Abstract

Facial Expression Recognition is an innovation which utilizes biometric markers to recognize feelings in human faces. This innovation is a supposition investigation device and can naturally distinguish the six fundamental or general articulations. In this paper, we investigate a deep learning vgg-16 network architecture for facial expression recognition with a 3D image data set which is divided into several classes of different emotions. This process aids in the recognition of emotion by the position and motion of facial muscles. FER has various steps in which data is pre-processed, features are extracted, and facial emotions are classified. The dataset used in this work is collected from the Kaggle repository, which consists of 35,287 samples, in which 28,079 were used for training and 7,178 were used for the purposes of validation.

1. Introduction

Three-dimensional (3D) face recognition has emerged as **an** **important** **area** **of** **​​biometrics** **because** **it** **is** **more** **accurate** and **robust** **than** **standard** 2D face **recognition.** This work presents a **new** **method** **for** 3D face recognition **that** **uses** facial **information** and **prioritizes** 2D depth images using transfer learning.

**Emotion Detection** **is** perhaps **a** **broad,** complex, and useful **field** **of** **study** in biomedical engineering, health, education, neuroscience, **psychology,** and **many** **other** **fields.** The ability of computers to **distinguish,** understand **and** **follow** human **expressions** is **an** **important** **area** **of** **​​research,** **particularly** **in** **the** **human-computer interactions** **(HCI)** **.** Suppose during online mode of study, a student’s acceptance of material will enhance if the computer understands the state of a student’s expression.

**We** believe that **this** **study** **will** **also** help us in the field of **recommendation** **systems.** **For** **example,** **if** we **use** a music **application,** if it **can** recognize our **facial expressions,** **it** will be able to **share** songs **according** **to** our **taste and** **thoughts.**

The proposed model **uses** VGG 16 to extract **distinctive** features and **learn** facial **muscles.** Initially, **an** **extensive** **database** **of** various facial expressions **was** collected **by** kaggle. To facilitate the recognition process, 3D **face** data is converted **to** **a** 2D depth **image,** reducing the complexity of **background** analysis while preserving depth **data.**

The main steps of the proposed **process** **are** as follows:

• The VGG16 **architecture** **is** **specially** designed to **ensure** **that** the **face** **of** a given **image** **is** **at** **the** **latency-tolerable** **limit** **of** **real-time** **data,** **and** **its** **accuracy** **is** **calculated.** **Up** **to** **92%** **of** **Set** in **Kaggle** **data** .

• The proposed architecture was implemented using open-source software (e.g. various Python libraries, OpenCV, Keras , Tensorflow), which makes the model more flexible to the proposed system.

• The VGG16 **architectural** model is **designed** and **trained** **using** the kaggle dataset, **and** then **the** image dataset generator **is** **used** **to** **generate** **a** **dataset** for **real-time** **search** **theory.** The **data** **we** created **was** trained **with** the VGG16 model and the **actual** image or **images** **were** **divided** into 6 **emotions.**

**Literature Review :**

In their daily life, people come across various situations and environments and the way the human body reacts to its immediate surroundings is by exhibiting emotions. The very mood of a person gets triggered by their emotional response. Their response is acknowledged by the transitions in their facial expressions. Acknowledging a person’s emotion before presenting them with something new can be of a great significance to us as it might help us predict the response we may expect from a person.

Consider an instance whereby a child is sad after having suffered from an injury while at play. The sadness that they would exhibit would be evident from the structure of their facial muscles. Now, if we were to offer them an ice-cream, we may expect the child to feel happy. Thus, the aim of this paper is to highlight how we may be able to train machine models to detect the emotions and feelings of a human being by investigating a deep learning vgg-16 network architecture. We also use Convolutional Neural Network (CNN) for marking out patterns on faces in real-time videos or photographs as and when required. This aids the model in detecting emotions even with the help of images and pictures.

