from the first and last convolution blocks, as shown in Figures 4 and 5, respectively. In visualization, all feature maps of the first and last blocks of convolution layers are visualized as 8 by 8 square images. The result shows that some feature maps are focusing on the foreground, whereas the others are focusing on the background. According to the visualization result, the feature maps of the first convolution layer still have the input shapes, and we can guess what feature maps look like by looking at the visualization of feature maps. However, visualization of the last block of the convolution layer shows that a deeper layer of deep neural network extracts more general features rather than the shallow layers for classification. It implies that if we want to get more general deep representations, we have to use the result from deeper layers. Therefore, we compared the discrimination capabilities of fully connected layers, namely 'fc6' and 'fc7', adopted by Parkhi et al. [9].

Figure 4. Visualization of Feature Maps from Convolution Layer (1st block)

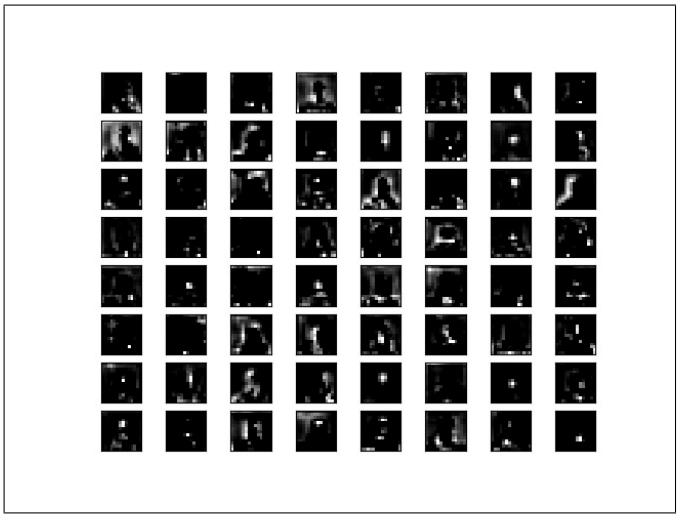

Figure 5. Visualization of Feature Maps from Convolution Layer (5th block)

For the evaluation of the model, we used 8,100 images for training and 2,100 images for testing. DAiSEE has 16 million frames from 112 users, which supports deep learning models. However, our purpose is to reuse the extracted features from the pre-trained model, such as VGG-16, for transferring the prior information and saving training time instead of learning from scratch. Therefore, we restricted our dataset to be small based on our purpose, and 0.005% samples of the original dataset are randomly selected from the video sequences. The randomized numbers of training images and validation are shown in Table I.

TABLE I. SAMPLES SET STRUCTURE

Train Labels	*# of samples*	*Validation Labels*	*# of samples*
Label 0	300	**Label 0**	**0**
Label 1	**0**	Label 1	600
Label 2	3600	Label 2	900
Label 3	4200	Label 3	600

Randomized selection of samples from the dataset cannot fairly cover all the labels. According to Table I, there are no samples sequences of engagement label '1' for training and label '0' for validation. So that, randomized samples from both sets are combined into a set and utilized leave-one-out of the v-fold cross-validation method [15]. It splits the dataset into a 'v' number of subsets and uses one subset for validation. The remaining subsets are for training the model. We repeated the process up to 10-fold. 9 fold is used for training the model, and the remaining is for validated the model. The results of 10-fold cross-validation are shown in Table II. We obtained an accuracy of around 50% for both training and validation.

TABLE II. LOSS AND ACCURACY OF FINE-TUNING VGG-16 MODEL WITH 10-FOLD CROSS VALIDATION

# of epoch	*train_loss*	*train_accuracy*	*val_loss*	*val_accuracy*
1	1.1204	0.4625	0.9317	0.4725
2	0.9471	0.4654	0.9250	0.4422
3	0.9398	0.4657	0.9091	0.4471
4	0.9309	0.4690	0.8864	0.4588
5	0.9233	0.4696	0.9083	0.4657
6	0.9246	0.4620	0.9006	0.4588
7	0.9223	0.4618	0.8804	**0.4784**
8	0.9188	0.4669	0.8933	0.4824
9	0.9114	**0.4747**	0.9353	0.4745
10	0.9168	0.4703	0.8839	0.4765

