Pre-trained Deep Learning Models for Facial Emotions Recognition

Atanas Atanassov
Dept.of Computer Science
University of Chemical Technology and Metallurgy
Sofia, Bulgaria
Email: naso@uctm.edu

Dimitar Pilev
Dept.of Computer Science
University of Chemical Technology and Metallurgy
Sofia, Bulgaria
Email: pilev@uctm.edu

***Abstract*— The present study is devoted to a comparative analysis of known pre-trained models of deep learning neural networks used to facial emotions recognition. The aim of the study is to select appropriate lightweight models to be used for offline facial emotions recognition of the students during their semestrial learning. The basic facial emotions can be grouped as positive (happiness and surprise), neutral, and negative (fear, anger, sadness and disgust). Based on these three groups of emotions corrective actions by adapting and personalizing the lecture material are undertaken.**

Keywords—facial emotions recognition, deep learning, CNN, pre-trained models

I. INTRODUCTION

Facial emotions are one of the most informative means of assessing the attitude of a person to the environment, a topic, object, commercial product and more. Currently, technologies such as facial recognition, fingerprint recognition, voice command recognition or the integration of voice virtual assistants such as Siri, Alexa and Google Assistant are part of the software of almost all smartphones or devices available on the market. Part of these technologies based on deep neural networks are intended for image and human face recognition and a specific part of them to facial emotions recognition (FER). In recent years, there has been a rapid development of convolutional neural networks (CNN) for facial recognition and facial emotion recognition. Thanks to the efforts of the scientific community, many trained models of neural networks for FER such as DeepFace, VGG16, VGG19, ResNet, etc., are available for non-commercial use. In the present study, an analysis of several existing FER models was made in order to select an appropriate model to be used to analyze students' emotions during their studies. As on the basis of their emotions (positive, neutral or negative) actions to personalize and individualize the lecture material to the particular student are performed. The structure of this report is as follows: in section 2 the emotions and their application in different subject areas are discussed; section 3 describes the process of recognizing facial emotions through CNN; the most common datasets for FER are discussed in section 4; in section 5 includes an analysis of trained models for FER; section 6 compares the results of predicting the emotions of students from 2 selected models; conclusions have been drawn in section 7.

II. FACIAL EMOTIONS

Emotions are related to our perceptions and interactions with the world around us. They are reflections of our social contacts, of the processes in our memory and of our attitude towards various objects. We express our emotions in different ways: through different movements and gestures of the head and limbs, facial expressions, through the intonation of our voice and others. Our face is one of the most expressive indicators of our emotions. Through various facial expressions involving the eyes, eyelids, eyebrows, lips, nose and chin, we show specific emotions and their nuances. The main emotions are happiness (joy), anger, surprise, fear, sadness and disgust, as well neutral. In 1978 P. Ekman and W. V. Friesen [1] proposed a system for coding facial expressions, named Facial Action Coding System (FACS). The system classifies the facial expressions on the base of elementary components called Action Units (AUs). Each AU has a number (AU1 or AU35) and is related to one or more of face muscles. For example, “Fig. 1.” the happiness [2] can be summarized as a combination of cheek raiser (AU6) and lip corner puller (AU12).

Fig. 1. Happiness is combination of AU6 + AU12

Through the analysis of facial emotions people 's behavior can be predicted and recommendations or suggestions for the use of services and products can be generated. Facial emotion analysis can be used in medicine and psychology to detect mental and other illnesses; in police practice whether a suspect is lying or not; in modern cars, whether the driver is tired or not, and whether his gaze is focused on the road; in products advertisements - what is the consumer's attitude to a product; in the film industry - to improve some scenes from the film in order to enhance a specific emotion in the viewer, etc.

III. FACIAL EMOTIONS RECOGNITION

Facial emotions recognition or facial extraction recognition is a part of the task of facial image recognition. Convolutional neural networks known as Deep neural networks (DNN) are applied for FER. Initially, these networks are trained with large number of photos of human faces (datasets) in order to recognize (classify) human facial emotions. For example, some datasets include several thousand pictures of the faces of different people, other datasets include several millions pictures. The pictures are taken from different angles at various levels of illumination and different camera settings (aperture, sharpness, balance, resolution, etc.). Pictures in the datasets are labeled with information about the emotion presented in each face. Usually, one dataset contains training, testing and validation

978-1-7281-9308-3/20/$31.00 ©20XX IEEE

Authorized licensed use limited to: Rutgers University. Downloaded on May 20,2021 at 13:49:46 UTC from IEEE Xplore. Restrictions apply.

subsets of pictures. After training phase, the CNN are checked with testing subset of images in order to estimate their accuracy or whether they correctly predict the emotions. Next, the trained CNN can be applied for verification/prediction of emotions of not labeled pictures.