Our paper emphasizes on training a model in detecting six major emotions of humans from their facial expressions and change in the muscles in their face. The emotions are namely happiness, sadness, fear, surprise, disgust, and neutrality. Attempts have been made to help the machine comprehend the emotions by the transitions of the structure of the cheeks and lips and the position of the eyes. This way, the machine would be able to infer the possible emotional response that the person is experiencing from being exposed to their immediate environment or situation.

**Methodology :**

First, the facial expression dataset undergoes preprocessing, which includes splitting it into training and testing sets, resizing all images to a uniform size, and normalizing pixel values. Then, transfer learning with VGG16 comes into play. A pre-trained VGG16 model is used to extract features from the data. The model's weights are frozen to preserve its learned features, while the final layers are replaced with new ones specifically designed for emotion detection. Next, the modified VGG16 model is trained on the preprocessed dataset. During training, a loss function like categorical cross-entropy helps measure the difference between the model's predictions and the actual emotions.

Categorical cross entropy loss is calculated as :

Loss = - Σ (y\_true \* log(y\_predicted))

Where,

y-true : is the one-hot encoded vector representing the true emotion label.

y\_predicted : is the model's predicted probability distribution for each emotion.

An optimizer like Adam iteratively adjusts the weights in the newly added layers to minimize this loss. Hyperparameters like learning rate and batch size are also tuned to optimize the model's performance. Finally, the trained model is evaluated on the testing set using metrics like accuracy, precision, and recall to assess its ability to correctly classify emotions.

Accuracy = (True Positives + True Negatives) / (Total Samples)

Additionally, confusion matrices can be employed to visualize the model's performance and identify any biases towards specific expressions. Optionally, the methodology can be extended for real-time applications by using an image dataset generator to create synthetic data and deploying the trained model for real-time processing of video streams or captured images.

2. Related work

Transfer learning is a self learning process which helps us to solve several important tasks. The primary goal of transfer learning in machine learning is to develop lifelong machine learning architectures that retain and reuse knowledge acquired from past experiences, as highlighted in a paper by Pan, Sinno Jialin, and Qiang Yang in 2010[3] which we studied through google scholar. This approach emphasizes the need for machines to continually build upon their knowledge base[4].

An important technique in transfer learning is multi-task learning, as proposed by Bozinovski and Fulgosi in 1976. This method involves simultaneously learning multiple tasks, even when they are not directly related.

Zhenye Li , Hongyan Zou model on face recognition using transfer learning showed the process of image preprocessing. [1]These experiments have demonstrated promising results for classification tasks, including applications in facial emotion recognition (FER).

To optimize classification accuracy while reducing computational complexity, transfer learning can be employed. By leveraging pre-trained network architectures trained on high-resolution datasets, knowledge can be transferred and adapted to smaller datasets without sacrificing generalization ability. This enables efficient learning on new tasks without the need to start from the beginning. Transfer learning, especially when combined with CNN architectures, holds great potential for enhancing FER applications and other classification tasks.

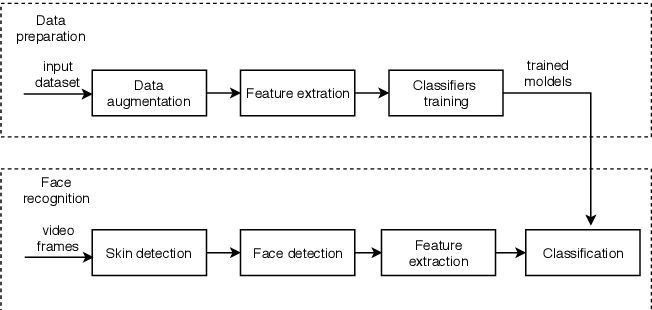


Fig. 1 Data Augmentation in FER using Transfer Learning

A critical step in machine learning is feature extraction, which entails turning raw data into understandable representations to facilitate more effective and precise model learning. The dimensionality of the data is decreased by choosing pertinent characteristics and eliminating unimportant ones, improving computing efficiency and generalization. Principal Component Analysis (PCA), Histogram of Oriented Gradients (HOG), and Convolutional Neural Networks (CNNs) are common methods for feature extraction. These techniques are essential for improving machine learning algorithms' performance and enabling them to be used for a variety of tasks, from image identification to natural language processing.

