Facial Expression Recognition System Development

Authors

AMMOURI Yassire , ALLAL Yahia, AIT ICHOU Mustapha

LRIT Laboratory, Associated Unit with CNRST (URAC 29), Mohammed V University, Rabat, Morocco.

Abstract

Facial Expression Recognition (FER) is a critical research area with applications spanning artificial intelligence, social communication, Human-Computer Interaction (HCI), and psychology. This study presents a novel hybrid model combining Deep Convolutional Neural Networks (DCNN) and Haar Cascade architectures to classify facial expressions into seven distinct categories. Utilizing the FER-2013 dataset, the imbalanced data is addressed through Random Over Sampling to enhance model performance.

The proposed DCNN incorporates multiple convolutional layers, ReLU activation functions, and various kernels to improve feature extraction and filtering depth. Additionally, the Haar Cascade model assists in real-time facial feature detection. The training and validation processes are accelerated using GPU computation, and pre-processing and data augmentation techniques are applied to boost training efficiency and classification accuracy.

Keywords: Facial Emotion Recognition, Deep Learning, CNN, DCNN, Haar Cascade, Human-Computer Interaction, FER-2013, Data Augmentation, GPU Computation, Image Processing.

1 Introduction

Humans possess a natural ability to understand facial expressions. In real life, humans express the emotions on their faces to show their psychological state and disposition at a time and during their interactions with other people. However, the current trend of transferring cognitive intelligence to machines has stirred up conversations and research in the domain of Human-Computer Interaction (HCI) and Computer Vision, with a particular interest in Facial Emotion Recognition and its application in human computer collaboration, data-driven animation, human-robot communication, etc.

Emotions are physical and instinctive, instantly prompting bodily reactions to threats, rewards, and other environmental factors. Responses to these factors are primarily based on objective measurements such as pupil dilation (eye-tracking), skin conductance (EDA/GSR), voice analysis, body language analysis, gait analysis, brain activity (fMRI), heart rate (ECG), and facial expressions. A notable ability for humans to interpret emotions is crucial to effective communication; hypothetically, 93% of efficient conversation depends on the emotion of an entity. Hence, for ideal Human-computer interaction (HCI), a high-level understanding of the human emotion is required by machines.

Emotions are a fundamental part of human communication, driven by the erratic nature of the human mind and the perception of relayed information from the environment. There are varied emotions that inform decisionmaking and are vital components in individual reactions and psychological state. Contemporary psychological research observed that facial expressions are predominantly used to understand social interactions rather than the psychological state or personal emotions. Consequently, the credibility assessment of facial expressions, which includes the discernment of genuine (natural) expressions from postured (deliberate/volitional/deceptive) expressions, is a crucial yet challenging task in facial emotion recognition. This research will focus on educating objective facial parameters from real time and digital images to determine the emotional states of people given their facial expressions [1].

2 Literature Review

2.1 Facial Emotion Recognetion Dataset (FER-2013)

The paper highlights the FER-2013 dataset, which was created in 2013 for facial emotion recognition tasks.The data consists of 48x48 pixel grayscale images of faces. The faces have been automatically registered so that the face is more or less centred and occupies about the same amount of space in each image. The images are classified into seven different emotions: anger, disgust, fear, happiness, neutrality, sadness and surprise. The images come from a Google search, which means that they come from a variety of sources, including animated characters. Initially, the dataset was divided into training images (28,709) and public test images (3,589). After the end of a competition, a further 3,589 images were added for private testing. Researchers are using the FER-2013 data distribution in different ways, for example by using the public test images as part of the training set, validation set or test set. Some examples of images from the FER-2013 data set are also presented in Figure 1.

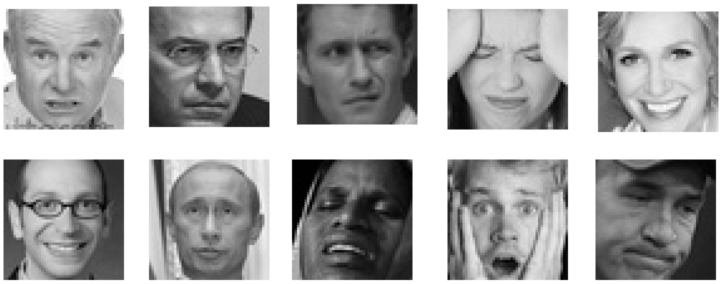


Figure 1: FER-2013 Sample Training Set Images.

