Communications on Applied Nonlinear Analysis
ISSN: 1074-133X
Vol 31 No. 2s (2024)

Efficient Facial Emotion Detection through Deep Learning Techniques

Priti Singh[1], Hari Om Sharan[2], C.S. Raghuvanshi[3]

[1,2,3]Department of Computer Science & Engineering, Faculty of Engineering and Technology, Rama University, Uttar Pradesh, Kanpur

preetirama05@gmail.com*[1], sharan.hariom@gmail.com[2], drcsraghuvanshi@gmail.com[3]

Article History:

Received: *25-03-2024*

Revised: *15-05-2024*

Accepted: *29-05-2024*

Abstract:
Smart facial emotion detection represents a captivating realm of inquiry that has found applications across diverse sectors such as defense, healthcare, and human-machine interfaces. Researchers are diligently exploring methods to encode, decode, and even obfuscate facial cues to refine algorithmic predictions. Leveraging a combination of deep learning algorithms and Cognitive Internet of Things (CIoT), efforts are underway to bolster efficiency in response to the rapid evolution of this technology. This study aims to distill recent advancements in smart facial expression recognition utilizing deep learning algorithms while pioneering novel approaches to emotion detection. The burgeoning Internet of Things landscape has underscored a deficiency in technological infrastructure within current automated intelligent services, rendering them ill-equipped to cater to industrial demands. The gradual augmentation of Internet of Things technologies tailored for intelligent environments has inadvertently led to delays and diminished market efficacy. Deep learning stands out as a cornerstone in myriad applications and experimental setups. Addressing this challenge necessitates the formulation of emotionally intelligent methodologies within the framework of deep learning, thereby invigorating Internet of Things initiatives, as elucidated by recent strides in facial emotion detection applications.

Keywords: Deep learning, facial emotion recognition, Cognitive Internet of Things.

1. Introduction

Emotional intelligence is the ability to understand and interpret thoughts and emotions. Various cues such as gestures, facial expressions, voice tone, and verbal communication can aid in recognizing individual emotions. When we speak of emotional intelligence, we inherently refer to the comprehension and expression of feelings. It involves not just recognizing emotions but also managing them effectively. As per [1], emotional intelligence entails assimilating emotions into one's cognition, understanding, and regulating them within oneself and others.

A smart ecosystem, as defined, is a realm where diverse smart technologies continually strive to enhance the quality of life for its inhabitants. Smart workplaces, a component of this ecosystem, aim to replace hazardous tasks, manual labor, and repetitive actions with automated solutions, benefiting people from all walks of life.

The characteristics of smart environments can be broadly categorized as follows:

1. Centralized management of computers, facilitated, for instance, by utilizing power line transmission networks.
2. Integration of computer networking, middleware, and wireless connectivity to conceptualize interconnected digital realms.

3. Gathering and dissemination of data through sensor networks.
4. Enhancement of service quality through intelligent systems.
5. Capacity for anticipation and decision-making.

Deep learning models offer a solution to the challenge of managing extensive datasets, as they possess the ability to autonomously identify pertinent features, requiring minimal intervention from programmers. In scenarios involving copious inputs and outputs, deep learning algorithms are instrumental.

Scholars assert that the fundamental objective of deep learning is to engineer algorithms capable of emulating the human brain. Deep learning, being a subset of artificial intelligence, aims to replicate human behavior. The implementation of deep learning relies on Neural Networks, which draw inspiration from biological neurons, essentially mimicking brain cells. A collective array of deep learning computational methodologies, founded on artificial neural networks, is recognized as deep learning.

The progression towards the Internet of Things (IoT) is propelled by the convergence of internet technology and smart devices or smartphones. Smart devices, equipped with sensors, actuators, and communication capabilities, aim to enhance daily life by generating and consuming information, thereby optimizing processes and reducing costs across various domains. However, there remains a pressing need for novel mechanisms to enhance cognitive control over the IoT. Distinguishing itself from traditional personal computers, smart machines exhibit user-independent adaptability, autonomously organizing and adjusting to their surroundings. To enable informed decisions and adaptive responses, logical sensors and actuators are imperative.

