

# Facial Emotion Image Classification based on Face orientation

by

Zhansaya Sovetbek

Submitted to the Department of Data Science  
in partial fulfillment of the requirements for the degree of

Master of Science in Data Science

at the

NAZARBAYEV UNIVERSITY

April 2023

© Nazarbayev University 2023. All rights reserved.

Author .....  
Department of Data Science  
7 May, 2023

Certified by .....  
Martin Lukac  
Associate Professor at Nazarbayev University, PhD  
Thesis Supervisor

Accepted by .....  
Vassilios Tourassis  
Dean, School of Engineering and Digital Sciences, PhD



# Facial Emotion Image Classification based on Face orientation

by

Zhansaya Sovetbek

Submitted to the Department of Data Science  
on 7 May, 2023, in partial fulfillment of the  
requirements for the degree of  
Master of Science in Data Science

## Abstract

Facial emotion recognition has received increasing attention in recent years due to its potential applications in various fields such as human-computer interaction, security, and healthcare. In this context, the orientation of a face has been identified as an important factor affecting the accuracy of facial emotion recognition.

Emotions can be in different forms, this thesis will focus on facial emotional expressions. The importance of facial emotion recognition is crucial in daily life because it can be used to help people in case of emergency or for quick crime prevention.

Facial Emotion recognition could be used in various applications such as HCI, driver warning systems, automated tutoring systems, picture and video retrieval, and smart surroundings.

Two methodological approaches were used in this research: the baseline model and the proposed model. All two models classify the face orientation directions and facial emotions. The models will use Hopenet to identify head pose direction angles such as pitch, yaw, and roll then to determine one of the directions, namely, forward, left, right, up, and down.

Pre-trained models such as MobileNetV3-small, ResNet-18, GoogleNet, and others will be used to classify emotions and find the connection between facial emotion classification and head pose orientation.

Thesis Supervisor: Martin Lukac

Title: Associate Professor at Nazarbayev University, PhD



# Acknowledgments

I would like to express my sincerest gratitude to my thesis supervisor Professor Martin Lukac for his guidance, expertise, and unwavering support throughout my academic journey. His encouragement and mentorship have been invaluable in shaping my growth and development.

I would also like to thank my family for their unwavering love, encouragement, and support. Their belief in me and my goals has been a constant source of inspiration, and I am grateful for their unwavering presence in my life.



# Contents

<b>1</b>	<b>Introduction</b>	<b>13</b>
1.1	History of Emotion Recognition . . . . .	13
1.2	Motivation . . . . .	14
1.3	Problem statement . . . . .	15
<b>2</b>	<b>Related works</b>	<b>17</b>
2.1	The Facial Emotion Detection Problems . . . . .	17
2.2	Datasets . . . . .	18
2.3	Models . . . . .	19
2.3.1	Models on Affectnet . . . . .	19
2.3.2	Pre-trained models . . . . .	22
2.3.3	Loss Function . . . . .	27
2.3.4	Pitch, Yaw, Roll . . . . .	28
<b>3</b>	<b>Methodology</b>	<b>33</b>
3.1	Methodological approach . . . . .	33
3.2	Preprocessing . . . . .	33
3.2.1	Imbalanced Dataset . . . . .	33
3.2.2	Balanced Dataset . . . . .	36
3.3	Baseline model . . . . .	37
3.4	Proposed model . . . . .	40
3.5	Performance Evaluation Metrics . . . . .	42

<b>4</b>	<b>Results</b>	<b>45</b>
4.1	Baseline model results . . . . .	45
4.2	Proposed model results . . . . .	49
4.3	Baseline model vs Proposed model . . . . .	52
<b>5</b>	<b>Conclusion</b>	<b>55</b>



# List of Figures

1-1	Different facial emotions . . . . .	14
2-1	Different models used on Affectnet and their accuracy . . . . .	19
2-2	Generated results by VGG-FACE model using StarGAN and Ganimation	20
2-3	The MT-ArcRes network that has been trained with the ArcFace loss	21
2-4	Visualization of children's emotions predicted by EfficientNet-B2 . . .	22
2-5	MobileNet-V3 small Architecture . . . . .	24
2-6	MobileNetV3 main block . . . . .	25
2-7	Head Orientation with Pitch, Yaw and Roll . . . . .	29
3-1	Thesis Workflow . . . . .	34
3-2	Extended Thesis Workflow . . . . .	35
3-3	Bad samples of datasets, that can not be used . . . . .	35
3-4	Pitch, yaw, roll calculation and visualization for each direction (pitch- green, roll-red, yaw-blue) . . . . .	37
3-5	Neural Network Architecture . . . . .	41
3-6	Neural Network Accuracy Diagram . . . . .	42
3-7	Neural Network Loss Diagram . . . . .	43
4-1	Baseline model Confusion matrix for MobileNetV3-small . . . . .	46
4-2	Baseline model Loss Diagram for MobileNetV3-small . . . . .	47
4-3	Baseline model Accuracy Diagram for MobileNetV3-small . . . . .	48
4-4	Proposed model Accuracy Diagram for MobileNetV3-small . . . . .	49
4-5	Proposed model Accuracy Diagram for MobileNetV3-small . . . . .	50

4-6	Proposed model Accuracy Diagram for MobileNetV3-small . . . . .	51
4-7	Visualisation of results of mostly classified directions by emotions . .	53

# List of Tables

2.1	VGG-Face model comparison with other models . . . . .	20
2.2	Retrained multi-task networks with ArcFace loss, for expression recognition . . . . .	21
2.3	Accuracy for the AffectNet Validation Set . . . . .	22
3.1	Raw Imbalanced Affectnet distribution . . . . .	34
3.2	Percentage Distribution of Raw Imbalanced dataset . . . . .	36
3.3	Pitch, Yaw, Roll values for Anger images for direction "Down" . . . .	38
3.4	Overall Direction Classification by each model . . . . .	40
3.5	Confusion Matrix . . . . .	43
4.1	Overall Emotion Classification by each pre-trained model . . . . .	45
4.2	Direction by Emotion true labeled table for Baseline Model . . . . .	47
4.3	The highest percentage of direction per each emotion(by column) . .	48
4.4	The highest percentage of direction per emotion (by row) . . . . .	48
4.5	Overall Emotion Classification by each pre-trained model . . . . .	49
4.6	Direction by Emotion true labeled table for Proposed Model . . . . .	51
4.7	The highest percentage of direction per each emotion(by column) . .	52
4.8	The highest percentage of direction per emotion (by row) . . . . .	52



# Chapter 1

## Introduction

### 1.1 History of Emotion Recognition

With the rise of facial recognition technology, there has been a growing interest in developing algorithms that can accurately detect and interpret human emotions. Facial Emotion Recognition is a technology that examines emotions through images and videos [1].

The computer should decode the human facial emotion and find which emotion it refers to. Starting from the early 1800's scientists did much research on facial emotion decoding. Based on physiological engagement, William James identified four primary emotions in 1890: fear, sadness, love, and fury [2].

Over 1 century, the list of emotions was enlarged to 27 different types of emotion. The researchers from the University of California, Berkeley, in 2017 found different types of emotion: admiration, adoration, aesthetic appreciation, amusement, anger, anxiety, awe, awkwardness, boredom, calmness, confusion, craving, disgust, empathic pain, entrancement, excitement, fear, horror, interest, joy, nostalgia, relief, romance, sadness, satisfaction, sexual desire, and surprise [3].

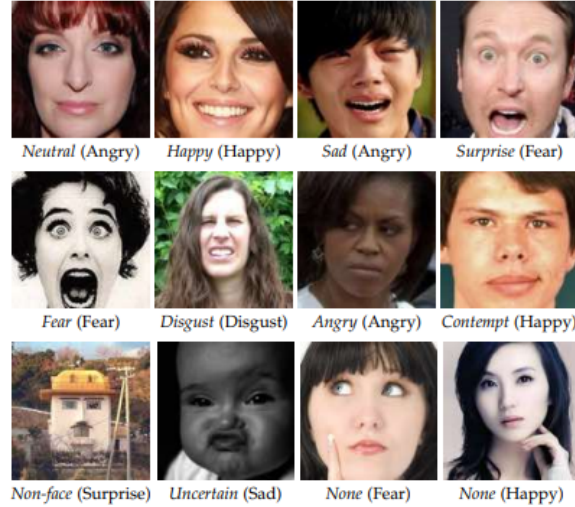


Figure 1-1: Different facial emotions

## 1.2 Motivation

Currently, the ability to use a computer to evaluate a detected facial image has increasingly gained popularity as artificial intelligence and computer technology have advanced.

Nowadays, most banks, police, and other governmental structures use face detection technologies for identity verification. However, the potential of these technologies is wider, than it is expected. The capability to accurately interpret human emotions from facial expressions has the ability to revolutionize many industries, including healthcare, marketing, and customer service.

