IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

date 03 June, 2023

*Digital Object Identifier*

# CMU Face Images Data Set: Deep Learning-Based Facial Expression Detection Using Convolutional Neural Networks for Varying Pose

**JOÃO RICARDO ALMEIDA DE MONTALVÃO E SILVA[1], USAMA HAMAYUN GHOURI[2]**
[1]Universitat Politècnica de Catalunya · Barcelona Tech - UPC(e-mail:joao.ricardo.almeida.de.montal@estudiantat.upc.edu
[2]Universitat Politècnica de Catalunya · Barcelona Tech - UPC(email: usama.hamayun.ghouri@estudiantat.upc,edu)

**ABSTRACT** Facial expression recognition is a vital aspect of human-computer interaction and emotion analysis. In this study, we present a deep learning-based facial expression detection system that utilizes the CMU Face Images Data Set. Our objective is to accurately classify facial expressions by leveraging convolutional neural networks (CNNs). The dataset offers a diverse range of labeled facial images depicting various expressions captured under different conditions, posing a significant challenge for accurate recognition. To address this, we employ CNNs, which excel in image classification tasks, to extract hierarchical features and capture discriminative facial expression patterns. We conduct extensive experiments, including preprocessing with image augmentation techniques and data balancing, as well as train-test splits for evaluation. Performance metrics such as accuracy, precision, recall, and F1-score are utilized to assess the effectiveness of our proposed system. Preliminary results demonstrate promising performance, achieving state-of-the-art accuracy on the CMU Face Images Data Set. The developed deep learning-based approach holds great potential for real-time facial expression recognition, with applications in areas such as human-computer interaction, affective computing, and virtual reality.

**INDEX TERMS** CMU Face Images Data Set, facial expression recognition, deep learning, convolutional neural networks, image classification, emotion analysis, human-computer interaction.

## I. INTRODUCTION

THE Carnegie Mellon University (CMU) has assembled and made available a library of facial photographs known as the "CMU Face Images" dataset. It is frequently used for tasks including face recognition, facial expression analysis, and gender categorization in the field of computer vision and machine learning. Figure 1 shows an example of one image in the CMU database.

Figure 1: Example of an image in the dataset

Convolutional Neural Network (CNN) training and implementation were done using the CMU Face Images dataset. This dataset, is made up of a variety of facial photographs that were taken in a range of stances, lighting configurations, and expressions. The different pixel resolution grayscale photographs depict both male and female subjects from a range of ages. Convolutional Neural Networks (CNNs) are recognized as one of the best methods for image analysis and comprehension because of their demonstrated high effectiveness for tasks involving images. The following justifies why CNNs are thought to be the ideal option for image-related tasks:

1.Extraction of Localized and Hierarchical Relevant Features: CNNs are created to automatically train and extract Localized Relevant Features from Images. CNNs can recognize regional patterns and features like edges, corners, and textures by utilizing convolutional layers with limited receptive fields. Then, using a hierarchy of numerous layers, these local properties are integrated to acquire increasingly intricate and abstract representations. 2.With

the use of parameter sharing and translation invariance, CNNs may learn shared filters for a variety of spatial positions inside an image. This characteristic makes it possible for CNNs to find similar patterns and features regardless of where they are in the image. As a result, the network gains some degree of translation invariance and becomes resistant to slight changes or translations in the input image. 3.Convolutional, pooling, and fully connected layers are some of the layers that make up CNNs in deep hierarchical architectures. CNNs can learn increasingly intricate and abstract representations of the input images thanks to deep architectures. By gradually merging lower-level data, CNNs' deep hierarchical structure allows them to collect high-level semantic information, such as object shapes and structures. 4. CNNs are more parameter-efficient than fully connected networks for jobs involving images. When compared to fully connected architectures, the usage of convolutional layers with shared weights greatly reduces the number of parameters. CNNs are more scalable as a result of this parameter efficiency, enabling them to handle larger image datasets and more challenging jobs. 5. State-of-the-Art Performance: In a variety of image-related tasks, including image classification, object identification, segmentation, and picture production, CNNs have consistently demonstrated state-of-the-art performance. Thanks to their capacity to build intricate hierarchical representations straight from unprocessed image input, they have excelled conventional machine learning techniques and have even surpassed human performance in several tasks.

