Institute for Visualization and Interactive Systems

University of Stuttgart
Universitätsstraße 38
D–70569 Stuttgart

InfoTech Research Project Nr. 3440612

Facial Emotion Recognition behind Facial Masks

Nidhi Joshi

Course of Study: Information Technology

Examiner: Prof. Dr. Andreas Bulling

Supervisor: Dominike Thomas, M.Sc.

Commenced: 01. July 2021

Completed: 31. December 2021

CR-Classification: I.7.2

Abstract

The onset of a pandemic condition owing to Covid-19 has introduced new obstacles in Facial Emotion Recognition (FER), particularly due to the requirement to wear facial masks. In light of this issue, the goal of this study is to train the Attention-based Convolutional Neural Network (ACNN) developed by Yong Li et al. and train the model on artificially synthesised masked datasets derived from existing emotion datasets namely AffectNet, RAF-DB and CK+. A hypothesis is proposed with the notion that emotions namely Neutral, Happy, Angry and Surprise are better perceived than Sad, Fear and Disgust emotions. The experiments and results acquired following a successful training and testing regime of the model, indicate that the presented hypothesis is correct. Although the test accuracies obtained in the study do have further scope to be improved. Many difficulties arise as a result of information loss caused by face masks, which may be solved by creating additional datasets collected under controlled and lab circumstances.

Contents

List of Figures

List of Tables

List of Abbreviations

ACNN Attention mechanism based Convolutional Neutral Network. 7

CK+ Extended Cohn-Kanade Dataset. 18

FAN Facial Alignment Network. 24

FER Facial Emotion Recognition. 13

FLSRC Fusion of Local Sparse Representation Classifiers. 17

gACNN global–local based ACNN. 18

GG-Unit Global-Gated Unit. 27

HOG Histogram of Oriented Gradients. 18

JAFFE JApanese Female Facial Expression. 17

kNN k-Nearest Neighbor. 18

LBP Local Binary Pattern. 17

LDA Linear Discriminant Analysis. 18

LGBP Local Gabor Binary Pattern. 17

MUG Multimedia Understanding Group. 18

pACNN patch based ACNN. 18

PCA Principal Component Analysis. 17

RAF-DB Real-world Affective Faces Database. 18

ROI Region Of Interest. 18

RPCA Robust Principal Component Analysis. 17

SRC Sparse Representation based Classification. 17

SVM Support Vector Machine. 18

WLD Weber Local Descriptor. 17

1 Introduction

Facial expressions are the primary and fundamental ways to convey human emotions. People can have in depth interactions with one another by the virtue of facial expressions. Many efforts are being undertaken to entirely automate Facial Emotion Recognition (FER) discipline via the use of machine learning and deep learning technologies that evaluate human emotions for a variety of purposes. Several companies such as Kairos[12], InSight[09] have come up with FER algorithms to leverage digital video as primary data that aims to test digital content, advertising, movie trailers, and TV programs [Jes19] for consumer's reactions/reviews. Additionally FER is also employed in driver emotion monitoring system (Naqvi et al. [NAR+20]), criminal investigation (Stanković et al. [SNO+15]), sentiment analysis on social media (Tanna et al. [TDS+20]) and many more.

Nevertheless, the onset of the pandemic crisis caused by Covid-19, has obligated everyone to wear facial mask to help curb the spread of infection. Hence there is a need to implement an algorithm that can recognize the facial emotion behind the facial mask. Greco et al. [GSVV21] concludes in their work that more research is required to improve the efficacy of existing techniques of emotion recognition under partial occlusion. Following the same requirement, an algorithm is implemented in this project that is trained on images of masked emotional faces. It concentrates and learns on the obstructed and unobstructed portion of the face and classifies the emotion according to the standard emotions accepted in FER studies i.e., Neutral, Happy, Angry, Surprise, Sad, Fear and Disgust. The emotion images are obtained from different datasets that are augmented by artificially overlaying facial mask on the subject in the image. Subsequently, the results of the model's performance and a detailed discussion on the experiments conducted to evaluate the model are presented. Furthermore, based on the observations, a conclusion is drawn regarding the model's efficacy to classify distinct emotions and obstacles encountered leading to such outcomes are summarized.

1.1 Hypothesis

As face masks cover major portion of the face, predominantly below the nose, it renders quite a bit of setback on the emotion recognition [GES21]. Consequently, emotions that mainly depend on lower region of the face such as Sad, Fear or Disgust have significantly inaccurate predictions as studied by Marini et al. [MAP+21]. For this reason, the following hypothesis is proposed:

Facial expressions such as Neutral, Happy, Angry and Surprise are perceived better than Sad, Fear and Disgust under facial mask.

Preliminary analysis of the expressions under facial masks done for this project, have led to this hypothesis. The main points analyzed iterating over each expression suggest that for the-

- 'Happy' expression: the subject's eyes are frequently observed to be squinted.
- 'Surprise' expression: the distance between eyes and the eyebrows is elongated.
- 'Angry' expression: the eyebrows tend form a V-shaped arch and size of glabella is diminished.
- 'Neutral' expression: there are not many changes in structure of eyes, eyebrows and regions surrounding it.

The distinct features for each expressions as mentioned above are extracted for training the model. Such trained model is tested and experimented under different conditions to prove this hypothesis.

1.2 Skeleton of the Report

In upcoming chapters, the following topics will be discussed:

Chapter 2 – Related Work: Here, the past work by different authors and their results are discussed along with how their work influences this project.

Chapter 3 – Implementation: The implementation of the project that includes dataset creation, feature extractions and classification of the images to emotional category are presented.

Chapter 4 – Results and Discussion: After the implemented model is trained over seven and four emotions on multiple datasets, the results obtained are presented and the observations discussed. The discussion on original authors work in comparison with this project is carried out.

Chapter 5 – Conclusion: In this section, a final conclusion on the effectiveness of the project's implementations and results are discussed along with future scope and improvement for the project.

2 Related Work

There has been a lot of work done in FER for partially occluded images. A well-known survey conducted by Zhang et al. [ZVTC18] covers major achievements in this field while discussing all possible approaches. One of the studies done in this field was by Shane Cotter [Cot11] who proposed a new classification method termed as Sparse Representation based Classification (SRC). It dynamically determines the set of local regions that best classifies the facial expression by using the representation of error characteristics from each local region. This algorithm is termed as Fusion of Local Sparse Representation Classifiers (FLSRC). In an experiment where the mouth region is occluded, the FLSRC recognition rate was 93.4% compared to 72.8% using SRC with all available pixels and rates $< 60\%$ using Gabor or Principal Component Analysis (PCA) based methods. FLSRC significantly outperforms other methods when large facial regions are occluded. For example, when approximately 25% of the image is occluded, FLSRC gives a recognition rate of 85% compared to $< 65\%$ for other methods.

