International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

An Optimal approach of Emotion Detection Using Deep Learning

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

Submitted: 30/12/2023 **Revised**: 06/02/2024 **Accepted**: 14/02/2024

Abstract*:* Smart facial emotion detection is a fascinating field of research that has been presented and implemented in several fields, including defense, health, and human-machine interfaces. Researchers in this area are focusing on strategies to encrypt, decode and erase facial expressions to enhance algorithm prediction. Various deep learning algorithms and cognitive internet of thing (CIoT) are being used to improve efficiency due to the exponential development of this technology. The aim of this work is to provide a summary of recent work on smart facial expression recognition using deep learning Algorithm and define new approach of emotion detection. Due to the exponential growth of the Internet of Things, current internet of thing-based technologies around automated intelligent services lacks technological resources, meaning that they would be unable to meet the needs of industrial services. The incremental enrichment of internet of thing technology for smart environments has resulted in technology delays and decreased market productivity. Deep learning is one of the most used machine learning techniques in a variety of applications and experiments. Designing an emotional intelligent approach and deep learning that will inspire internet of thing is a pressing need to solve this problem, according to Recent Application in facial Emotion detection.

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

1. Introduction

Emotional intelligence refers to the capacity to perceive and comprehend thoughts and moods. To recognize individual emotions or emotions, gestures, facial expressions, voice, or conversation, writing, and other tools may be used. When we hear the term "emotional intelligence," we immediately associate it with impulses and the communication of emotions. Enthusiastic knowledge, which is a form of understanding, requires the ability to comprehend and control emotions. According [1] to the enthusiastic perspective, "the best of accepting and spreading intelligence, absorbing it into thinking, knowing and dominating it, and possessing the capacity to manage it in yourself and others."

"A small world where different forms of smart technologies are constantly working to make the lives of its people more comfortable," according to the definition of a smart ecosystem. People from all walks of life will benefit from smart workplaces, which seek to replace risky jobs, manual labour, and routine activities with automated agents [2].

Smart environments are broadly classified to have the following features.

1. Computers may be managed from a central location, for example, using power line transmission networks.
2. Merge computer networking, middleware, and wireless connectivity to create a picture of wired worlds.
3. Data collection and distribution from sensor networks
4. Intelligent Systems Improve Service Quality
5. Predictive and decision-making capacity

Deep learning models are particularly useful for dealing with the dimensionality problem because they can focus on the right features on their own, with just a little support from the programmer. When there are a vast number of inputs and outputs, deep learning algorithms are used.

According to the researchers, "the aim of deep learning is to create such an algorithm that can replicate the brain" since deep learning is based on machine learning, which is a branch of artificial intelligence, and the objective of artificial intelligence is to imitate human behaviour. Deep learning is implemented with the help of Neural Networks, and the idea behind the motivation of Neural Network is the biological neurons, which is nothing but a brain cell [3].

A community of artificial neural network-based deep learning computational approaches for learning feature hierarchies is known as deep learning."

Deep learning is accomplished using deep networks, which are just neural networks with several hidden layers.

The Internet of Things (IoT) is the next evolution of development, helped by the internet and smart thins or smartphones. Smart devices are objects with appropriate sensors and actuators, as well as communication technologies that aim to assist people in their daily lives by producing and consuming information, such as lowering costs and increasing optimization in any application field. However, a novel mechanism to improve cognitive ability to control the IoT is still needed. Smart machines and

[1, 2, 3] *Department of Computer Science & Engineering, FET, Rama University, Kanpur 209217, INDIA*
** Corresponding Author: [3]drcsraghuvanshi@gmail.com*

personal computers are not the same thing; they are user-independent, enabling them to adapt to a variety of situations; they are self-sufficient, assisting them in organizing themselves in accordance with their environment and adapting accordingly to incidents occurring all around them. To make the best decisions and adapt to what is going on around them, they'll need logical sensors and actuators [4].

The Internet of Things (IOT) is a large and ubiquitous network that has been widely used around digital intelligent services and has a wide variety of possible applications and opportunities for future intelligent infrastructure. There are already Machine Learning demands, ranging from web page ranking to collaborative scanning to image or voice recognition. Medicine, oil and gas, education, electricity, weather fore-casts, and the stock market are only a few of the fields where machine learning is making huge strides. As a result of these developments, Machine Learning is changing not only technology but also the lives of everyday people. An ordinary person has developed into a computer savvy human, a gadget man, thanks to Machine Learning. Current IOT, on the other hand, is clearly lacking in intelligence, implying that it cannot be used for industrial service application requirements [5]. In comparison, the modern Internet of Things is based on pre-defined architectures and models. It lacks modern intelligence systems, such as emotion detection, and it cannot keep up with industry's increasing productivity demands. By integrating emotional intelligence into a modern paradigm of emotional intelligence, it is possible to create a new model of emotional intelligence using IOT.