The Internet of Things (IoT) represents an extensive and pervasive network underpinning digital intelligent services, offering myriad applications and prospects for future intelligent infrastructure. Machine learning finds widespread application across diverse domains, revolutionizing fields such as medicine, education, and finance. The evolution of Machine Learning has transformed ordinary individuals into adept users, profoundly influencing everyday life. Despite its omnipresence, contemporary IoT infrastructure lacks intelligence, impeding its utilization for industrial service applications. Conversely, a modernized IoT framework incorporating emotional intelligence can revolutionize industries by facilitating tasks such as emotion detection. By integrating emotional intelligence into IoT frameworks, a paradigm shift towards emotional intelligence-driven IoT models can be achieved.

The Cognitive Internet of Things (CIoT) represents an IoT paradigm equipped with cognitive capabilities, collaborating to optimize efficiency and cognitive prowess. CIoT leverages existing network architectures, analyzes surface data, makes decisions, and executes adaptive tasks to bolster network capabilities. Given the intricate nature of the Internet of Things, characterized by interconnected networks reliant on knowledge exchange and information processing technologies, integrating emotional intelligence poses a considerable challenge.

In light of the aforementioned discourse, emotional intelligence emerges as a crucial component in IoT-related frameworks, enabling comprehension of human affective states and behavior based on

physiological signals. Emotional intelligence endeavors encompass arduous tasks such as amassing extensive data to discern and monitor various emotional phases over an extended period.

2. Literature Review

Paul Ekman's seminal research [7] on emotion recognition identified six primary emotions: happiness, sadness, anger, surprise, fear, and disgust, culminating in the development of FACS. Later, the inclusion of 'neutral' expanded most human recognition datasets to encompass seven fundamental emotions.

Scholars examine two main categories of landmark detection methods: regression-based [8], [9], [10], and model-based techniques [11], [12], [13]. While regression-based approaches estimate landmark positions from facial appearance, model-based methods capture both the shape and appearance of landmarks. Nonetheless, landmark estimation may falter under specific conditions such as extreme out-of-plane rotations, low scale, or notable variations in face bounding boxes.

Historically, emotion recognition employed a two-stage machine learning methodology involving feature extraction and classification, utilizing techniques like SVM and neural networks. Handcrafted features such as HOG [14][15], LBP[16], Gabor wavelets[17], and Haar features[18] were prevalent for facial expression recognition. However, these methods demonstrated limitations when confronted with more complex datasets featuring intra-class variability and challenging image conditions like partial faces or occlusions.

Some researchers explore the distinction between emotion classification and expression regression tasks. Emotion classification entails categorizing images or videos into discrete sets of facial emotion classes or action units [19][20], often leveraging facial landmarks. Recently, deep networks [21][22][23] have supplanted landmark position estimation.

In multimedia concept modeling [24], [25], and computer vision, emphasis is placed on modeling objects [25], scenes [26], and activities[27]. Furthermore, efforts have been made to model unconventional concepts such as image aesthetics and emotions. These models are trained using SVM and other methodologies.

Extensive research has delved into the transfer of learned deep representations across tasks in an unsupervised context [28], [29]. However, these models have primarily operated on small datasets and have achieved only moderate success. To counteract the challenge of limited training data, unsupervised pre-training followed by supervised fine-tuning and supervised pre-training using a concept-bank approach [30] have been proposed and proven effective in computer vision and multimedia contexts. Recently, supervised pre-training followed by domain-adaptive fine-tuning has emerged as an efficient paradigm for scarce data.

While prior studies have advanced emotion recognition, they lack a method to prioritize crucial facial regions for detection. In this study, we introduce an approach utilizing an attentional convolutional network to address this gap.

3. Methodology

Our project aims to utilize Convolutional Neural Networks (CNNs) for real-time analysis of facial expressions. Through extensive experimentation, we have adopted this specific deep learning

approach as it effectively addresses various challenges in facial expression detection. CNNs offer a significant advantage in enabling "end-to-end" learning directly from raw input data, thereby minimizing the reliance on physics-based simulations or other preprocessing methods [31].

Automated analysis of facial emotional states and identification holds considerable value, particularly in the context of identifying, assessing, and supporting vulnerable individuals such as those with psychiatric disorders, individuals under significant mental stress, and children with limited emotional regulation.

During our project, we encountered difficulty in distinguishing between the emotions of fear, surprise, and disgust while working with the Fer2013 dataset. Consequently, we opted to classify them collectively as 'surprise' among the five emotions considered, which also include happiness, sadness, anger, and neutrality [32]. The Fer2013 dataset, sourced from Kaggle, comprises spontaneous facial expressions captured under diverse and challenging conditions, including variations in lighting, head movements, and inherent differences in facial attributes across demographics such as race, age, gender, facial hair, and glasses.