3 Proposed Methodology

With the need for real-time object detection, several object detection architectures have gained wide adoption by most researchers. However, the Hybrid Architecture put forward by this research uses both the Haar Cascade Face Detection, a popular facial detection algorithm proposed by Paul Viola and Michael Jones in 2001, and the CNN Model together. In Figure 2 below, the CNN architecture initially needs to extract input pictures of 48x48x1 (48 wide, 48 high, 1 color channel) from dataset FER-2013. The network starts with an input layer equal to the input data dimension of 48 by 48. It also consists of seven concatenated convolutional layers parallelly processed via ReLU activation functions to upscale accuracy and obtain facial features of images flawlessly, as shown in Figure 2. Sub-models for extracting features share this input and have the same kernel size. The outputs from these feature extraction sub-models are flattened into vectors, concatenated into one lengthy vector matrix, and transmitted to a fully connected layer for evaluation before a final output layer permits classification. A detailed step of the methodology is described in the architecture below

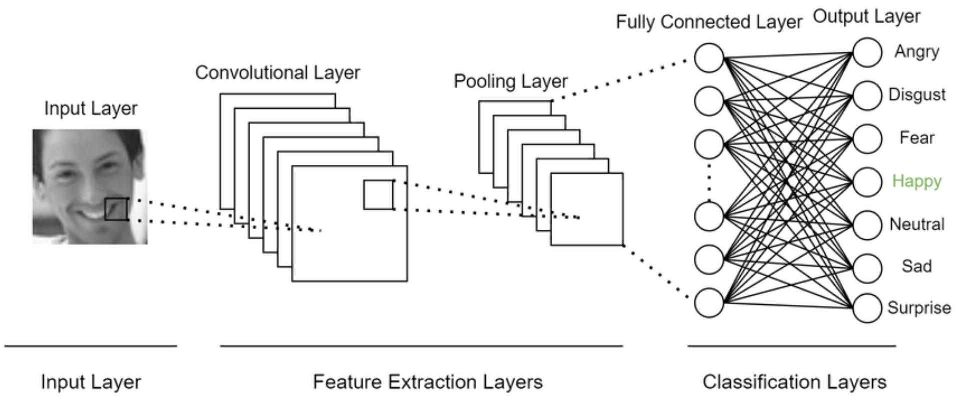


Figure 2: Proposed CNN Model Structure.

As seen in Figure 2 above, the proposed CNN model consists of a batch normalization layer followed by 7 convolutional layers with independent learnable filters (kernels), each with sizes of [3x3], and a local contrast normalization layer to remove the average from the neighborhood pixels. It also consists of a max-pooling layer to reduce the spatial dimension of the image, ensuring an increased processing pace and flatten and dense layers. It is followed by a completely connected layer and SoftMax for classifying seven emotions. A dropout of 0.5 for decreasing over-fitting was applied to the fully connected layer, and all layers encompasses rectified linear units (ReLU) activation function. After that, concatenation of two comparable models is linked to a SoftMax output layer that can classify the seven target emotions.

3.1 Data Preparation

```
from tensorflow.keras.preprocessing.image import ImageDataGenerator

# G n rateur d’images pour les donn es d’entra nement avec augmentation des donn es
train_datagen = ImageDataGenerator(
    rescale=1./255,
    rotation_range=30,
    shear_range=0.3,
    zoom_range=0.3,
    horizontal_flip=True,
    fill_mode=’nearest’  # Remplir les pixels manquants apr s transformation
)
# G n rateur d’images pour les donn es de validation (sans augmentation des donn es)
validation_datagen = ImageDataGenerator(rescale=1./255)

# G n rateur pour les donn es d’entra nement
train_generator = train_datagen.flow_from_directory(
    train_data_dir,
    color_mode=’grayscale’,  # Les images sont en niveaux de gris
    target_size=(48, 48),
    batch_size=64,
    class_mode=’categorical’,
    shuffle=True  # M langer les donn es   chaque  poque
)
# G n rateur pour les donn es de validation
validation_generator = validation_datagen.flow_from_directory(
    validation_data_dir,
    color_mode=’grayscale’,  # Les images sont en niveaux de gris
    target_size=(48, 48),
    batch_size=64,
    class_mode='categorical',
    shuffle=True)  # M langer les donn es   chaque  poque
```