For example, Emotion Recognition could be used to determine, does a human need medical help. Sometimes people could have pain, seizures, and other signs of illnesses, and the first medical aid should be applied immediately. The camera by using Emotion Recognition could recognize illness signs and call an ambulance. The time of waiting for help plays an important role in saving a life. Therefore, this process could be automatized.

Also, Emotion Recognition could be used to understand people with emotional disabilities. Better understanding and interpretation of emotions could prevent future misunderstandings among people. For example, a person with an emotional disability

may have problems with facial expressions, and other people may understand his/her behavior not correctly. So, by using Emotion Recognition the conflicts between people could be resolved.

In addition, Emotion Recognition could help to detect any suspicious movements in public places. For instance, in airports understanding the facial emotion of people could prevent serious crimes, such as drug dealing, human trafficking, terrorism and etc.

Solving this problem requires advanced machine-learning techniques, innovative algorithms, and a deep understanding of human emotions.

### 1.3 Problem statement

The task of classifying facial emotions in images based on face orientation is a complex challenge in computer vision, machine learning, and deep learning. The utilization of techniques for identifying emotions through surveillance cameras presupposes particular circumstances concerning the positioning of a person's face in various angles in relation to the camera.

For example, it involves accurately recognizing the emotions portrayed in an image while considering the angle and position of the face, as such variations can affect the accuracy of recognition.

Thus, to tackle this problem, different machine learning models should be checked for facial emotion image classification with different orientations.

The main goal of this research is to explore state-of-the-art recognition works and investigate the impact of both facial orientation and emotional category on the accuracy of recognition.







# Chapter 2

## Related works

### 2.1 The Facial Emotion Detection Problems

Firstly, **Large intra-class variances** brought on by elements like lighting and position changes, occlusion, and head movement make it difficult to identify facial emotions in uncontrolled conditions. Two crucial criteria often determine a face emotion detection system's accuracy:

- 1) extraction of facial characteristics that are resilient to intra-class differences (such changes in stance) yet are unique for diverse emotions
- 2) design of a classifier capable of identifying various facial expressions from noisy and unreliable data (e.g., illumination changes and occlusion)[4].

Next, the main problem is **Facial Emotion Feature Extraction**. In feature extraction and facial expression identification, precise localization of a facial feature is crucial. However, in practical applications, it is challenging to find the face feature exactly because of the variation in the facial form and the quality of the database image [5].

Similar to the challenge of **Image segmentation** from specific image areas, it is difficult to extract the feature of facial emotions. **Model-based segmentation** and **Image-based segmentation** are the two types of segmentation techniques.

The "**Live-wire approach**" and other image segmentation techniques are greater at determining feature boundaries than they are at recognizing boundaries, which makes it hard to extract feature regions by matching object areas over the whole image [6] [7].

## 2.2 Datasets

**AffectNet** is a new face expression database that was created by gathering and annotating facial photos. AffectNet has more than 1 million facial photos that were gathered from the Internet using 1250 emotion-related keywords in six different languages and three major search engines. A categorical model was used to manually annotate the existence of seven distinct facial expressions

("happy", "sad", "surprise", "fear", "disgust", "anger", "contempt") with "neutral" expression as well as the degree of valence and arousal in around half of the recovered photos (440K) (dimensional model). The research on automated facial expression identification in two separate emotion models is made possible by AffectNet, which is by far the biggest library of facial expressions, valence, and arousal in the wild. To categorize pictures in the category model and forecast the strength of valence and arousal, two baseline deep neural networks were employed [8].

**The Pointing'04** head-position image database, a collection of images, was created to help with head pose estimate studies in the field of computer vision. Researchers and academicians all over the world utilize the database, which was developed by Dr. Bernd Heisele, a computer vision expert. 15,000 images of people's faces from various perspectives, with various lighting and facial expressions, are included in the database. The photos were taken using a high-quality digital camera, and they have labels on them that describe the subject's head stance, including where their head is located in three dimensions and which way they are facing [20].

**The AFLW2000-3D** is a dataset for computer vision that includes 2,000 3D

annotated face photos. Researchers from the Max Planck Institute for Informatics and Saarland University in Germany published this dataset in 2017.

The photos in the collection are of different persons and were taken under varied lighting and expression-related circumstances. The left and right eye centers, the left and right mouth corners, the tip of the nose, and the chin are among the six important facial landmarks whose 3D coordinates are included in the annotations for each image [23].

## 2.3 Models

### 2.3.1 Models on Affectnet

Most of the research on Facial Emotion Recognition is done using the Affectnet dataset. Therefore, the following models are used in emotion recognition using this dataset.

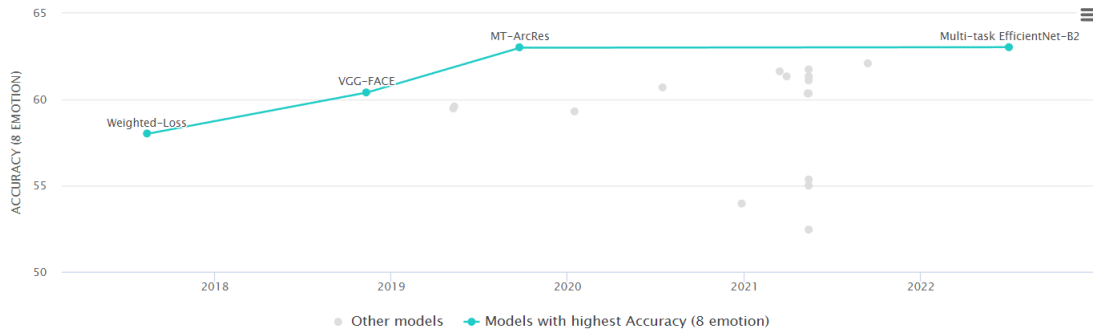


Figure 2-1: Different models used on Affectnet and their accuracy

Firstly, the Weighted-Loss model presented by Affectnet authors in 2017 showed the lowest accuracy, reaching 58%. The loss function for each class was weighted in the weighted-loss technique according to their relative share in the training dataset. In other words, the loss function severely penalizes the networks when they incorrectly identify instances from classes that are underrepresented while only lightly punishing them when they incorrectly categorize examples from classes that are highly represented.

Secondly, the VGG-FACE model presented by [9] showed 60.4% which is slightly higher than the Weighted-Loss model. It can be seen that the VGG-FACE network outperformed AffectNet’s database baseline model.

Table 2.1 confirms that the network trained using the proposed methodology(VGG-FACE model) outperformed all other networks. This improvement was significant across all evaluation criteria, as compared to the VGG-FACE baseline network, and was spread across the VA space. The reason for this improvement can be attributed to the large number of synthesized images used during training, which helped the network to train and generalize better. This was necessary because the training set had an insufficient representation of many ranges, which the synthesized images helped to overcome.

In Figure 2-2, it can be seen the visual representation of the VGG-FACE model using StarGAN, and Ganimation, which were presented in Table 2.1.

Networks	CCC	CCC	P-CC	P-CC	SAGR	SAGR	MSE	MSE
	Valence	Arousal	Valence	Arousal	Valence	Arousal	Valence	Arousal
AlexNet [42]	0.60	0.34	0.66	0.54	0.74	0.65	0.14	0.17
the VGG-FACE baseline	0.50	0.37	0.54	0.48	0.65	0.60	0.19	0.18
VGG-FACE trained using StarGAN	0.55	0.42	0.58	0.49	0.74	0.73	0.17	0.16
VGG-FACE trained using Ganimation	0.56	0.45	0.59	0.51	0.74	0.74	0.15	0.16
VGG-FACE trained using the proposed approach	0.62	0.54	0.66	0.55	0.78	0.75	0.14	0.15

