Vidhyayana - ISSN 2454-8596
An International Multidisciplinary Peer-Reviewed E-Journal
www.j.vidhyayanaejournal.org
Indexed in: ROAD & Google Scholar

Convolutional Neural Networks for Facial Expression Recognition

Jaimin Jani
Research Scholar
Department of Computer Applications,
Sabarmati University, Ahmedabad

Prof. Subhashchandra Desai
Director, School of Pure and Applied Sciences
Department of Computer Applications,
Sabarmati University, Ahmedabad

Vidhyayana - ISSN 2454-8596
An International Multidisciplinary Peer-Reviewed E-Journal
www.j.vidhyayanaejournal.org
Indexed in: ROAD & Google Scholar

Abstract:

Human facial expressions are a kind of communication that are frequently utilised to convey emotions. People are paying more attention to facial expression recognition (FER) technology as human-computer interface technology advances. Furthermore, humans have made some progress in the field of FER. We looked at the evolution of FER in this research, including VGGNet, ResNet, GoogleNet, and AlexNet. In addition, we looked at various CNN (Convolutional Neural Network) concepts, and we chose FER2013 as the dataset to consider. FER2013 is one of the most significant databases of human faces. We also made several improvements based on the original FER methodology. The best accuracy value we got by training the FER2013 dataset in various revised techniques was 0.6424. Finally, we generated and summarised the study's progress and shortcomings. Facial Expression

Keywords: Recognition, CNN; FER2013; VGGNet, ResNet, GoogleNet, and AlexNet

Vidhyayana - ISSN 2454-8596
An International Multidisciplinary Peer-Reviewed E-Journal
www.j.vidhyayanaejournal.org
Indexed in: ROAD & Google Scholar

Introduction:

Faces are considerably more than identifiers for individuals. Furthermore, people's facial expressions are the most direct manner of expressing their intuitive intuition. The main goal of Facial Expression Recognition (FER) is to categorise facial expressions into several categories, such as joy, fear, sadness, contempt, anger, surprise, and so on. Because it is difficult for humans to identify the subtleties of others' facial expressions, the machine becomes an important tool for capturing individual facial emotion during face-to-face interactions. Because facial expressions are so important in human connection, the ability to do FER automatically via computer vision opens up a whole new world of possibilities in fields like human-computer interaction and data analytics [1]. Face detection technology has been employed in a variety of areas, including medical, e-learning, monitoring, entertainment, law, and marketing, because to its high accuracy. A face recognizer, for example, might be deployed at the front entry of a building for automatic access control. They could be used to improve user authentication security in ATMs by identifying faces rather than requiring passwords [3]. The basic principle behind face detection is to locate the faces in a picture using bounding boxes, as shown in Figure 1.

Vidhyayana - ISSN 2454-8596
An International Multidisciplinary Peer-Reviewed E-Journal
www.j.vidhyayanaejournal.org
Indexed in: ROAD & Google Scholar

Figure 1 shows the detection of a human face. All of the persons in this photograph are denoted by a red bounding box. The image above comes from the FER2013 dataset.

Face detection is followed by face recognition, which is an identifying system based on human facial information. Face recognition is made up of various steps: face picture detection, preprocessing, extraction of facial features, and matching and identification of face images. The system must always input a sequence of face photos with unknown identities, then output a set of similarity scores indicating the human's identity. For example, because humans cannot remember all of the faces of criminals, machines have a big impact on face identification when it comes to recognising those who are on the criminal blacklist. Many algorithms can tackle face detection, which is the first stage in facial emotion identification. The first type uses statistical approaches to turn two-dimensional human faces into one-dimensional characteristics.

This category includes eigenfaces and adaboosting. Another option is to extract facial features and send them to a classifier to recognise facial expressions. Color-based skin

detection is a good example. The focus of this paper is on face expression recognition. Feature extraction can be divided into two categories: feature Permission is granted without charge to make digital or hard copies of all or part of this work for personal or classroom use, provided that copies are not created or disseminated for profit or commercial gain, and that copies bear this notice and the whole citation on the first page. Copyrights for parts of this work that are not owned by ACM must be respected. It is permissible to abstract with credit. Otherwise, you'll need to get permission and/or pay a price to copy, republish, put on servers, or redistribute to mailing lists.