Investigations on Gabor filters, Local Binary Pattern (LBP) and Local Gabor Binary Pattern (LGBP) for feature extraction were carried out by Reza Azmi et al. [AY12]. Six basic facial expressions plus neutral pose were considered. The k-NN classifier with sum of absolute differences distance was used in classification phase. Basic occlusions occurring in real life were considered for JApanese Female Facial Expression (JAFFE) dataset [LKG98]. Experimental results show the effectiveness and high robustness of LGBP approach under a variety of occlusion conditions and provide useful insights about the effects of occlusion on FER. Using LGBP features the average accuracy 96.25% on non- occluded images, 88.77% on eyes occluded images, 92.78% on mouth occluded images, 89.18% on lower face occluded images and 90.17% on upper face occluded images was obtained.

Utilization of Weber Local Descriptor (WLD) for development of an emotion recognition framework robust to occlusions in facial expressions was demonstrated by Ramirez Cornejo et al. [RP18]. Initially, occluded facial expressions are recreated using Robust Principal Component Analysis (RPCA) extension technique. Then, WLD features are extracted from the facial expression representation, as well as LBP and

Histogram of Oriented Gradients (HOG). The feature vector space is reduced using PCA and Linear Discriminant Analysis (LDA). Finally, k-Nearest Neighbor (kNN) and Support Vector Machine (SVM) classifiers are used to recognize the expressions. Experimental results on three public datasets demonstrated that the WLD representation achieved competitive accuracy rates for occluded and non-occluded facial expressions compared to other available approaches. The proposed method yielded accuracy as much as 91.04% on Extended Cohn-Kanade Dataset (CK+), 92.86% on JAFFE and 90.12% on Multimedia Understanding Group (MUG) [APD10] for partial artificially synthesised occluded images from these datasets. On the other hand, it also showed outstanding accuracy for non-occluded images i.e., 93.13% on CK+,95.35% on JAFFE and 93.09% on MUG datasets.

The aforementioned studies include occlusions on the face but not particularly caused by facial mask at larger scale. Additionally, the feature extraction tasks are carried out using imaging science techniques such as LBP, LGBP or WLD and then using a single deep learning architecture for classification. These approaches require datasets that offer precise face and utilization of robust feature trackers adding extreme preprocessing of the datasets. On the other hand, the use of deep learning techniques to automatically extract discriminative feature patterns of facial expressions from the raw face data is highly beneficial especially for the available emotion datasets. Therefore, work by Yong Li et al. [LZSC19], who employ an ACNN that performs automatic feature extraction and classification of the emotion. In this paper, the authors propose a CNN with attention mechanism (ACNN) that can perceive the occluded regions of the face and focus on the most discriminative non-occluded regions. ACNN is a complete learning framework. It combines various representations of facial Region Of Interests (ROIs). Each representation is weighted using a proposed gate unit, which computes an adaptive weight from the region itself based on its unobstructedness and importance. In order to accommodate different ROIs, two versions of ACNN are presented: patch based ACNN (pACNN) and global–local based ACNN (gACNN). pACNN only considers local facial patches. gACNN combines patch-level local representations with global representations at the image level. The proposed ACNNs are tested on real and synthetic occlusion datasets with real-world occlusions, the two largest in-the-wild facial expression datasets: Real-world Affective Faces Database (RAF-DB) and AffectNet, and their modifications with synthesized facial occlusions. The experimental results show that the recognition accuracy on both non-occluded and occluded faces is 85.07% and 80.54% for models trained on the RAF-DB dataset, respectively, and 58.78% and 54.84% for models trained on the AffectNet dataset.

This project takes inspiration from the work by Yong Li et al. [LZSC19] and attempts to apply a similar model to a relatively unexplored area of FER i.e., recognizing facial emotion behind facial masks. Because their work employs an end-to-end framework for feature extraction and classification tasks, beneficial to utilize existing in-the-wild emotion datasets. The implementation details are covered in the Implementation section below.

3 Implementation

To deal with FER behind facial masks, a Convolution Neural Network with Attention Mechanism (ACNN) [LZSC19] is used, which mimics how humans recognize facial expressions. ACNNs essentially use gate units that learn an adaptive weight based on unobstructed-ness or importance of the region in the image.

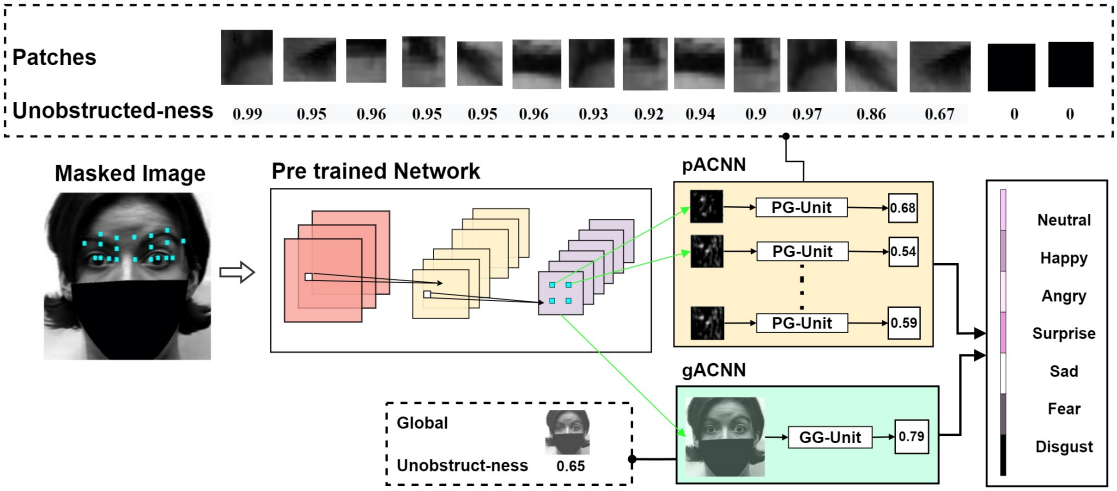


Figure 3.1: Illustration of the ACNN for occlusion-aware facial expression recognition by Yong Li et al.

The main idea of this method is depicted in Fig. 3.1. The network receives an occluded image of a face where the facial mask has inhibited the lower part of the face. The image is first fed to the first nine convolutional layers of the pretrained network of VGG-16 [SZ15] as the feature maps extractor. This choice is made because VGG-16 provides simple structure and excellent performance in object classification. From the VGG-16 net, feature maps of the image are obtained and select patches of the image are fed to one of the version of the ACNN i.e., pACNN which only pertains to local attention mechanism. It consists of Patch-Gated Unit (PG-Unit) that encodes each local patch as a weighted vector. A PG-Unit computes

the weight of each patch by an Attention Net while taking its obstructed-ness into account. On the other hand, the feature maps as representation of the entire image is fed to gACNN which focuses on local patches as well as whole image. It consists of Global-Gate Unit (GG-Unit) that encodes whole face as a weighted vector. Then both the weighted global facial features and the local representations are concatenated that serves as a representation of the entire occluded face. Lastly, to assign the face to one of the emotional categories, two fully connected layers are employed.