TABLE III. LOSS AND ACCURACY OF DEEP REPRESENTATIONS FROM 'FC6' DENSE LAYER

# of epoch	*train_loss*	*train_accuracy*	*val_loss*	*val_accuracy*
1	1.5751	0.4728	4.0148	0.4657
2	8.3836	**0.4814**	13.4151	0.4657
3	13.8168	0.4764	14.5101	0.4765
4	14.3863	0.4748	14.8962	0.4814
5	14.5621	0.4796	14.9514	0.4549
6	14.6212	0.4723	14.8178	**0.4941**
7	14.7216	0.4719	14.7799	0.4814
8	14.7823	0.4749	14.5372	0.4853
9	14.7558	0.4769	14.7691	0.4843
10	14.8142	0.4748	14.7804	0.4843

To determine which layers are suitable to extract more general features, we followed the usage from Parkhi et al. [9]. The fully connected layers, namely, 'fc6' and 'fc7', are considered as candidate layers. They are the last two layers before the final classification for deep representations. We utilized linear Support Vector Machine (SVM) to investigate the abilities of the facial expressions classification and compared the results from these two dense layers. The results are shown in Tables III and IV for 'fc6' and 'fc7' layers, respectively. Based on

 ISBN: 978-1-61208-761-0

TABLE IV. LOSS AND ACCURACY OF DEEP REPRESENTATIONS FROM 'FC7' DENSE LAYER

# of epoch	*train_loss*	*train_accuracy*	*val_loss*	*val_accuracy*
1	11.1773	0.4644	13.8237	0.4716
2	12.5062	0.4667	13.7523	0.4637
3	12.4247	0.4700	13.5395	0.4696
4	12.4178	0.4666	13.6752	**0.4892**
5	13.6556	0.4664	14.6021	0.4716
6	13.9989	0.4658	14.2492	0.4824
7	13.9343	**0.4745**	14.1655	0.4657
8	14.0053	0.4686	14.2087	0.4745
9	13.9698	0.4690	14.1892	0.4706
10	13.9838	0.4667	14.3439	0.4853

the results, deep representations from 'fc6' dense layers can extract more general features. Therefore, they are used in the next section.

C. Clustering of Peak and Neutral Frames

According to our proposed method, we obtained deep representations for engagement estimation from the 'fc6' layer of the VGG-16 model. We discovered that the detection of peak and neutral frames are essential to assist in the estimation process. In order to determine the peak and neutral frames, firstly, all of the frames are clustered into two groups. A first group is a peak-like group, and the second one is a neutral-like group. In this case, the number of clusters is known, numerically 2, which categorizes the frames into peak and neutral. Therefore, the k-means method is used for the clustering task because of its simplicity and optimum if the number of k is known.

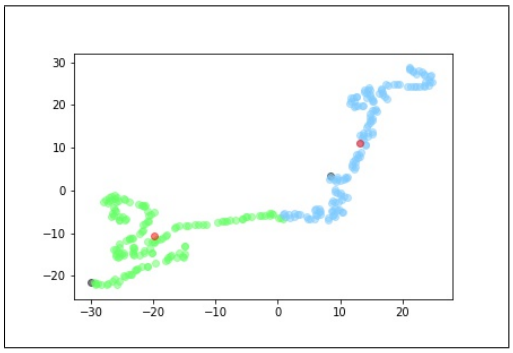

Figure 6. k-means cluster for peak and neutral frames where red: centers, black: sample, green: cluster 0, blue: cluster 1

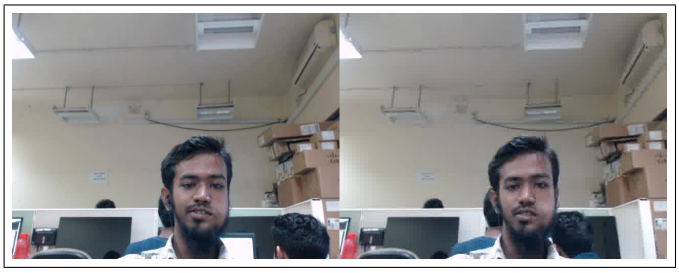

Figure 7. Samples from k-means cluster 0 (left) and cluster 1 (right)