Table 2.1: VGG-Face model comparison with other models

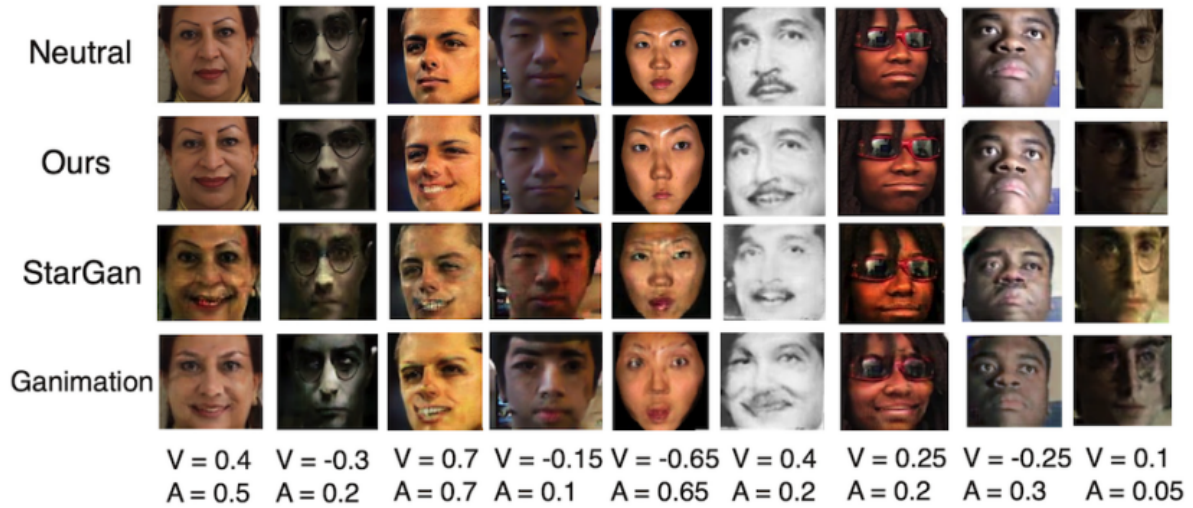


Figure 2-2: Generated results by VGG-FACE model using StarGAN and Ganimation

Thirdly, the MT-ArcRes model revealed better results, namely, 63% in facial emotion recognition in Figure 2-1. MT-ArcRes model showed the highest result in comparison with other models, as can be seen in Table 2.2. This model was introduced by Kollias et al [10] later in 2019.

Figure 2-3 shows the Multi-Task-ArcFace-Residual (MT-ArcRes) CNN architecture, which employs only residual units and it does not contain VGG-Face’s layers.

Databases	MT-ArcRes	MT-ArcVGG	FT-MT-VGG	MT-VGG	AlexNet	DLP-CNN	VGG
AffectNet	0.63	0.62	0.59	0.54	0.58	-	-
RAF-DB	0.75	0.76	0.71	0.61	-	0.74	-
IMFDB	0.55	0.56	0.51	0.42	-	-	-
FER2013	0.8	0.79	0.78	0.76	-	-	0.75

Table 2.2: Retrained multi-task networks with ArcFace loss, for expression recognition

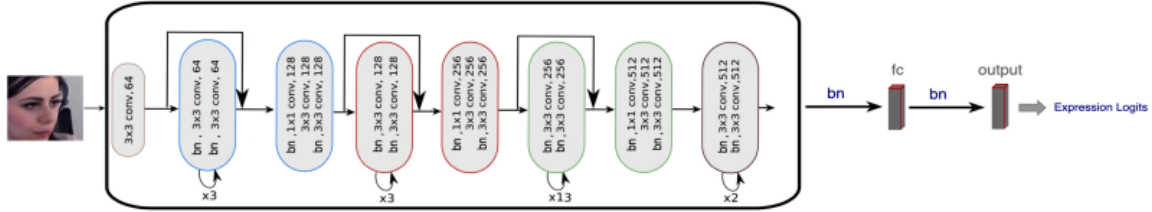


Figure 2-3: The MT-ArcRes network that has been trained with the ArcFace loss

Finally, the current leader in facial emotion recognition is Multi-task EfficientNet-B2 in Figure 2-1, which showed 63.03%, was presented by Savchenko et al [11]. Multi-task EfficientNet-B2 was trained in two stages with (1) pre-training on face recognition, and (2) fine-tuning on emotion classification. It is crucial to note that the CNNs were trained on faces that were cropped and without borders, meaning that the majority of the backdrop, hair, etc., is not visible.

Since learned facial characteristics are better suited for emotional analysis, face recognition performance may somewhat decline as a result. The last layer of the network pre-trained on VGGFace2 may be thought of as an extractor of facial characteristics since the new head (fully-connected layer with C outputs and softmax activation) replaces the final layer of the network.

In Table 2.3 it can be seen that EfficientNet-B2 reached the highest accuracy result

on 8 classes and 7 classes classifications on the AffectNet dataset. Figure 2-4 displays both the anticipated emotional categories and a simple-to-understand GradCAM visualization of the CNN’s determination.

Model	Accuracy of 8 classes , %	Accuracy of 7 classes, %
Baseline (AlexNet)	58.0	-
Deep Attentive Center Loss	-	65.20
Distilled student	61.60	65.40
DAN	62.09	65.69
TransFER	-	66.23
EmotionGCN	-	66.46
MobileNet-v1	60.20	64.71
EfficientNet-B0	61.32	65.74
EfficientNet-B2	63.03	66.34

Table 2.3: Accuracy for the AffectNet Validation Set

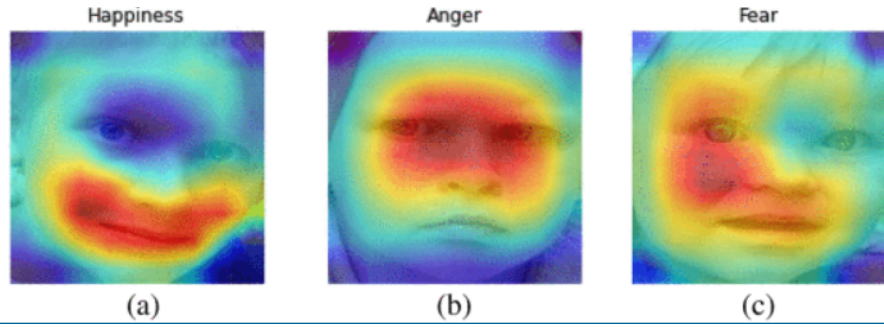


Figure 2-4: Visualization of children’s emotions predicted by EfficientNet-B2

### 2.3.2 Pre-trained models

Pre-trained models are machine learning models that have already been trained on large datasets, allowing developers to use them for various tasks without needing to train them from scratch. Pre-trained models can be used for image recognition, object detection, classification, and other tasks.

1) **MobileNet**: This is a pre-trained model for image classification that is optimized for mobile devices. It has a small footprint and low latency, making it ideal for applications where computational resources are limited.

A. G. Howard et al. [14] presented a new type of convolutional neural network (CNN) designed specifically for mobile devices. Traditional CNNs are computationally intensive and require large amounts of memory, which can be problematic for mobile devices with limited processing power and storage capacity. MobileNets, on the other hand, use a lightweight architecture that is optimized for mobile devices, achieving high accuracy with significantly fewer parameters and computation.

MobileNets has key architectural features such as depthwise separable convolutions, which break down the traditional convolution operation into two simpler operations, and linear bottlenecks, which reduce the dimensionality of intermediate feature maps. MobileNets could perform well on a range of computer vision tasks, including image classification, object detection, and semantic segmentation while requiring less computation and memory than traditional CNNs.

By reducing the computational and memory requirements of CNNs, MobileNets enable a wider range of mobile vision applications, from real-time object recognition to augmented reality and more.

MobileNetV3-Small is built using a combination of depthwise separable convolutions, which split the convolutional operation into a depthwise and a pointwise convolution, and linear bottlenecks, which help reduce the number of parameters and improve the efficiency of the model.

Additionally, the architecture of MobileNetV3-Small in Figure 2-5 includes a series of inverted residual blocks with linear bottlenecks and a squeeze-and-excitation module, which helps the model focus on the most informative features. Figure 2-6 shows MobileNetV3 main block in detail the sequence of operations, including a depthwise separable convolution, a non-linear activation function, and a linear bottleneck layer. Compared to other pre-trained models, MobileNetV3-Small is smaller in size and faster in inference, making it well-suited for use in mobile and embedded devices. However, its smaller size also means that it may not perform as well as larger models on more complex tasks, and may require additional training or fine-tuning for specific use cases.

2) **GoogleNet:** GoogleNet is a pre-trained model that uses an inception module



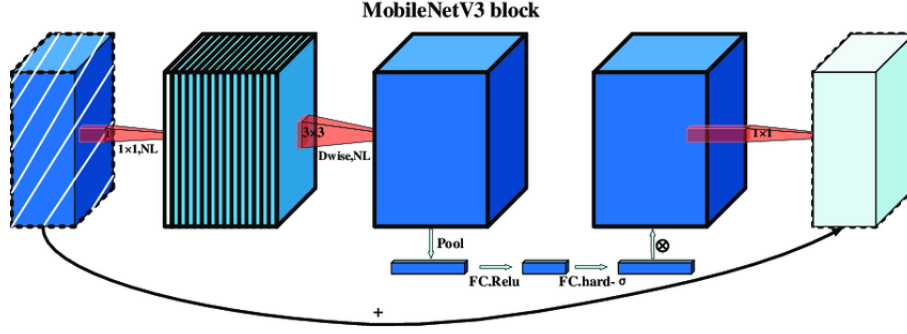


Figure 2-5: MobileNet-V3 small Architecture

to increase the efficiency of the network and it was designed by Google for the ImageNet Large Scale Visual Recognition Challenge in 2014 [15].