Facial expression recognition (FER) techniques based on deep learning frameworks mitigate the reliance on face-physics-based models and preprocessing methods by enabling direct learning from input images within the processing pipeline [33]. The key phases of our methodology are as follows:

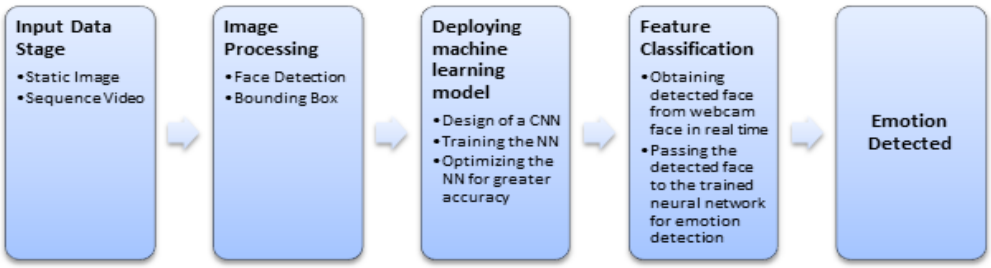

Figure1. Step for Emotion Detection

Humans have seven basic emotions (BE): happiness, surprise, indignation, sad-ness, fear, disgust, and neutral. Compound emotions (CE) are created when two basic emotions are mixed. Du et al. [12] identified 22 human emotions, including seven BE, twelve CE, and three others (appall, hate and awe). Micro movements (ME) are in-voluntary facial gestures that are minimal and unintentional. They have an uncanny capacity to reveal a person's true and hidden feelings in a short period of time [25].

Table1.Descriptions of facial muscles involved in the emotions Darwin considered universal

Emotion	Fear	Anger	Disgust	Contempt
Darwin's Facial Description	• Eyes open • Mouth open • Lips retracted • Eyebrows raised	• Eyes wide open • Mouth compressed • Nostrils raised	• Mouth open • Lower lip down • Upper lip raised	• Turn away eyes • Upper lip raised • Lip protrusion • Nose wrinkle

Emotion	Happiness	Surprise	Sadness	Joy
Darwin's Facial Description	• Eyes sparkle • Mouth drawn back at corners • Skin under eyes wrinkled	• Eyes open • Mouth open • Eyebrows raised • Lips protruded	• Corner of mouth depressed • Inner corner of eyebrows raised	• Upper lip raised • Nose labial fold formed • Orbicularis • Zygomatic

3.1 Data Preprocessing

The model demonstrates improved performance in predicting emotions when the images remain unaltered in terms of rotation and flipping. All manipulations are feasible with the Kaggle dataset. The dataset can be acquired and stored as a NumPy array with the dimensions: Samples * (Number * 48 * 48 * 48 * 1). Each image is a variant of (48 * 48 * 1) in some form. Although we maintain it in the array for convenience, you have the option to organize it into folders for data augmentation using the Image Data Generator class from keras preprocessing image.

3.2 Model Architecture & Functionality

Utilizing various CNN designs, the CNN (Convolutional Neural Networks) model architecture achieves outstanding accuracy. This architecture is employed to categorize concepts into four groups, and the resultant models are amalgamated to offer five emotion categories.