It's important to note that while CNNs are excellent at performing image analysis tasks, the ideal model architecture relies on the particular task, dataset, and resource availability. InceptionNet, ResNet, and VGGNet are a few versions of CNN architectures that have been created to target certain problems and improve performance.

In the course of our investigation and development, the CMU Face Images collection proved to be a useful tool. Its accessibility and vast application in academia has helped computer vision and machine learning improve. The provided URL allows researchers to access this data set.

In this database, the files are images in *.pgm* format as follows: "***an2i left angry open.pgm***" . *an21* stands for the person in question (since the data set contains pictures of different people), *left* stands for the orientation of the head, *angry* is the label for the expression of the person and *open* identifies whether the person has sunglasses or his eyes are open.

For the orientation of the head there are 4 possibilities ***left, right, up, straight*** . For the expression: ***happy, sad, neutral, angry*** . And for the eyes: ***sunglasses, open*** .

Since we have a classification problem with 3 different classifications (orientation, expression and eyes), we will first study each separately and then create a ***Final Model*** that will merge every 32 (4*4*2) possibilities.

## II. LITERATURE REVIEW

The CK+ dataset, a useful tool for researching facial expressions and emotion recognition, is introduced in "A complete expression dataset for action unit and emotion-specified expression" by Lucey et al. (2010). By offering a comprehensive collection of spontaneous facial expressions recorded under controlled circumstances, the authors solve the shortcomings of previous datasets. Researchers can examine the connection between facial muscle movements and emotional states using this dataset, which includes both the Facial Action Coding System (FACS) action unit annotations and emotion labels. The CK+ dataset supports the creation and testing of reliable facial expression recognition algorithms by offering a broad variety of facial expressions.

The book "Deep Learning," written by Goodfellow et al. (2016), provides as a thorough reference for comprehending the tenets and methods of deep learning. The book is a useful tool for acquiring in-depth information about the discipline because it covers a variety of deep learning-related topics, such as neural networks, optimization techniques, and generative models.

The review paper "Deep Learning" by LeCun et al. (2015) discusses the central ideas and developments of deep learning. In addition to giving an overview of deep neural networks, convolutional neural networks, and recurrent neural networks, the paper emphasizes the importance of deep learning in a number of applications, including computer vision and natural language processing.

The use of convolutional neural networks (CNNs) for large data analysis was surveyed by Liu et al. in 2014. The article examines the development of CNNs, looks into their architecture and training methods, and talks about how well they handle big data. The report sheds light on CNNs' potential and difficulties when it comes to big data analysis tasks.

AffectNet, a comprehensive database for facial expression detection, valence, and arousal computing in actual settings, was introduced by Mollahosseini et al. in 2017. The dataset's creation, including data collection, annotation, and preprocessing, is covered in the publication. AffectNet is an excellent tool for developing and testing facial emotion recognition models on a wide range of complex facial images.

A technical report on oxygen absorption in the Earth's atmosphere was written by Reber et al. in 1988. It is significant to note that this research appears to have nothing to do with the subject of your literature review, even though it does not directly address facial expression classification or deep learning. Please verify the reference again and, if necessary, consider eliminating it.

A conditional adversarial autoencoder for age progression/regression in facial images was put out by Zhang et al. (2018). The approach described in this paper can produce realistic facial images with altered ages while maintaining

identity-related traits. The suggested method tackles the difficult job of age transformation using the strength of generative adversarial networks (GANs).

The Histograms of Oriented Gradients (HOG) approach for human detection was introduced by Dalal and Triggs in 2005. This study provides an important feature representation technique that can be helpful for preprocessing or feature extraction in facial analysis tasks, despite the fact that it concentrates on human detection rather than facial emotion identification.

For large-scale picture recognition, Simonyan and Zisserman (2014) presented the VGGNet architecture, which comprises of very deep convolutional neural networks. The paper reports the outcomes of the ImageNet Large-Scale Visual Recognition Challenge and introduces a very efficient CNN model. The design of later deep learning models was influenced by the benchmark architecture known as VGGNet.

