Emotion Recognition Based on Facial Expressions Using Convolutional Neural Network (CNN)

Sabrina Begaj
Department of Computer Engineering
Epoka University
Tirana, Albania
sbegaj@epoka.edu.al

Ali Osman Topal
Department of Computer Engineering
Epoka University
Tirana, Albania
aotopal@epoka.edu.al

Maaruf Ali
Department of Computer Engineering
Epoka University
Tirana, Albania
mali@epoka.edu.al

***Abstract*— Over the last few years, there has been an increasing number of studies about facial emotion recognition because of the importance and the impact that it has in the interaction of humans with computers. With the growing number of challenging datasets, the application of deep learning techniques have all become necessary. In this paper, we study the challenges of Emotion Recognition Datasets and we also try different parameters and architectures of the Conventional Neural Networks (CNNs) in order to detect the seven emotions in human faces, such as: anger, fear, disgust, contempt, happiness, sadness and surprise. We have chosen iCV MEFED (Multi-Emotion Facial Expression Dataset) as the main dataset for our study, which is relatively new, interesting and very challenging.**

Keywords*— *Deep Learning, Facial Expression Recognition, Data Preprocessing, Convolutional Neural Network, CNN, Facial Emotion Recognition, FER, Image Recognition

I. Introduction

Facial expressions are the main natural source for human beings to communicate their own emotional state to other people. Many studies [1][2] have proven that more than 50% of our feeling are shown directly with our facial expressions, which is a ten times bigger percentage when compared with emotions expressed by the inflection of spoken words. Since we live in the networked Age of Technology, it is becoming increasingly necessary for everyday life to have intelligent monitoring. For example, cameras and assistive robots need to understand what is going on with human emotions. Expression recognition is very easy to guess for humans, but is an extremely difficult task for even the smartest of AI technologies. Automatic recognition of emotions has a lot of challenges, starting with the categorisation of emotions and up to a greater study from psychologists and their collaboration with scientists.

Facial Emotion Expressions (FER) are very useful in different situations in real life because they give important insights for a particular person. That is why they are studied in detail and applied in many systems. FER can be applied in different applications such as: medical treatments, human resources, police investigation, education (on students while they are in a lecture, customer service, journalism (during interviews) and many others. In the early 20th Century, Paul Ekman and Friesen [3] defined six basic emotions, which are anger, disgust, fear, happiness, sadness and surprise, which are studied in the majority of papers.

II. Literature Review

There are different methods that are used for Face Expression Recognition in the last 20 years but usually they are separated into two main methods: Conventional or Traditional FER Approach and Deep Learning-based approach.

A. Datasets

In Facial Emotion Recognition, many kinds of datasets are utilised with each having their own characteristics. Scientists need to use big datasets in order to carefully research automatic recognition of facial expressions. The datasets have been created by them, but they should be supervised by psychologists who understand human emotion recognition better and must give their input. Unfortunately, most datasets are limited, however, the scientific community have been trying passionately to create new, extended and useful ones during these last few years.

In the majority of studies, two-dimensional static images have been used. While in others, sequences from 2D videos have been studied, since they show expressions in many dimensions with rarer still a few that use 3D modelling. The degree to which FER is successful is highly affected by some certain circumstances such as the: ambient light, background, posture of the subject, etc. For this matter, the acquired images are separated in two main groups: “images in the lab” and “images in the wild”. Which one is chosen depends on the purpose of the study and the methods employed for analysis. The prevailing most used datasets are currently: CK+ [4], MMI [5], MPI [6], JAFFE [7], KDEF [8]. To obtain better results, not only in FER, but in image recognition in general, data augmentation is applied. This method is applied in training data by adding some modifications to the image, such as cropping images in different coordinates, rotating in different degrees, scaling in random numbers, shearing, etc.

One technique used for analysing the data is the usage of ‘Action Units’. They are based on the FACS (Facial Action Coding System) systems which classifies the movements of the muscles of the human face while showing a certain expression.

B. Traditional Approach

Hand-Crafted or Traditional approaches are highly dependant on manual feature engineering. The researcher at first needs to process the image in order to find the most appropriate feature extraction and classification methods for the target dataset.