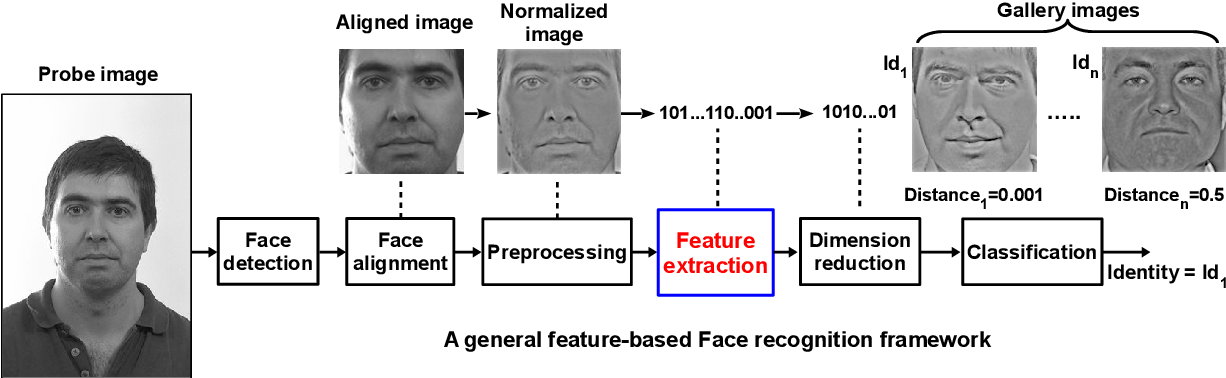


Fig. 2 Feature extraction in FER

3.Dataset

This dataset contains folders pertaining to different expressions of the human face, namely , Surprise, Anger, Happiness, Sad, Neutral, Disgust, Fear.

The folders are split into two super-folders, Training and Testing, so that it can become easier for the end user to configure any model using this data.

The training set consists of 28,079 samples in total with the testing set consisting of 7,178 samples in total. The data consists of 48x48 pixel grayscale images of faces. The faces have been automatically registered so that the face is more or less centered and occupies about the same amount of space in each image.

3.1.VGG16

VGG-16 (Visual Geometry Group 16) is a deep convolutional neural network architecture widely used in computer vision tasks. Developed by the Visual Geometry Group at the University of Oxford, VGG-16 gained significant popularity after its participation in the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) in 2014.

The VGG-16 architecture is characterized by its simplicity and uniformity, consisting of 16 convolutional layers, followed by fully connected layers for classification. Each convolutional layer has a small 3x3 filter size with a stride of 1 and a padding of 1, ensuring that the spatial resolution of the feature maps is preserved. The network's depth is a key aspect, contributing to its success in capturing intricate features in images.

The architecture's building blocks involve stacking multiple convolutional layers, each followed by a Rectified Linear Unit (ReLU) activation function and max-pooling layer. The max-pooling operation helps downsample the feature maps, reducing the spatial dimensions while retaining essential information.

VGG-16 achieved remarkable performance on various image classification benchmarks, showcasing its ability to learn hierarchical representations of visual data. Despite its effectiveness, VGG-16 is computationally expensive due to its large number of parameters. As a result, training and inference with VGG-16 can be resource-intensive, limiting its use in real-time applications and on resource-constrained devices.

To address the computational complexity, researchers have proposed optimized versions of VGG, such as VGG-19, which includes 19 layers with a similar architecture to VGG-16. Additionally, advancements in the field of deep learning have led to the emergence of more efficient architectures like ResNet, Inception, and EfficientNet.

In conclusion, VGG-16 is a significant milestone in the development of deep learning models for computer vision tasks. Its uniform architecture and depth allow it to capture intricate features effectively, making it suitable for various image classification challenges. However, with the advent of more efficient models, VGG-16's usage has somewhat diminished. Nevertheless, it remains an essential reference point in the evolution of deep neural networks and their applications in computer vision.