The whole FER process or pipeline includes following steps: face detection and cropping, preprocessing of cropped image, deep feature learning and deep classification [3]. Face detection and cropping are related to find human face or faces in the big picture and to extract (crop) it/them. The face alignment preprocessing step based on face landmarks [4] can be used to reduce variations of face rotation and scale. During the face alignment the background around the face and other non-face elements are removed. Next preprocessing steps are normalization of illumination and facial pose. Variations in the illumination and contrast of the face picture may influence the features extraction, that's way the normalization algorithms are performed in order to reduce this noise. On the base of facial landmarks and application of pose normalization algorithms the frontal face is generated. Other preprocessing steps are related to scaling the cropped face image to the input size required by the CNN and with the conversion of RGB image to BW format. Not all of described above preprocessing steps are obligatory. For example, one can skip the face alignment step and this will reduce slightly the prediction accuracy. At the same time, applying the alignment in Google deep leaning face recognition model FaceNet [5] increases the prediction accuracy with 1% up to 99.63%. Deep learning CNN does the steps of feature learning and classification. As mentioned above the CNN trained with large set of labeled face images expressing different facial emotions and then its accuracy is verified with another testing dataset. Then it can be used for prediction of emotions on unknown facial images. Usually, every CNN for FER "Fig. 2." contains groups of convolution, activation and pooling layers (CONV, RELU and POOL) used for feature learning and at the end of network there is one or more fully connected layers (FC) intended to emotions classification.

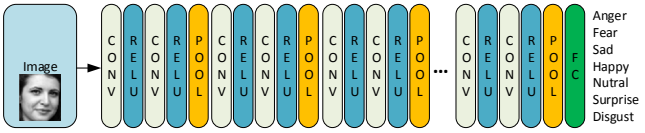

Fig. 2. Typical archtecture of CNN for facial emotions recognition

The convolution layer calculates the output of neurons connected to the inputs of the input layer. Each convolution layer neuron calculates a scalar product between its weights and the input values of the local area to which it is bound. Starting from the input image size, each convolution layer reduces the image size depending on applied filter size. Normally, the filter size is 3x3 and it used to find some specific shapes or pixels at certain positions on the image. The image is scanned horizontally and vertically with the filter and convolved feature is calculated by multiplying (or logical AND) filter and same size frame in the image "Fig. 3.". Many different filters can be applied at the same time on a specific convolution layer, so many convolved features can be extracted at once on this layer. The activation layer applies rectifier-linear-unit (ReLU) activation function that produce fast output to the next pooling layer. The pooling layers additionally reduce the features size applying max or average filters on the its input data. In most cases the filter size is 2x2 as given on "Fig. 4.". The fully connected layer calculates class grades from 1 to 7 related to the mentioned seven facial emotions (happy, anger, surprise, fear, sad, disgust and neutral). Usually, for the classification of emotions at the fully connected layer the Support Vector Machines (SVM) algorithm applied.