This approach can be divided into three main steps:

i. **image pre-processing**: where the main goal is to eliminate the irrelevant information of the input image and enhance the detection ability of the important characteristics.

978-1-7281-8488-3/20/$31.00 ©2020 IEEE

ii. **feature extraction**: which is used as a process to get some information or data from the images which are useful (symbols, vector, etc.). Here the most used ones are the Local Binary Pattern and Gabor feature extraction.

iii. **Expression/emotion classification**: which is used to predict the emotion in the human face. The most accurate classifiers in Facial Emotion Detection are the SVM (Support Vector Machine), kNN (k-Nearest Neighbours), Adaboost (Adaptive Boosting) and Bayesian algorithms.

C. *Deep Learning Approach*

In the Deep Learning Approach, there are three main steps that are followed: pre-processing, Deep Feature Learning and Deep Feature Classification Pre-processing the image - being a very important step. Usually we have noticed that this phase includes face detection using mostly the Viola-John algorithm (with Haar Cascade), face alignment, normalisation (illumination, pose) and augmentation (scaling, rotating, colours, noises, etc.).

Then Deep Feature Learning and Classification takes place. A number of techniques like CNN (Convolutional Neural Network) [9], DBN (Deep Belief Network) [10], DAE (Deep AutoEncoder) [11], RNN (Recurrent Neural Network) and GAN (Generative Adversarial Network) [13] have been created and researched on.

CNN configurations are still popular and the state of art when it comes to image pre-processing, especially in emotion recognition. Some of them are Region based CNNs (R-CNN) [14], Faster R-CNNs [15] and 3D CNNs [16].

Table 1, below, shows the main characteristics, such as the number of layers, Dropout, Inception, DA (Data Augmentation) and BN (Batch Normalization) of the most used and best performing CNN algorithms in FER.

TABLE I. FER CHARACTERISTICS OF CNN ALGORITHMS.

Year	**2012**	**2014**	**2014**	**2015**
No. of Layers	5+3	13/16 + 3	21+1	151+1
DA	Yes	Yes	Yes	Yes
Dropout	Yes	Yes	Yes	Yes
Incept.	No	No	Yes	No
BN	No	No	No	Yes
Used in	[21]	[22], [23]	[24], [25]	[26], [27]

The accuracy of each algorithm or approach is dependant on the data. For example, CNN is used in both the FER2013 and CK+ datasets but with very different accuracies. In the CK+ case, CNN gives an accuracy of 96.62% while with the FER2013 dataset, CNN only gave an accuracy of 72.10%. Based on how challenging the data can be, the difference of accuracies for example with the CK+ dataset, which is considered an "easy" dataset, can be improved further to 98.57% using a combination of both DBN + MPL. However, the highest accuracy achieved in SFEW is only 51.72% using CNN (ACNN).

Based on the studies conducted so far, the following conclusions can be made:

The Traditional Approach when compared to the Deep Learning Approaches is less dependant on the hardware and the type of data being processed. In this approach, feature extraction and classification are treated as completely separate steps which have to be applied manually, so it is not possible to execute them simultaneously. In contrast, the Deep Learning Approach can be done simultaneously. For more challenging tasks like FER under wild environmental conditions (natural settings, outdoors for example), conventional approaches are rarely applied since feature extraction for complex datasets is still very challenging and an impediment.

III. THE DATASETS AND CHALLENGES (ICV MEFED)

The choice of selecting the dataset proved difficult initially. The goal was to undertake a complete study of the matter from working with the raw photo quality image up to training and testing the neural network. The choice upon further research was narrowed down to these three datasets: FER2013, AffectNet and iCV MEFED. The last one was selected due to the fact that it offered completely raw images, was relatively new, the subjects had been supervised by psychologists and that the dataset had not been exhaustively explored nor tested by researchers.

This dataset is created by iCV Research Lab in Estonia and it consists of 125 different subjects who act 50 emotions each in front of a Canon 60D camera. Each image sample is 5184×3456 pixels in size taken under the same lighting condition with a green background. The subjects were trained by psychologists to express their emotions effectively. There are seven basic emotional states plus a Neutral pose: i) Anger, ii) Contempt, iii) Disgust, iv) Fear, v) Happiness, vi) Sad, vii) Surprise and N-Neutral. However, the creators of this dataset intended to have compound (mixed) emotions for example: angrily surprised. In the dataset there were 115 subjects with 50 emotions consisting of five images per emotion.

