Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 192 (2021) 951–960

Procedia Computer Science

www.elsevier.com/locate/procedia

25th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems

Student Behavior Recognition in Classroom using Deep Transfer Learning with VGG-16

Taoufik Ben Abdallah[a,*], Islam Elleuch[a], Radhouane Guermazi[b]

[a]*MIR@CL Laboratory, Faculty of Economics and Management, Sfax University, Tunisia*

[b]*Saudi Electronic University, Riyadh, Kingdom of Saudi Arabia*

Abstract

Tracking numerous students' behavior by observing and questioning them is a difficult task. Therefore, several methods based on automatic facial expression recognition have been proposed to capture and make a summary of students' behavior in the classroom. However, these methods cannot guarantee an effective classification due to the lack of huge datasets in this field. To improve students' behavior identification from video sequences, we propose in this paper a new approach based on deep transfer learning. Our approach pre-trains the model on a facial expression dataset. Then, it transfers the model to classify students' behavior. Experimental results confirm that our approach ensures a preferment students' behavior classification.

Keywords: Student behavior classification; facial expression recognition; deep transfer learning

1. Introduction

Communication is the act of transmitting information, knowledge, thoughts, ideas, or emotions between two or more persons. This interpersonal communication includes concealed messages, expressed through tone of voice, gestures, and facial expressions. According to Mehrabian [1], 55% of non-verbal messages are recognized by facial expressions. These latter are considered as one of the more important tools of human communications. Therefore, research on facial expressions is a crucial issue that affects various areas of science such as psychology, medicine,

* Corresponding author. Tel.: +216-26-445-012
E-mail address: taoufik.benabdallah@fsegs.rnu.tn

10.1016/j.procs.2021.08.098

behavioral science, and computer science. In order to enhance the recognition of facial expressions, researchers have significantly focused on the field of development of *Automatic Facial-Expression Recognition* (AFER) [2]. AFER can be widely applied to several applications like terrorists, health care, security, education, etc. Especially, students' behavior management in the classroom based on AFER devoted researchers [3]. These researches focused on the facial expressions of the students and the role they play during the face-to-face sessions. Indeed, smart classrooms with computer systems are capable to track students, recognize faces to interpret their gestures and facial expressions made by the students. The aim is to identify and interpret the comprehension level displayed by these expressions that helps the teachers to improve their teaching style and thus, keep the students interested and enthusiastic during the sessions. Besides, a teacher can use students' facial expressions as relevant sources of feedback. This helps the teachers to decide whether to slow down, speed up, or improve the teaching methods. Over the past few decades, many researchers focused on the AFER research area.

The literature on this field exhibits a diversity of methods including face detection and tracking, facial feature detection, and classification [4]. These methods can be divided into three categories: (i) appearance-based, (ii) motion-based, and (iii) deep learning-based. The appearance-based methods describe the changes in texture on a face by furrows, wrinkles, and bulges. Several representations were used to recognize the facial expressions such as the Gabor filters [5], the Discriminant Tensor Subspace Analysis [6], the Local Binary Patterns and its variants [7], the Pyramid of Local Binary Pattern [8], and Completed Local Binary Pattern [9]. However, the main disadvantage of these methods is the high dimension of the feature space. Several motion-based methods measures movement and orientations of the facial component like the Histogram of Oriented Optical Flow [10], and Histogram of Image Gradient Orientation on Three Orthogonal Planes [11]. Nevertheless, one of their biggest issues is their performance decay due to noise like brightness, non-aligned face, and the subtle movement of expressions.