In order to implement this methodology, the following steps are carried out, details of which are discussed in impending sections.

- Synthesising a dataset by overlaying an artificial mask on the emotional faces in the images.
- Implementing the model: pACNN and gACNN that extracts features from the dataset in two distinct ways and classifies the images. The specifics of both ACNN versions will be discussed.
- Finally training and testing of the model on the datasets according to chosen hyperparameters.

3.1 Synthesis of Masked Dataset

There are no emotion datasets with facial mask as a cause for partial occlusion, therefore, for this project, the existing emotion datasets are modified to synthesize datasets with facial mask as a cause for partial occlusion. To achieve this goal, the images from existing emotion datasets namely AffectNet[MHM17], RAF-DB[LDD17] and CK+[LCK+10] are utilized. Synthesis of masked dataset is performed in two folds:

1. to detect the faces in the images crop them per the faces' dimension using Dlib toolkit
2. to overlay artificial mask. Specific facial landmarks are detected and selected which eventually define the shape of the face masks. The shape of the artificial mask is chosen such that it resembles the real-life face masks as closely as possible, which covers the nose and the mouth as shown in Fig. 3.2

Figure 3.2: Overlaying an artificial mask on sample image from CK+ dataset

The model is evaluated on both in-the-wild datasets (AffectNet and RAF-DB) and in-the-lab dataset (CK+). These datasets are chosen because they consist highest number of images with respect to the style of dataset (in-the-wild/in-the-lab dataset) with class labels as required for this project, eventually good for training the model. Details of these datasets are given below:

AffectNet

The dataset AffectNet consists of more than 1 million images with class labels and annotations included. Despite this number, only 35000 images are used which are evenly distributed among each category of emotion (5000 each) for training and similarly 3500 images for testing. The images are in-the-wild style where the faces are in uncontrolled environment and collected directly from the internet. Few images from the raw dataset already contain occlusions caused due to sunglasses, hair or hands on the face etc. which can cause inaccuracies in the results, discussed in further sections.

RAF-DB

RAF-DB is a large-scale facial expression dataset with around 30000 great-diverse facial images downloaded from the internet. Out of these many images, 15339 images are made available, not balanced for each class of emotion. The class imbalance in the dataset is taken care by a random sampler. The images in this dataset are also in-the-wild style and of great variability in subjects' age, gender & ethnicity, head poses, lighting conditions, occlusions such as glasses, facial hair or self-occlusion and post-processing operations namely filters, special effects, etc.

CK+

The extended version of Cohn-Kanade dataset known as CK+, is a popular comprehensive dataset for facial expression benchmark tests. It contains 593 video sequences recorded from 123 subjects. All the frames from each video sequences are

selected, which resulted in 5876 images. The subjects in the images of this dataset are in a lab controlled environment and facing the camera resulting in no alignment of the faces. Fortunately, CK+ does not include occluded faces. These set of images are also imbalanced in each class of emotion, thus random sampler to the rescue.

3.2 Feature Extraction

Following the creation of the masked datasets, feature extraction and classification is performed using the ACNN model. The feature extraction task is presented in this section. As mentioned earlier two version of ACNN are employed for it: pACNN and gACNN.

3.2.1 pACNN

pACNN is intended to concentrate on discriminative and representative patches at the local level. pACNN is made up of two main schemes: region decomposition and occlusion perception.

Region Decomposition

Since the images in hand are occluded by a facial mask, the prominent regions of the face relevant for emotion recognition are the regions near the eyes, the brows and the forehead. Recognizing facial expressions requires the localization and encoding of expression-related parts [ZLY+12]. To extract these facial parts relevant for an expression, first 68 facial landmark points as shown in Fig. 3.3 are detected using the method by Bulat and Tzimiropoulos [BT17]. It illustrates Facial Alignment Network (FAN), a state-of-the-art neural network for landmark localization that is trained and evaluated on existing 2D and 3D face alignment datasets. After the standard 68 points are detected, 20 select points are recomputed that cover the fine informative regions of the face. Then multiple patchs of image surrounding these 20 landmarks, according to the positions of each subject's facial landmarks are extracted.

The selection of 20 landmark points is quite simple. Since the eyes and brows area is the most informative region now, the only distinguishing features are the landmarks on the eyes and brows, between both the eyes and brows, the glabella, and the root of the nose. As the perpendicular distance between the eyes and brows as well as

Figure 3.3: Facial landmarks detected on a masked image from CK+[LCK+10] dataset

the area around the glabella vary depending on the expressions, custom landmarks are opted to obtain patches from these regions. With this concept:

- 10 landmarks from the ones in between eyes and brows and 1 on the root of the nose are directly obtained from list of 68 landmarks shown as encircled points in Fig. 3.4. These single landmarks are indexed at 17,19,21,22,24,26,36,39,42,45 and 27.
- 9 landmarks are derived from the above chosen 11 landmarks. These are essentially the landmarks obtained as mid points of landmarks indexed at [21,22],[21,39],[22,42],[37,41],[38,40],[43,47],[44,46],[19,37] and [24,44]. The resulting landmarks are shown in Fig. 3.4 represented with yellow triangular symbols.

Occlusion Perception

Patch-Gated Units are the fundamental units in the pACNN that automatically perceives the blocked facial patches. One such unit is shown in Fig. 3.5 in which the cropped local feature maps are given to two convolution layers. Each patch-specific PG-Unit learns region-specific patterns without lowering the spatial resolution and preserving the information. The final feature maps are then split into two groups. The input feature maps are encoded as vector-shaped local features say ψ in the first

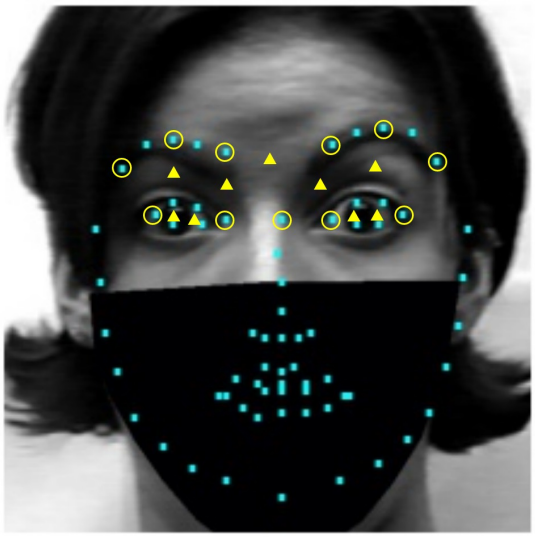


Figure 3.4: Region decomposition of the face

branch. The Attention Net in the second branch estimates a scalar weight to reflect the relevance of the local patch. The calculated weight is subsequently applied to the local feature.