GoogleNet is a multi-stage network, in which a variety of convolutional filters were applied in parallel at each stage, followed by pooling, and concatenated to form a rich feature representation for the input data.

This architecture allows for deeper and more efficient learning of complex visual patterns, leading to state-of-the-art performance on image classification tasks such as the ImageNet challenge.

Additionally, several optimizations, including dimensionality reduction and the use of 1x1 convolutions, were used to reduce computational costs and improve performance.

3) **ResNet**: This is a pre-trained model that uses residual connections to allow for deeper networks while avoiding the vanishing gradient problem. It was designed by Microsoft [16] and won the ImageNet Large Scale Visual Recognition Challenge in 2015.

It was observed that as the depth of a neural network increased, the training error would eventually begin to increase rather than decrease, indicating that the network was becoming too difficult to optimize. They proposed the use of residual connections, which allowed for the training of much deeper networks by allowing information to bypass some layers.

4) **VGG**: This is a pre-trained model that uses a deep convolutional neural network to achieve high accuracy in image recognition tasks. K. Simonyan et al. [17] proposed a novel deep learning architecture called VGGNet for image recognition tasks. The

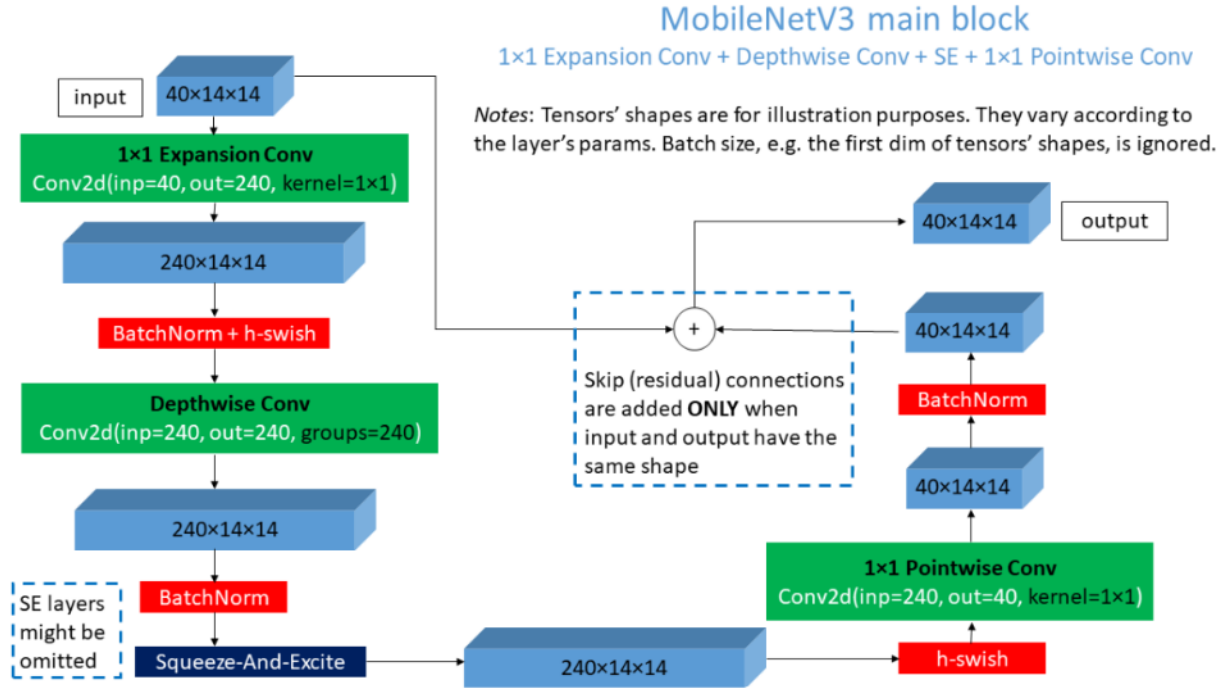


Figure 2-6: MobileNetV3 main block

authors address the problem of improving the performance of convolutional neural networks (CNNs) by increasing the depth of the network, which has been shown to improve accuracy.

VGGNet is a deep CNN architecture that achieves state-of-the-art performance on the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) 2014 dataset. The model, with up to 19 layers, outperforms previous CNN models with fewer layers, including the 2012 AlexNet model, which was the previous winner of the ILSVRC.

5) **AlexNet**: Alex Krizhevsky et al [21] presented one of the first convolutional neural networks to achieve state-of-the-art performance on the ImageNet dataset, which won the ImageNet Large Scale Visual Recognition Challenge in 2012.

A large-scale image classification system, which is called AlexNet, and was trained on a dataset of over one million images from the ImageNet dataset. There were several key architectural innovations that made AlexNet's CNN particularly effective, including the use of Rectified Linear Units (ReLU) for activation functions, local response normalization, and overlapping pooling. They also employed data augmentation techniques to prevent overfitting and improve generalization.

6) **AdaBoost**: This is a pre-trained model that is used for classification tasks. It is an ensemble learning method that combines multiple weak learners into a strong classifier. AdaBoost was presented by Freund et al [22] and is commonly used in computer vision and often used in object detection tasks.

The concept of "online learning," where the learning algorithm receives input data in sequential order and makes predictions or decisions without having access to the entire dataset, was presented. Then a decision-theoretic framework for online learning was developed, where the learning algorithm minimizes a loss function that measures the difference between the predicted output and the true output.

The AdaBoost algorithm uses a set of weak learners (i.e., classifiers that perform slightly better than random guessing) and combines them in a way that improves their overall accuracy. AdaBoost assigns weights to each training example based on its difficulty and trains a sequence of weak learners on the weighted examples. The final prediction is then obtained by combining the predictions of all the weak learners.

7) **HopeNet** is a deep learning model that is pre-trained to predict the 3D head pose of a person from a single 2D image or a video frame. The model is based on a convolutional neural network (CNN) architecture and was developed by Angelos Nicolaou, et al. in 2017 [12] [13].

The model takes as input a single image or a video frame and produces three continuous outputs corresponding to the pitch, yaw, and roll angles of the person's head. These angles represent the rotation of the head along the X, Y, and Z axes, respectively. Hopenet achieves state-of-the-art performance on several benchmark datasets, including the AFLW2000-3D dataset.

Hopenet is trained using a large-scale dataset of annotated images and videos of human faces. The training data is augmented to account for different lighting conditions, head poses, and facial expressions. The model is trained using a combination of regression and classification losses, with the goal of minimizing the angular error between the predicted and ground-truth head poses [26] .

Hopenet has several applications, including human-computer interaction, virtual and augmented reality, and driver monitoring systems. For example, it can be used to

control the movement of virtual characters in video games or to enable gaze tracking in virtual reality headsets. It can also be used to monitor the attention and fatigue levels of drivers by tracking their head movements.

### 2.3.3 Loss Function

CrossEntropyLoss is a commonly used loss function in machine learning, particularly in classification problems. It is a measure of the difference between the predicted and actual probability distributions of the target class.

In simple terms, CrossEntropyLoss is a measure of how well a model is able to predict the correct class for a given input. It is often used in deep learning models that are trained using stochastic gradient descent, where the goal is to minimize the loss function during training.

The CrossEntropyLoss is defined as follows:

$$L(y, \hat{y}) = -\sum_{i=1}^n y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)$$

where  $y$  is the true probability distribution of the target class,  $\hat{y}$  is the predicted probability distribution of the target class, and the summation is taken over all classes.

The CrossEntropyLoss is a measure of the amount of information needed to represent the true distribution using the predicted distribution. The smaller the CrossEntropyLoss, the better the model is at predicting the correct class.

It is important to note that the CrossEntropyLoss is not only used for binary classification problems but also for multi-class classification problems. In the case of multi-class classification, the predicted probability distribution  $\hat{y}$  is typically represented using a softmax function, which converts the output of the model into a probability distribution.

In summary, CrossEntropyLoss is a loss function commonly used in machine learning for classification problems. It measures the difference between the predicted and actual probability distributions of the target class and is often used in deep learning models trained using stochastic gradient descent.

### 2.3.4 Pitch, Yaw, Roll

**Pitch, yaw, and roll** are terms used to describe the orientation or rotation of an object, such as an aircraft, a spacecraft, or a camera. In head pose estimation, pitch, yaw, and roll values are used to describe the orientation of the human head in three-dimensional space.

**Pitch** refers to the rotation of the head around the X-axis, which runs horizontally from left to right. A positive pitch value indicates that the head is tilted downwards towards the chest, while a negative pitch value indicates that the head is tilted upwards towards the ceiling.