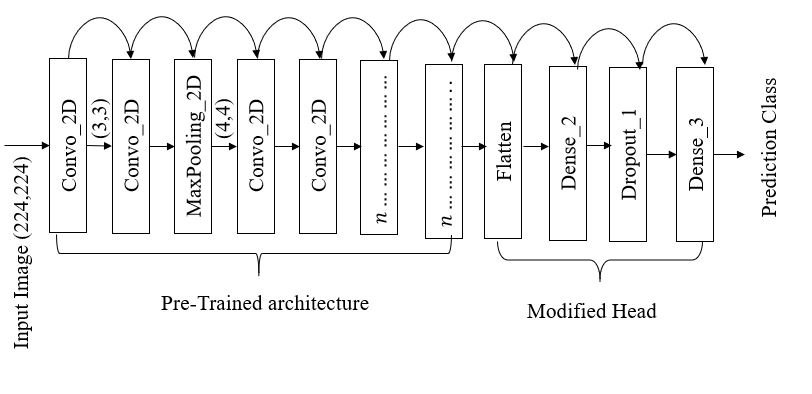


Fig. 3 Layers of Vgg16 model

# **Architecture**

* A fixed size of (224 \* 224) RGB image was given as input to this network which means that the matrix was of shape (224,224,3).
* The only preprocessing that was done is that they subtracted the mean RGB value from each pixel, computed over the whole training set.
* Used kernels of (3 \* 3) size with a stride size of 1 pixel, this enabled them to cover the whole notion of the image.
* spatial padding was used to preserve the spatial resolution of the image.
* max pooling was performed over a 2 \* 2 pixel windows with stride 2.
* This was followed by Rectified linear unit(ReLu) to introduce non-linearity to make the model classify better and to improve computational time as the previous models used tanh or sigmoid functions which proved much better than those.
* implemented three fully connected layers from which first two were of size 4096 and after that a layer with 1000 channels for 1000-way *ILSVRC* classification and the final layer is a softmax function.

**Advantages of 3D Face Recognition model :**

The described approach for 3D face recognition and emotion detection using a VGG16-based model and transfer learning offers several advantages:

* **Improved Accuracy in 3D Face Recognition:** Utilizing 3D facial information enhances the accuracy of face recognition systems compared to traditional 2D approaches. The depth information provides additional cues for distinguishing between individuals, making the system more robust and accurate.
* **Transfer Learning for Efficiency:** Transfer learning, using a pre-trained VGG16 model, allows the system to benefit from knowledge gained on a large dataset (such as Kaggle) for a related task. This results in faster convergence during training and can improve the overall performance of the model.
* **Latency-Tolerable Real-Time Processing:** The specially designed VGG16 architecture ensures that the recognition process operates within latency-tolerable limits for real-time data. This is crucial for applications where timely responses are required, such as human-computer interactions or security systems.
* **High Accuracy in Emotion Detection:** The use of VGG16 for extracting distinctive features and learning facial muscles contributes to high accuracy in emotion detection. Achieving up to 92% accuracy on the Kaggle dataset demonstrates the effectiveness of the proposed approach in recognizing and categorizing facial expressions.
* **Flexibility with Open-Source Software:** Implementing the proposed architecture using open-source software, including Python libraries (e.g., OpenCV, Keras, TensorFlow), enhances flexibility. This allows for easy customization, adaptation, and integration with other systems or platforms.
* **Reduction in Background Complexity:** Converting 3D face data to 2D depth images simplifies the analysis of the background while preserving essential depth information. This reduction in complexity can contribute to improved efficiency and speed in recognition processes.
* **Application in Human-Computer Interaction and Education:** The system's ability to understand and respond to human expressions has valuable applications in various fields, including biomedical engineering, health, education, neuroscience, psychology, and human-computer interaction (HCI). In online education, the recognition of a student's expression can potentially enhance the learning experience by adapting content based on the student's emotional state.
* **Potential for Recommendation Systems:** The study suggests potential applications in recommendation systems, such as personalized music recommendations based on facial expressions. If implemented in a music application, the system could tailor song suggestions to the user's emotional state, providing a more personalized and engaging user experience.
* **Real-Time Search Theory Implementation:** The utilization of image dataset generators for real-time search theory demonstrates a practical application of the proposed architecture. This implementation can be particularly beneficial in scenarios where quick and accurate recognition of facial expressions is essential.