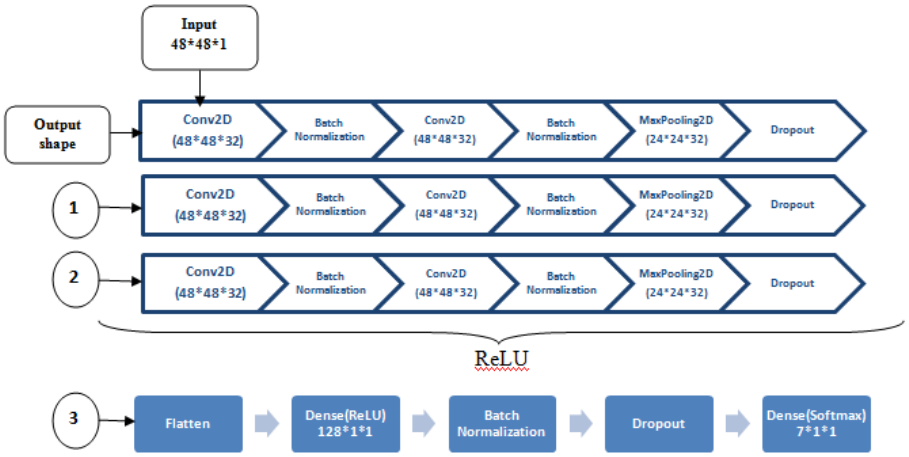

Figure2. CNN Model Architecture

The following GIF (graphics interchange format) uses the original image as input and the OpenCV module to turn it into a 48*48*1 shape to show how the CNN model functions. In this work, The CNN model applies all three blocks to the input image using the Conv2D, Batch Normalization, MaxPooling2D, and Dropout layers. Pooling layers minimised the size of a picture in half by only using the most important components from each block (which is shown by a slightly blurred image and reduced size, just for better visualization). In conv2d, this layer creates a convolution kernel that is convolved with the layer input to produce a tensor of outputs. If use bias is True, a bias vector is created and added to the outputs. Finally, if activation is not None, it is applied to the outputs as well. Batch Normalization is a normalization technique done between the layers of a Neural Network instead of in the raw data. It is done along mini batches instead of the full data set. It serves to speed up training and use higher learning rates, making learning easier. MaxPooling2D works by selecting the maximum value from every pool. Max Pooling retains the most prominent features of the feature map, and the returned image is sharper than the original image. The Dropout layer is a mask that nullifies the contribution of some neurons towards the next layer and leaves unmodified all others. We can apply a Dropout layer to the input vector, in which case it nullifies some of its features; but we can also apply it to a hidden layer, in which case it nullifies some hidden neurons. Dropout layers are important in training CNNs because they prevent overfitting on the training data. If they are not present, the first batch of training samples influences the learning in a disproportionately high manner. This, in turn, would prevent the learning of features that appear only in later samples or batches.

4. Results

We have work on 7-class (happy, angry, disgust, neutral, sad, fear, surprise) classification using the above model:

VGG can use a relatively small architecture of 3-by-3 convolution features to attain impressive accuracy in image classification. The number associated with each VGG model is the number of total depth layers, the majority of those being convolutional layers. The most widely used VGG models are VGG-16 and VGG-19, which are the two models that we chose for our study.

Despite being among the best CNN models at both object detection and image classification, VGG does have a few drawbacks which can make it challenging to use. Due to its robustness, VGG can be slow to train; the initial VGG model was trained over a period of weeks on a state of the art Nvidia GPU. Additionally, when VGG was utilized in the ILSVRC, the size of the weights used caused VGG to use a substantial amount of bandwidth and disk space.

ResNet is a type of CNN that was designed to improve image classification accuracy by allowing for the addition of more layers. This architecture enables the neural network to learn more complex features and achieve better performance. However, adding too many layers can cause a decrease in accuracy.

Inception is a convolutional neural network designed to address the challenge of varying salient parts of images. Instead of going deeper, Inception goes wider by using filters with multiple sizes. Inception v2 and Inception v3 were proposed to im-prove efficiency and accuracy, with Inception v3 introducing 7-by-7 convolutions and adjustments to auxiliary classifiers. Different versions of Inception have differences in image classification.

Table2. Accuracies for different models and CNN variants

Clas-ses	VGG 16	VGG 19	Incep-tionV3	Rest-Net50	Xcep-tion	CN N1	CN N2	CN N3	CN N4	CN N5
	64.30	58.55	35.76	44.34	40.45	66.45	66.65	64.55	65.28	64.80

Over-fitting is often seen in preset models like VGG's, Xception, etc. This describes the usage of CNNs. The accuracy vs. epochs plot for the CNN architecture can be seen in the graph below.

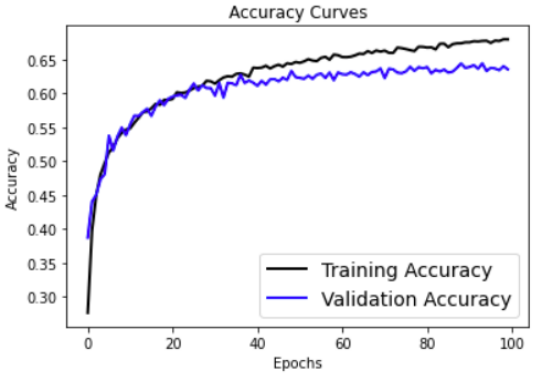

Figure3. Accuracy Vs Epochs graph for CNN model

Using the same CNN architecture, the four-class classification accuracy (happy, neu-tral, furious, and sad) increases from 64.46% to 74.68%.