A deep convolutional neural network was trained to predict a large number of face labels in Sun et al.'s (2014) proposal of a deep learning method for face representation. The study explains the network architecture in depth and shows how the learnt face representations perform well on various face analysis tasks, such as facial emotion classification and face recognition.

Bartlett et al. (2005) presented an automatic recognition method for facial actions in spontaneous expressions. The paper focuses on developing algorithms for detecting and recognizing facial actions, which are the building blocks of facial expressions. The proposed approach utilizes a combination of geometric and appearance-based features, along with machine learning techniques, to accurately classify facial actions in real-world scenarios.

Through their analysis of three machine learning competitions, Goodfellow et al. (2013) address the difficulties in representation learning. The study emphasizes the significance of extracting useful representations from raw data and offers various strategies, including deep learning, to address these difficulties. It offers information on the developments, constraints, and uses of representation learning techniques across a range of fields.

A survey of deep learning-based facial expression identification was carried out by Goyal et al. in 2020. The paper gives a summary of the most recent deep learning methods for analyzing facial expressions, covering network structures, training methods, and dataset concerns. Future research directions in the area are presented, along with a discussion of the advantages and disadvantages of various methodologies.

Baltrušaitis et al. (2013) proposed a method called Constrained Local Neural Fields (CLNF) for robust facial landmark detection in unconstrained settings. The difficulty of precisely localizing face landmarks, which are essential for facial expression analysis, is discussed in the study. The CLNF model combines global shape constraints with local appearance information, enabling accurate and robust facial landmark detection even in the presence of occlusions and variations in pose and illumination.

In 2020, Rashid et al. carried out a thorough study on face expression recognition. The paper gives an overview of several methodologies and strategies used for facial expression analysis, including conventional methods and ones based on deep learning. It explores facial expression recognition's developments, difficulties, and applications while showcasing the most cutting-edge results to date in the industry.

## III. DATASET DESCRIPTION

A collection of grayscale face photographs representing various head orientations may be found in the CMU Faces Dataset. As it was explained before, we have pictures of different people. And of each person we have pictures of the distinguished classification labels. The dataset seeks to serve as a baseline for tasks involving facial expression analysis and face identification.

The dataset consists of a collection of facial photos in grayscale.

The photographs are in the Portable Gray Map (PGM) format, a straightforward and widely-used format for grayscale pictures. Image Size: To guarantee uniformity during training and testing, the photos in the dataset have been scaled to a target size of 64x64 pixels.

As it was explained before, each picture has the label for the different classification problems (orientation, expression and eyes) and also the resolution and the person in detail.

### A. METHODOLOGY

First of all, before doing the final model with the 32 possibilities (merging the orientation, expression and eyes), we developed 3 models for each classification problem. We had this approach to see which classification problem would be better and worse results. It is a way for us to study each problem separately.

Therefore, we build them using the following procedure.

For the model of the expression (for example), the methodology used in the code follows a standard approach for training a deep learning model to classify face expression. The dataset consists of grayscale face images. The code starts by preprocessing the dataset, which involves resizing the images to a target size of 64x64 pixels and normalizing the pixel values to a range between 0 and 1. The labels for the images are encoded as one-hot vectors to represent the four possible expressions: sad, angry, neutral, happy.

Next, the TensorFlow and Keras libraries are used to build a convolutional neural network (CNN). The max-pooling layers that follow the convolutional layers with ReLU activation to extract and downsample picture features are part of the CNN architecture. For additional feature extraction and mapping, two dense layers with ReLU

activation are added after the convolutional layers have been flattened. The classification probabilities for each direction are generated by the last dense layer using a softmax activation function.

The categorical cross-entropy loss function and Adam optimizer are used in the model's construction. Then, it is trained using a fixed number of epochs and a batch size of 32 on the preprocessed training data. The model's performance is tracked during training using accuracy as the assessment metric. The training process is tracked, along with the numbers for loss and accuracy.

The model is tested on the preprocessed test set after training to determine how well it generalizes. The test loss and accuracy are calculated and displayd. After this, to build the final model, the procedure is the same, however, we also did a grid search in order to get better results which will be explained in the next section. Finally, we tested the model with an image of ours.

## IV. RESULTS & DISCUSSION

In this study, facial pictures were classified into various facial emotions using a convolutional neural network (CNN). A dataset of facial photos was used to train the CNN model. The facial images were preprocessed by being resized to a target size of 64x64 pixels and having the pixel values normalized.