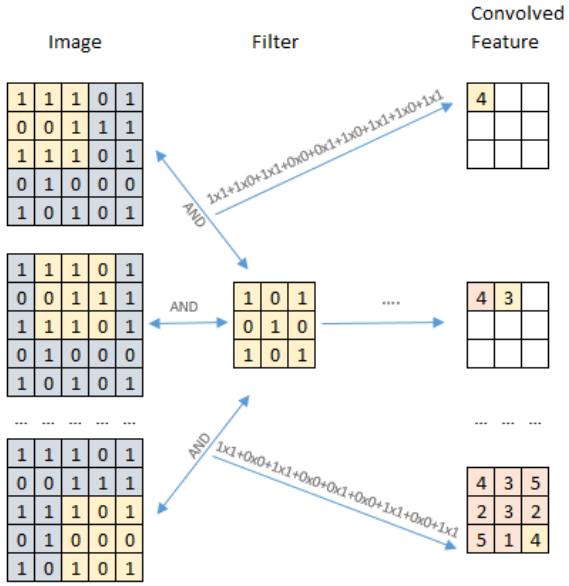

Fig. 3. Calculation of convolved feature ising 3x3 filter

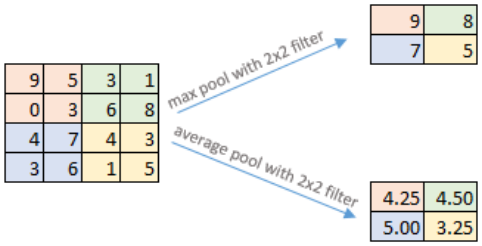

Fig. 4. Max pooling or average pooling calculations ising 2x2 filter

IV. FER DATASETS

There are many FER datasets with various number of facial emotion images taken in different face positions, with different pixel size and colors (BW or RGB). The usual number of emotions is 7, but some datasets doesn't include neutral or disgust emotion. Here the concise information about discussed in this paper datasets is presented. One of most frequently used dataset is FER-2013. It includes 35887 labeled BW images with size 48x48 pixels. Images divided in 3 subsets: - 28709 for training; 3589 for testing and 3589 for validation. Number of emotions in FER-2013 is 7. The Extended Cohn-Kanade Dataset (CK+) includes 327 video sequence of 118 people with 7 labeled emotions. Images are BW 640x490 pixel sized. Only the first and last images from each sequence are labeled on the base of FACS. The Radboud Faces Database (RaFD) dataset contains 8040 images taken from 67 persons. Labeled emotions are 8: - anger, contempt, disgust, fear, happiness, sadness, surprise and neutral. The format of the images is 681x1024 RGB. The Karolinska Directed Emotional Faces (KDEF) dataset contains 4900 images taken from 70 people. The number of emotions is 7. Images are in RGB format 562 * 762 pixels. The Multimedia

Authorized licensed use limited to: Rutgers University. Downloaded on May 20,2021 at 13:49:46 UTC from IEEE Xplore. Restrictions apply.

Understanding Group (MUG) dataset contains 1462 image sequences of 86 people with length from 50 to 160 RGB images with resolution 862x862 pixels. Emotions number is 6 and does not include neutral one.

V. Analysis of Pre-trained FER models

Short analysis of available pre-trained FER model has been done in this section. It includes non-commercial models developed on Python through TensorFlow, Keras, PyTorch or other frameworks and libraries. Most of these models are based on known models for face recognition as VGG16, VGG19, VGG-Face, Google FaceNet, OpenFace, Facebook DeepFace and DeepID. Mentioned models "Table I" provide high accuracy in face recognition (up to 99%) but the accuracy of their demographic option for facial recognition varies widely [6] from 40% to 75%.

TABLE I. Comparission of FER Models

Model	Parameters	Image size	Output	Accuracy
VGG-Face	145,002,878	224x224x3	2622	90.5
Facenet	22,808,144	160x160x3	128	99.63
OpenFace	3,743,280	96×96x3	128	97.3
DeepFace	137,774,071	152x152x3	8631	97.35
DeepId	400,000	55x47x3	160	96.05

This can be explained by the fact that these models are trained with large datasets for face recognition (up to several million images) but when they are trained with FER datasets (around 20-35 thousand images) the accuracy is normally to be lower. Also the classification algorithms for FER [7,8] differ from those for face recognition which are based on cosine or Euclidian similarity between face features or face embedding's.

Many of the models use the architecture of Visual Geometry Group (VGG) and its variations VGG16, VGG19 or VGG-Face. VGG is a pre-trained model with more than 150 million weight parameters. VGG [9] is deep learning CNN with 37 layers including 13 convolutions plus 13 ReLU layers forming five convolution groups. At the end of each group, there is one max-pooling layer. Filters applied on convolution layers are 3x3 and these on max-pooling layers are 2x2 with stride 2. Last 3 layers of the network are fully connected.