The Internet of Things (IOT) with cognitive capacity and an enabling tool that collaborates to optimize efficiency and brainpower is known as Cognitive Internet of Things. CIOT can identify existing network architectures, analyze surface data, make decisions, and perform adaptive tasks to increase network capability. Since the Internet of Things is a massive network of interconnected networks that depend on knowledge exchange, interconnecting, sensing, and information processing technologies, it is difficult to introduce.

Given the above, emotional intelligence is clearly needed for IOT-related structures to understand human affective states or behaviour based on six physiological signals: optimistic, negative, disgust, sadness, tension, and annoyance [6]. Emotional intelligence depicts daunting activities such as collecting a vast volume of data from a subject to inspire and observe one of eight enthusiastic phases of human development over a period of several weeks.

2. Review of Literature

Paul Ekman's [7] work on emotion recognition, where he identified six principal emotions (happiness, sadness, anger, surprise, fear, and disgust) and later developed FACS. Neutral was later included in most human recognition datasets, resulting in seven basic emotions.

Researchers discusses two categories of landmark detectors: regression based [8], [9], [10] and model-based techniques [11], [12], [13]. Regression based methods estimate landmark locations from facial appearance, while model-based methods model both the shape and appearance of landmarks. However, landmark estimation can fail under certain conditions such as extreme out-of-plane rotations, low scale, or significant differences in face bounding boxes.

Previous emotion recognition techniques used a two-step machine learning approach involving feature extraction and classification using methods such as SVM and neural networks. Hand-crafted features like HOG [14] [15], LBP [16], Gabor wavelets [17], and Haar features [18] were popular for facial expression recognition. However, these techniques showed limitations with more challenging datasets containing intra-class variation and difficult image conditions such as partial faces or occlusion.

Some researchers discuss the difference between emotion classification and expression regression tasks. Emotion classification involves classifying images or videos into discrete sets of facial emotion classes or action units [19] [20], often using facial landmarks. Recently, deep networks [21] [22] [23] have been used instead of estimating landmark positions.

Concept modeling in multimedia [24], [25] and computer vision, with a focus on modeling objects [25], scenes [26], and activities [27]. There is also work on modeling non-conventional concepts such as image aesthetic and emotions. The models are trained using SVM and other methods.

The transfer of learned deep representations across tasks has been studied extensively in an unsupervised setting [28], [29]. However, these models have been limited to small datasets and have only achieved modest success. To address the problem of insufficient training data, unsupervised pre-training followed by supervised fine-tuning and supervised pre-training using a concept-bank paradigm [30] have been proposed and proven successful in computer vision and multimedia settings. Recently, supervised pre-training followed by domain-adaptive fine-tuning has been shown to be an efficient paradigm for scarce data.

Previous works have improved emotion recognition but lacks a way to focus on important face regions for detection. In this work, we propose approach using an attentional convolutional network to address this problem.

3. Methodology

The aim of our project is to use CNN to investigate real-time facial expressions. After a large amount of experience, this is a particular form of deep learning methodology that provides us with the solution to several problems in facial expression detection. The most significant benefit of CNN is that it allows for "end-to-end" learning directly from the input source, eliminating, or greatly reducing the need for physics-based simulations and/or other pre-processing techniques [31].

Automatic facial emotional status analysis and automatic identification are particularly useful for identifying, evaluating, and maintaining healthy fragile individuals such as patients with psychiatric illnesses, people with vital mental weight, and children with little self-control.

We found it difficult to differentiate between the emotions fear, surprise, and disgust while using the Fer2013 dataset for our project, so we grouped them together as surprise in the 5 emotions we considered, which also included happy, sad, rage, and neutral [32]. The Fer2013 dataset was created by Kaggle and reflects real-life, spontaneous facial expressions created under several difficult conditions, such as changing lighting, head movements, and differences in facial features due to race, age, gender, facial hair, and glasses.

Deep learning-based facial expression recognition (FER) techniques reduce de-pendency on face-physics-based models and other pre-processing techniques by allowing lengthwise learning to occur directly from the input images in the pipeline [33]. The phases of the process are as follows:

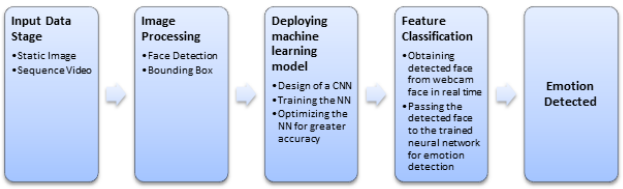

Fig. 1. Step for Emotion Detection

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].

Table 1. Descriptions of facial muscles involved in the emotions Darwin considered universal.