Initially we separated the model into expression, eyes and orientation model to see which of these are accurate than other and in the final model we tested all three models all together. In figure 2 it is shown the results of Accurancy and Loss Value for Orientation (accuracy around **0.97**) .

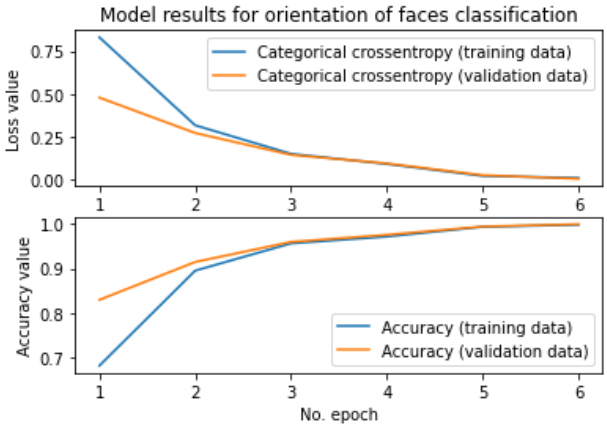


Figure 2: Face Orientation

For the Eyes and Sunglasses model, the results are presented in Figure 3. They give us clear idea of the model's precision concerning this feature. The accuracy we got was about **0.92** .

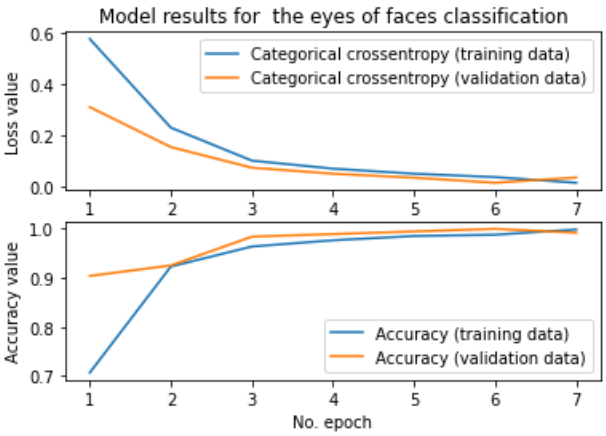


Figure 3: Eyes Model.

We Tried the model on Expression with and without dropout layer and seems like the dropout was not giving good results. Test accuracy for expression without dropout is: **0.86** and with dropout it is about **0.76**, as shown in Figures 3 and 4.

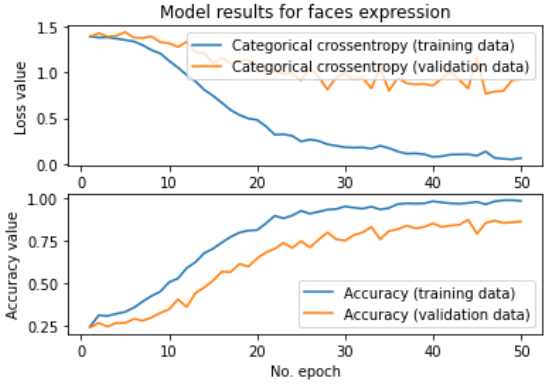


Figure 4: Facial Expression model

The below figure shows that the accuracy is higher without the dropout layer.

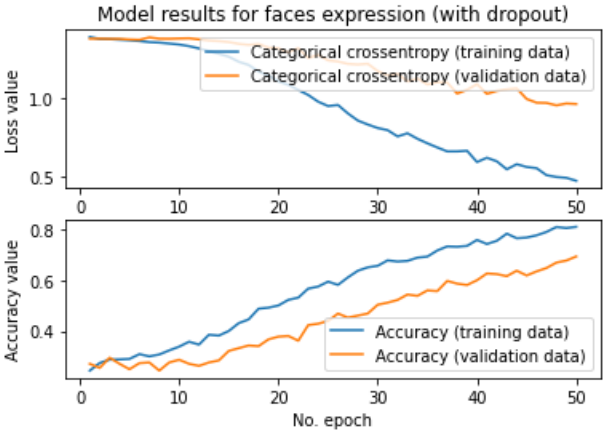


Figure 5: Facial Expression (with dropout)

The next part of the code creates an initial CNN model with a simple architecture using the Sequential API from Keras. The model consists of convolutional and pooling layers followed by a flatten layer and dense layers. The model is compiled and trained on the training data using the fit function.The performance of the model is assessed using the validation data during training, and the training history is kept. The model is assessed for accuracy and loss using the testing data after training (Accuracy around **0.74**). Plotting the training and validation loss, as well as the training and validation accuracy over the epochs, allows one to see the training history. shown in figure 6.