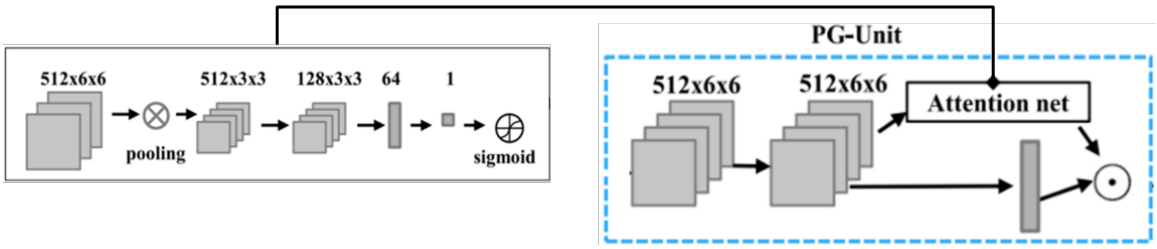


Figure 3.5: Structure of pACNN: A single unit of PG-Unit from pACNN

The Attention Net is made up of a pooling operation, one convolution operation, two inner productions, and a sigmoid activation. The sigmoid activation yields a weight between [0, 1] say α, with 1 signifying the most conspicuous unobstructed patch and 0 denoting the entirely occluded patch. Finally, the vector-shaped local feature ψ is multiplied with α, yielding a weighted feature say φ given by :

$$\varphi = \psi \cdot \alpha \tag{3.1}$$

As 20 custom landmarks are selected for each image, the pACNN structure consists of 20 units of PG-Units which process the cropped patch and weighted differently according to its occlusion conditions or importance.

3.2.2 gACNN

Along with pACNN, another mechanism is needed which can weigh the image as a whole facilitating to include complementary information that may have been disregarded in pACNN. In VGG16 net, the feature maps of the entire face are encoded from layer 'conv4_2' to layer 'conv5_2' which are 9th and 12th layer of the VGG net respectively. Subsequently the encoded region with the size of 512 x 14 x 14 are obtained based on the 512 x 28 x 28 feature maps. Global-Gated Unit (GG-Unit) is the basic framework of gACNN that weighs the global facial representation as shown in Fig. 3.6. The first branch of the GG-Unit algorithm encodes the input feature maps into a vector-shaped global representation. The second branch is made up of an Attention Net that learns a scalar weight to reflect the global face representation's contribution. The calculated weight is subsequently applied to the global representation in similar fashion done in pACNN.

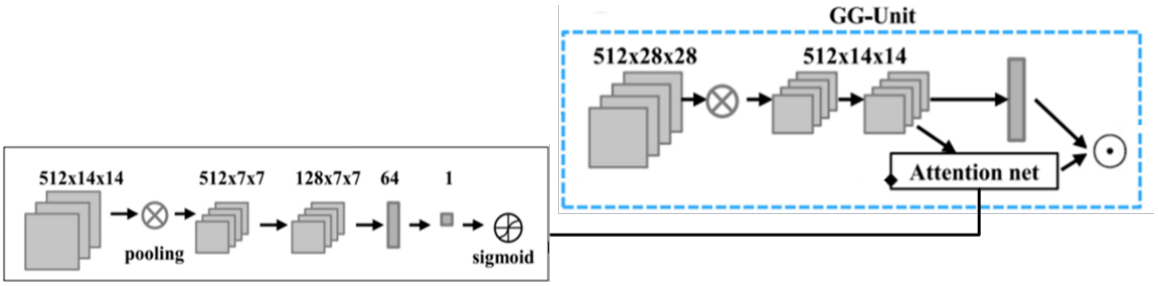


Figure 3.6: Structure of gACNN

3.3 Model Implementation

The extracted features from ACNN structure provides the weighted global facial features concatenated with weighted local facial features resulting as a representation of the occluded face. Now for classification of the image to one of the emotional categories, two fully connected layers are employed. The schematic representation of the entire model with intricate details are shown in Fig. 3.7.

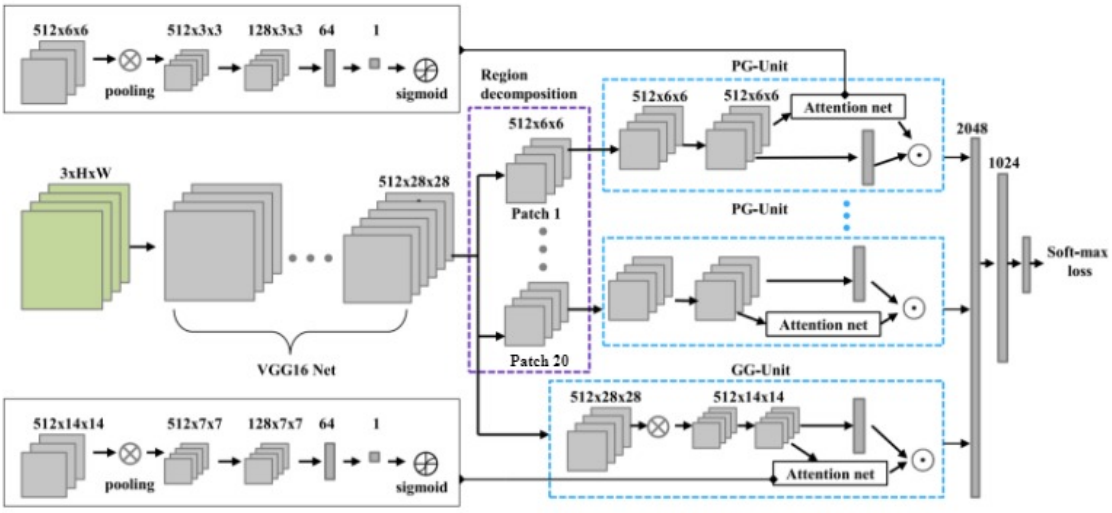


Figure 3.7: Structure of the ACNN model: Integration of pACNN with gACNN for feature extraction along with Classification layers

The Fig. 3.7 illustrates the framework of the ACNN. As described previously, ACNN takes a facial image as the input and encodes the image with VGG-16 Net. Then for pACNN the feature maps from the last convolution layer are cropped into 20 local patches through a region decomposition scheme. Each patch is then processed by PG-Unit. Each PG-Unit encodes a patch by a vector-shaped feature and estimates how informative the patch is through an Attention net. For gACNN, the full feature maps are encoded and weighed as a representation vector through a GG-Unit. The cross entropy loss (softmax loss) is attached at the end. Parameters in the overall network are learned by minimizing the cross entropy loss. The training & experimental conditions and results obtained after training the model are discussed in next chapter Results and Discussion.

4 Results and Discussion

The experimental assessments of the ACNN are provided in this section. First, the experimental conditions are outlined and the results are presented. Next, the findings of these experiments are discussed and compared to that with the original authors.

4.1 Experimental Conditions

The model was trained and tested on both unmasked and masked images of the aforementioned datasets. The purpose for training the model on unmasked photos of the dataset was because these results are representative of how the model would function optimally, i.e. in the absence of occlusion. This clarifies the model's performance in ideal circumstances. It is now easy to compare these findings to those obtained when the model is trained on the masked images of the datasets.