Listing 1: Script du programme

3.2 Model overview

The model begins with an input layer for grayscale images of size 48 x 48. It then has four convolutional layers that use ReLU activation and with filter sizes of 32, 64, 128 and 256, respectively. The second, third, and fourth convolutional layers are followed by MaxPooling and Dropout layers to minimize dimensionality and avoid overfitting. After being flattened, the output from these layers is routed via a Dropout layer and a fully linked layer with 512 units and ReLU activation. The final output layer divides the images into seven classes using softmax activation. The model measures accuracy during training and is constructed using the Adam optimizer and categorical cross-entropy loss function.

```
from keras.models import Sequential
from keras.layers import Dense,Dropout,Flatten,Conv2D,MaxPooling2D

model = Sequential()
model.add(Conv2D(32, kernel_size=(3, 3), activation='relu',
    input_shape=(48,48,1)))

model.add(Conv2D(64, kernel_size=(3, 3), activation='relu'))
model.add(MaxPooling2D(pool_size=(2, 2)))
model.add(Dropout(0.1))

model.add(Conv2D(128, kernel_size=(3, 3), activation='relu'))
model.add(MaxPooling2D(pool_size=(2, 2)))
model.add(Dropout(0.1))

model.add(Conv2D(256, kernel_size=(3, 3), activation='relu'))
model.add(MaxPooling2D(pool_size=(2, 2)))

model.add(Dropout(0.1))

model.add(Flatten())
model.add(Dense(512, activation='relu'))
model.add(Dropout(0.2))

model.add(Dense(7, activation='softmax'))
model.compile(optimizer = 'adam', loss='categorical_crossentropy',
    metrics=['accuracy'])
```

Listing 2: Script du programme

4 RESULTS

After training the model for 100 epochs, we observed notable improvements in both accuracy and loss metrics, as depicted in Figures 3. The model achieved a training accuracy of approximately 60% and a validation accuracy of approximately 61%, indicating a relatively stable and well-generalized performance on the validation dataset

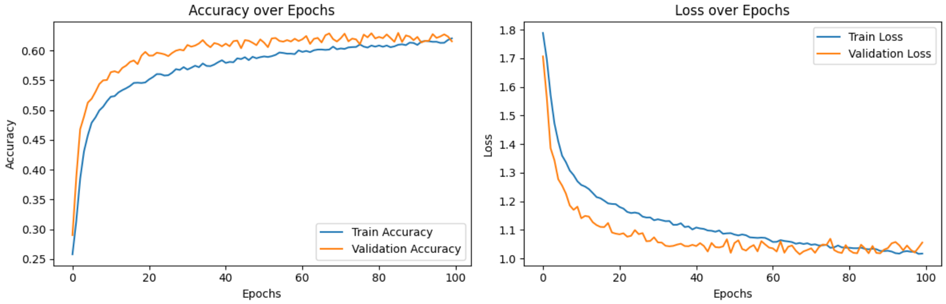


Figure 3: Model Accuracy and Loss

In the left image in the Figure 3, accuracy over Epochs shows a consistent upward trajectory in both training and validation accuracy, reflecting the model's effective learning and generalization capabilities. This indicates that our approach, including class weighting to address imbalances, has successfully enhanced performance on misclassified emotions.