Dealing with such a high-volume dataset was very challenging. 74 GB of data are huge to work with when using an average internet speed of 15 Mbps to download and 4 Mbps to upload. It would take 40 mins per Gigabit. In this situation a Virtual Machine (VM) such as Microsoft Azure Portal was used. This is where most of the upload/download functions were executed. The only difficulty was moving the data back and forth between the VM and my computer, but a solution to that was using RDP (Remote Desktop Protocol).

A. *Pre-processing and Challenges*

As Fig. 1 shows, the images are quite big with more than 66% of them having unnecessary information. That is why cropping of the face part was needed.

Fig. 1. A preview of the iCV MEFED (Multi-Emotion Facial Expression Dataset).

For this, two methods were employed: the Viola-Jones [27] algorithms for face detection and cropping; the second one was detecting and cropping it manually. The main purpose of this was to understand the difference, the impact that it would have in accuracy latter. Regarding the time, using the Haar Cascade algorithm was a very fast and efficient method, while cropping manually one by one took a considerable amount of time. But when it comes to accuracy, some of the image were cropped completely wrong using the algorithm since it was mistaking the neck for being part of the face. While in some other cases, when the face was not completely straight, it was not detected at all. Three such examples are shown in Fig. 2.

Fig. 2. Inaccurate face detection and cropping using Viola-Jones algorithm.

The next step was to convert to grayscale and resize the images. From 5184×3456 pixels, different scalings were tried such as 255×255, 150×150, 170×190, 140×180 pixels. For the end results, the last dimensions were chosen. The images themselves were clear, there were no illumination, colour problems, noise or any other similar information which would interfere with the final result.

One of the key parts was to label the seven basic emotions of the subject images manually. By default, this dataset had labels of two mixed emotions, for example, sadly surprised or surprisingly sad. The chosen process was to select the dominating emotion of the faces of the subjects. After labelling the images in a separate .csv file, another .csv file with image data on it was created.

After different experimentations, these ways of storing the data were chosen in order to make it ready to feed to the network:

-*one folder to csv* (Excel stores only 16,384 columns, our data was bigger), that is why it took a lot of time to process. It is recommended for images smaller than 128×128 pixels.

- *one folder to many folders*, it is time consuming and prone to mistakes separating images into different folders. If they are labelled, there are ready programs that can be used to separate them, but using Python is recommended.

- *from one folder to HDF5* is recommended when there is a huge amount of data.

The dataset was diverse in age, race and gender as shown in Fig. 3. The dataset also had different illumination, position and distance as shown in Fig. 4.

Fig. 3. Subject from different age, race, gender.

Fig. 4. Subject in different illumination, position and distances.

Another notable challenge faced, which had an impact in the algorithm is the fact that some subjects did not show a decent difference from one emotion to another. This issue brought difficulties in emotion classification since they might overlap. Fig. 5 shows subject with little to no difference in emotions.

Fig. 5. Subjects with little to no difference in emotions.

In some subjects, it was very difficult to detect the emotions, even by human eye judgement. One example is shown in Fig. 6, below.

Fig. 6. Different emotions not expressed properly.

Since the dataset had 50 emotions and only 7 emotions with the same number of images were required, relabelling was needed. At first, 4/5 images were deleted for every emotion by leaving only the most decent one and then relabelling according to the 7 emotion cases. It was ensured to have a similar distribution of data, 250 images for an emotion. At the end of the data pre-processing stage, the result was as shown in Fig. 7.

Data Augmentation was needed since the dataset was not large enough to properly train a CNN. Online or real-time Augmentation was used.

