engineering proceedings

MDPI

Proceeding Paper

Facial Expression Recognition Using Pre-trained Architectures †

Resmi K. Reghunathan [1,*], Vineetha K. Ramankutty [1], Amrutha Kallingal [1] and Vishnu Vinod [2]

[1] Department of Computer Science, CHRIST University, Bangalore 560029, India; vineetha.kr@christuniversity.in (V.K.R.); amrutha.k@christuniversity.in (A.K.)
[2] Software Engineer, Idea Elan India Pvt. Ltd., Hyderabad 500082, India; vishnuvinod277@gmail.com
* Correspondence: resmi.kr@christuniversity.in
† Presented at the 2nd Computing Congress 2023, Chennai, India, 28–29 December 2023.

Abstract: In the area of computer vision, one of the most difficult and challenging tasks is facial emotion recognition. Facial expression recognition (FER) stands out as a pivotal focus within computer vision research, with applications in various domains such as emotion analysis, mental health assessment, and human–computer interaction. In this study, we explore the effectiveness of ensemble methods that combine pre-trained deep learning architectures, specifically AlexNet, ResNet50, and Inception V3, to enhance FER performance on the FER2013 dataset. The results from this study offer insights into the potential advantages of ensemble-based approaches for FER, demonstrating that combining pre-trained architectures can yield superior recognition outcomes.

Keywords: pre-trained architectures; facial expression recognition; FER2013; CNN

1. Introduction

Facial expression recognition (FER) is gaining more importance in computer vision these days. It is used for the analysis and classification of a given facial expression. FER can be used in fields like robotics, security, driving assistance, mental health disorder prediction, lie detectors, etc. [1,2]. With the progress of deep learning, FER technology achieves a remarkable increase in recognition accuracy compared to traditional methods.

Because of its capacity to extract image information, Convolutional Neural Networks (CNNs) have been widely employed for image classification tasks, particularly in FER. However, there may be certain difficulties in training a CNN model for FER.

Overfitting on uncertain inputs, for example, may result in mislabeled outputs. Furthermore, a high percentage of inaccurate labels during the early stages of optimization can prevent the model from converging. Transfer learning is one of the deep learning methods that has gained much attention in the field. It employs a pre-trained CNN to solve an issue that is similar to one that the CNN was trained to solve in the first place. Pre-trained models are commonly utilized in FER research [3–5].

2. Related Works

In the last decade, researchers have turned to deep learning instead of machine learning because of its great automatic recognition capability. This section describes several known studies in FER utilizing deep learning.

Yolcu et al. [3] proposed the detection of essential parts of the face using three CNNs of the same architecture to detect the eyebrow, mouth, and eye. They developed a system for monitoring neurological disorders using facial expressions. The experiment conducted on the RafD dataset reports an accuracy of 94.44%. Li et al. [4] introduced a new CNN for facial occlusion problems. VGGNet network and a CNN called ACNN are used to train two databases, AffectNet and RAF-DB. A recognition accuracy of 80.54% and 54.84% is reported on RAF-DB and AffectNet, respectively.

Citation: Reghunathan, R.K.; Ramankutty, V.K.; Kallingal, A.; Vinod, V. Facial Expression Recognition Using Pre-trained Architectures. *Eng. Proc.* **2024**, *62*, 22. https://doi.org/10.3390/engproc2024062022

Academic Editors: Geetha Ganesan, Xiaochun Cheng and Valentina Emilia Balas

Published: 22 April 2024

Zahara et al. [5] used CNN based on Raspberry Pi for facial expression recognition. The experiment reported a recognition accuracy of 65.97% on the FEI 2013 dataset. It includes face detection, feature extraction, and face emotion recognition. In [6], pre-trained architectures like VGG, inception, and ResNet are experimented with to find emotion recognition on the FER2013 dataset. A maximum accuracy of 75.2% was reported on ensembles of modern deep CNNs. Pre-trained architectures GoogleNet and AlexNet were examined for emotion recognition in [7]. GoogleNet reported a maximum accuracy of 65.20% on the FER2013 dataset. Liu et al. [8] used several subnets, each of which is a compact CNN model for emotion recognition. A single subnet achieved an accuracy of 62.44% of the FER2013 dataset.