Deep learning-based methods have increasingly been developed to handle the challenging factors for AFER [12]. These methods rely on artificial neural networks of multiple layers. The neural networks allow high-level feature detection and facial expressions classification into a unified process. The *Convolutional Neural Network* (CNN) and the *Recurrent Neural Network* (RNN) are the two most basic deep neural network architectures, which approved their robustness in extracting features regardless of the translation, rotation, and scale invariance of facial expressions [13]. Yolcu *et al.,* [14] used three CNN with identical architecture each one detects a portion of the face such as eye, mouth, and eyebrow. In addition, Kim et *al.,* proposed a spatial-temporal architecture with a combination between CNN and the Long Short-Term Memory (LSTM), which is the improved version of RNN, to detect high-level spatial-temporal features at different expression states [15]. The authors of [16] also proposed a *Spatial-temporal Convolutional with Nested LSTM* (STC-NLSTM) architecture based on three deep learning sub-network. The 3DCNN extracts the spatial and temporal features. The Temporal T-LSTM follows the 3DCNN to preserve the temporal dynamics, and then the Convolutional C-LSTM models the multi-level features. Moreover, Wang *et al.,* [17] extract features from each frame of video sequences using the *Transferring Long-term Convolutional Neural Network* (TLCNN). Each one of the above-mentioned deep learning-based methods is applied to train the model in separation on a specific feature space and same distribution. If features and/or distribution change, the model needs to be rebuilt from the scratch. Thus, it is a difficult process to collect associated data and rebuild the model. In such cases, using the pre-trained models is the best alternative. A pre-trained model is a model trained on other problems as a starting point, and so can solve a similar problem. Some popular pre-trained deep learning models used for AFER include VGG-16 [18], VGG-19 [18], AlexNet [19], Inception-V3 [20], ResNet-50 [21], etc. Furthermore, deep learning-based methods require a large amount of data to obtain insightful high-level features and avoid the problem of overfitting. In this context, researchers used the deep transfer learning that re-uses a pre-trained model knowledge for another task. Therefore, this method restricts the problem of insufficient samples in the facial expression datasets, thereby enhancing the AFER performance [22].

To take on the challenges of the related works, we propose in this paper a novel approach to classify students' behavior in the classroom environment based on *Automatic Facial-Expression Recognition* (AFER). This approach has to identify multiple students' faces independently of face shooting as well as the intensity of facial expressions. We start by training a first model to recognize facial expressions using a large dataset. To do this properly, we propose to adopt the architecture of VGG-16 model that not only extracts complex features from facial expression but also meets accurate results. After that, we transfer the knowledge learned from AFER task (*i.e.,* first model) to students' behavior classification task (*i.e.,* second model) in order to mitigate the problem of an insufficient dataset.

The remainder of this paper is organized as follows. Section 2 describes the proposed approach of student behavior classification. Section 3 presents the used datasets and defines the experimental conditions. Section 4

shows the experimental results and discusses the effectiveness of our proposal. Finally, Section 5 concludes the paper and gives some perspectives.

2. Proposed Approach

We propose to build an approach for classifying student behavior in classroom based on facial expressions. Our aim is to identify whether the students are understanding or not understanding the lesson. Initially, we pre-trained our model (Model α) on the facial expression from video sequences of the large dataset 1. Then, we transferred and fine-tuned our model to classify students' behavior in classroom (Model β) using video sequences of the small dataset 2. Fig. 1 illustrate an overview of the approach proposed. It consists of three main steps: (i) face detection and tracking, (ii) facial expression recognition, and (iii) student behavior classification. We detail each step of our proposed approach in the following sub-sections.

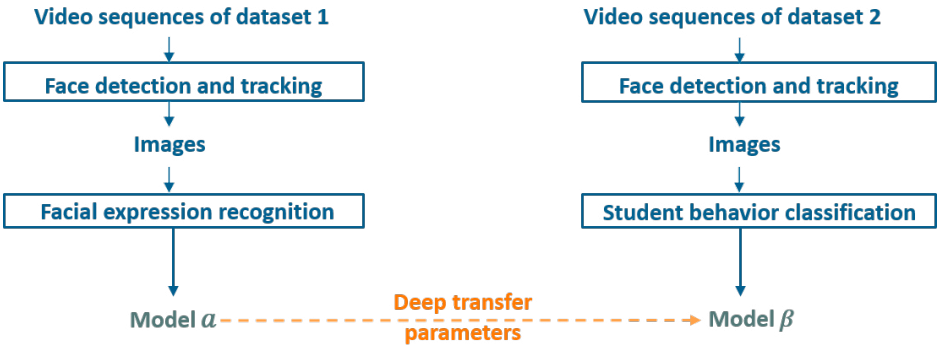

Fig. 1. An overview of the students' behavior classification approach

2.1. Face Detection and Tracking

Face detection is a process of detecting faces that appear in a given video sequence or image. In order to detect faces, we use the Haar Cascade algorithm. This algorithm is a machine learning object detection algorithm applied to identify objects in an image or video and based on the concept of features suggested by Viola and Jones [23]. The importance of this algorithm lies in the accurate detection of one or more faces. In our work, we apply Haar Cascade algorithm to the first frame of a video using the *Open Computer Vision Library* (OpenCV). Then, we track the detected faces in each of the subsequent frames using dlib library.