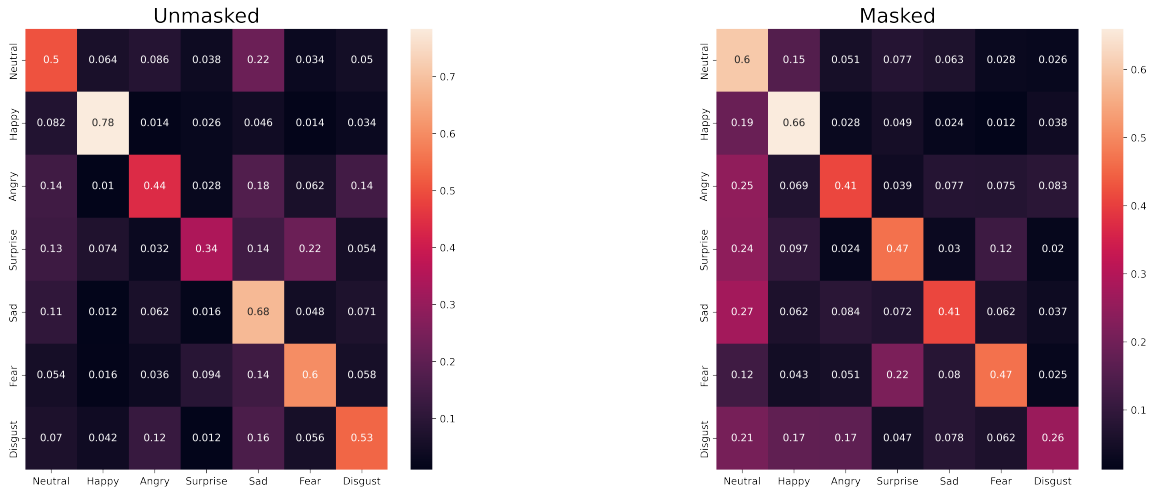
Total of 133,991,004 trainable parameters were obtained. The training and validation sets were in 8:2 ratio. The training was optimized by cross entropy loss function with learning rate of 0.95 at the first epoch, subsequently, the learning rate was increased geometrically. The training was stopped from over-fitting by early stopping mechanism [Pre12] after three consecutive loss values with increasing trend were observed.

For training and testing, a variety of batch sizes were investigated. The accuracy of testing rose when the batch size was reduced. As a result, it was determined that a training batch size of 5 and a testing batch size of 10 were sufficient. To further investigate the outcomes and performance, a confusion matrix is used as the evaluation metric since it accurately represents the model's classification accuracy for each emotion. The following findings were achieved under these experimental conditions.

4.2 Results

This section presents the confusion matrix and testing accuracies obtained after model training for both versions of each dataset: Unmasked and Masked images, that were evaluated over both 7 and 4 emotions separately. Upon gathering the data, the observations are discussed in the next section 4.3.

4.2.1 For 7 standard expressions

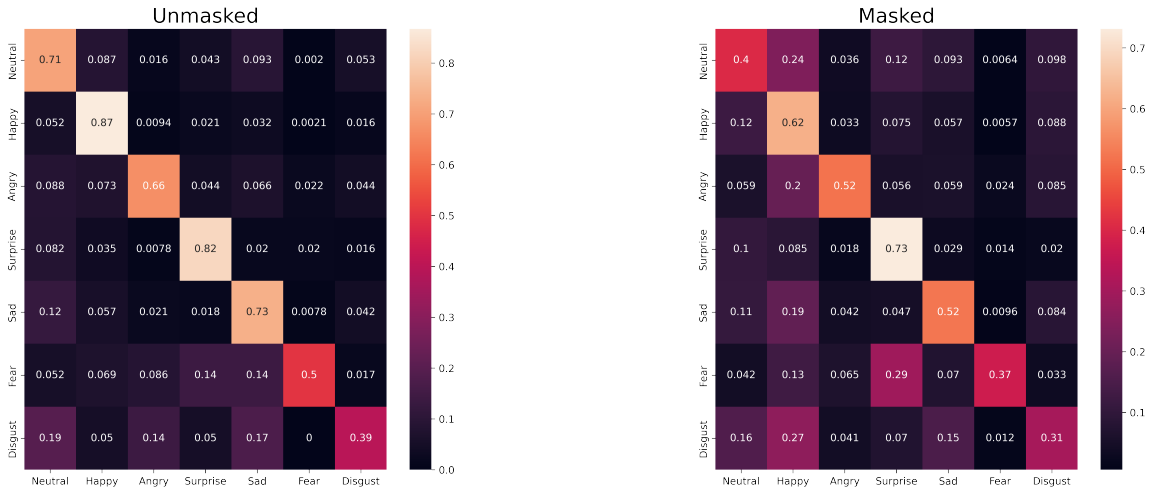


(a) Unmasked Images **(b)** Masked Images

Figure 4.1: Confusion matrix for **AffectNet** dataset for 7 emotions

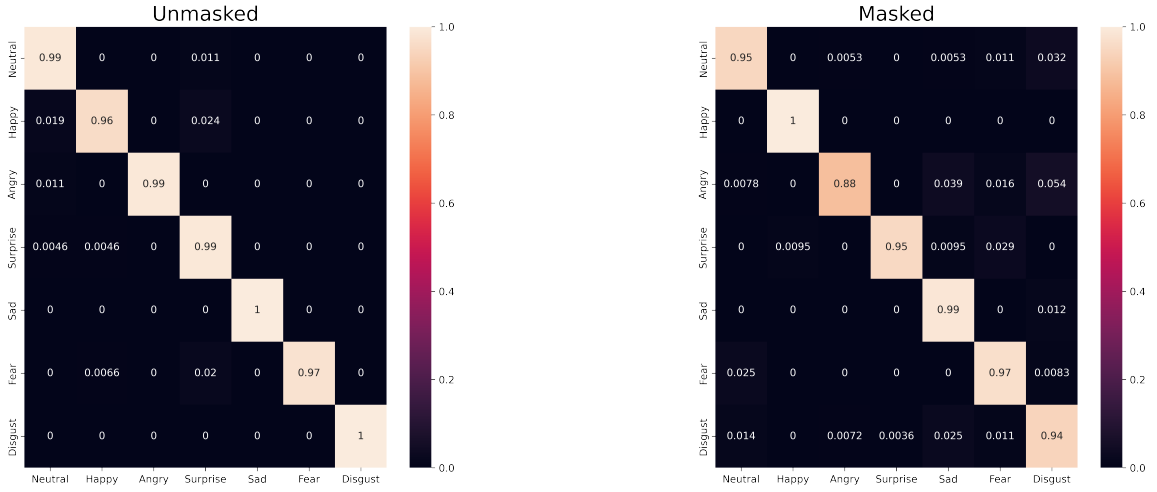
The confusion matrices obtained for AffectNet dataset when model is evaluated over 7 standard emotions is shown in Fig. 4.1. The classification of unmasked images as shown in Fig. 4.1a is comparatively sophisticated though the emotion ‘Surprise’ is identified the least number of times, unlike the classification of masked images (Fig. 4.1b) which is uneven and the emotion ‘Happy’ is accurately identified. This indicates that the model does behave rather inefficiently for masked images of the dataset than the unmasked images. Similar case runs for RAF-DB dataset as shown in Fig. 4.2. The Fig. 4.2a reveals that the emotions are reasonably categorized with the exception of the emotions ‘Fear’ and ‘Disgust’ which are poorly defined, although the emotion ’Happy’ is accurately detected yet again. In contrast, the classification of masked images is crude although now, the emotion ‘Surprise’ is recognized correctly,

refer Fig. 4.2b. For in-the-lab dataset that is CK+, the categorization of emotions is quite clean for both masked and unmasked images as shown in Fig. 4.3