However, deep representations which are obtained from the 'fc6' layer of VGG-16 have 4,096 dimensions. It is not feasible to perform clustering as the data are in the form of a high dimension. Therefore, before the clustering method is performed, it is crucial to make dimension reduction in dealing with this high dimensional data. The dimension reduction technique is performed by using t-Distributed Stochastic Neighbor Embedding (t-SNE), which is a non-linear feature extraction method. The non-linear feature extraction method calculates the probability distributions of similarity of points in both high and low dimensions. It also minimizes the difference between these two distributions by using Kullback-Leibler divergence [16]. The reason for choosing the t-SNE method is because of the non-linear property of deep representations. In the VGG-16 model, Rectified Linear-Unit (relu) activation units are used to select the suitable neurons in the network to propagate the data. Besides, relu is a limited type of non-linear function due to the selection of maximum value between 0 and x (input). It serves as a linear function if the inputs are negative values, whereas it will be a non-linear function for positive inputs. Therefore, after applying non-linear activation units to filter the neuron. As a result, the output feature maps have non-linear properties. It is comprised of 4,096 dimensions. These dimensions are reduced from 4096 to 2, and the clustering results are shown in Figure 6. In addition, the samples from the experimental result of two clusters are reconstructed and visualized in Figure 7. In Figure 7, after k-means clustering, the first frame of a video sequence is classified as cluster 0. Generally, facial expressions of the student at the beginning of the lectures seem to be relaxed and have neutral expressions. As the lecture continued, based on their engagement levels, facial expressions of the students changed. We discovered that in cluster 1, the student is looking at the camera and seems to be interested in the lecture. Therefore, the frames in the cluster 1 are classified as peak frames, whereas the others in the cluster 0 are classified as neutral frames.

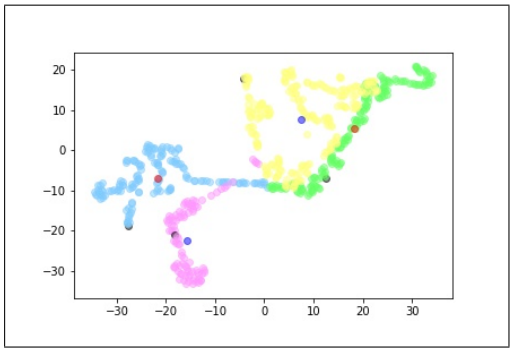

Figure 8. k-means cluster for peak and neutral frames where red or dark blue: centers, black: samples, green and yellow: cluster 0, blue and pink: cluster 1

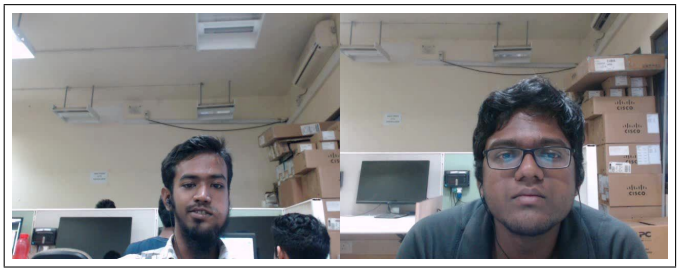

Figure 9. Samples from k-means cluster 0

The way for expressing internal feelings depends on people and their characteristics. Each person expresses six basic

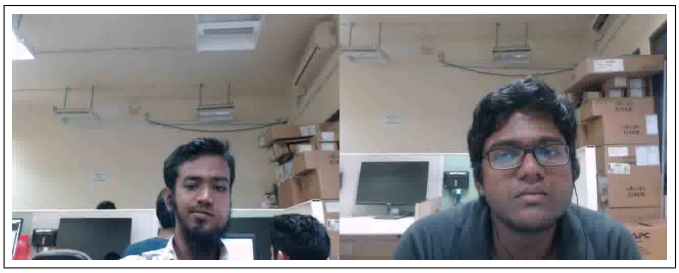

Figure 10. Samples from k-means cluster 1

emotions in different ways. For example, the first person shows happiness with a smiling face, whereas the second one expresses with laughing. In an alternative word, it is conditioning on individual differences. Therefore, individual differences among different people are also considered in clustering for peak-like and neutral-like groups. The same procedures are performed, and the experimental results are shown in Figure 8. After clustering, peak and neutral frames are grouped together based on their respective features. Besides, the samples from each cluster are also shown in Figures 9 and 10 for both cluster 0 and cluster 1. Based on the visualization result, persons from cluster 0 in Figure 9 are looking at the camera directly and seem to be listening to their lectures. Their faces express normal expressions compared to samples from cluster 1. However, in Figure 10, the person in the right image shows that his eyes were looking away from the camera and did not concentrate on the lectures. Therefore, according to their facial expressions, cluster 1 can be classified as a group of peak frames, whereas cluster 0 can be classified as a group of neutral frames.

V. Conclusion and Future Work

To conclude our study, we discovered that the deep learning-based methods require a large scale of labeled data. However, in some problems, obtaining the desired dataset for training is impossible. In order to overcome the dataset limitation, the transfer learning technique is performed to train the proposed model. The process consists of two steps. In the first stage, fine-tuning the pre-trained model is performed. Moreover, the benefit of the VGG-16 model, such as a higher face recognition rate, is used to assist in engagement estimation and to reduce the process of computational time and resources while training the model. In our experimental result, we achieved the result of fine-tuning up to 50% of accuracy. Although the results still have gaps to make improvements. Fortunately, we discovered that these results could be compared with the ones that are obtained from explicitly training for engagement estimation without using transfer learning.