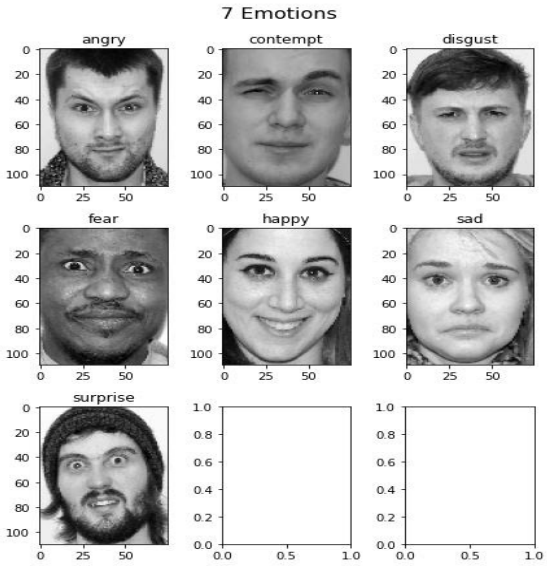

Fig. 7. Data samples used to feed the network.

IV. CNN Model and Results

The implemented first CNN network had four Convolutional Layers, four Max Pooling, One dropout and two Fully Connected Layers. In total, the model had 899,718 parameters.

At first the image is passed through the model and the Conv2D filter is passed through the image then ReLu (rectified linear activation function) was applied. The pooling layer (MaxPooling2D) takes the maximum values from the convolution layer and reduces the dimensions of the image. Then Flatten and Dropout layer is applied. Fig. 8 shows the result with 39 epochs with a batch size of 64.

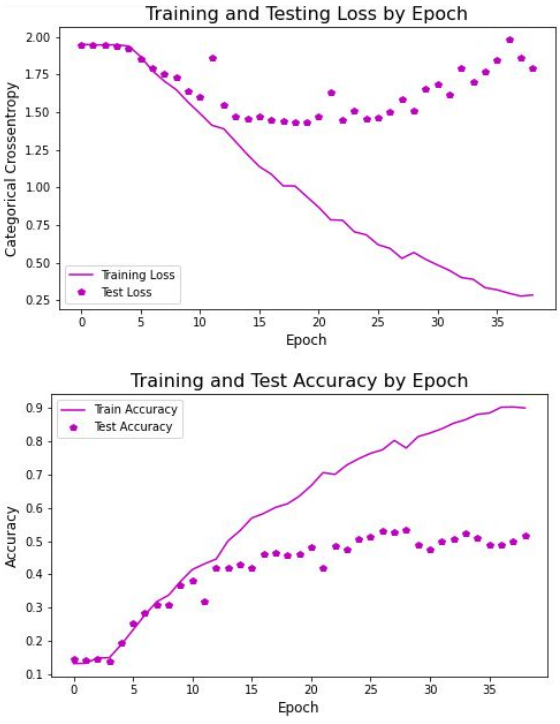

Fig. 8. Model 1 Evaluation without data augmentation.

According to the evaluation the training data performed better than the testing, exhibiting overfitting. After epoch 25, the accuracy score is higher for the training data, meaning the proposed model makes better predictions on the training data than the testing data. Trying the same model but with data augmentation, gave better results, as shown in Fig. 9. At the third experiment a pre-trained VGG16 model was applied to our dataset with image size 250×250 pixels.

Our models predicted the emotions by showing a certain percentage for each emotion/label and then the highest weight defined the final expression as it is shown on Fig. 10. Table II shows the sample of two first images for probabilities (from 1 to 0) that the model predicted for the seven emotions.

TABLE II. Predicted Probabilities for the Seven Emotions.

	angry	contempt	disgust	fear	happy	sad	surprised
0	0.751891	0.006519	0.011989	0.001366	0.000003	0.228214	0.000019
1	0.038393	0.001027	0.531945	0.217244	0.000093	0.189987	0.021311

We had different results for each emotion.