In [10] several architectures of CNN for FER are proposed. Some of them based on the available pre-trained models as VGG16, VGG19, ResNet, DenseNet, MobileNet and one is authors' development. The authors' architecture contains four convolution groups with size 64, 128, 512 and 512. Each group includes convolution, batch normalization, ReLU and max-pooling layers. At the end of the CNN there are 3 fully connected layers with size 256, 256 and 7, where 7 relates to the output of 7 emotions. The input layer provides 48x48 BW images and the output. The authors compared the accuracy of their own model trained from scratch to the mentioned pre-trained models integrated in developed architectures. For training, validation and test they used Kaggle 2013 dataset. The test accuracy of VGG16 and VGG19 based architecture is 67% and 66% respectively, followed by 65% of authors' architecture. Architectures based on ResNet, DenseNet and MobileNet have accuracy 62%, 49% and 53%.

Proposed in [11] deep learning CNN based on transfer learning approach. It uses pre-trained with ImageNet dataset VGG19 model for facial emotions recognition. The model additionally is fine-tuned with JAFFE and CK+ datasets. The architecture, respectively, the number of convolutional layers of VGG19 is identical to VGG discussed above in this section. In current transfer learning CNN, the features from each 5 max-pooling layer of VGG19 are extracted and combined with those from first fully connected layer. Then principal component analysis (PCA) applied for dimensionality reduction of learned features and finally for feature selection a Linear SVM is used. As result detected facial emotion produced. Proposed system includes some preprocessing steps in order to prepare face images to the required by VGG19 input layer format of 224x224 pixels. The test accuracy results are 92% for CK+ dataset and 88% for JAFFE.

Proposed in [12] CNN contains one input 165x165 RGB layer, 4 convolutional, 3 max-pooling and 2 fully connected layers. Each of the first 3 convolutional layers are followed by max-pooling layer. The last FC layer is softmax layer producing as output 6 emotions (without disgust emotion). The CNN is trained and fine-tuned with VGG model and additionally is tuned with Ck+, KDEF, RaFD and MUG datasets. Authors state that obtained accuracy on MUG is more than 87%.

Next pre-trained CNN model "Table II" for FER architecturally based on VGG. It has 5 groups of convolutional layers with batch normalization layers between internal convolutional layers. Each group ends with max-pooling and dropout layer. The number of convolutional layers in first group is 2, in second group is 3 and in the last three groups is 4. Input layer provides BW images with size 48x48 pixels and last two layers are fully connected and their output is 7 basic emotions. The number of weights of this model is more than 13 million. We selected this model for validation of students' facial emotions discussed in next section and for convenience called it "Model A".

TABLE II. Model A - CNN based on VGG16

Layer type	Size	Batch normalization
Input	48x48x1	
Conv1_1 to Conv1_2	48x48x64	Conv1_1 - Max_Pooling1
Max_Pooling1 + Dropout1	24x24x64	
Conv2_1 to Conv2_3	24x24x128	Conv2_1 - Max_Pooling2
Max_Pooling2 +Dropout2	12x12x128	
Conv3_1 to Conv3_4	12x12x256	Conv3_1 - Max_Pooling3
Max_Pooling3 + Dropout3	6x6x256	
Conv4_1 to Conv4_4	6x6x256	Conv4_1 - Max_Pooling4
Max_Pooling4 + Dropout4	3x3x256	
Conv5_1 to Conv5_4	3x3x512	Conv5_1 – Conv5_4
Max_Pooling5 + Dropout5	1x1x512	
Fully Connected	1x512	
Output (FC)	7 emotions	

The DeepFace framework developed by S. Serengil [13] uses ensemble of 5 pre-trained models. The models are described in Table I. Presented accuracy in the table is for face recognition, not for FER. The framework provides

Authorized licensed use limited to: Rutgers University. Downloaded on May 20,2021 at 13:49:46 UTC from IEEE Xplore. Restrictions apply.

demographic functionality including detection of facial emotions, age, gender and race. From the source code of this model we succeed to extract the pre-trained CNN model and features intended only for FER. The architecture of this CNN is given in Table III.

TABLE III. MODEL B - CNN BASED DEEPFACE

Layer type	Size
Input	48x48x1
Conv1	44x44x64
Max_Pooling1	20x20x64
Conv2	18x18x64
Conv3	16x16x64
Average_Pooling2	7x7x64
Conv4	5x5x128
Conv5	3x3x128
Average_Pooling2	1x1x128
Fully Connected	128
Fully Connected + Dropout1	1024
Fully Connected + Dropout2	1024
Output (FC)	7 emotions

It has 3 convolutional groups with 5 convolutional layers. First group is simple only one convolutional layer followed by max-pooling layer. Next two groups have 2 convolutional layers followed by one average-pooling layers. Input layer provides BW images with size 48x48 pixels. Output layers are 4 fully connected layers with 2 dropout layers in between. Last output layer returns as result 7 emotions. This model is simpler then Model A and the number of its weight parameters is 1.485000. We called it "Model B" and used for FER analysis in next section. Both models are trained and verified using the FER-2013 dataset. The accuracy "Fig. 5." for Model A is almost 70% and for Model B 57%. An additional verification of the considered models was made with CK + dataset. The accuracy "Fig. 6." of both models is 56%. The accuracy for model B is the same for both datasets, which emphasizes the importance of the datasets used in the learning process.