**Yaw** refers to the rotation of the head around the Y-axis, which runs vertically from top to bottom. A positive yaw value indicates that the head is turned to the right, while a negative yaw value indicates that the head is turned to the left.

**Roll** refers to the rotation of the head around the Z-axis, which runs perpendicular to the face. A positive roll value indicates that the head is tilted towards the right shoulder, while a negative roll value indicates that the head is tilted towards the left shoulder.

Together, these three values provide a complete description of the head orientation in three-dimensional space, allowing for accurate tracking of the position and movement of the head. Head pose estimation is used in a variety of applications, including computer vision, virtual reality, and augmented reality.

We can determine an object's rotational angles by measuring the circular feature's rotational angles using perspective projection.

Figure 2-7 illustrates that by rotating the head around the X, Y, and Z axes, the position of the head can be expressed using the corresponding Euler angles, namely  $\theta_x$  (Pitch),  $\theta_y$  (Yaw), and  $\theta_z$  (Roll) [25].

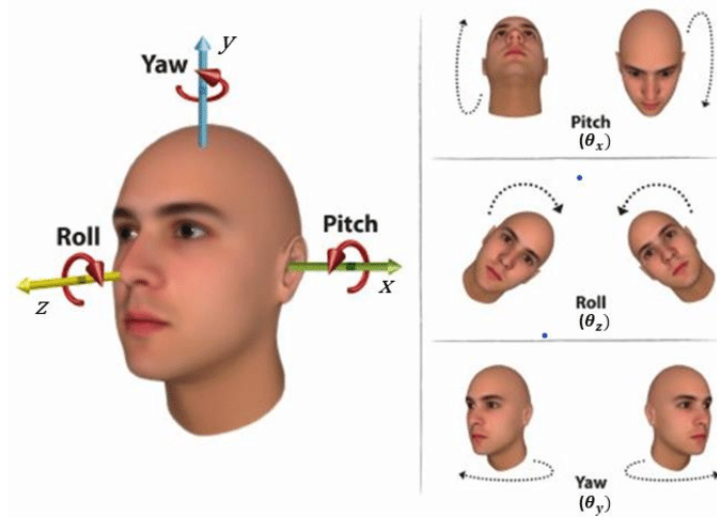


Figure 2-7: Head Orientation with Pitch, Yaw and Roll

The point's new coordinates will result from rotating it by an angle of  $\theta_x$  around the X-axis:

$$(x_x y_x z_x) = R_x \cdot (xyz)^T \quad (2.1)$$

where

$$R_x = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & \cos\theta_x & -\sin\theta_x & 0 \\ 0 & \sin\theta_x & \cos\theta_x & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad (2.2)$$

Similarly, when the point undergoes rotation around the Y and Z axes by angles of  $\theta_y$  and  $\theta_z$ , respectively, the point's coordinates will be changed accordingly.

$$(x_y y_y z_y) = R_y \cdot (xyz)^T \quad (2.3)$$

and

$$(x_z y_z z_z) = R_z \cdot (xyz)^T \quad (2.4)$$

where

$$R_y = \begin{bmatrix} \cos\theta_y & 0 & \sin\theta_y & 0 \\ 0 & 1 & 0 & 0 \\ -\sin\theta_y & 0 & \cos\theta_y & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad (2.5)$$

and

$$R_z = \begin{bmatrix} \cos\theta_z & -\sin\theta_z & 0 & 0 \\ \sin\theta_z & \cos\theta_z & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad (2.6)$$

By combining the values (2.2, 2.5, 2.6) to represent a point's rotation along all axes, the resulting coordinates of the point can be obtained in following equation:

$$(x_{xyz}y_{xyz}z_{xyz})^T = R_x R_y R_z \cdot (xyz)^T = R \cdot (xyz)^T \quad (2.7)$$

where

$$R = \begin{bmatrix} r_{11} & r_{12} & r_{13} & 0 \\ r_{21} & r_{22} & r_{23} & 0 \\ r_{31} & r_{32} & r_{33} & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad (2.8)$$

The rotation matrix is referred to as R, and it is possible to determine the Euler angles  $\theta_x$ ,  $\theta_y$ , and  $\theta_z$  through calculation.

$$\theta_x = \tan^{-1} \frac{r_{32}}{r_{33}} \quad (2.9)$$

$$\theta_y = -\tan^{-1} \frac{r_{31}}{\sqrt{r_{32}^2 + r_{33}^2}} \quad (2.10)$$

$$\theta_z = \tan^{-1} \frac{r_{21}}{r_{11}} \quad (2.11)$$

Furthermore, the translation of a point in 3D space can be achieved using the translation matrix, which is given by -

$$T(d_x, d_y, d_z) = \begin{bmatrix} 1 & 0 & 0 & d_x \\ 0 & 1 & 0 & d_y \\ 0 & 0 & 1 & d_z \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad (2.12)$$

where  $d_x$ ,  $d_y$ ,  $d_z$  are the displacement of any point along the x, y, z-axis respectively. An open-source 3D computer graphics software, which is called Blender [27], offers a transformation matrix that combines the three rotation and translation matrices into  $TR_xR_yR_z$ .

Individual Euler rotation of yaw, pitch and roll can be calculated with equations( 2.9, 2.10, 2.11).







# Chapter 3

## Methodology

### 3.1 Methodological approach

The objective of this thesis is to examine algorithms for emotion classification and determine if the accuracy of these algorithms is influenced by facial orientation and emotion category.

This research entails analyzing various models that have been presented to the scientific community and comparing their performance in relation to emotion categorization and facial orientation.

The results will be displayed as the number of correct predictions, the number of predictions made, and the percentage value. Figure 3-1 presents the workflow for this study, which involves Preprocessing step, then yaw, pitch, roll values calculation, training the model, testing the model, and finally making an analysis of the results. Figure 3-2 shows the extended version of General Thesis Workflow in more details.

### 3.2 Preprocessing

#### 3.2.1 Imbalanced Dataset

Firstly, the lighter version of AffectNet was taken to work with, because the whole dataset is very large(122 GB). Using annotations, the images were divided into folders,

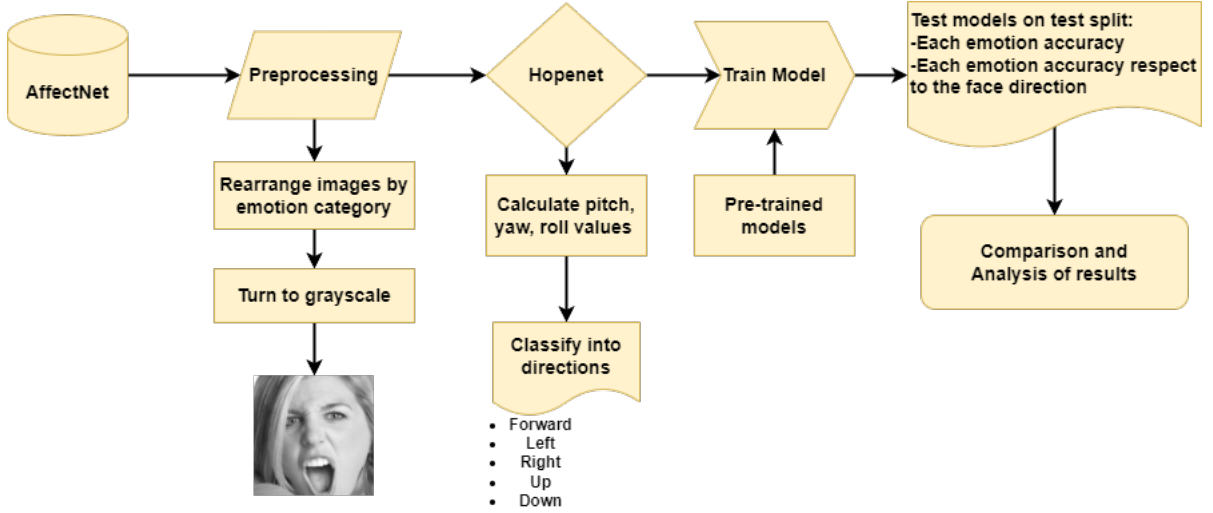


Figure 3-1: Thesis Workflow

namely, into 8 emotion categories as Anger, Contempt, Disgust, Fear, Happy, Neutral, Sad, and Surprise. The number of images of emotions such as Contempt, Disgust was much smaller in comparison with other emotions. Therefore, in order to balance the number of images per emotion, only over one-tenth(32622) of Affectnet(291,651) is used in this research,

In table 3.1 you can see a raw distribution of 32622 images, where the number of images per emotion is almost equal.