Figure 2 shows the extraction of a geometric feature and a method based on overall statistical features.

Although facial recognition is no longer a technical barrier, computers are still unable to recognise human expression rapidly. Computers that can understand human emotion will be able to provide greater service to people. Picture acquisition, image preprocessing, feature extraction, and classification are the primary phases. In this research, we assume that the first two processes have been achieved, and we will concentrate on face expression identification using CNN (Convolutional Neural Network).

Vidhyayana - ISSN 2454-8596
An International Multidisciplinary Peer-Reviewed E-Journal
www.j.vidhyayanaejournal.org
Indexed in: ROAD & Google Scholar

Figure 2: shows the Mona Lisa. On the left side, a bounding box denotes the woman's face, as well as the five sense organs. It can be recognised using the geometrical characteristics method. Right side: the woman's face is marked by a bounding box, which can be identified by the procedure based on general statistical features. We can use facial expression recognition to try to solve a problem that has perplexed humans for thousands of years: Is the woman in this painting a happy person?

Furthermore, in the last year, CNN and DNN (Deep Convolutional Neural Network), which are feature-based approaches for facial recognition, have become popular algorithms. A CNN model has a convolutional layer, a pooling layer, a fully connected layer, and an output layer. In CNN, nonlinear activation functions such as Sigmoid, Tanh, and others are used. It has an input layer, an output layer, and numerous hidden layers in the DNN model.

With the advent of face recognition in recent years, more datasets have been created for users to utilise, such as Kaggle, ORL, and FERET. Kaggle is a website that allows developers and scientists to hold machine learning competitions and share code. The Kaggle dataset will be the subject of this paper.

We concentrated our efforts in this paper on using CNN to address the FER problem. To recognise face expressions, we used multiple architectures such as VGG16, ResNet, and GoogLeNet. The following is a breakdown of the paper's structure. First, Section 2 provides an overview of relevant works. Section 3 then introduces the precise procedures. Section 4 contains the exact experimental results. Finally, in Section 5, we came to some conclusions.

2. CONNECTED WORK

Because the accuracy of facial expression will affect the results of the classification of facial expression, which is the next step in the extraction of facial expression, it is necessary to accurately extract the features of facial expression in the image of human expression during the process of facial expression feature extraction.

Despite the fact that there have been several studies on the history and evolution of face recognition, we have discovered that there are several ways that may be employed in

Vidhyayana - ISSN 2454-8596
An International Multidisciplinary Peer-Reviewed E-Journal
www.j.vidhyayanaejournal.org
Indexed in: ROAD & Google Scholar

face recognition, including face recognition based on geometrical feature points extraction and eigenface.

The geometrical features-based method is commonly used to extract the placement of facial organs as classification characteristics. The goal, according to Roberto Brunelli and Tomaso Poggio, is to get relative location and other data of distinguishing features like eyes, mouth, nose, and chin [8]. This function is old and boring, but there are still two major flaws with geometrical features-based functions: the first is that the weighting coefficients in the energy function are difficult to summarise and can only be discovered through experience. Another downside is that the process of optimising the energy function takes a long period. On the other hand, feature point identification technology is not precise, and processing is time-consuming.

Eigenface, which was first developed in 1991 by Matthew Turk and Alex Pentland, has become one of the most widely used algorithms in recent years. It is well-known for being both simple and effective. To extract the information contained in a face image, a straightforward way is to record the variation in a group of face photographs, independent of any evaluation of features, and then encode and compare individual face images using this information [5].

To put it another way, eigenface's core principle is to transform a face image from pixel space to another space, then calculate similarity in that space. To obtain the composition of human faces, the eigenface employs the PCA function. Otherwise, we want to get the eigenvalue decomposition of the covariance matrix of human face photos and the accompanying eigenvectors from the training set. Feature faces refer to all of the eigenvectors that we obtain during computation.