An ensemble-based approach using AlexNet, ResNet, and VGG16 on the FER2013 dataset is employed in [9] to achieve an accuracy of 71.27%. SVM is used for classification. Facial expression recognition using deep architectures is presented in [10,11].

3. Pre-trained Architectures

AlexNet, a neural network architecture in the field of deep learning, marked a significant milestone in the advancement of computer vision tasks, particularly image classification. AlexNet's architecture consisted of eight layers, with five convolutional layers followed by three fully connected layers.

Google's Inception v3, a convolutional neural network (CNN), is an example of a sophisticated architecture created to tackle image classification problems. The inception modules form the foundation of the architecture of Inception v3. By using alternative kernel sizes for concurrent convolutional operations, these modules enable the network to simultaneously record features at various scales. Through the integration of multiple parallel pathways, Inception v3 improves its recognition of complex structures and patterns in images.

The key innovation of ResNet50 lies in its approach to deep learning through residual connections. Unlike traditional architectures, ResNet introduces skip connections or shortcuts, allowing the network to learn residual functions. This alleviates the vanishing gradient problem and facilitates the training of extremely deep networks by enabling the direct flow of information through the network.

4. Methodology

This study uses the concept of transfer learning. In transfer learning, a CNN trained for one application can be reused for another application. Building deep learning models from the bottom requires high resources and a large amount of data. This can be minimized by using the concept of transfer learning. CNN is tested in three ways: First, features extracted from all pre-trained models by removing the fully connected classification layers are fed to SVM for classification. In the second approach, all the other layers in pre-trained networks are frozen, and only the SoftMax layer is changed to 7. In the third approach, an ensemble-based approach using model averaging is used for prediction. The proposed methodology adopted in this work is shown in Figure 1.

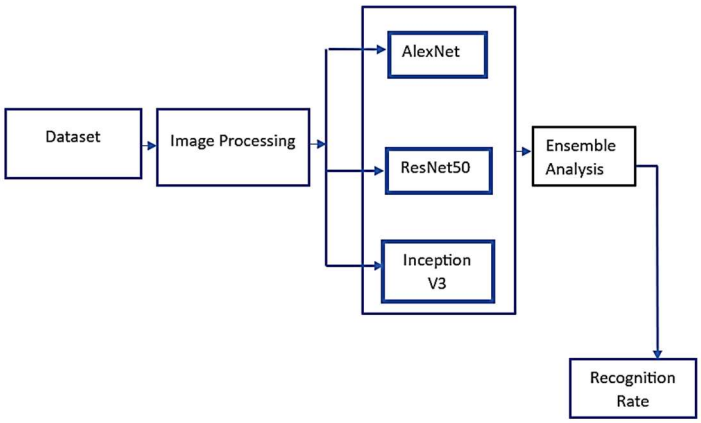

Figure 1. Proposed methodology.

5. Experiments

The model is made in Python, leveraging the deep learning framework TensorFlow, with Keras as a pivotal element. The experimental arrangement is conducted on hardware equipped with an Intel Core i5 processor and 8 GB of RAM.

5.1. Dataset

FER2013 [Facial Expression Recognition 2013], a publicly available facial expression dataset, is used to train the model [12]. The database exclusively comprises grayscale images, all of which have been size-normalized to dimensions of 48 × 48 pixels. This dataset encompasses around 30,000 images, each associated with one of seven distinct emotions: happiness, anger, surprise, fear, disgust, sadness, and neutrality. The images include both posed and unposed headshots. Sample images from the FER2013 database are shown in Figure 2.

Figure 2. Sample images from FER2013.

5.2. Face Detection Using Haar Cascade

Face detection using a Haar cascade is an efficient method proposed by [13] for detecting human faces from real-time video or images. It is an approach where a cascade function is used to train a large number of positive and negative images, and it is then used to detect images later. The algorithm uses features detected from the edge or line proposed by [13]. Face detection is first performed from live video using the Haar cascade for validation work. From the detected face, expressions are evaluated using the trained model.