Direction	Anger	Contempt	Disgust	Fear	Happy	Neutral	Sad	Surprise
Forward	1504	1334	1454	2052	1661	1823	1554	1769
Left	411	310	371	361	347	397	391	412
Right	367	251	337	281	300	380	387	376
Up	788	399	750	1058	553	706	972	1005
Down	1194	1323	709	507	1356	913	896	663
<b>Overall</b>	<b>4264</b>	<b>3617</b>	<b>3621</b>	<b>4259</b>	<b>4217</b>	<b>4219</b>	<b>4200</b>	<b>4225</b>

Table 3.1: Raw Imbalanced Affectnet distribution

### Grayscale conversion

Secondly, all images were converted into Grayscale in order to reduce the computational power 3 times and preserve the relevant features [24].

### Three dimensions of movement

Thirdly, images from which pitch, yaw, and roll, using Hopenet, which is trained on

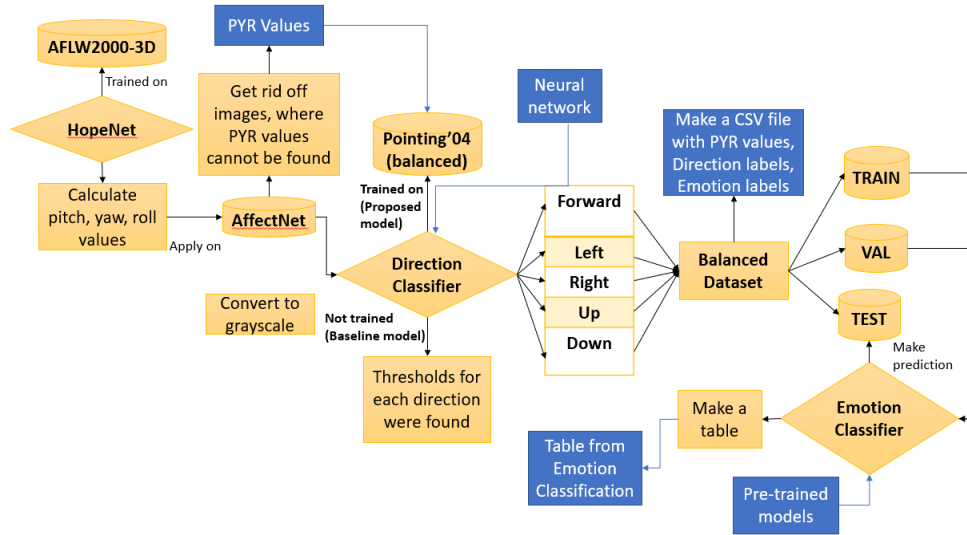


Figure 3-2: Extended Thesis Workflow

the AFLW2000-3D dataset, were calculated, and those images, where they cannot be found were removed from our main dataset. The samples of wrong images could be seen in Figure 3-3.



Figure 3-3: Bad samples of datasets, that can not be used

Finally, the images were classified through a manual algorithm into 12 directions, then were grouped into 5 directions(Forward, Left, Right, Up, Down). Then files in CSV format were created to write the image paths and their labels. There are two CSV files one for training and one for testing.

### 3.2.2 Balanced Dataset

As you can see in Table 3.2 most of the images are in the "Forward" direction. The emotion classifier could become biased, if we do not balance the dataset.

Direction	Anger	Contempt	Disgust	Fear	Happy	Neutral	Sad	Surprise
Forward	<b>35%</b>	<b>37%</b>	<b>40%</b>	<b>48%</b>	<b>39%</b>	<b>43%</b>	<b>37%</b>	<b>42%</b>
Left	10%	9%	11%	8%	8%	10%	10%	10%
Right	9%	7%	10%	7%	7%	9%	9%	9%
Up	18%	11%	20%	25%	13%	17%	23%	24%
Down	28%	36%	19%	12%	33%	21%	21%	15%
<b>Overall</b>	<b>4264</b>	<b>3617</b>	<b>3621</b>	<b>4259</b>	<b>4217</b>	<b>4219</b>	<b>4200</b>	<b>4225</b>

Table 3.2: Percentage Distribution of Raw Imbalanced dataset

Therefore, for Baseline and Proposed models, the datasets were balanced per direction and each emotion.

In the **Baseline model**, 3200 (80 images per direction) images were used for training, 1600 (40 images per direction) images for validation, and 800 (20 images per direction) images for testing.

In the **Proposed model**, 8840 (221 images per direction) images were used for training, 1200 (30 images per direction) images for validation, and 1200 (30 images per direction) images for testing.

#### Additional dataset

The dataset **Pointing'04** by [18] [19] was used in the **Proposed Model** for training the Direction Classifier.

The dataset consists of 30 folders(2790 images), where every 2 folders contain images of one person from different angles.

The name of each image file is presented in "personne01100-90+0.jpg" format, where the last 5-6 symbols are angles.

The images were separated into "Forward", "Left", "Right", "Up", and "Down" according to angles and grouped together.

There are 570 images for "Down", 540 images for "Forward", 570 images for "Left", 540 images for "Up", and 570 images for "Right".

But in order for the classifier not to become biased, the dataset was balanced, and each direction contained 540 images.

All image names, pitch, yaw, roll values, and direction labels were written in a CSV file.

### 3.3 Baseline model

#### HopeNet

Using HopeNet, the pitch, yaw, and roll values were calculated and written in the text file for each emotion category. The pitch, yaw, and roll values are in tensor type. The angles could be visualized as in Figure 3-4. Hopenet uses MTCNN for face detection and facial landmarks on images and ResNet50 to create a model.

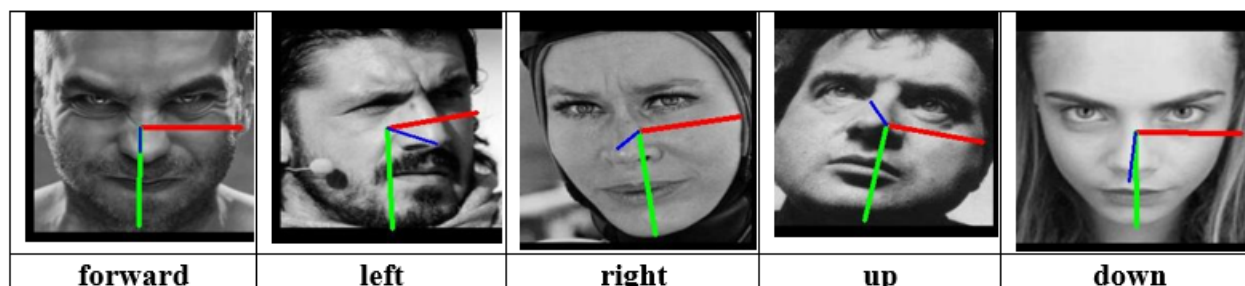


Figure 3-4: Pitch, yaw, roll calculation and visualization for each direction (pitch-green, roll-red, yaw-blue)

#### Manual algorithm

By analyzing the results after HopeNet and looking and comparing images manually, the algorithm was created for direction classification from the pitch, yaw, and roll values, where the threshold for each direction was found through calculation and manual classification since the needed external dataset exactly for this task was hard to find.

Let's look to the Table 3.3 where there are FileNameID and pitch, yaw, roll values for each image, which is classified as "Anger". In this example, I selected 5 images, which is looking "Down" direction. Then pitch, yaw, roll values were calculated.

After that, we will look which of pitch,yaw, roll values have more common values, in other words, we are looking for the threshold.

As you can see the pitch values have common range of values from -24 to -28. So, range from -24 to -28 of pitch values is threshold for direction "Down".

I do the same procedure manually for each direction, finding the thresholds. You can see the part of pseudo-code of Manual Algorithm below, where thresholds for each directions are set and each image is labeled with one of the directions(forward, left, right, up, down).