CNN is a deep learning method for picture categorization and recognition that was created by Google. In addition, it is currently frequently employed in the field of FER. CNN, as a deep learning architecture, can reduce model complexity and extract picture data accurately [7].

Vidhyayana - ISSN 2454-8596
An International Multidisciplinary Peer-Reviewed E-Journal
www.j.vidhyayanaejournal.org
Indexed in: ROAD & Google Scholar

3. TECHNIQUES

As shown in Figure 3, the original structure has six steps: input image, training data, template library, feature extraction, comparison, and output result. However, after combining the procedures of template library, feature extraction, and comparison to facial expression recognition, the reduced structure used in this research only contains four steps, as shown in Figure 4. It will considerably improve efficiency and cut down on operating time.

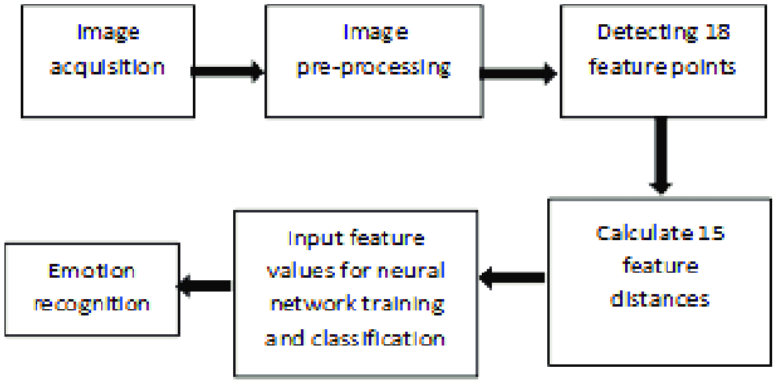

Figure 3: The original face expression recognition methodology. Figure 4 shows the face expression identification approach we took in this paper. The original procedure was simplified.

There are only four steps in this flow diagram: Face expression recognition – input image – training data – output result

There have been several outstanding CNN networks in the growth of CNN network structure, including AlexNet, Vgg, GoogLenet, Resnet, and Densenet.

3.1AlexNet

Input	Conv	Conv	Pool	Conv	Pool	FC	FC	Soft-Max

Figure 5. The structure of Alexnet

Vidhyayana - ISSN 2454-8596
An International Multidisciplinary Peer-Reviewed E-Journal
www.j.vidhyayanaejournal.org
Indexed in: ROAD & Google Scholar

Alexnet was trained using ImageNet, a dataset with more than 1.2 million images for classification [9], as illustrated in Figure 5. AlexNet features five convolutional layers and three fully linked layers, demonstrating CNN's utility in a complicated model. It will also output the 1000*1 vector and transfer the 1000-class SoftMax classifier, after which the classification results will be obtained. AlexNet differs from the first traditional technique, LeNet, in that it has certain new features. AlexNet, for example, uses ReLu as an activation function, and ReLu is now commonly utilised in CNN topologies. AlexNet has not only established deep learning's dominance in computer science, but has also aided deep learning's advancement in the field of face expression recognition. We changed the SoftMax layer in our work to produce seven classes.

3.2VGG-net

Figure 6. The VGG-net structure.

As illustrated in Figure 6, the last three levels of the full connection layer have the same structure as a series of VGG-nets, and the total structure has five sets of convolutional layers, followed by the Max pool layer. Based on prior architectures, VGG-net has improved. Throughout the convolutional layers of the VGG-net, 3*3 filters are used.

A ReLU activation function was applied to each convolutional layer [2]. VGG-net trains the simple system of level A initially, then reuses the weight of that network's network to initialise numerous complicated models later, accelerating convergence. Third, the convolutional kernel of 1*1 is introduced to minimise computation in the VGG-convolutional net's structure.