After the phases of face detection and tracking, we extract the target face of each frame in the video sequence. Then, we resize the obtained images to 48 × 48 pixels resolution, and we convert them to grayscale level.

2.2. Facial expressions recognition

In this paper, we adopt one of the pre-trained models *Visual Group Geometry* (VGG-16) [18] to extract features from each frame of video sequences. However, we modify the VGG-16 architecture with a simpler number of layers. Our aim is to increase the accuracy of the classification results compared to the original VGG-16 architecture.

The modified VGG-16 model consists of 8 convolutional layers, 4 max-pooling layers, 1 flatten layer, 1 dropout layer, and 2 fully connected dense layers. The ELU activation function is applied in all the layers in order to achieve high accuracy. In addition, the SAME padding is applied in each convolutional layer to ensure that the output has the same size as the input. We illustrate the precise structure of the new VGG-16 architecture in Fig. 2.

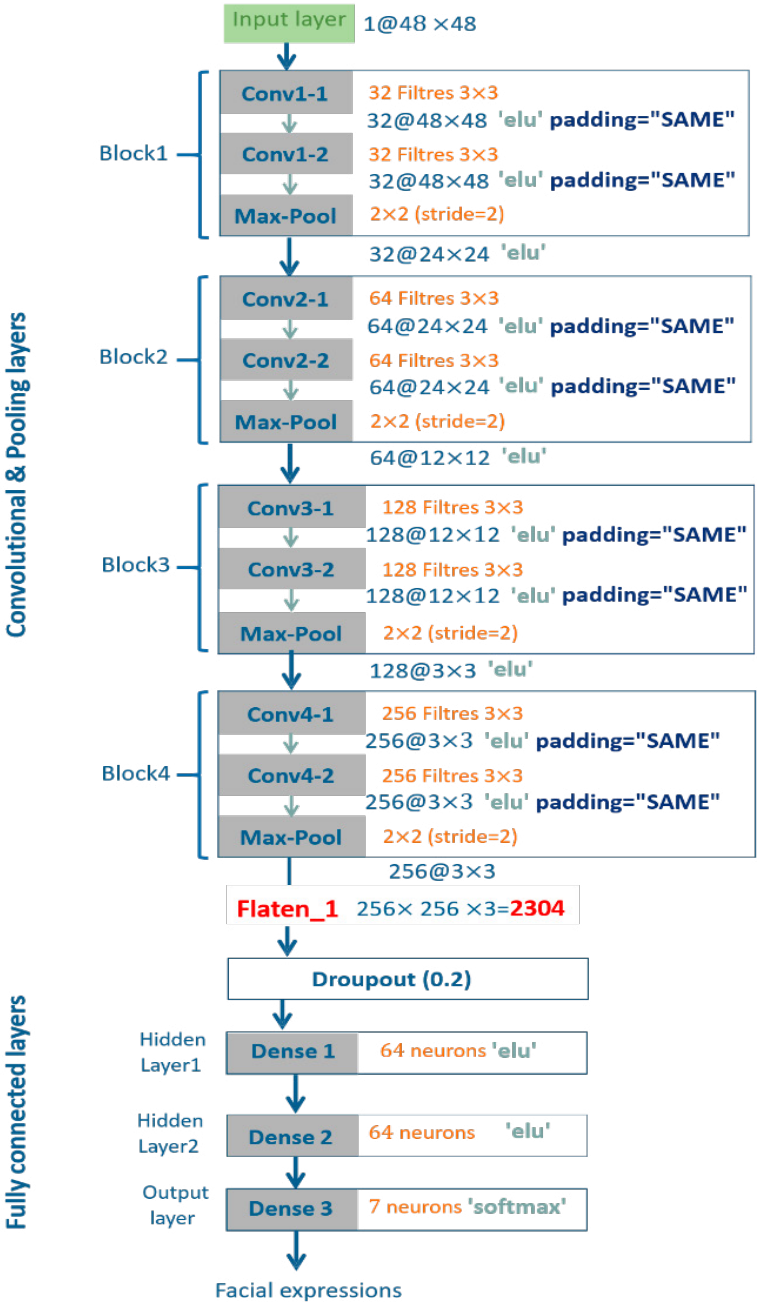

Fig. 2. The architecture of the modified VGG-16 model

- The first and second convolutional layers (Conv1-1 and Conv1-2) are comprised of 32 feature kernel filters. The size of the filter is 3×3. When the input image frame passes into the first and the second convolutional layer, its dimensions are changed to 48x48x32. Then, the resulting output is passed to the max-pooling layer with a stride of 2.