<b>FileNameID</b>	<b>252606</b>	<b>157352</b>	<b>145637</b>	<b>71072</b>	<b>57738</b>
<b>pitch</b>	-24.1261	-28.0626	-28.7395	-24.2781	-24.1895
<b>yaw</b>	3.7839	-3.3803	-1.5367	7.9262	-3.3464
<b>roll</b>	-4.1922	-0.9985	-0.8966	-3.1061	-4.8848

Table 3.3: Pitch, Yaw, Roll values for Anger images for direction "Down"

```

pathAnger ←
for x,y in Anger.items() do
  if y[1] < y[0] and y[1] < y[2] and y[1] < 0 and y[1] < -15 then
    if y[0] > 0 and y[2] < 0 then
      if y[0] < 3 then
        pathAnger[x] ← "left"
      else
        pathAnger[x] ← "up"
    else if y[0] < 0 and y[2] < 0 then
      if y[0] + y[2] > -15 then
        pathAnger[x] ← "left"
      else
        pathAnger[x] ← "down"
    else
      pathAnger[x] ← "left"
  else if y[1] > y[0] and y[1] > y[2] and y[1] > 0 and y[1] > 5 then

```



```

if  $y[0] + y[2] > -9$  and  $y[0] < 0$  and  $y[2] < 0$  then
     $pathAnger[x] \leftarrow "right"$ 
else
    if  $y[0] + y[2] > 0$  then
         $pathAnger[x] \leftarrow "up"$ 
    else if  $y[0] + y[2] < 0$  then
         $pathAnger[x] \leftarrow "down"$ 
    else
         $pathAnger[x] \leftarrow "right"$ 
else
    if  $y[0] > 0$  and  $y[2] > 0$  then
        if  $y[1] < 0$  then
             $pathAnger[x] \leftarrow "forward" \dots$ 

```

Here "Anger" is dictionary, which contains the path of images of folder "Anger" and its pitch, yaw, roll values. "x" is the paths of images, when "y" is array, which contains pitch,yaw,roll values("y[0]" is pitch value, "y[1]" is yaw value, "y[2]" is roll value). "pathAnger" is array, which will save labels of directions (forward, left, right, up, down) for each image.

### Emotion Classification

After direction classification, pre-trained models MobileNetV3-small, GoogleNet,ResNet-18, VGG-16, AlexNet, and AdaBoost were used to predict the accuracies for each Emotion Category. Then Confusion Matrix was calculated and the path of each truly labeled image was found to identify each direction's accuracy.

For both direction and emotion classifications Cross-entropy loss function was used to measure the difference between two probability distributions.

### 3.4 Proposed model

In the proposed model, the size of the dataset was enlarged to 11240 images(*AffectNet*). Then pitch, yaw, and roll values for each image of the Pointing'04 balanced dataset were calculated using **HopeNet**. The pitch, yaw, roll values are in tensor type. The values were converted into float numbers. The angles could be visualized as in Figure 3-4.

#### Direction Classification

The direction labels from "Forward", "Left", "Right", "Up", and "Down" were converted to numeric values, such as 0-"Forward", 1-"Left", 2-"Right", 3-"Up", 4-"Down". We do not need images anymore, because we have their pitch, yaw, and roll values. Therefore, all values were converted into numeric ones.

Model	Accuracy
Neural Network	85%
Googlenet	84%
MobileNetV3-small	83%
ResNet-18	82%

Table 3.4: Overall Direction Classification by each model

Since, as you can see in Figure 3.4, Neural Network showed the highest result, it was used for Direction Classification based on pitch, yaw, roll values. In Figure 3-5, you can see the architecture of Neural Network. The Neural Network was specifically designed for this task.

This model has 3 fully connected layers (also known as Linear layers), each followed by a batch normalization layer, ReLU activation function, and dropout layer. The output of the last layer is fed to the final linear layer with `out_features=5` which gives the output for the 5 classes( forward, left, right, up, down directions).

Note that the input features to the model have `in_features=3`, which means that the model expects input data with 3 features(pitch, yaw, roll values). Using Neural Network, directions are predicted for each image in our dataset.

In the Figure 3-6 and Figure 3-7, you can the performance of Neural Network: its

Accuracy Diagram and its Loss Diagram.

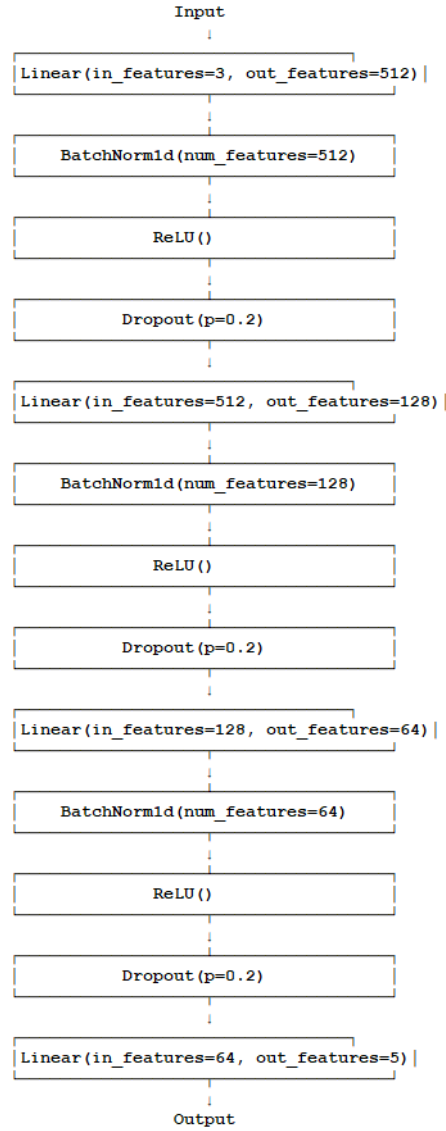


Figure 3-5: Neural Network Architecture

### Emotion Classification

The same pre-trained models as in the Baseline model were used in the Proposed model, namely, MobileNetV3-small, GoogleNet, ResNet-18, VGG-16, AlexNet, and AdaBoost were used for emotion classification in order to compare the results in the future.

For both categories of direction and emotion, two probability distributions were compared using the Cross-Entropy Loss Function.

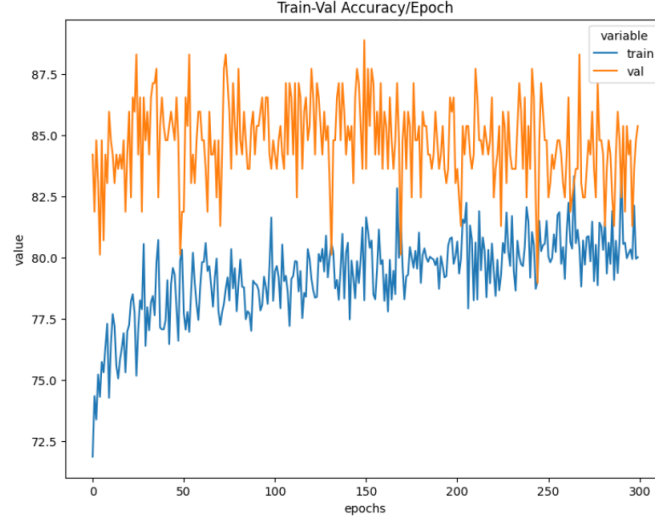


Figure 3-6: Neural Network Accuracy Diagram

### 3.5 Performance Evaluation Metrics

The performance of the baseline model and proposed model will be evaluated through three key metrics: Accuracy, F1 score, and confusion matrix.

Firstly, **Accuracy** is a measure of how often the model correctly predicts the outcome. It's calculated as the number of correct predictions divided by the total number of predictions.

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (3.1)$$

Next, **F1 score** is a metric used to evaluate the performance of a classification model. It is calculated as the harmonic mean of precision and recall, where precision is the ratio of true positives to the total predicted positives, and recall is the ratio of true positives to the total actual positives.

The F1 score can be calculated using the following formula:

$$F1 = 2 \times \frac{precision \times recall}{precision + recall} \quad (3.2)$$

where precision and recall are calculated as follows:

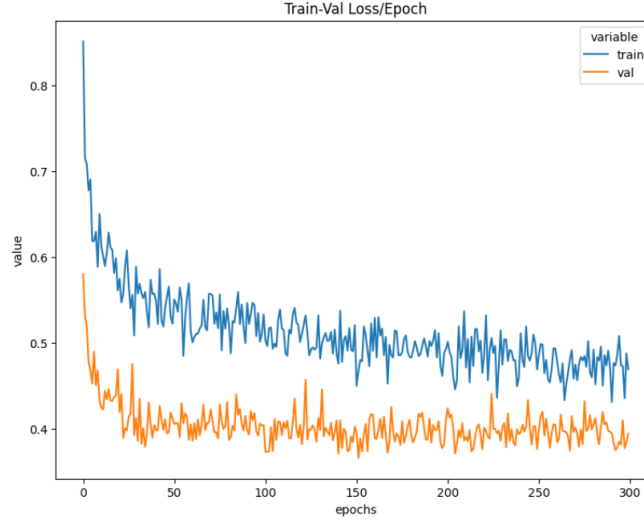


Figure 3-7: Neural Network Loss Diagram

$$precision = \frac{TP}{TP + FP} \quad (3.3)$$

$$recall = \frac{TP}{TP + FN} \quad (3.4)$$

Here, TP (true positives) is the number of correctly classified positive instances, FP (false positives) is the number of incorrectly classified positive instances, and FN (false negatives) is the number of incorrectly classified negative instances.

Finally, the **Confusion matrix** is a table that shows the number of true positives, true negatives, false positives, and false negatives for a given model. It's a powerful tool for understanding the performance of the model and identifying areas for improvement.