(a) Unmasked Images **(b)** Masked Images

Figure 4.2: Confusion matrix for **RAF-DB** dataset for 7 emotions



(a) Unmasked Images **(b)** Masked Images

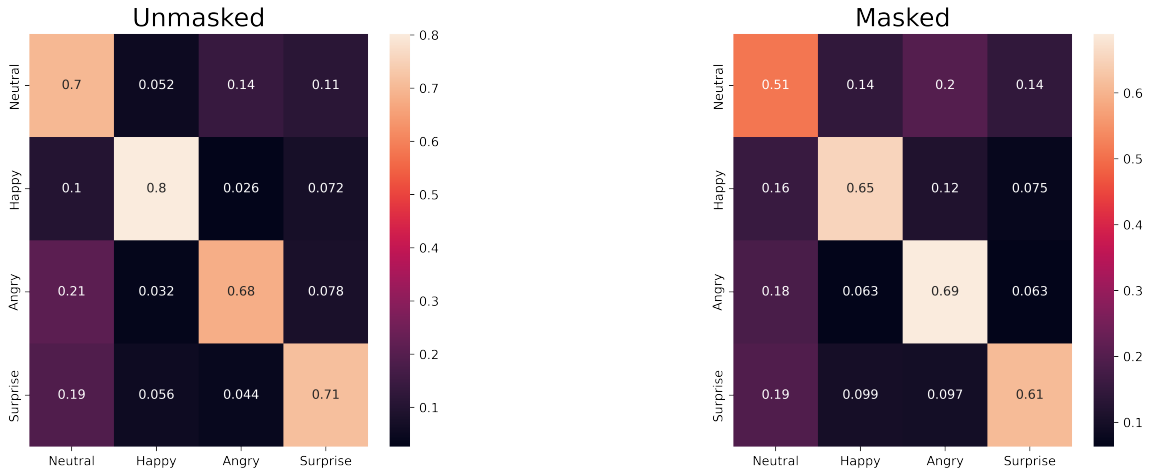
Figure 4.3: Confusion matrix for **CK+** dataset for 7 emotions

Corpora	Unmasked Images	Masked Images
AffectNet	55.47	46.84
RAF-DB	75.78	60.49
CK+	98.11	95.13

Table 4.1: Test accuracy (%) for **AffectNet**, **RAF-DB** and **CK+** datasets evaluated over 7 standard emotions

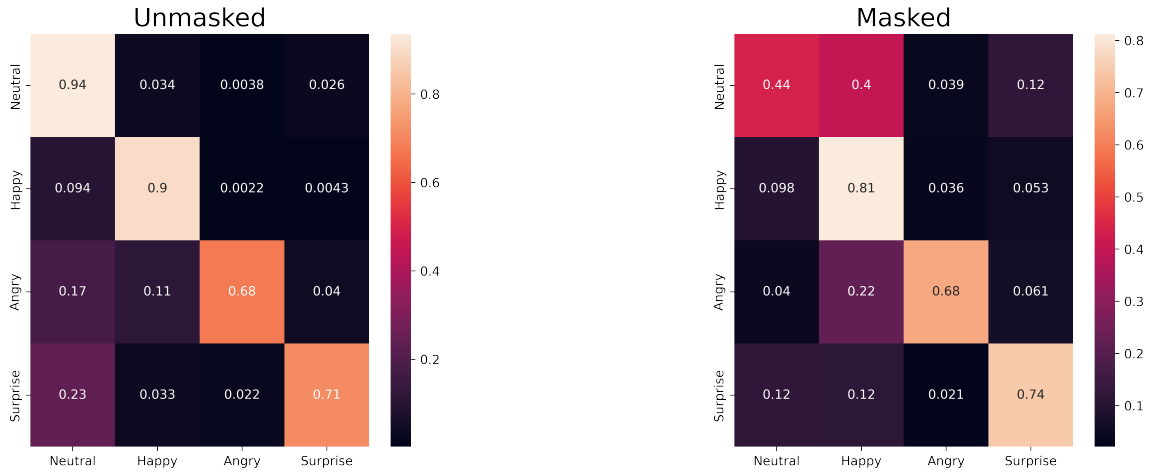
4.2.2 For 4 expressions

When the model is assessed over 4 emotions i.e. 'Neutral', 'Happy', 'Angry' and 'Surprise', the confusion matrices derived for the AffectNet dataset are displayed in Fig. 4.4. The classification of unmasked images as shown in Fig. 4.4a is even and the emotion 'Happy' is determined the most number of times. The classification of masked images (Fig. 4.4b) is comparatively dispersed with the emotion 'Angry' being accurately identified and the detection of the emotion 'Neutral' has quite vague. This indicates that the model does find it difficult for masked images of the dataset than the unmasked images.