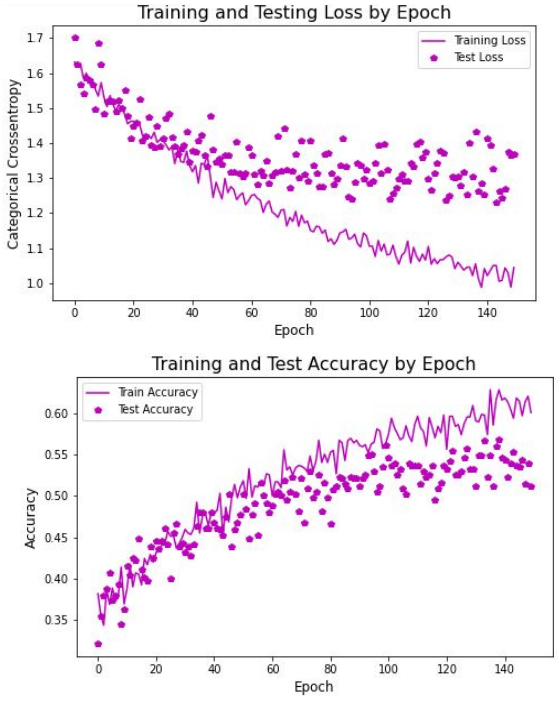

Fig. 9. Model 1 Evaluation with data augmentation.

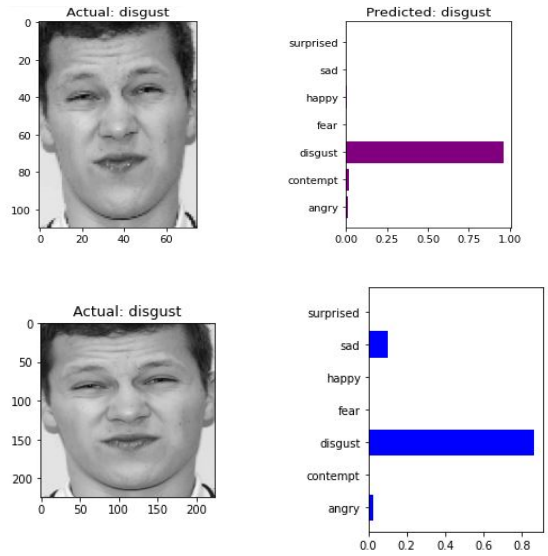

Fig. 10. Prediction of actual emotion First Model (disgust): 96.0%. Second Model (disgust): 86.64%.

As a result, the algorithm has performed best in detecting happy faces as it is shown in our Confusion Matrix in Fig. 13 and performed worst in detecting contempt. The most problematic case is confusing fear with surprise as it is shown in Fig. 11, which actually we expected this to happen due to their similarity, where in both cases the eyebrows are risen up and the eyes are wide open. The second most problematic is confusing contempt with sadness, where again facial expression are very similar to each other. We notice that the algorithm performs very well in recognising the difference between surprised and angry emotions as it is shown in Fig. 12.

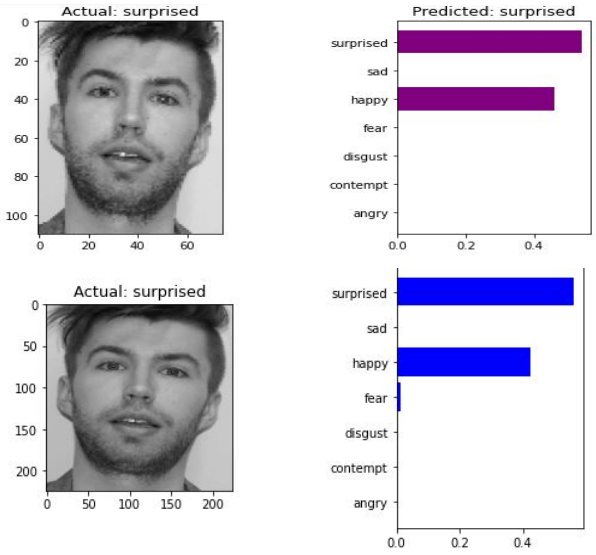

Fig. 11. Prediction of actual emotion First Model 55.17% and Second Model: 56.52%.

As Fig. 12 shows, these emotions are overpredicted: angry, disgust, fear and surprised, whereas these emotions are underpredicted: contempt and sadness.