Vidhyayana - ISSN 2454-8596
An International Multidisciplinary Peer-Reviewed E-Journal
www.j.vidhyayanaejournal.org
Indexed in: ROAD & Google Scholar

3.3 GoogleNet

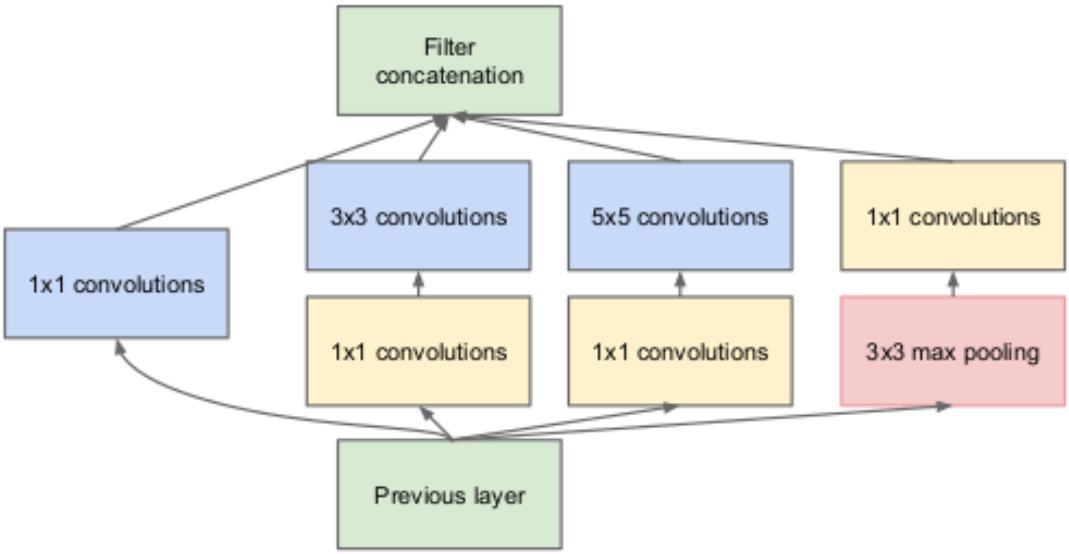

Figure 7: The Inception Structure of GoogleNet.
GoogleNetis compositing by many inception structures

One notable feature of GoogLeNet is that it is constructed very deeply, as demonstrated in Figure 7. GoogLeNet's optimization approaches are deserving of additional investigation.

GoogLeNet, for example, has taken a modular approach to standardising the results, making it easy to alter.

Furthermore, average pooling was utilised to replace the entire connectional layer at the end, resulting in a 0.6 increase in the success rate.

Vidhyayana - ISSN 2454-8596
An International Multidisciplinary Peer-Reviewed E-Journal
www.j.vidhyayanaejournal.org
Indexed in: ROAD & Google Scholar

3.4 Resnet

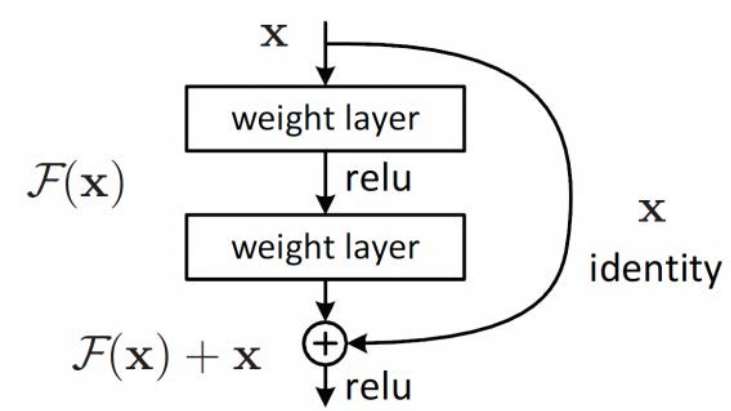

Figure 8:The Structure of Resnet

ResNet was proposed in 2015 and took first place in the ImageNet competition. In the computer vision world, ResNet with hundreds or even thousands of layers has become the most successful image recognition model [4].