(a) Unmasked Images **(b)** Masked Images

Figure 4.4: Confusion matrix for **AffectNet** dataset for 4 emotions

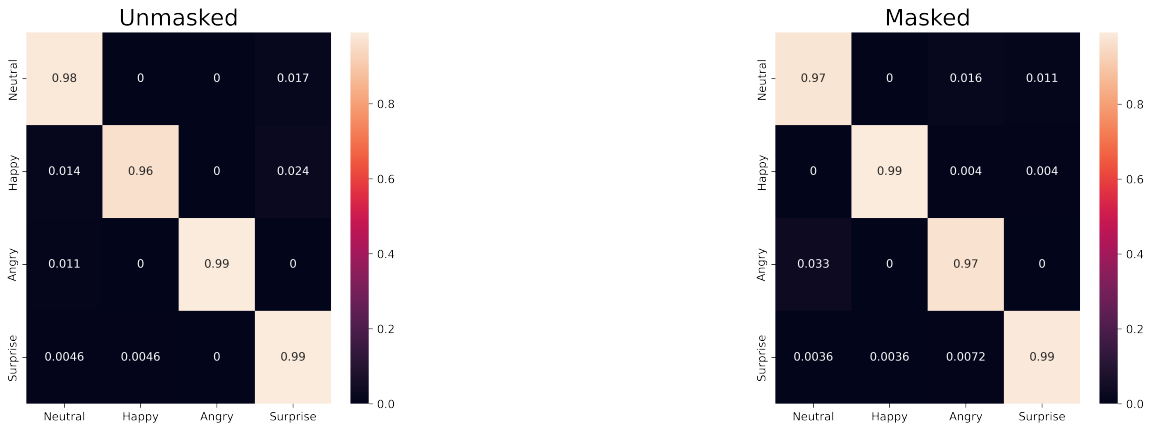


(a) Unmasked Images

(b) Masked Images

Figure 4.5: Confusion matrix for **RAF-DB** dataset for 4 emotions

Similarly, for RAF-DB dataset the emotions predicted are distinct for unmasked and scattered for masked images as shown in Fig. 4.5. The classification of unmasked images as shown in Fig. 4.5a is even, though the emotions 'Angry' and 'Surprise' are identified the lesser number of times than the emotions 'Neutral' and 'Happy'. In contrast, the classification of masked images is relatively crude, refer Fig. 4.5b. The emotion 'Neutral' is weakly classified and is confused with 'Happy' emotion although, rest of the emotions are fairly accurate in categorization. For in-the-lab dataset that is CK+, the categorization of emotions is near perfection, for both masked and unmasked images as shown in Fig. 4.6



(a) Unmasked Images **(b)** Masked Images

Figure 4.6: Confusion matrix for **CK+** dataset for 4 emotions

Corpora	Unmasked Images	Masked Images
AffectNet	72.28	61.55
RAF-DB	86.43	71.66
CK+	99.01	98.20

Table 4.2: Test accuracy (%) for **AffectNet**, **RAF-DB** and **CK+** datasets evaluated over 4 emotions

4.3 Discussion

The results obtained for 7 standard emotions generally suggests the emotions are sparsely classified as observed from the confusion matrices (Fig. 4.1, Fig. 4.2, Fig. 4.3). Although the emotion that was classified accurately across all datasets is 'Happy'. Furthermore, as seen in Table 4.1, testing accuracies are in the lower range. For in-the-wild datasets such as AffectNet and RAF-DB, model performance is good for RAF-DB dataset. Under ideal conditions (trained on unmasked images) the test accuracy is 75.8% and when trained on masked images the test accuracy is 60.5%. In comparison with AffectNet dataset which consists of highest number of images, the ideal performance is 55.5% and under occluded conditions, it performs poorly with 46.9% as test accuracy. When it comes to the in-lab dataset: CK+, a dataset

with the least number of images, the results are outstanding. Even under occluded conditions of the images, the model performs 95.1%. Although the cross-dataset evaluation i.e, training the model on CK+ dataset and testing with AffectNet or RAF-DB dataset, resulted in poor performance.

According to the confusion matrices with reference to Fig. 4.4, Fig. 4.5, Fig. 4.6, the findings obtained for four emotions imply that the emotions are categorised better, except for masked images of RAF-DB dataset. Here, the performance of the model for AffectNet dataset, has evidently improved. The emotions are classified distinctly as compared to the performance when model is trained for 7 standard emotions. Similar trend is seen for RAF-DB dataset as well. But interestingly, the performance of model for CK+ dataset has almost reached the perfect classification streak. Moreover, the testing accuracy has reformed for 4 emotions as shown in Table.4.2. The model performance among in-the-wild datasets is good for the RAF-DB dataset, yet again. The test accuracy is 86.43% under optimal conditions (trained on unmasked pictures) and 71.66% when trained on masked images. As compared to the Affectnet dataset, which has the most images, the optimum performance is 72.28%, and in obstructed conditions, it performs badly with a test accuracy of 61.55% yet better than that for 7 emotions. This pattern of improving test performance shows that the hypothesis stated at the outset is correct.

From the above discussion while referring to the results from Table.4.1 and Table.4.2, it can be seen that the overall performance of the model has advanced when trained over 4 emotions. The in-depth analysis of the confusion matrices for 7 and 4 emotions also suggest that the classification is more discrete for the latter.

4.3.1 Comparison

Apart from the comparison among the obtained results, it is also ideal to bring in the results of the original authors of ACNN model. In their work, experimental conditions for testing the model's performance under occluded conditions slightly differ. The occlusions considered were synthetic in nature and not particularly concentrated on occlusion due to face masks. The synthetic occlusions created as shown in Fig. 4.7, still provide some details about the lower part of the face. As a result, their outcomes are superior with testing accuracy of 54.84% and 80.54% for masked images of AffectNet and RAF-DB datasets respectively. Moreover the main idea of the paper was not about facial emotion recognition behind facial masks rather general occlusion on the face was considered. Because the problem is so new, there hasn't been much research done on facial expression recognition behind

facial masks that is comparable to the implementation and results of this project. The complete summary of the results and possible conclusions are discussed in next section 5

Figure 4.7: Synthetic occlusions on images included in the paper by Yong Li et al.

5 Conclusion

After obtaining the results and discussing the significant observations, a conclusion for the findings is drawn in this section. The possible challenges to such outcomes have been gathered.

5.1 Challenges

The implemented model performs well when trained on in-the-lab dataset and comparatively inadequate when trained on in-the-wild dataset. Possible explanations for the decreased range of accuracy in masked images of in-the-wild datasets include:

Scarcity in corpora: Despite the fact that at least 10,000 images from each in-the-wild datasets were utilized for training, the dispersion of the images across each emotion is significantly less.

Misaligned images: The images are acquired while the subject is actually 'in-the-wild'. If the subject in the photograph is facing away, it is almost guaranteed that not much meaningful information may be retrieved (Fig.5.1), eventually leading to poor learning.

Occluded images: Occlusion from hair, sunglasses, or hands is already visible in a number of photos (Fig.5.1). Overcoming these obstructions, in addition to the occlusion caused by the face mask, might result in poor model performance.

Figure 5.1: Variety of Images from in-the-wild datasets

5.2 Summary and Future Scope

Furthermore, the experiments conducted over 7 and 4 emotions lead to the conclusion that the perception of Neutral, Happy, Angry, and Surprise emotions is simpler for faces wearing face masks. Use of only these emotions is critical in daily life yet for simple yes/no or like/dislikes assessments these emotions as a response can be exploited. Despite the fact that the presented results have a certain credence toward what emotions are clearly detected, the overall performance is not as expected. The model must be trained on larger datasets with a large number of clear images. Of course, improvements in the domain of FER under partial occlusion are still required because other better alternative techniques out there can learn the features significantly better for in-the-wild datasets.

Bibliography

[09] *Insight Now by nViso*. https://www.nviso.ai/en/insights-now. Insights Now by nViso is a software that can be used to recognize human emotions by analyzing facial expressions and eye movements using 3D imaging technology. 2009 (cit. on p. 13).

[12] *Kairos*. https://www.kairos.com/. Kairos is a computer vision software that claims to identify and verify faces in videos and photos. 2012 (cit. on p. 13).