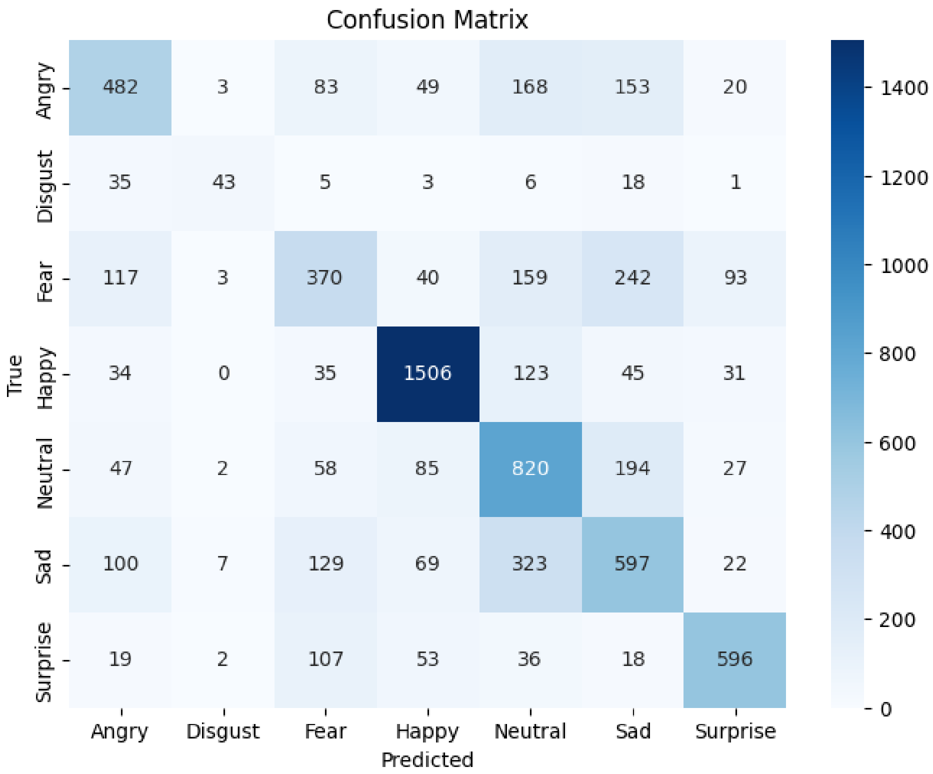
In the right image in the Figure 3,Loss over Epochs depicts a clear downward trend in both training and validation loss. The significant initial decrease in training loss, followed by a more gradual decline, indicates effective error minimization. The validation loss mirrors this pattern, reinforcing the model's robustness and error reduction over successive epochs.

Overall, as the number of epochs increases, the model's accuracy improves steadily while loss values decrease, demonstrating a well-trained model that balances learning from training data and generalizing to new, unseen data.

In the confusion matrix, the actual (True) classes are compared against the predicted classes, providing a detailed breakdown of the model's performance across different emotions. Each cell in the matrix represents the count of instances where the true emotion (rows) matches the predicted emotion (columns).

- **Diagonal Cells:** Represent correct predictions. For instance, 482 "Angry" and 1506 "Happy" instances were correctly classified.
- **Off-Diagonal Cells:** Show misclassifications. For example, 242 "Fear" instances were misclassified as "Neutral."

This matrix highlights the model's strengths and areas for improvement, allowing us to measure overall accuracy and identify specific confusion patterns.



True \ Predicted	Angry	Disgust	Fear	Happy	Neutral	Sad	Surprise
Angry	482	3	83	49	168	153	20
Disgust	35	43	5	3	6	18	1
Fear	117	3	370	40	159	242	93
Happy	34	0	35	1506	123	45	31
Neutral	47	2	58	85	820	194	27
Sad	100	7	129	69	323	597	22
Surprise	19	2	107	53	36	18	596

Figure 4: Confusion Matrix

5 CONCLUSION

This report proposed a CNN model for Real-time Facial Expression Recognition based on seven emotional classes (‘neutral’, ‘happy’, ‘sad’, ‘angry’, ‘surprised’, ‘fear’, and ‘disgusted’). The model demonstrated strong generalization and classification performance. The process began with acquiring a variety of well-classified, high-quality databases. Facial regions were detected, cropped, and converted into grayscale images to remove unnecessary information. To combat overfitting, image data augmentation was employed, increasing the number and variation of training images.

Focusing on achieving the highest accuracy for the seven emotions, we noted that the ’disgust’ emotion had the fewest images. Despite extensive data augmentation, rotation, and optimizer changes, accuracy fluctuated. However, after applying the Random Oversampling technique to balance the dataset, a significant rise in accuracy and decrease in loss were observed, making ’disgust’ easier to identify.

The CNN model achieved an 61% test accuracy using the FER2013 dataset. Future improvements could include utilizing facial landmarks detection and alignment to further enhance accuracy. Additionally, adapting the model for real-world application, decreasing misclassified emotions through pipeline model improvements, and employing multi-label classification to handle images with multiple possible emotion labels, will further refine the model’s performance.

References

[1] Ozioma Collins Oguine, Kanyifeechukwu Jane Oguine, Hashim Ibrahim Bisallah, Daniel Ofuani, "Hybrid Facial Expression Recognition (FER2013) Model for Real-Time Emotion Classification and Prediction"

Abbreviations

DCNN - Deep Convolutional Neural Network
CNN - Convolutional Neural Network
FER - Facial Emotion Recognition
GPU - Graphical Processing Unit
HCI - Human Computer Interaction