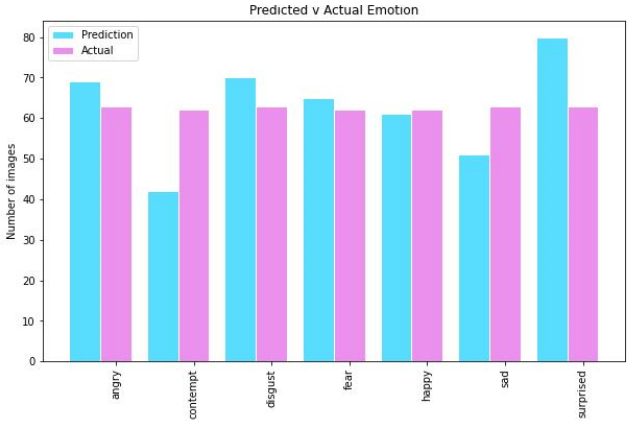

Fig. 12. Graphical representation of the comparison between actual and predicted emotions.

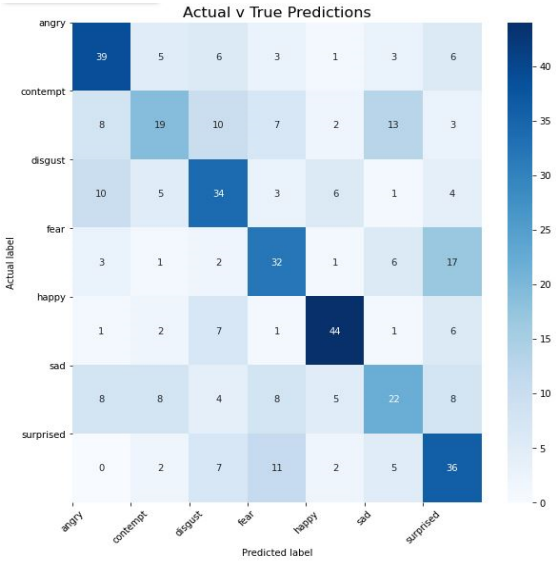

Fig. 13. Confusion Matrix of the results.

V. Future Work

Our aim in this work were to go in detail in every step of Deep Learning, starting from finding the dataset and ending up analysing the results. We separated the work into two main parts: Dataset and the construction of the CNN. As a future work we would suggest using the same dataset but trying a different approach by taking into consideration "Action Units" to detect as features the movement of the muscles of the face and then to feed the CNN. In order to have fast results and to perform many experiments with different parameters, a machine with proper parameters (especially GPU) is recommended.

References

[1] C. Darwin, The expression of the emotions in man and animals. D. Appleton And Company, New York, USA, 1899. Available: https://www.gutenberg.org/files/1227/1227-h/1227-h.htm

[2] Y.-I. Tian, T. Kanade and J. F. Cohn, "Recognizing action units for facial expression analysis", in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 2, pp. 97-115, Feb. 2001, doi: 10.1109/34.908962.

[3] P. Ekman, "Lie Catching and Microexpressions", in C. Martin (ed.), The Philosophy of Deception, Oxford Scholarship Online, 2009. DOI:10.1093/acprof:oso/9780195327939.003.0008

[4] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar and I. Matthews, "The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression", *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*, San Francisco, CA, 2010, pp. 94-101, doi: 10.1109/CVPRW.2010.5543262.

[5] M. Pantic, M. Valstar, R. Rademaker, and L. Maat, "Web-based database for facial expression analysis", *2005 IEEE International Conference on Multimedia and Expo*, Amsterdam, 2005, pp. 5 pp.-, doi: 10.1109/ICME.2005.1521424.

[6] M. Valstar and M. Pantic, "Induced disgust, happiness and surprise: an addition to the mmi facial expression database," in Proc. 3rd Intern. Workshop on EMOTION (satellite of LREC): Corpora for Research on Emotion and Affect, 2010. Available: http://www.cs.nott.ac.uk/~pszmv/Documents/MMI_spontaneous.pdf

[7] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba, "Coding facial expressions with Gabor wavelets", *Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition*, Nara, 1998, pp. 200-205, doi: 10.1109/AFGR.1998.670949.

[8] J. M. Susskind, A. K. Anderson, and G. E. Hinton, "The Toronto face dataset", Department of Computer Science, University of Toronto, Toronto, ON, Canada, Technical Report UTML TR 2010-001, 2010.