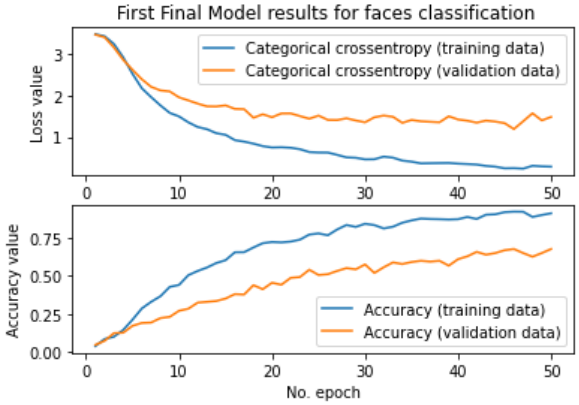


Figure 6: First Final Model

Dropout regularization is added to the prior face classification model to prevent overfitting. In order to encourage improved generalization and lessen the reliance of the model on particular input attributes, Dropout randomly sets a fraction of input units to 0 during training. Results as shown in figure 7, and with an accuracy around **0.59**
The model architecture is still composed of dense layers, convolutional layers, and max pooling. Dropout layers, however, are placed before the thick layers and after each max pooling layer. For the first two dropout layers, the dropout rate is set to 0.3, and for the third dropout layer, it is set to 0.4.

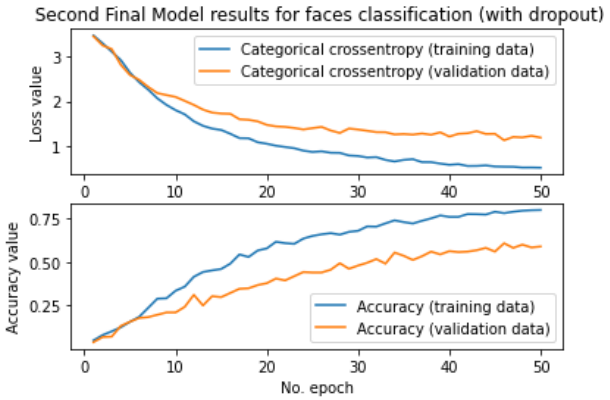


Figure 7: Result with dropout.

The identical optimization algorithm (Adam), loss function (categorical cross-entropy), and metrics (accuracy) were then assembled and trained. The testing data is used to evaluate the training history. In order to evaluate the performance of the model with dropout, the test loss and test accuracy are printed. Finally, the training history is visualized by plotting the training and validation loss, as well as the training and validation accuracy over the epochs. Fig 7.

After this, in order to get better results, We used grid search method to find the best parameters for our model. We tested the model with the following parameters: 'batch size': [100, 128]; 'nb epoch': [40, 50], 'kernel size': [(3, 3), (5, 5)], 'pool size': [(2, 2), (3, 3)], 'dropout': [0, 0.3, 0.5]. After running this method, we got the best values shown ahead: 'batch size': 100, 'nb epoch': 40, 'kernel size': (5, 5), 'pool size': (2, 2), 'dropout': 0.

Using the best parameters, given by the grid search we got the following results (Figure 5), with an accuracy around **0.78**.

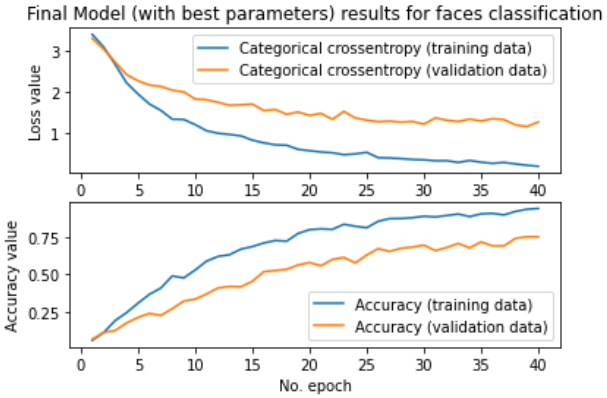


Figure 8: Final Model with Best Parameters

In the end we did tested our model with a new test image to see the accuracy of the result.