- The third and fourth convolutional layers (Conv2-1 and Conv2-2) are composed of 64 feature kernel filters and the size of the filter is 3×3. These two layers are followed by a max-pooling layer with stride 2, and thus, the resulting output is reduced to 24x24x64.
- The fifth and sixth convolutional layers (Conv3-1 and Conv3-2) use 128 feature maps with kernel size 3×3. The resulting output is reduced to 12x12x128. A max-pooling layer with a stride 2 follows the two convolutional layers.
- The seventh and eighth convolutional layers (Conv4-1 and Conv4-2) are with kernel size 3×3. These sets of convolutional layers have 256 kernel filters and are followed by a max-pooling layer with a stride of 2.
- The flatten layer (flatten_1) converts the data into a 1-dimensional array with a size of 256×256×3 (2304).
- The dropout layer is a mask that nullifies some neurons towards the next layer and leaves unmodified all others in order to prevent overfitting on the training data. The dropout rate is set to 0.2.
- The two fully connected are hidden layers (Dense 1 and Dense2) of 2304 units followed by a softmax output layer.

2.3. Student behavior classification

Deep learning heavily relies on huge quantities of data particularly labeled ones. However, these data are sometimes very expensive and the collected ones are not sufficient. To alleviate the dependency of deep learning accuracy on the size of the dataset and to maximize the use of existing little datasets, we opt to use deep transfer learning. This technique overcomes the isolated learning paradigm and can transfer learning from preceding tasks to new tasks. In fact, it re-uses the learned weights from a pre-trained model from a huge dataset, such as ImageNet. Then, it applies those weights to retrain the remaining layers or fine-tune the network.

In this paper, we unleash the power of deep transfer learning by using the adopted pre-trained model VGG-16 as an effective feature extractor to classify student behavior, *i.e.,* understanding or not understanding, even with fewer videos from the second dataset. To do so, we exploit the model learned in the facial expression recognition task based on the first dataset and we transfer the learned weights to classify the students' behavior in the classroom, as shown in Fig. 3.

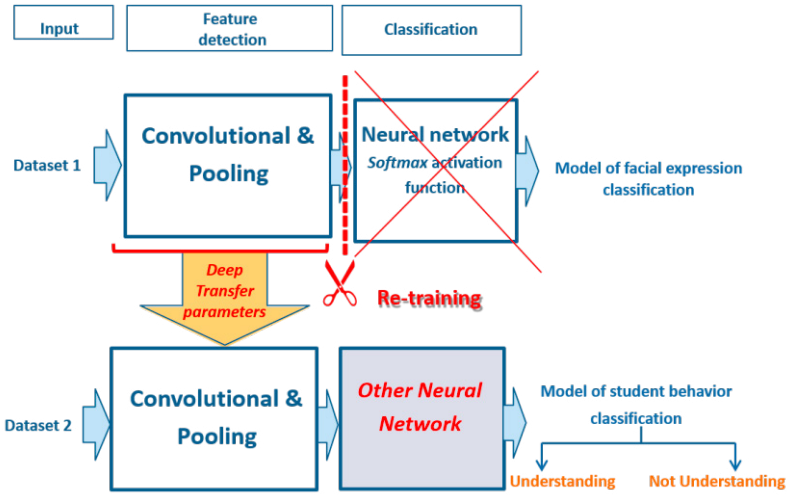

Fig. 3. Deep Transfer Learning for students' behavior classification

Moreover, we fine-tune the model of the students' behaviour classification by using data augmentation technique in order to improve the accuracy and eliminate the over-fitting problem. The idea of this technique is to add small variations without damaging the central object to increase the size of data used for the training model. To do so, we use various parameters like rotation, saturation, gaussian blur, horizontal stretch and desaturation.

3. Experiments

In this section, we first describe the chosen datasets to validate the proposed approach. We detail also the experimental set up. Then, we present the conducted experimental series and results.