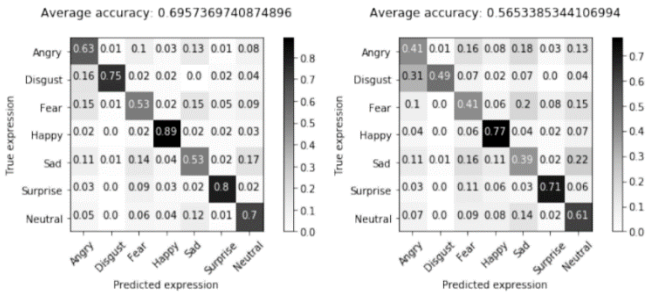

Fig. 5. Confusion matrices of Model A (on the left) and Model B (on the right) on FER-2013 dataset

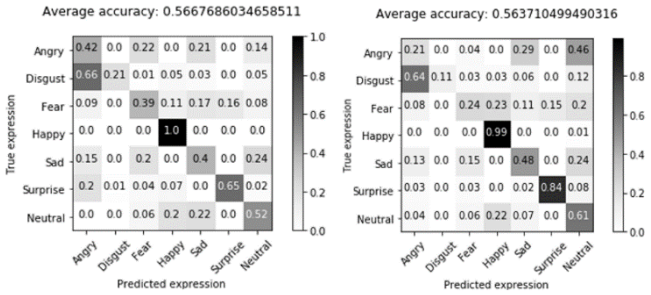

Fig. 6. Confusion matrices of Model A and Model B, CK+ dataset

VI. RESULTS VERIFYING STUDENTS' EMOTIONS

The verification of models A and B has been done with our private dataset with images collected during the Summer Poster Session at UCTM. 520 pictures of more than 200 regular or PhD students are taken before and after nomination of 120 papers divided in 5 scientific sections. On "Fig. 7." a small part of the dataset's images are provided.

Fig. 7. Some pictures of our private dataset

The results of the verification of facial emotions recognition of Model A are presented from "Fig. 8." to "Fig. 12.".

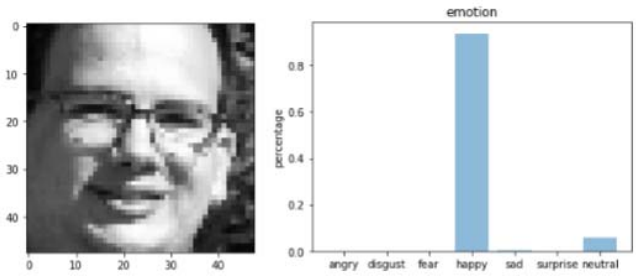

Fig. 8. Model A - Happy

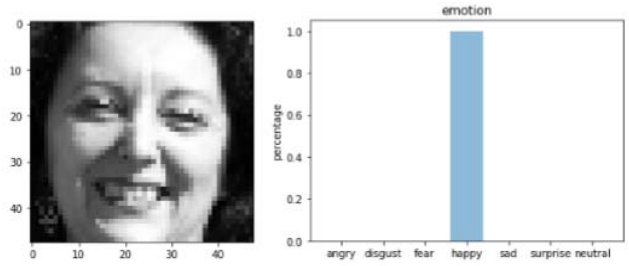

Fig. 9. Model A - Happy

Authorized licensed use limited to: Rutgers University. Downloaded on May 20,2021 at 13:49:46 UTC from IEEE Xplore. Restrictions apply.