Figure eight. ResNet as a thumbnail. ResNet's structure is depicted in Figure 8. The basic idea behind this function is that the input data will first pass through a layer with a tiny output of 1*1, then to a layer of 3*3, and finally to a layer of 1*1 to handle a larger number of features. Furthermore, as compared to traditional CNNs like VGG, ResNet requires less parameters. Furthermore, ResNet can be more detailed without the issue of gradient dispersion. At the time, ResNet was seen to be a significant improvement. In the next articles, this method will be built and proven.

4. RESULTS OF EXPERIMENTATION

4.1Dataset

The dataset we're working with is named Emotion Recognition Dataset (FER2013) and it's available on Kaggle. It's a human expressions dataset that includes 35887 distinct face photos with expressions that fall into seven categories. International Conference on Machine Learning released FER2013.

Vidhyayana - ISSN 2454-8596
An International Multidisciplinary Peer-Reviewed E-Journal
www.j.vidhyayanaejournal.org
Indexed in: ROAD & Google Scholar

This dataset is separated into three categories: private test, public test, and training, with a total of 28709 training datasets.

To train the model, we employed various approaches such as AlexNet, GoogLeNet, VGGNet, and ResNet, and the results are displayed in the next part.

4.2 Results

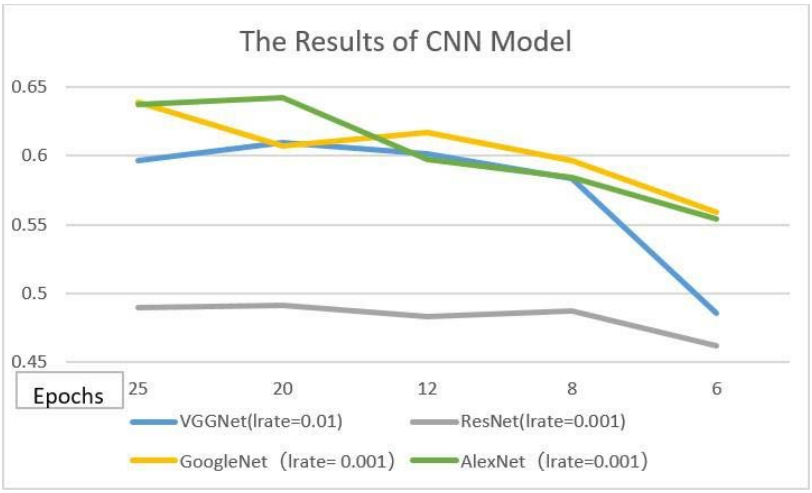

Figure 9 shows the CNN model's output. We created a curve chart using the upgraded CNN models (VGGNet, ResNet, GoogleNet, and AlexNet) so that the changes in precisions could be shown.

Figure 9 shows the five epochs we chose: 25, 20, 12, 8, and 6. Yichuan Tang [6] computed the top accuracy result on FER2013 in machine learning contests, which is 0.7116. Except for ResNet's accuracy, the average of the other approaches is above 0.55 to 0.6, indicating that our methods produce good outcomes. Furthermore, the accuracy of VGGNet and AlexNet increases from 25 to 20, with the accuracy peaking when the epochs equal 20. One of the causes could be overfitting. In other words, the number of epochs alone cannot predict the final accuracy. Furthermore, ResNet's overall results are not promising. ResNet can reach state-of-the-art performance on many additional tasks, while being mostly a collection of shallow networks. Furthermore, it is likely that we lack sufficient data to train a decent ResNet model. We can produce some data in the future by flipping and cutting the same image for FER to expand the dataset size.

Vidhyayana - ISSN 2454-8596
An International Multidisciplinary Peer-Reviewed E-Journal
www.j.vidhyayanaejournal.org
Indexed in: ROAD & Google Scholar

Table 1: Accuracy findings utilising upgraded models.