3.1. Datasets

Several publicly datasets are available in the literature for facial expression recognition. These datasets include basic expressions and contain a large number of images collected from the real world to avail the evaluation of the deep learning algorithms. In our work, we have evaluated the performance of the facial expression recognition proposal using the *Facial Expression Recognition 2013* (FER-2013) dataset[†].

FER-2013 is a large-scale and unconstrained dataset gathered automatically by the Google image search API. It is an open source dataset and contain 35.887 images divided into three sets: 28709 training images, 3589 validation images and 3589 test images. All images are grayscale of 48 × 48 pixels in size. The images are labeled by seven expressions: anger, disgust, fear, happiness, neutral, sadness, surprise. The samples of the FER-2013 database are shown in Fig. 4.

Fig. 4. Some samples of FER-2013 dataset

In the literature, datasets on the classification of students' behavior in the classroom are not extensive. As a result, we chosen to collect our own real-world dataset, that we called *Student Behavior Classification 2020* (SBC-2020). **SBC-2020** is a small private video sequences dataset collected from traditional lessons at a Tunisian faculty. In total, 48 students participated in this study and gave informed consent for participation. Average age of the students was 20 years old and 60% were female. All students attended the same lesson: some of them showed the ease of understanding and others showed difficulty in understanding. Fig. 5 shows some examples of SBC-2020 dataset.

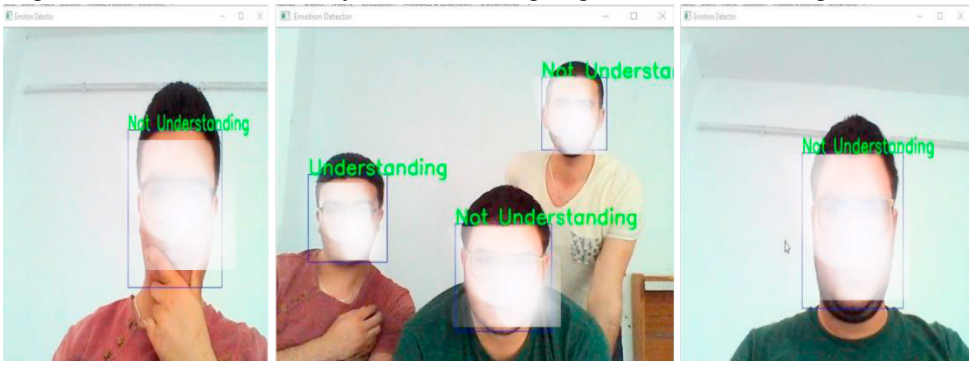

Fig. 5. Examples of SBC-2020 dataset

[†] https://datarepository.wolframcloud.com/resources/FER-2013

The student behavior monitoring system is directly connected to the camera network. Therefore, we recorded 96 videos and we extracted 1240 frames from them. Each frame contained a maximum of 10 students. Table 1 displays the characteristics of training and validation sets for our collected dataset. We classified our SBC2020 into two classes: (1) understanding, and (2) not understanding.

Table 1. The details of our collected dataset

	Number of participants	Number of videos	Number of frames	Number of frames "Understanding"	Number of frames "Not Understanding"
Train	44	88	1120	500	620
Validation	4	8	120	50	70

3.2. Experimental set up

The primary purpose of the experiments is to assess the proposed approach performance. We distinguish three series of experiments. In the first series of experiments, we leveraged the adopted pre-trained model VGG-16 which is previously trained on a large dataset with various classes to classify and evaluate the facial expression recognition model. In the second series of experiments, we performed a comparison between the new VGG-16, the original VGG-16 and VGG-19 architectures. In both series of experiments, we used the video sequences of the FER-2013 dataset. In the third series of experiments, we exploited deep transfer learning to classify students' behavior and we improved the accuracy using data augmentation technique. For this series, we used the video sequences of our collected dataset SBC-2020.

We have implemented web and desktop applications of face detection and tracking, facial expression recognition, and students' behavior classification using python keras library with TensorFlow back end. In addition, we performed all of our experiments on Google Colaboratory, which provides a runtime fully configured for deep learning with free access to a robust GPU [24].

We list the hyper-parameters used in the three series of experiments in Table 2.

Table 2. The applied hyper-parameters

Hyper-parameter	Value
Initial learning rate	0.001
Number of epochs	30
Momentum	0.9
Batch size	32

As a validation metrics, we have calculated the accuracy and the loss rates in training and validation sets. Accuracy measures the proportion of true results amongst the evaluated set, and loss captures the difference between the predicted value and the true value.