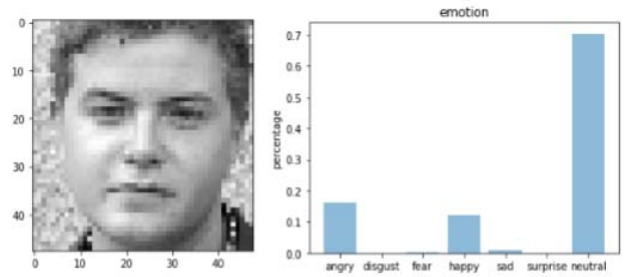

Fig. 10. -Model A - Neutral

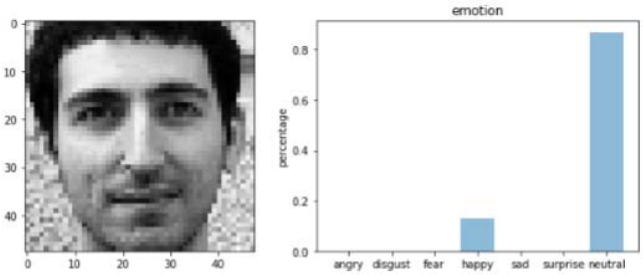

Fig. 11. Model A - Neutral

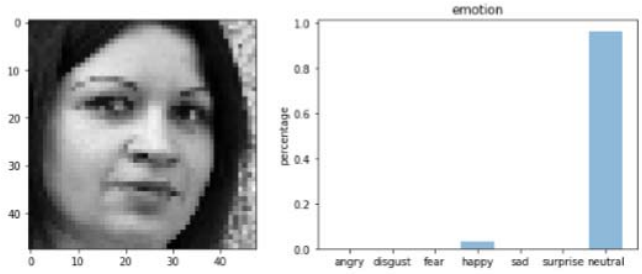

Fig. 12. Model A – Neutral

The results of the verification of Model B are given from "Fig. 13." to "Fig. 17.". As can be seen in absolute value both models give similar results. An exception is shown in "Fig. 12." and "Fig. 17.", where Model A defines the emotional state as Neutral and model B as Sad (with a possibility of neutral, anger and fear).