6. Results and Discussion

Due to the small and unbalanced nature of the FER2013 dataset, the application of transfer learning can enhance the model's accuracy. Pre-trained models based on transfer learning were explored. Initially, a preprocessing step is imple×mented to resize all images in the FER2013 dataset to color images.

Each of the three models has undergone distinct rescaling procedures owing to variations in their required input image sizes. The AlexNet, ResNet50, and InceptionV3 models

necessitate input images sized at 227 × 227, 224 × 224, and 139 × 139 (Figure 3), respectively. To accommodate the design specifications of models that expect images with three input channels, a grayscale image comprising only one channel is expanded by duplicating the grayscale information across the remaining two channels. This modification ensures that the image format aligns with the requirements of all three models. Following this, zero-mean normalization is applied to standardize inputs for each mini-batch, promoting a stable learning process in subsequent layers. Across all models, the stochastic gradient descent (SGD) optimizer was utilized, with a learning rate set at 0.0001. Batch sizes were defined as 128, and the training process extended to 100 epochs.

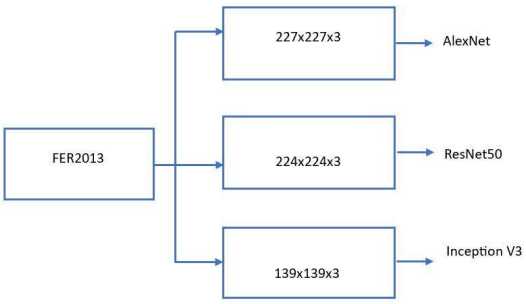

Figure 3. Preprocessing.

Table 1 shows Rank1 accuracy using different models using data with and without augmentation.

Table 1. Accuracy using different methods.

	Rank-1 Accuracy [%]	
Model	**Without Data Augmentation**	**With Data Augmentation**
AlexNet	71.30%	71.76%
Resnet50	72.45%	72.89%
Inception V3	72.23%	72.67%
Ensemble-based	73.21%	73.56%

Figure 4 shows the output of expression recognition using sample inputs, and Figure 5 shows expression recognition using live video.

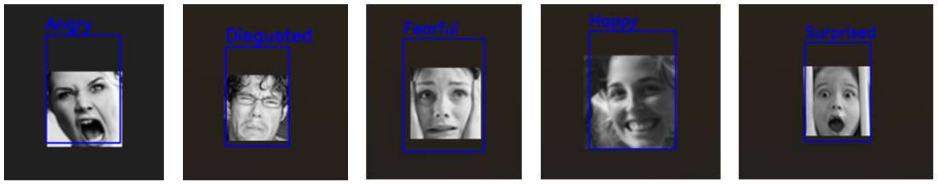

Figure 4. Expression recognition on sample input images.

Figure 5. (**a**–**d**) Expression recognition using live video.

Extensive research in deep learning was conducted using the FER2013 dataset, and Table 2 compares the current study and previous research on the same dataset.

Table 2. Comparison of accuracy using the FER2013 dataset with existing methods.

Method	Accuracy Rate
CNN [8]	62.44%
GoogleNet [7]	65.20%
CNN based on Raspberry Pi [5]	65.97%
Inception [6]	71.60%
ResNet 18 [6]	72.40%
VGG [6]	72.70%
AlexNet + VGG16 + ResNet + SVM (Ensemble) [9]	71.27%
CNN + ResNet50 + Inception V3 (Ensemble) [14]	72.3%
AlexNet + Inception V3 + ResNet50 (Ensemble)	73.56%

7. Conclusions

Facial expression recognition is still a challenging task in real-time scenarios. This paper presents a study on FER using ensemble-based pre-trained architectures. FER2013, the publicly available database, was used for experimentation. The current study is limited to the FER2013 dataset. Several other datasets with a large number of samples are also available. This study should be expanded by incorporating additional datasets and utilizing advanced deep-learning architectures.

Author Contributions: Conceptualization, R.K.R. and V.V.; methodology, A.K.; software, V.V.; validation, R.K.R., A.K. and V.K.R.; formal analysis, R.K.R.; investigation, V.V.; resources, R.K.R. and A.K.; data curation, V.K.R.; writing—original draft preparation, R.K.R.; writing—review and editing, A.K.; visualization, V.V.; supervision, R.K.R. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: Data are contained within the article.