4. Results and discussions

In this section, we present and discuss the results of the different series of experiments.

4.1. Evaluation of the facial expression recognition

In this sub-section, we evaluate the performance of the facial expression recognition task using the new architecture of the modified VGG-16. Fig. 6 plots the performance of the network training and validation processes for 30 epochs.

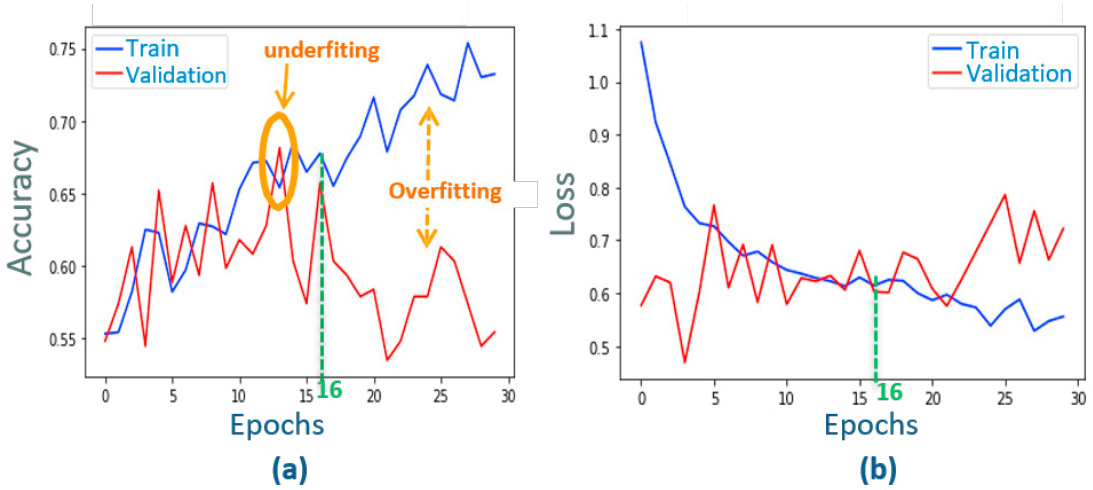

Fig. 6. Train and validation (a) accuracy and (b) loss metrics using the adopted VGG-16 for facial expression recognition.

In the Fig. 6, we see that the results of the model built using the adopted VGG-16 on facial expression recognition are good. The maximum learning rate attain at epoch 16. Therefore, the measured accuracy is 68.3 % (*cf.* Fig. 6 (a)), and the obtained loss rate is 0.61 (*cf.* Fig. 6 (b)). Indeed, the head movements, the illumination, the pose, and the background variation in the FER-2013 dataset enhance the problems of facial expression recognition.

4.2. Comparative Study for facial expression recognition

We have conducted a comparison of facial expression recognition task based on the new VGG-16 with other pre-trained models: VGG-19 and VGG-16. The difference between the three neural network architecture is shown in Table 3.

Table 3. Comparison between VGG-19, VGG-16 and the modified VGG-16

Layer	VGG-19	VGG-16	New VGG-16
Image Input Size	224×224 pixel	224×224 pixel	48×48 pixel
Convolutional Layer	16	13	8
Filter Size	64, 128, 256 and 512	64 and 128	32, 64, 128 and256
ReLU/ELU	18 ReLU	5 ReLU	13 ELU
Max Pooling	5	5	4
FCL	3	3	2
Dropout	0.5	0.5	0.2
Softmax	1	1	1

Fig. 7 shows validation accuracy for VGG-19, VGG-16 and the modified VGG-16. The first model built using VGG-19 gives test accuracy of 59.8 %. The second one using VGG-16 achieve accuracy of 65.03 %. The results show that the modified VGG-16 has better accuracy compared to VGG-16 and VGG-19 with an accuracy of 68.3 %. This is due to the reduction of the number of layers and the chosen applied hyper-parameters.