Emotion	Fear	Anger	Disgust	Contempt
Darwin's Facial Descrip-tion	• Eyes open • Mouth open • Lips re-tracted • Eye-brows raised	• Eyes wide open • Mouth com-pressed • 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 Descrip-tion	• Eyes sparkle • Mouth drawn back at corners • Skin under eyes wrin-kled	• Eyes open • Mouth open • Eye-brows raised • Lips pro-truded	• Corner of mouth de-pressed • Inner corner of eye-brows raised	• Upper lip raised • Nose labial fold formed • Orbicularis • Zygomatic

3.1 Data Preprocessing

When the pictures are neither rotated nor flipped, the model performs better at predicting emotion. With the Kaggle dataset, they are all doable. The dataset may be obtained and saved in a NumPy array with the following dimensions: Samples * (Number * 48 * 48 * 48 * 1) every image is a variation of (48 * 48 * 1) in some way. For ease of usage, we retain it in the array, but you may add it to folders for data augmentation by using the Image Data Generator class from keras preprocessing image.

3.2 Model Architecture & Working

Due to the use of several CNN designs, the CNN (Convolutional Neural Networks) model architecture provides exceptional accuracy. The same architecture is used to split concepts into four groups, and the models are combined to provide five categories for emotions.

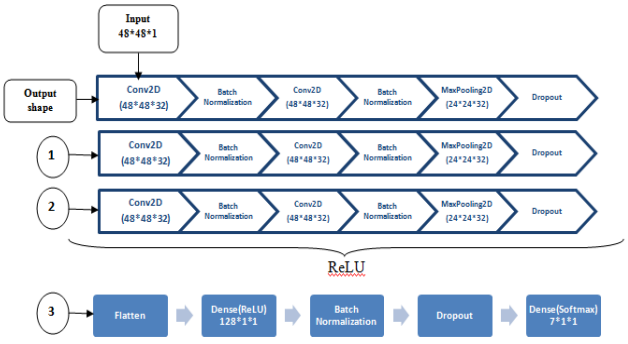

Fig. 2. CNN Model Architecture

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 mod-el 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 minimized 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 to-wards 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 improve 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.

Table 2. 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	66.4 5	66.6 5	64.5 5	65.2 8	64.8 0

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.

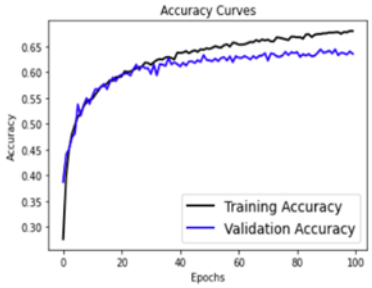

Fig. 3. Accuracy Vs Epochs graph for CNN model

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, sur-prised, furious, and neutral). This method of combining different CNNs produces top-notch results while using the same CNN architecture.

5. Conclusion

These results are favorable. The device's great precision and quick reaction time make it suited for most practical applications. When faces are initially extracted using the OpenCV module and then emotion detection is performed, the system performs better for types of data. In this work, CNN model applies three blocks to the input image using the Conv2D, Batch Normalization, MaxPooling2D, and

Dropout layers. Pooling layers minimized the size of a picture in half by only using the most important components from each block which help to reduced size just for better visualization without change the image quality. For seven-class emotion recognition and for four-class emotion recognition, respectively, the accuracy of suggested CNN architectures is 65.58% and 74.58%. With the ensemble approach, the models accuracy greatly increased, providing 74.3% accuracy and a 0.753 F1-score for five class classifications. In future we add more emotion and emotion parameter to better result and prediction and work on other most accurate datasets for improve the accuracy.

References

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

[2] M. Alhussein, G. Muhammad, M.S. Hossain, S.U. Amin, "Cognitive IoT-cloud integra-tion for smart healthcare: case study for epileptic seizure detection and monitoring", 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 com-puting", 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. De-mestichas, 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 recognition 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 Auto-matic 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 psy-chology review 22.6: 789-832, 2002.

[19] C. Fabian Benitez-Quiroz, R. Srinivasan, and A. M. Martinez. Emotionet: An accurate, 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 convolu-tional 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 vec-tors. In International Conference on Multimedia and Expo, 2003

[26] Genevieve Patterson and James Hays. Sun attribute database: Discovering, annotating, 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 international confer-ence on Machine learning, pages 759–766. ACM, 2007.

[29] GrégoireMesnil, Yann Dauphin, Xavier Glorot, Salah Rifai, YoshuaBengio, Ian J Goodfel-low, Erick Lavoie, Xavier Muller, Guillaume Desjardins, David Warde-Farley, et al. Unsuper-vised 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 annotations (version 1.0). 2006.

[31] F. Khan(2018 Dec. 10), "Facial Expression Recognition using Facial Landmark Detection and Feature Extraction via Neural Networks"(Online), Department of Electronics and Commu-nication 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 Intelligence 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 Struc-tured Learning for Facial Action Unit Intensity Estimation", IJACSA (Online), Available: https://ibug.doc.ic.ac.uk/media/uploads/documents/deep-structured-learning.pdf