For a second stage of the model, elimination of face identity information is performed. All frames are divided into two groups: peak-like frames and neutral-like frames. So to dealing with this problem, k-means clustering is performed. The frame is clustered based on individual differences because of the differences in expressing their internal emotions. According to our result, k-means clustering shows that it performed well to divide the same people characteristic into the same group. We also obtained the peak-like and neutral-like frames clusters based on individual differences. In the future, we will improve our proposed model to handle more levels of facial expressions on online learning engagement.

Acknowledgment

This work was supported by IMAGICA GROUP. We would like to express our sincere gratitude to their support during our research project.

References

[1] R. W. Larson and M. H. Richards, "Boredom in the middle school years: Blaming schools versus blaming students," American journal of education, vol. 99, 1991, pp. 418–443, ISSN: 0195-6744.

[2] T. R. Almaev and M. F. Valstar, "Local gabor binary patterns from three orthogonal planes for automatic facial expression recognition," in 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction. IEEE, 2013, pp. 356–361, ISSN: 0769550487, URL: https://ieeexplore.ieee.org/ [accessed: 2020-02-23].

[3] W.-L. Chao, J.-J. Ding, and J.-Z. Liu, "Facial expression recognition based on improved local binary pattern and class-regularized locality preserving projection," Signal Processing, vol. 117, 2015, pp. 1–10, ISSN: 0165-1684.

[4] Z. Luo, L. Liu, J. Chen, Y. Liu, and Z. Su, "Spontaneous smile recognition for interest detection," in Chinese Conference on Pattern Recognition. Springer Singapore, 2016, pp. 119–130, ISBN: 978-981-10-3002-4 , URL: https://link.springer.com/ [accessed: 2020-02-23].

[5] V. Mayya, R. M. Pai, and M. M. Pai, "Automatic facial expression recognition using dcnn," Procedia Computer Science, vol. 93, 2016, pp. 453–461, ISSN: 1877-0509.

[6] D. K. Jain, P. Shamsolmoali, and P. Sehdev, "Extended deep neural network for facial emotion recognition," Pattern Recognition Letters, vol. 120, 2019, pp. 69–74, ISSN: 0167-8655.

[7] X. He and W. Zhang, "Emotion recognition by assisted learning with convolutional neural networks," Neurocomputing, vol. 291, 2018, pp. 187–194, ISSN: 0925-2312.

[8] J. Chen, R. Xu, and L. Liu, "Deep peak-neutral difference feature for facial expression recognition," Multimedia Tools and Applications, vol. 77, 2018, pp. 29 871–29 887, ISSN: 1380-7501.

[9] O. M. Parkhi, A. Vedaldi, A. Zisserman et al., "Deep face recognition." in bmvc, vol. 1, no. 3. British Machine Vision Association, 2015, p. 1–12, URL: https://ora.ox.ac.uk/ [accessed: 2020-02-23].

[10] M. Sabri and T. Kurita, "Facial expression intensity estimation using siamese and triplet networks," Neurocomputing, vol. 313, 2018, pp. 143–154, ISSN: 0925-2312.

[11] G. Koch, R. Zemel, and R. Salakhutdinov, "Siamese neural networks for one-shot image recognition," in ICML deep learning workshop, vol. 2. Lille, 2015, URL: http://www.cs.toronto.edu/ [accessed: 2020-02-23].

[12] X. Wang and A. Gupta, "Unsupervised learning of visual representations using videos," in Proceedings of the IEEE International Conference on Computer Vision, December 2015, pp. 2794–2802, URL: http://openaccess.thecvf.com/ [accessed: 2020-02-23].

[13] A. Gupta, A. D'Cunha, K. Awasthi, and V. Balasubramanian, "Daisee: Towards user engagement recognition in the wild," CoRR, vol. abs/1609.01885, 2016.

[14] "FFmpeg," URL: https://www.ffmpeg.org/ [accessed: 2020-02-23].

[15] P. Burman, "A comparative study of ordinary cross-validation, v-fold cross-validation and the repeated learning-testing methods," Biometrika, vol. 76, 1989, pp. 503–514, ISSN: 1464-3510.

[16] S. Kullback and R. A. Leibler, "On information and sufficiency," The annals of mathematical statistics, vol. 22, 1951, pp. 79–86, ISSN: 0003-4851.

 ISBN: 978-1-61208-761-0