The ensemble approach yields good results because it trains the model separately for each emotion class before merging it to make predictions based on test data. With an accuracy rate of 74.3% for the five emotion categories, the model surpassed the top CNN model for both the four and the seven emotion categories (sad, happy, surprised, furious, and neutral). This method of combining different CNNs produces top-notch results while using the same CNN architecture.

5. Conclusion

These findings indicate positive outcomes. The device's exceptional precision and rapid response make it well-suited for a wide range of practical applications. Utilizing the OpenCV module for initial face extraction followed by emotion detection enhances system performance across various data types. In this study, a CNN model is employed with three blocks comprising Conv2D, Batch Normalization, MaxPooling2D, and Dropout layers. The pooling layers effectively reduce image size by half while preserving essential components, enhancing visualization without compromising quality. The proposed CNN architectures achieve accuracies of 65.58% and 74.58% for seven-class and four-class emotion recognition, respectively. Incorporating an ensemble approach significantly boosts model accuracy, resulting in 74.3% accuracy and a 0.753 F1-score for five-class classifications. Future endeavors will focus on expanding the range of emotions and refining emotion parameters to yield improved results and predictions, along with exploring more accurate datasets to further enhance accuracy.

Refrences

[1] M. Chen, F. Herrera, K. Hwang,"Cognitive computing: architecture, technologies and intelligent applications", IEEE Access, vol 6, pp. 19774–19783, Jan 2018.

[2] M. Alhussein, G. Muhammad, M.S. Hossain, S.U. Amin, "Cognitive IoT-cloud in-tegration for smart healthcare: case study for epileptic seizure detection and moni-toring", Mob NetwAppl, vol. 23, pp. 1624–1635, Sep 2018.

[3] S. Gupta, A.K. Kar, A. Baabdullah, A.A. Wassan, Al. Khowaiter, "Big data with cognitive computing: a review for the future", International Journal of Information Management, vol. 69, pp. 78–89, Oct 2018.

[4] H. Xu, W. Yu, D. Grifth, N. Golmie., "A survey on industrial internet of things: a cyber-physical systems perspective", IEEE Access, vol. 6, pp. 78238–78259, Dec. 2018

[5] A. Sheth, "Internet of things to smart IoT through semantic, cognitive, and perceptual computing", IEEE Intelligent Systems, vol. 31(2), pp. 108–112, Mar.-Apr. 2016.

[6] P. Vlacheas, R. Giafreda, V. Stavroulaki, D. Kelaidonis, V. Foteinos, G. Poulios, P. Demestichas, A. Somov, A.R. Biswas, K. Moessner, "Enabling smart cities through a cognitive management framework for the internet of things", IEEE Communications Magazine, vol. 51, pp. 102–111, June 2013.

[7] Ekman, Paul, and Wallace V. Friesen. ”Constants across cultures in the face and emotion.” Journal of personality and social psychology 17.2: 124, 1971.

[8] X. P. Burgos-Artizzu, P. Perona, and P. Dollar. Robust face landmark ´ estimation under occlusion. In Proc. Int. Conf. Comput. Vision, pages 1513–1520. IEEE, 2013.

[9] D. E. King. Dlib-ml: A machine learning toolkit. J. Mach. Learning Research, 10(Jul):1755–1758, 2009.

[10] Y. Wu, T. Hassner, K. Kim, G. Medioni, and P. Natarajan. Facial landmark detection with tweaked convolutional neural networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017.

[11] T. Baltrusaitis, P. Robinson, and L.-P. Morency. Openface: an open ˇ source facial behavior analysis toolkit. In Winter Conf. on App. of Comput. Vision, 2016.

[12] A. Zadeh, T. Baltrusaitis, and L.-P. Morency. Convolutional experts ˇ constrained local model for facial landmark detection. In Proc. Conf. Comput. Vision Pattern Recognition Workshops, 2017.

[13] X. Zhu, Z. Lei, X. Liu, H. Shi, and S. Li. Face alignment across large poses: A 3D solution. In Proc. Conf. Comput. Vision Pattern Recognition, Las Vegas, NV, June 2016.

[14] Hough, Paul VC. ”Method and means for recognizing complex patterns.” U.S. Patent 3,069,654, issued December 18, 1962.

[15] Shan, Caifeng, Shaogang Gong, and Peter W. McOwan. ”Facial expression recogni-tion based on local binary patterns: A comprehensive study.” Image and vision Computing 27.6: 803-816, 2009.