**Limitations :**

While the described approach for 3D face recognition and emotion detection using a VGG16-based model has several advantages, it also comes with certain limitations and disadvantages:

* **Dependency on Quality of Training Data:** The effectiveness of the model heavily relies on the quality and diversity of the training dataset. If the dataset is biased, lacks diversity, or does not represent real-world scenarios adequately, the model may struggle to generalize to new, unseen data.
* **Limited Generalization to Unseen Expressions or Individuals:** Emotion detection models may struggle with generalizing to expressions or individuals not well-represented in the training data. Unseen or rare expressions may lead to misclassifications, limiting the model's applicability in dynamic and diverse settings.
* **Sensitivity to Lighting Conditions:** Facial recognition systems, including those based on depth information, can be sensitive to variations in lighting conditions. Changes in lighting may affect the quality of the captured images, potentially leading to recognition errors or decreased performance.
* **Hardware and Processing Requirements:** Deep learning models, especially those with architectures like VGG16, can be computationally intensive. Real-time processing may require powerful hardware, and the deployment of such models on resource-constrained devices could be challenging.
* **Challenges in 3D Data Acquisition:** While the proposed method converts 3D face data to 2D depth images, the initial acquisition of accurate 3D data can be challenging and may require specialized hardware. This could limit the practicality and accessibility of the system in certain environments.
* **Inherent Challenges in Emotion Detection:** Emotion detection itself is a complex task, as human expressions are highly nuanced and context-dependent. The model may struggle with subtle or ambiguous expressions, leading to inaccuracies in emotion classification.
* **Ethical and Privacy Concerns:** Facial recognition technologies raise ethical concerns related to privacy and surveillance. The deployment of such systems should consider and address potential misuse, unauthorized access, and the implications of collecting and storing facial data.
* **Limited Adaptability to Cultural Differences:** Emotions and expressions can vary across cultures, and a model trained on a specific dataset may not generalize well to individuals from diverse cultural backgrounds. This limitation could impact the model's accuracy and reliability in a global context.
* **Difficulty in Handling Occlusions:** Facial occlusions, such as wearing glasses or having facial hair, can pose challenges for recognition systems. The model may struggle to accurately identify individuals with partial facial obstructions, affecting overall performance.
* **Challenge in Handling Real-Time Constraints:** While the architecture is designed for real-time processing, achieving low latency in all scenarios, especially under challenging conditions or with large-scale deployments, can be a complex task. Real-world applications may require further optimization for speed and efficiency.

References

1. **Michael J. Lyons, Shigeru Akamatsu, Miyuki Kamachi, Jiro Gyoba. Coding Facial Expressions with Gabor Wavelets, 3rd IEEE International Conference on Automatic Face and Gesture Recognition, pp. 200-205 (1998).**
2. **Z. Zhang, M.J. Lyons, M. Schuster, S. Akamatsu, Comparison between geometry-based and Gabor-wavelets-based facial expression recognition using multi-layer perceptron, in: IEEE International Conference on Automatic Face & Gesture Recognition (FG), 1998.**
3. **Pan, Sinno Jialin, and Qiang Yang. (2010) “A Survey on Transfer Learning”, IEEE**

**Transactions on Knowledge and Data Engineering 22.10 (2010): 1345-359.**

1. **Chukwuemeka C Atabansi et al 2021 J. Phys.: Conf. Ser. 1873 012033,”Transfer Learning Technique with VGG-16 for Near-Infrared Facial Expression Recognition”.**
2. **Zhenye Li 1,†, Hongyan Zou 1,2,\*,†, Xinyan Sun 1, Tingting Zhu 1 and Chao Ni 1,”3D Expression-Invariant Face Verification Based on Transfer Learning and Siamese Network”**