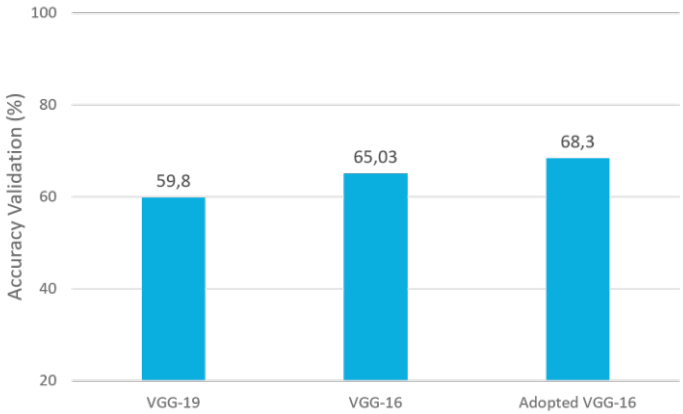

Fig. 7. Validation accuracy for different classification models.

4.3. Evaluation of students' behavior classification

We compare the performance of the proposed approach without and with the image augmentation technique according to the best epoch, the accuracy, and loss metrics for the SBC2020 dataset. The experimental results are shown in Table 4.

Table 4. Statistical results of students' behaviour classification model

Model	Best epoch	Accuracy	Loss
Students' behavior classification model without image augmentation	21	70.5 %	0.73
Students' behavior classification model with image augmentation	18	79.4 %	0.56

According to Table 4, we see that the accuracy rate of the original images of the video sequences of the SBC2020 dataset is 70.5% with a loss rate of 0.73 and a best epoch of 21. In addition, the accuracy rate of the augmented images is 79.4% with a loss rate of 0.56 and a best epoch of 18. Therefore, we see that data augmentation technique improve the performance of our model using our SBC2020 dataset. To conclude, our proposed approach achieves satisfactory results to classify students' behavior in the classroom environment.

5. Conclusion

Identifying the good or bad behaviors of a large number of students directly in the real environment is considered a big challenge. So several studies have proposed methods based on automatic facial expression recognition to classify students' behavior in the classrooms. Nevertheless, the existing datasets in this field are whether private for each faculty or are small, which cannot meet accurate classification results. To address this challenge, we have proposed a new students' behavior approach based on deep transfer learning. To do so, our model is firstly pre-trained, using the modified VGG-16, on the facial expression from the images of the FER-2013 dataset. Then, the learned model was transferred and fine-tuned to classify students' behavior using the collected SBC2020 dataset. In addition, the data augmentation techniques are applied to increase the SBC2020 dataset. From the experimental results, we conclude that augmented image results are better than the results obtained from the original images. Moreover, an accuracy of 79.4% is reached, demonstrating the effectiveness of the proposed approach to determine whether students are understanding or not understanding the lessons.

In our future work, we intend to further improve our approach by studying the impact of the different hyper-parameters of the CNN architecture on the performance of our proposal. We also plan to collect more video sequences from the traditional lessons to expand the collected SBC2020 dataset, and thus enhance the students'

behavior classification rate. Furthermore, we intend to set up a real-time application to monitor students' behavior in faculties' classrooms.

References

[1] Albert Mehrabian. (1968) "Communication without words." *Psychology Today* **2** (**4**), 53–56

[2] Kumari Jyoti, R. Rajesh, and K. M. Pooja. (2015) "Facial expression recognition: A survey." *Procedia Computer Science* **58** (**2015**): 486–491.

[3] Ngoc Anh, Bui and Tung Son, Ngo and Truong Lam, Phan and Phuong Chi, Le and Huu Tuan, Nguyen and Cong Dat, Nguyen and Huu Trung, Nguyen and Umar Aftab, Muhammad and Van Dinh, Tran. (2019) "A computer-vision based application for student behavior monitoring in classroom." *Applied Sciences* **9** (**22**): 4729–4746.

[4] Takalkar Madhumita, Xu Min, Wu Qiang and Chaczko Zenon. (2018) "A survey: facial micro-expression recognition." *Multimedia Tools and Applications* **77** (**15**): 19301–19325.

[5] Zhang Peng, Ben Xianye, Yan Rui, Wu Chen and Guo Chang (2016) "Micro-expression recognition system." *Light and Electron Optics* **127** (**3**): 1395–1400.

[6] Wang Su-Jing, Chen Hui-Ling, Yan Wen-Jing, Chen Yu-Hsin and Fu Xiaolan. (2014) "Face recognition and micro-expression recognition based on discriminant tensor subspace analysis plus extreme learning machine." *Neural processing letters* **39** (**1**): 25–43.

[7] Wang Yandan, See John, Phan Raphael C-W, Oh Yee-Hui. (2014) "Lbp with six intersection points: Reducing redundant information in lbp-top for micro-expression recognition." *Asian conference on computer vision,* 525–537.