[16] Chen, Junkai, Zenghai Chen, Zheru Chi, and Hong Fu. ”Facial expression recognition based on facial components detection and hog features.” In International workshops on electrical and computer engineering subfields, pp. 884-888, 2014.

[17] Whitehill, Jacob, and Christian W. Omlin. ”Haar features for facs au recognition.” In Automatic Face and Gesture Recognition, 2006. FGR 2006. 7th International Conference on, pp. 5-pp. IEEE, 2006.

[18] Edwards, Jane, Henry J. Jackson, and Philippa E. Pattison. ”Emotion recognition via facial expression and affective prosody in schizophrenia: a methodological review.” Clinical psychology review 22.6: 789-832, 2002.

[19] C. Fabian Benitez-Quiroz, R. Srinivasan, and A. M. Martinez. Emotionet: An accu-rate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. In Proc. Conf. Comput. Vision Pattern Recognition, pages 5562–5570, 2016

[20] S. Zafeiriou, A. Papaioannou, I. Kotsia, M. Nicolaou, and G. Zhao. Facial affect“in-the-wild”. In Proc. Conf. Comput. Vision Pattern Recognition Workshops, pages 36–47, 2016.

[21] R. Kosti, J. M. Alvarez, A. Recasens, and A. Lapedriza. Emotion recognition in context. In Proc. Conf. Comput. Vision Pattern Recognition, 2017.

[22] G. Levi and T. Hassner. Emotion recognition in the wild via convolutional neural networks and mapped binary patterns. In Int. Conf. on Multimodal Interaction, pages 503–510. ACM, 2015.

[23] K. Zhang, L. Tan, Z. Li, and Y. Qiao. Gender and smile classification using deep convolutional neural networks. In Proc. Conf. Comput. Vision Pattern Recognition Workshops, pages 34–38, 2016.

[24] M. Naphade, J. Smith, J. Tesic, S.-F. Chang, W. Hsu, L. Kennedy, A. Hauptmann, and Curtis J. Large-scale concept ontology for multimedia. In IEEE Multimedia, 2006.

[25] J.R. Smith, M. Naphade, and A. Natsev. Multimedia semantic indexing using model vectors. In International Conference on Multimedia and Expo, 2003

[26] Genevieve Patterson and James Hays. Sun attribute database: Discovering, annotat-ing, and recognizing scene attributes. In Computer Vision and Pattern Recognition. IEEE, 2012.

[27] Yanwei Fu, Timothy M Hospedales, Tao Xiang, and Shaogang Gong. Attribute learning for understanding unstructured social activity. In European Conference on Computer Vision. Springer, 2012.

[28] Rajat Raina, Alexis Battle, Honglak Lee, Benjamin Packer, and Andrew Y Ng. Self-taught learning: transfer learning from unlabeled data. In Proceedings of the 24th in-ternational conference on Machine learning, pages 759–766. ACM, 2007.

[29] GrégoireMesnil, Yann Dauphin, Xavier Glorot, Salah Rifai, YoshuaBengio, Ian J Goodfellow, Erick Lavoie, Xavier Muller, Guillaume Desjardins, David Warde-Farley, et al. Unsupervised and transfer learning challenge: a deep learning approach. In ICML Unsupervised and Transfer Learning, pages 97–110, 2012.

[30] Lyndon Kennedy and Alexander Hauptmann. Lscom lexicon definitions and annota-tions (version 1.0). 2006.

[31] F. Khan(2018 Dec. 10), “Facial Expression Recognition using Facial Landmark De-tection and Feature Extraction via Neural Networks”(Online), Department of Elec-tronics and Communication Engineering, NIT Karnataka, Mangalore, India, IJACSA, Available: https://www.groundai.com/project/facial-expression-recognition-using-facial-landmark-detection-and-feature-extraction-on-neural- networks/

[32] S. Mishra, G.R.B. Prasada, R.K. Kumar, G. Sanyal (2018 Dec. 10), “Emotion Recognition Through Facial Gestures — A Deep Learning Approach”, Mining In-telligence and Knowledge Exploration” (Online), Available: https://link.springer.com/chapter/10.1007/978-3-319-71928-3_2

[33] R. Walecki, O. Rudovic, V.Pavlovic, B. Schuller, M. Pantic (2018 Dec. 10), “Deep Structured Learning for Facial Action Unit Intensity Estimation”, IJACSA (Online), Available: https://ibug.doc.ic.ac.uk/media/uploads/documents/deep-structured-learning.pdf