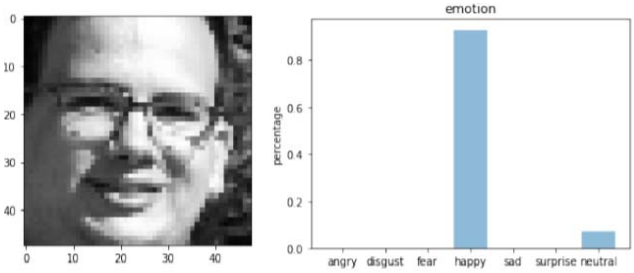

Fig. 13. Model B - Happy

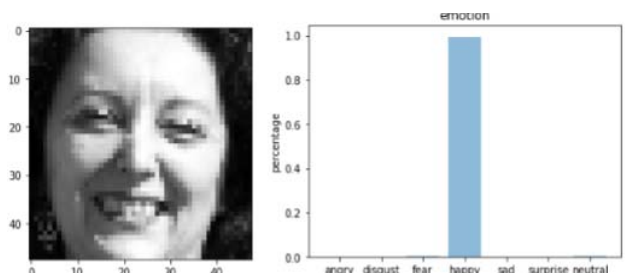

Fig. 14. Model B - Happy

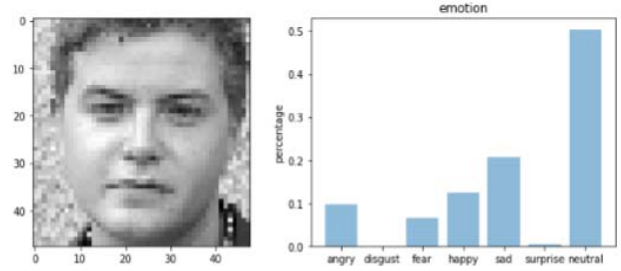

Fig. 15. Model B - Neutral

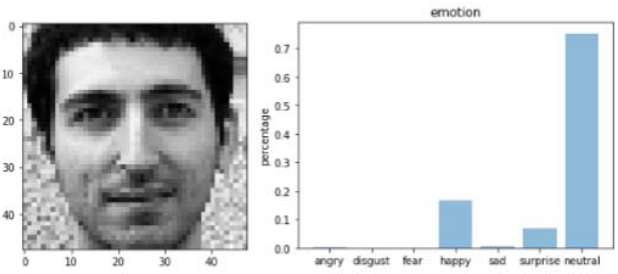

Fig. 16. Model B - Neutral

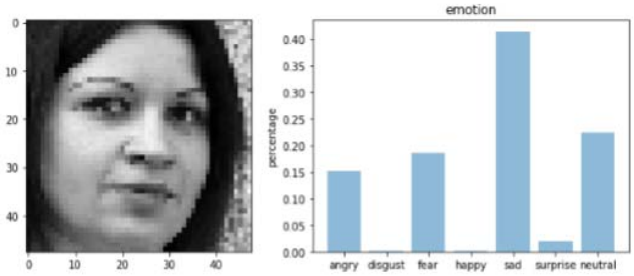

Fig. 17. Model B - Sad

Upon careful reading of the image, we can conclude that even a man may have some difficulty in determining the emotional state of a young woman - angry, sad, scared or even happy.

The expression is a typical example of the so-called Poker Face term coming from the game of the same name, in which it is extremely important that the player does not issue his available cards through his emotional state. Figures 18 to 21 show the results of testing the two models for emotional recognition with Poker Face facial expressions.

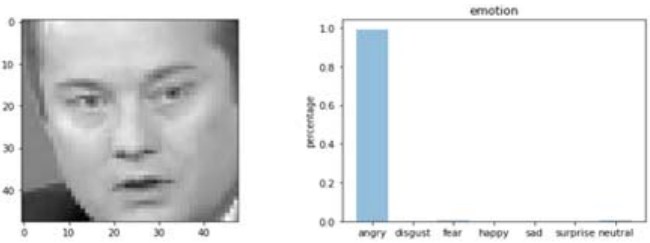

Fig. 18. Model A - Poker Face expression (Angry)

Authorized licensed use limited to: Rutgers University. Downloaded on May 20,2021 at 13:49:46 UTC from IEEE Xplore. Restrictions apply.

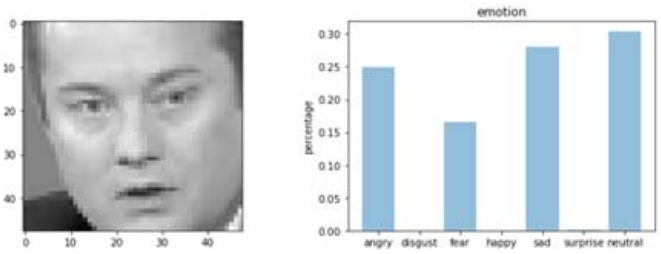

Fig. 19. Model B - Poker Face expressions (Neutral, Sad and Angry)

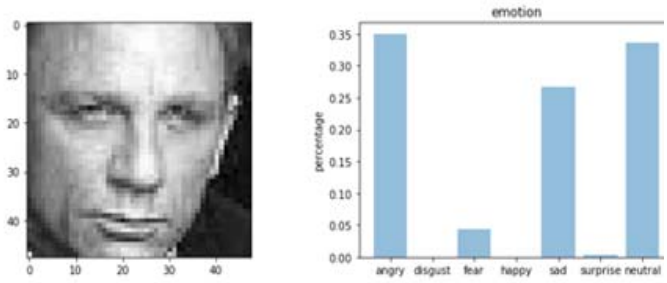

Fig. 20. Model A - Poker Face expressions (Angry,Neutral, Sad and Fear)

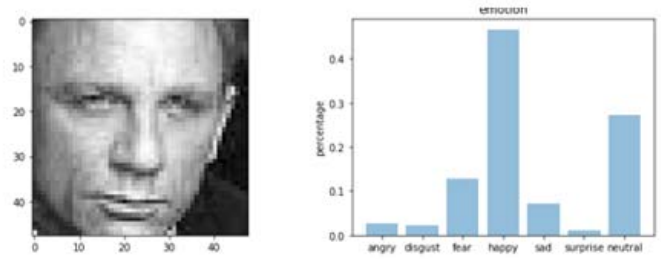

Fig. 21. Model B- Poker Face expressions (Happy,Neutral, Fear and Sad)

No matter how well the models for emotional recognition are constructed and trained, based on a given expression, there are situations in which it is impossible for them correctly to determine the respective emotion. This is the reason why most of the proposed models achieve accuracy in the range of only 50% - 70%. In such situations, in real life, in order to be able to determine an emotional state, people use additional information (if they have it) such as the character and mental resilience of the person, the environment conditions, long-term observation and more.