[8] Taoufik Ben Abdallah, Radhouane Guermazi, and Mohamed Hammami. (2018) "Towards Micro-expression Recognition Through Pyramid of Uniform Temporal Local Binary Pattern Features." *International Conference on Intelligent Systems Design and Applications*, 629–640.

[9] Cao, Nhan Thi, An Hoa Ton-That, and Hyung-Il Choi. (2016) "An effective facial expression recognition approach for intelligent game systems." *International Journal of Computational Vision and Robotics* **6** (**3**): 223–234.

[10] Goh Kam Meng, Ng Chee How, Lim, Li Li, and Sheikh Usman Ullah. (2020) "Micro-expression recognition: an updated review of current trends, challenges and solutions." *The Visual Computer* **36** (**3**): 445–468.

[11] Li Xiaobai, Hong Xiaopeng, Moilanen Antti, Huang Xiaohua, Pfister Tomas, Zhao Guoying and Pietikainen Matti. (2017) "Towards reading hidden emotions: A comparative study of spontaneous micro-expression spotting and recognition methods." *IEEE transactions on affective computing* **9** (**4**): 563–577.

[12] Li Shan, and Weihong Deng. (2020) "Deep facial expression recognition: A survey." *IEEE Transactions on Affective Computing*, 1–20.

[13] Luqin Song. (2019) "A Survey of Facial Expression Recognition Based on Convolutional Neural Network." *IEEE/ACIS 18th International Conference on Computer and Information Science (ICIS)*, *IEEE*, 1–6.

[14] Yolcu Gozde, Oztel Ismail, Kazan Serap, Oz Cemil, Palaniappan Kannappan, Lever Teresa E and Bunyak Filiz. (2019) "Facial expression recognition for monitoring neurological disorders based on convolutional neural network." *Multimedia Tools Applications* **78** (**22**): 31581–31603.

[15] Kim, Dae Hoe, Wissam J. Baddar, and Yong Man Ro. (2016) "Micro-expression recognition with expression-state constrained spatio-temporal feature representations." *Proceedings of the 24th ACM international conference on Multimedia*, 382–386.

[16] Yu, Zhenbo and Liu, Guangcan and Liu, Qingshan and Deng, Jiankang. (2018) "Spatio-temporal convolutional features with nested LSTM for facial expression recognition." *Neurocomputing* **317** (**2018**): 50–57.

[17] Reddy Sai Prasanna Teja, Karri Surya Teja, Dubey Shiv Ram and Mukherjee Snehasis. (2019) "Spontaneous facial micro-expression recognition using 3D spatiotemporal convolutional neural networks." *International Joint Conference on Neural Networks (IJCNN), IEEE*, 1–8.

[18] Simonyan Karen, and Andrew Zisserman. (2014) "Very deep convolutional networks for large-scale image recognition." arXiv preprint arXiv:1409.1556, 1–14.

[19] Krizhevsky Alex, Ilya Sutskever, and Geoffrey E. Hinton. (2012) "Imagenet classification with deep convolutional neural networks." *Advances in neural information processing systems*, 1097–1105.

[20] Szegedy Christian, Vanhoucke Vincent, Ioffe Sergey, Shlens Jon and Wojna Zbigniew. (2016) "Rethinking the inception architecture for computer vision." *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2818–2826.

[21] He Kaiming, Zhang Xiangyu, Ren Shaoqing and Sun Jian. (2016) "Deep residual learning for image recognition." *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

[22] Tan Chuanqi, Sun Fuchun, Kong Tao, Zhang Wenchang, Yang Chao and Liu Chunfang. (2018) "A survey on deep transfer learning." *In International conference on artificial neural networks*, 270–279.

[23] Viola, Paul, and Michael Jones. (2001) "Rapid object detection using a boosted cascade of simple features." International Conference on Computer Vision and Pattern Recognition, IEEE, Kauai, HI, USA, 511–518.

[24] Carneiro Tiago, Da Nóbrega, Raul Victor Medeiros, Nepomuceno Thiago, Bian Gui-Bin, De Albuquerque, Victor Hugo C and Reboucas Filho Pedro Pedrosa. (2018) "Performance analysis of google colaboratory as a tool for accelerating deep learning applications." *IEEE Access* **6** (**2018**): 61677–61685.