Conflicts of Interest: Author Vishnu Vinod was employed by the company Idea Elan India Pvt. Ltd. The authors declared no potential conflicts of interest with respect to the research, authorship, and publication of this article.

References

1. Zeng, D.; Lin, Z.; Yan, X.; Liu, Y.; Wang, F.; Tang, B. Face2exp: Combating data biases for facial expression recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 20291–20300.
2. Li, S.; Deng, W. Deep facial expression recognition: A survey. *arXiv* **2018**, arXiv:1804.08348. [CrossRef]
3. Yolcu, G.; Oztel, I.; Kazan, S.; Oz, C.; Palaniappan, K.; Lever, T.E.; Bunyak, F. Facial expression recognition for monitoring neurological disorders based on convolutional neural network. *Multimed. Tools Appl.* **2019**, *78*, 31581–31603. [CrossRef] [PubMed]
4. Li, Y.; Zeng, J.; Shan, S.; Chen, X. Occlusion Aware Facial Expression Recognition Using CNN with Attention Mechanism. *IEEE Trans. Image Process.* **2019**, *28*, 2439–2450. [CrossRef] [PubMed]
5. Zahara, L.; Musa, P.; Wibowo, E.P.; Karim, I.; Musa, S.B. The Facial Emotion Recognition (FER-2013) Dataset for Prediction System of MicroExpressions Face Using the Convolutional Neural Network (CNN) Algorithm based Raspberry Pi. In Proceedings of the 2020 Fifth International Conference on Informatics and Computing (ICIC), Gorontalo, Indonesia, 3–4 November 2020; pp. 1–9. [CrossRef]
6. Pramerdorfer, C.; Kampel, M. Facial Expression Recognition using Convolutional Neural Networks: State of the Art. *arXiv* **2016**, arXiv:1612.02903.
7. Giannopoulos, P.; Perikos, I.; Hatzilygeroudis, I. Deep Learning Approaches for Facial Emotion Recognition: A Case Study on FER-2013. In *Advances in Hybridization of Intelligent Methods. Smart Innovation, Systems and Technologies*; Hatzilygeroudis, I., Palade, V., Eds.; Springer: Cham, Switzerland, 2018; Volume 85.
8. Liu, K.; Zhang, M.; Pan, Z. Facial Expression Recognition with CNN Ensemble. In Proceedings of the 2016 International Conference on Cyberworlds (CW), Chongqing, China, 28–30 September 2016; pp. 163–166. [CrossRef]

9. Jia, C.; Li, C.L.; Ying, Z. Facial expression recognition based on the ensemble learning of CNNs. In Proceedings of the 2020 IEEE International Conference on Signal Processing, Communications and Computing (ICSPCC), Macau, China, 21–24 August 2020; pp. 1–5. [CrossRef]
10. Li, S.; Xu, Y.; Wu, H.; Wu, D.; Yin, Y.; Cao, J.; Ding, J. Facial Expression Recognition In-the-Wild with Deep Pre-trained Models. In *Computer Vision–ECCV 2022 Workshops. ECCV 2022*; Lecture Notes in Computer, Science; Karlinsky, L., Michaeli, T., Nishino, K., Eds.; Springer: Cham, Switzerland, 2023; Volume 13806. [CrossRef]
11. Canal, F.Z.; Müller, T.R.; Matias, J.C.; Scotton, G.G.; de Sa Junior, A.R.; Pozzebon, E.; Sobieranski, A.C. A survey on facial emotion recognition techniques: A state-of-the-art literature review. *Inf. Sci.* **2022**, *582*, 593–617. [CrossRef]
12. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, June 26–July 1 2016. [CrossRef]
13. Viola, P.; Jones, M. Rapid object detection using a boosted cascade of simple features. In Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001, Kauai, HI, USA, 8–14 December 2001; pp. 1–9. [CrossRef]
14. Moung, E.G.; Wooi, C.C.; Sufian, M.M.; On, C.K.; Dargham, J.A. Ensemble-based face expression recognition approach for image sentiment analysis. *Int. J. Electr. Comput. Eng. (IJECE)* **2022**, *12*, 2588. [CrossRef]