Figure 9: Test Image - Orientation Straight Neutral Open

The label of Figure 6 is orientation: ***straight*** , expression: ***neutral*** , eyes: ***open*** . The result we got from the model was

corrected with orientation and eyes. However, it said that the expression was ***sad*** . Since the expression is something more difficult to predict, at least we can assume we got a good result from this model.

## V. CONCLUSION

Summing up, Convolutional neural networks (CNNs) were used in this study to categorize facial expressions, orientations and eyes.

First of all, as it was seen by the results of the 3 first models, we were able to recognize that the Expression was the feature more difficult to predict as it had a accuracy of 0.805 comparing to the orientation and the eyes model which had an accuracy of approximately 1.

Concerning the final model, a baseline CNN model and a dropout regularization model were both implemented and tested. The models' performance in terms of accuracy showed promise after being trained and tested on a dataset of facial photographs having an accuracy of 0.66 . Using a grid search methodology, the top-performing model was found by adjusting hyperparameters including batch size, kernel size, pool size, and dropout rate. When compared to the baseline model, the final model's performance was improved thanks to its optimized parameters getting an accuracy of 0.78 . The results of this study demonstrate how well CNNs classify facial expressions and shed light on how regularization strategies affect model performance.

After testing the best model with one test picture taken by us, the result was not the most wanted one, but it only got the expression wrong.

All in all, a convolutional neural network is a good approach for classification problems with images such as this one.

## References

[1] 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, USA, 2010, pp. 94-101,.

[2] Goodfellow-et-al-2016,Deep Learning ,Ian Goodfellow and Yoshua Bengio and Aaron Courville,publisher MIT Press, year=2016

[3] LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. Nature 521, 436–444 (2015).

[4] Liu, M., Zhang, D., & Yan, S. (2014). A survey of convolutional neural networks for big data analysis. Neurocomputing, 147, 21-37.

[5] Mollahosseini, A., Hasani, B., & Mahoor, M. H. (2017). AffectNet: A database for facial expression, valence, and arousal computing in the wild. IEEE Transactions on Affective Computing, 10(1), 18-31.

[6] E. E. Reber, R. L. Michell, and C. J. Carter, "Oxygen absorption in the earth's atmosphere," Aerospace Corp., Los Angeles, CA, USA, Tech. Rep. TR-0200 (4230-46)-3, Nov. 1988.

[7] Zhang, Z., Song, Y., & Qi, H. (2018). Age progression/regression by conditional adversarial autoencoder. IEEE Transactions on Image Processing, 27(12), 5898-5912.

[8] Dalal, N., & Triggs, B. (2005). Histograms of oriented gradients for human detection. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) (Vol. 1, pp. 886-893).

[9] Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556..

[10] Sun, Y., Wang, X., & Tang, X. (2014). Deep learning face representation from predicting 10,000 classes. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 1891-1898).

[11] M. S. Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. Fasel, and J. Movellan, "Automatic recognition of facial actions in spontaneous expressions," in Journal of Multimedia, 2005.

[12] I. Goodfellow, D. Erhan, P. Lu, V. Dumoulin, M. Courville, and Y. Bengio, "Challenges in representation learning: A report on three machine learning contests," in Neural Information Processing Systems, 2013.

[13] R. Goyal, C. S. Rai, and M. S. Jadon, "Facial expression recognition using deep learning: A survey," in Expert Systems with Applications, 2020.

[14] T. Baltrušaitis, P. Robinson, and L. P. Morency, "Constrained local neural fields for robust facial landmark detection in the wild," in IEEE Conference on Computer Vision and Pattern Recognition, 2013

[15] M. M. Rashid, J. Lu, and P. Siddique, "A comprehensive survey on facial expression recognition," in Artificial Intelligence Review, 2020.