Epochs VGG	Epochs VGG	Epochs VGG	Epochs VGG	Epochs VGG
25	0.5964	0.49	0.6391	0.6374
20	0.6098	0.4916	0.6073	0.6424
12	0.6017	0.4833	0.6171	0.5978
8	0.5836	0.4869	0.5966	0.5844
6	0.4858	0.4621	0.5594	0.5546

The highest-level data accuracy of AlexNet, GoogleNet, and VGGNet is above 0.6 using various methodologies.Furthermore, the best accuracy is roughly 65% of all prior FER2013 results.There are numerous reasons why precision is difficult to improve.

For example, missing labels in training sets, data sets with noise, and overfitting can all affect accuracy.Namely,thesameobjectcanlookvastlydifferent.We can tell if the items are the same even if their status has changed as humans.However,sometimesmachinescannot.All of the elements can have an impact on the outcome.In our work, we use AlexNet(epochs=20andlrate=0.001) as our best result, which is comparable to FER2013's best findings.

5.CONCLUSION

In this paper, we applied CNN with four different architectures\s for facial expression recognition. We began by studying the fundamental structures of CNN models and improving their accuracy.We also calculated and compared each model's accuracy.GoogleNet, ResNet, VGGNet, and AlexNet are the four CNN models that are assessed on the same database, FER2013.FER2013isoneofthe\sfamousdatasets.Because the data collection is so vast, there are some sounds in it.Generally, the limited results we got a real so suitable for/ so the general situations. Finally, we discovered that AlexNet has the best overall accuracy of 0.6424.Our findings demonstrate that CNN can produce some useful results on FER.

Vidhyayana - ISSN 2454-8596
An International Multidisciplinary Peer-Reviewed E-Journal
www.j.vidhyayanaejournal.org
Indexed in: ROAD & Google Scholar

We will continue to work on improving the accuracy of each CNNmodel in the future because FERisagoodmethodthathelpshumanbeingsin dailylifeinanyfields.

Furthermore, we will study additional elements that may have an impact on accuracy, such as various biases that may affect the FER2013, and we expect to find efficient solutions to the challenges.

References

[1] A. T. Lopes, E. D. Aguiar, and T. Oliveirasantos. A facialexpression recognition system using CNN. In Graphics,Patterns and Images, pages 273–280, 2015.

[2] B. E. Bejnordi, J. Lin, B. Glass, M. Mullooly, G. L.Gierach, M. E. Sherman, N. Karssemeijer, J. V. D. Laak,and A. H. Beck. Deep learning-based assessment of tumorassociatedstroma for diagnosing breast cancer inhistopathology images. In IEEE International Symposiumon Biomedical Imaging, pages 929–932, 2017.

[3] C. F. Bobis, R. C. Gonza´lez, J. Cancelas, I. A´ lvarez, andJ. Enguita. Face recognition using binary thresholding forfeatures extraction. In International Conference on ImageAnalysis and Processing, page 1077, 1999.

[4] H. Li, H. Li, H. Li, H. Li, and H. Li. Does resnet learn goodgeneral-purpose features? In International Conference onArtificial Intelligence, Automation and ControlTechnologies, page 19, 2017.

[5] I. J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M.Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, andD. H. Lee. Challenges in representation learning: A reporton three machine learning contests. Neural Netw, 64:59–63,2015.

[6] M. A. Imran, M. S. U. Miah, and H. Rahman. Facerecognition using eigenfaces. ProcCvpr, 118(5):586–591,2002.

[7] Shen, Dinggang, Guorong Wu, and Heung-Il Suk. “DeepLearning in Medical Image Analysis.” Annual review ofbiomedical engineering 19 (2017): 221–248. PMC. Web.25 June 2018.

[8] Y. Tu, S. Li, and M. Wang. Intelligent facial expressionrecognition system r&c-fer. In Intelligent Control andAutomation, 2008. Wcica 2008. World Congress on, pages2501–2506, 2008.

[9] Y. Zhang, F. Chang, L. I. Nanjun, H. Liu, and Z. Gai.Modified alexnet for dense crowd counting. (cii), 2017.