VII. Conclusions

In this paper the concise analysis of available pre-trained deep learning CNN models for facial emotions recognition has been done. It is intended to enhance the personalized learning of the students based on their facial emotions recorded during online lectures or exercises, which become very important during the Covid 19 pandemic. On the base of this analysis, two pre-trained models of CNN for FER were selected. One based on DeepFace CNN and another on VGG. We trained and verified the models on FER-2013 and CK+ datasets. The accuracy results of the models are commented in previous section and are given on “Fig. 5.” and “Fig. 6.”. Additionally, we verified the models using our private dataset with students’ images. In the considered dataset of images collected during the Summer Poster Session at UCTM, both models (Model A and Model B) achieve higher accuracy in determining emotions. This is due to the fact that most of the participants in the poster session are smiling, calm and satisfied, and the models used have about 10% higher accuracy in recognizing these types of emotions. This can be seen if we compare the accuracy of “Fig. 5.” and “Fig. 6.” with those of “Fig. 8” to “Fig. 17.”. This would not be the case at events where participants are put under pressure, stress, anger, indifference or dissatisfaction (e.g. job interviews, exam sessions, etc.). So our future work will be focused on revealing neutral and negative emotions such as fatigue, boredom, etc. in order to adapt the study material in order to easily perceive it and increase student performance, as well to assist lecturers in presenting the study material by changing at the pace of teaching. Additionally, we intend to apply some statistical [14] and BI [15] data processing in order to better the results of facial emotions recognition.

Acknowledgment

The research is funded by the Research Fund (FNI) of the Ministry of Education, under the project "Synergy between procedural philosophy and elements of artificial intelligence in the theory of education" № DN 15/9 from 11.12.2017.

References

[1] P. Ekman and W. Friesen, “Facial action coding system: a technique for the measurement of facial movement”. Palo Alto, Calif: Consulting Psychologists Press.,1978

[2] P. Ekman, “Facial action coding system (facs),” A human face, 2002

[3] B. Farnsworth, “Facial action coding system (facs) – a visual guidebook,” August 18th, 2019, https://imotions.com/blog/facial-action-coding-system/#main-action-units

[4] A. Geitgey, “Machine learning is fun! part 4: modern face recognition with deep learning,” Medium Corporation, 2016

[5] S. Serengil, “Face recognition with facenet in keras,” Sep. 2018 https://sefiks.com/2018/09/03/face-recognition-with-facenet-in-keras/

[6] C. Pramerdorfer and M. Kampel, “Facial expression recognition using convolutional neural networks: State of the Art,” arXiv preprint arXiv:1612.02903v1, Dec. 2016

[7] S. Li and W. Deng, “Deep facial expression recognition: a survey,” IEEE Transactions on Affective Computing 2020, in arXiv preprint arXiv:1804.08348v2, Oct. 2018

[8] S. Minaee and A. Abdolrashidi, “Deep-emotion: facial expression recognition using attentional convolutional network,” arXiv:1902.01019v1, Feb., 2019, in press.

[9] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition”, arXiv preprint arXiv:1409.1556v6, April 2015

[10] C. Wu and L. Chen, “Facial emotion recognition using deep learning,” arXiv, preprint arXiv:1910.11113v1, Oct. 2019

[11] A. Ravi, “Pre-trained convolutional neural network features for facial expression recognition,” arXiv preprint arXiv:1812.06387v1, Dec. 2018

[12] A. Fathallah, L. Abdi and A. Douik, “Facial expression recognition via deep learning,” IEEE/ACS 14th International Conference on Computer Systems and Applications, Tunisia, Nov. 2017 pp.745-750

[13] https://github.com/serengil/deepface

[14] C. Yu and C. Ko, “Applying FaceReader to Recognize Consumer Emotions in Graphic Styles”, 27th CIRP Design Conference, 2017 pp. 104-109

[15] F. Tomova, "Application of the MicroStrategy BI platform to calculate claim ratio for the insurance companies", Proceedings of XI-th international conference "Challenges in higher education and research in the 21st century", Heron Press, Sofia, Bulgaria, 2013, pp. 243-246

Authorized licensed use limited to: Rutgers University. Downloaded on May 20,2021 at 13:49:46 UTC from IEEE Xplore. Restrictions apply.