[9] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets", Neural computation, vol. 18, no. 7, July 2006, pp. 1527–1554, doi: https://doi.org/10.1162/neco.2006.18.7.1527

[10] [10] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks", *Science,* vol. 313, Issue 5786, 28 Jul 2006, pp. 504-507, doi: 10.1126/science.1127647

[11] A. Sherstinsky, "Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) Network", *Physica D: Nonlinear Phenomena*, vol. 404, March 2020, article id. 132306, doi: 10.1016/j.physd.2019.132306

[12] R. Girshick, J. Donahue, T. Darrell and J. Malik, "Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation", *2014 IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, OH, 2014, pp. 580-587, doi: 10.1109/CVPR.2014.81.

[13] S. Ren, K. He, R. Girshick and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks", in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137-1149, 1 June 2017, doi: 10.1109/TPAMI.2016.2577031.

[14] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in Advances in neural information processing systems, 2014, pp. 2672–2680

[15] S. Ji, W. Xu, M. Yang and K. Yu, "3D Convolutional Neural Networks for Human Action Recognition", in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 1, pp. 221-231, Jan. 2013, doi: 10.1109/TPAMI.2012.59.

[16] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks". Commun. ACM 60, 6 (June 2017), pp. 84–90, doi: https://doi.org/10.1145/3065386

[17] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition", https://arxiv.org/abs/1409.1556v6

[18] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions", *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, 2015, pp. 1-9, doi: 10.1109/CVPR.2015.7298594.

[19] K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition", *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, 2016, pp. 770-778, doi: 10.1109/CVPR.2016.90.

[20] S. Ouellet, "Real-time emotion recognition for gaming using deep convolutional network features", https://arxiv.org/abs/1408.3750v1

[21] G. Levi and T. Hassner, "Emotion Recognition in the Wild via Convolutional Neural Networks and Mapped Binary Patterns", in Proceedings of the 2015 ACM on International Conference on Multimodal Interaction (ICMI '15). Association for Computing Machinery, New York, NY, USA, Nov. 2015, pp. 503–510, doi: https://doi.org/10.1145/2818346.2830587

[22] H. Ding, S. K. Zhou and R. Chellappa, "FaceNet2ExpNet: Regularizing a Deep Face Recognition Net for Expression Recognition", 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017), Washington, DC, 2017, pp. 118-126, doi: 10.1109/FG.2017.23.

[23] X. Zhao, X. Liang, L. Liu, T. Li, Y. Han, N. Vasconcelos and S. Yan, "Peak-Piloted Deep Network for Facial Expression Recognition", in: Leibe B., Matas J., Sebe N., Welling M. (eds) Computer Vision – ECCV 2016. ECCV 2016. Lecture Notes in Computer Science, vol 9906. Springer, Cham. https://doi.org/10.1007/978-3-319-46475-6_27

[24] C. A. Corneanu, M. O. Simón, J. F. Cohn and S. E. Guerrero, "Survey on RGB, 3D, Thermal, and Multimodal Approaches for Facial Expression Recognition: History, Trends, and Affect-Related Applications", in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 38, no. 8, pp. 1548-1568, 1 Aug. 2016, doi: 10.1109/TPAMI.2016.2515606.

[25] P. Hu, D. Cai, S. Wang, A. Yao and Y. Chen, "Learning supervised scoring ensemble for emotion recognition in the wild", in Proceedings of the 19th ACM International Conference on Multimodal Interaction (ICMI '17). Association for Computing Machinery, New York, NY, USA, 2017, pp. 553–560, doi: https://doi.org/10.1145/3136755.3143009

[26] B. Hasani and M. H. Mahoor, "Facial Expression Recognition Using Enhanced Deep 3D Convolutional Neural Networks", *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Honolulu, HI, 2017, pp. 2278-2288, doi: 10.1109/CVPRW.2017.282.

[27] K. Vikram and S. Padmavathi, "Facial parts detection using Viola Jones algorithm", 2017 4th International Conference on Advanced Computing and Communication Systems (ICACCS), Coimbatore, 2017, pp. 1-4, doi: 10.1109/ICACCS.2017.8014636.