[APD10] N. Aifanti, C. Papachristou, A. Delopoulos. "The MUG facial expression database." In: *11th International Workshop on Image Analysis for Multimedia Interactive Services WIAMIS 10*. 2010, pp. 1–4 (cit. on p. 18).

[AY12] R. Azmi, S. Yegane. "Facial expression recognition in the presence of occlusion using local Gabor binary patterns." In: *ICEE 2012 - 20th Iranian Conference on Electrical Engineering* (2012), pp. 742–747. DOI: 10.1109/IranianCEE.2012.6292452 (cit. on p. 17).

[BT17] A. Bulat, G. Tzimiropoulos. "How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks)." In: *International Conference on Computer Vision*. Vol. 1. 6. 2017, p. 8 (cit. on p. 24).

[Cot11] S. F. Cotter. "Recognition of occluded facial expressions using a Fusion of Localized Sparse Representation Classifiers." In: *2011 Digital Signal Processing and Signal Processing Education Meeting, DSP/SPE 2011 - Proceedings* (2011), pp. 437–442. DOI: 10.1109/DSP-SPE.2011.5739254 (cit. on p. 17).

[GES21] F. Grundmann, K. Epstude, S. Scheibe. "Face masks reduce emotion-recognition accuracy and perceived closeness." In: *PLOS ONE* 16.4 (Apr. 2021), pp. 1–18. DOI: 10.1371/journal.pone.0249792. URL: https://doi.org/10.1371/journal.pone.0249792 (cit. on p. 14).

[GSVV21] A. Greco, A. Saggese, M. Vento, V. Vigilante. "Performance Assessment of Face Analysis Algorithms with Occluded Faces." In: *Pattern Recognition. ICPR International Workshops and Challenges*. Ed. by A. Del Bimbo, R. Cucchiara, S. Sclaroff, G. M. Farinella, T. Mei, M. Bertini, H. J. Escalante, R. Vezzani. Cham: Springer International Publishing, 2021, pp. 472–486 (cit. on p. 13).

[Jes19] A. de Jesus. "Artificial Intelligence in Video Marketing – Emotion Recognition, Video Generation, and More." In: https://emerj.com/ai-sector-overviews/artificial-intelligence-for-video-marketing-emotion-recognition-video-generation-and-more/, 2019 (cit. on p. 13).

[LCK+10] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, I. Matthews. "The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression." In: *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*. 2010, pp. 94–101. DOI: 10.1109/CVPRW.2010.5543262 (cit. on pp. 22, 25).

[LDD17] S. Li, W. Deng, J. Du. "Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild." In: *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*. IEEE. 2017, pp. 2584–2593 (cit. on p. 22).

[LKG98] M. Lyons, M. Kamachi, J. Gyoba. "The Japanese Female Facial Expression (JAFFE) Dataset." In: (Apr. 1998). DOI: 10.5281/zenodo.3451524. URL: https://doi.org/10.5281/zenodo.3451524 (cit. on p. 17).

[LZSC19] Y. Li, J. Zeng, S. Shan, X. Chen. "Occlusion Aware Facial Expression Recognition Using CNN With Attention Mechanism." In: *IEEE Transactions on Image Processing* 28.5 (2019), pp. 2439–2450. ISSN: 10577149. DOI: 10.1109/TIP.2018.2886767 (cit. on pp. 18, 19, 21).

[MAP+21] M. Marini, A. Ansani, F. Paglieri, F. Caruana, M. Viola. "The impact of facemasks on emotion recognition, trust attribution and re-identification." In: *Scientific Reports* 11.1 (2021), pp. 1–14. ISSN: 20452322. DOI: 10.1038/s41598-021-84806-5. URL: https://doi.org/10.1038/s41598-021-84806-5 (cit. on p. 14).

[MHM17] A. Mollahosseini, B. Hasani, M. H. Mahoor. "AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild." In: *IEEE Transactions on Affective Computing* PP.99 (2017), pp. 1–1 (cit. on p. 22).

[NAR+20] R. A. Naqvi, M. Arsalan, A. Rehman, A. U. Rehman, W.-K. Loh, A. Paul. "remote sensing Deep Learning-Based Drivers Emotion Classification System in Time Series Data for Remote Applications." In: (2020). DOI: 10.3390/rs12030587. URL: www.mdpi.com/journal/remotesensing (cit. on p. 13).

[Pre12] L. Prechelt. "Early Stopping — But When?" In: *Neural Networks: Tricks of the Trade: Second Edition*. Ed. by G. Montavon, G. B. Orr, K.-R. Müller. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 53–67 (cit. on p. 29).

[RP18] J. Y. Ramirez Cornejo, H. Pedrini. "Emotion Recognition from Occluded Facial Expressions Using Weber Local Descriptor." In: *International Conference on Systems, Signals, and Image Processing* 2018-June (2018). ISSN: 21578702. DOI: 10.1109/IWSSIP.2018.8439631 (cit. on p. 17).

[SNO+15] M. Stanković, M. Nešić, J. Obrenović, D. Stojanović, V. Milošević. "Recognition of facial expressions of emotions in criminal and non-criminal psychopaths: Valence-specific hypothesis." In: *Personality and Individual Differences* 82 (Aug. 2015), pp. 242–247. ISSN: 01918869. DOI: 10.1016/j.paid.2015.03.002 (cit. on p. 13).

[SZ15] K. Simonyan, A. Zisserman. *Very Deep Convolutional Networks for Large-Scale Image Recognition*. 2015. arXiv: 1409.1556 [cs.CV] (cit. on p. 21).

[TDS+20] D. Tanna, M. Dudhane, A. Sardar, K. Deshpande, N. Deshmukh. "Sentiment Analysis on Social Media for Emotion Classification." In: *2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS)*. 2020, pp. 911–915. DOI: 10.1109/ICICCS48265.2020.9121057 (cit. on p. 13).

[ZLY+12] L. Zhong, Q. Liu, P. Yang, B. Liu, D. Metaxas. "Learning Multiscale Active Facial Patches for Expression Analysis." In: vol. 45. June 2012, pp. 2562–2569. ISBN: 978-1-4673-1226-4. DOI: 10.1109/CVPR.2012.6247974 (cit. on p. 24).

[ZVTC18] L. Zhang, B. Verma, D. Tjondronegoro, V. Chandran. "Facial expression analysis under partial occlusion: A survey." In: *ACM Computing Surveys* 51.2 (2018), pp. 1–36. ISSN: 15577341. DOI: 10.1145/3158369. arXiv: 1802.08784 (cit. on p. 17).

All links were last followed on December 10th, 2021.

Declaration

I hereby declare that the work presented in this thesis is entirely my own and that I did not use any other sources and references than the listed ones. I have marked all direct or indirect statements from other sources contained therein as quotations. Neither this work nor significant parts of it were part of another examination procedure. I have not published this work in whole or in part before. The electronic copy is consistent with all submitted copies.

---

place, date, signature