The confusion matrix can be presented as follows in Table 3.5:

Here, TP is the number of true positives, FP is the number of false positives, FN is

	Actual Positive	Actual Negative
Predicted Positive	TP	FP
Predicted Negative	FN	TN

Table 3.5: Confusion Matrix

the number of false negatives, and TN is the number of true negatives.





# Chapter 4

## Results

### 4.1 Baseline model results

#### Emotion Classification

As you can see in Table 4.1 MobileNetV3-small showed the highest percentage among

Pre-trained model	Accuracy	F1 score
MobileNetV3-small	46%	0.4623
Googlenet	43%	0.4285
ResNet-18	17%	0.1395
VGG-16	13%	0.0554
Alexnet	12%	0.0278
AdaBoost	12%	0.0277

Table 4.1: Overall Emotion Classification by each pre-trained model

all pre-trained models. Therefore, this model is used in the Baseline model for emotion classification.

Below you can see the Confusion Matrix in Figure 4-1, the Loss Function Diagram in Figure 4-2, and Accuracy Diagram for Training and Validation in Figure 4-3.

The **Confusion Matrix** in Figure 4-1 shows true classified and false classified values. But, we will take into account only true positive values, that are located on the diagonal of the matrix, because we want to know which direction impacts most on true classified emotion. Therefore, we do not need true negatives, false positives



and false negatives of Confusion matrix.

For example, for emotion "Anger", there are 35 true classified images as "Anger". For emotion "Contempt"- 46, for emotion "Disgust"-44, for emotion "Fear"-39, for emotion "Happy"-70, for emotion "Neutral"-56, for emotion "Sad"-38, for emotion "Surprise"-37 true classified images.

The confusion Matrix shows that the emotion "Happy" has the highest number of true classified images in emotion classification, while "Anger" has the lowest one.

Table 4.2 is using true values of Confusion Matrix in Figure 4-1 and shows image

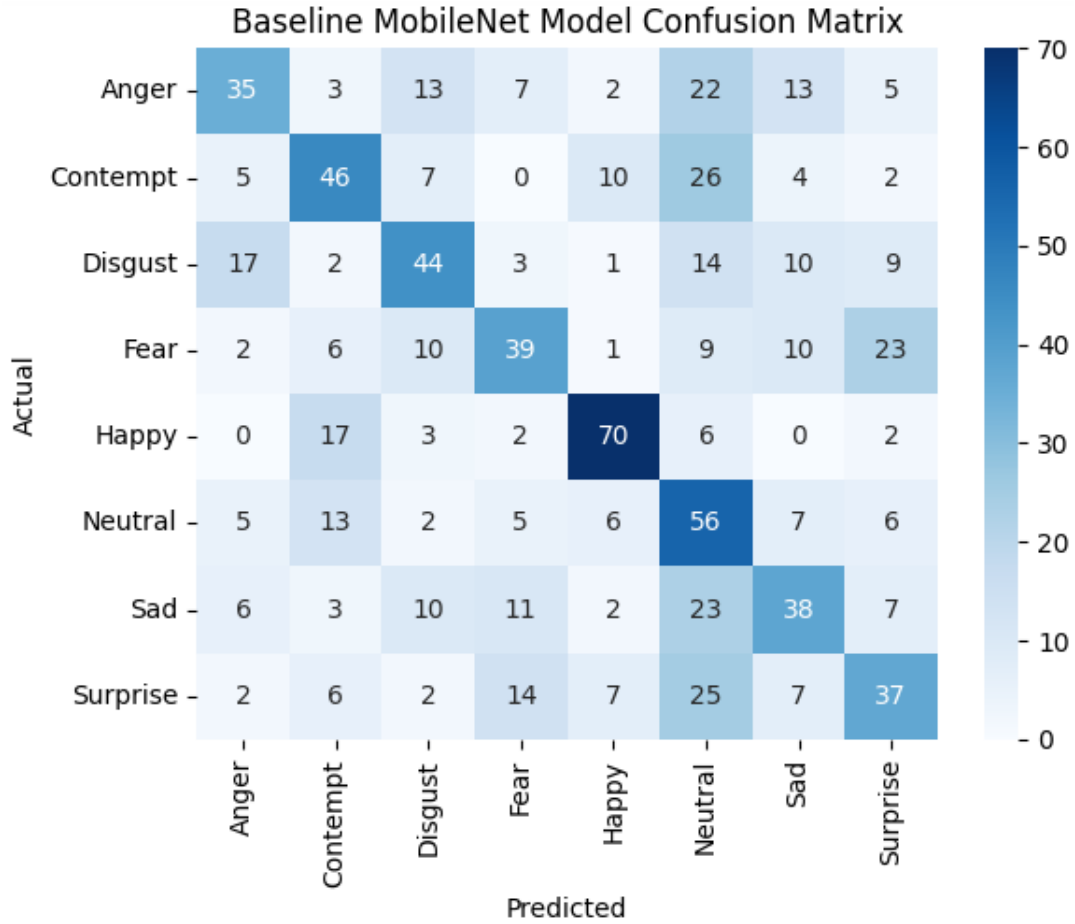


Figure 4-1: Baseline model Confusion matrix for MobileNetV3-small

distributions for each direction and for each emotion, considering the model was trained and tested on the balanced dataset.

It can be seen that in Table 4.2 emotion named "Happy" showed the highest

Direction	Anger	Contempt	Disgust	Fear	Happy	Neutral	Sad	Surprise	mean	var	std
Forward (160)	8	6	9	5	17	10	9	6	8.75	12.4375	3.526
Left(160)	5	13	9	9	13	13	6	8	9.5	9.0	3.0
Right(160)	10	13	9	7	12	11	9	9	10.0	3.25	1.802
Up(160)	5	7	8	11	14	10	8	6	8.625	7.4843	2.735
Down(160)	7	7	9	7	14	12	6	8	8.75	6.9375	2.633
<b>mean</b>	7.0	9.2	8.8	7.8	14.0	11.2	7.6	7.4			
<b>variance</b>	3.6	9.76	0.159	4.16	2.8	1.36	1.839	1.44			
<b>st deviation</b>	1.897	3.124	0.399	2.039	1.673	1.166	1.356	1.2			
<b>Emotion accuracy</b>	<b>35/100</b>	<b>46/100</b>	<b>44/100</b>	<b>39/100</b>	<b>70/100</b>	<b>56/100</b>	<b>38/100</b>	<b>37/100</b>			

Table 4.2: Direction by Emotion true labeled table for Baseline Model



Figure 4-2: Baseline model Loss Diagram for MobileNetV3-small

number of truly classified images among other emotion categories. Moreover, in this table, you can see the calculation of mean, variance, and standard deviation for each emotion category and for each direction.

For Emotions:

1)The highest variance was seen in "Contempt", while the lowest variance was seen in "Disgust".

2)The highest standard deviation was seen in "Contempt", while the lowest standard deviation was seen in "Disgust".

For directions:

The highest variance and standard deviation were seen in "Forward", while the lowest

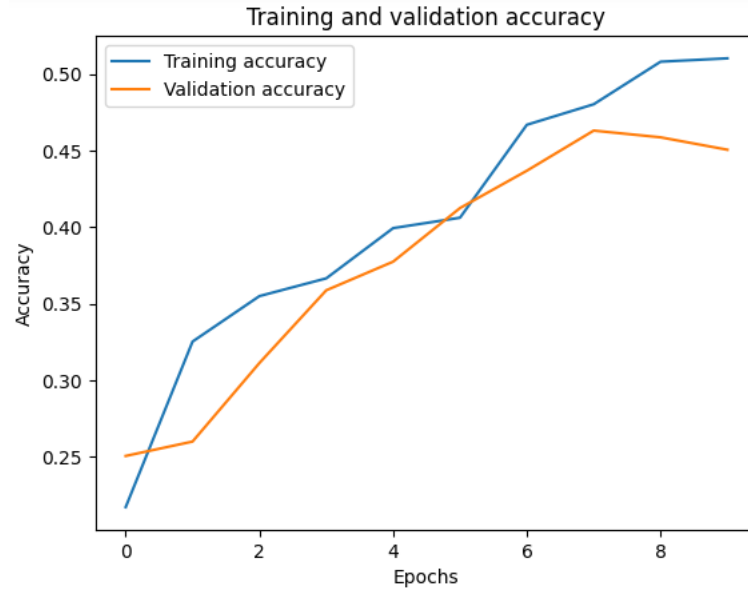


Figure 4-3: Baseline model Accuracy Diagram for MobileNetV3-small

variance and standard deviation were seen in "Right".

In the following tables 4.3 and 4.4 you can see the percentage distribution for each emotion and for each direction.

Direction	Anger	Contempt	Disgust	Fear	Happy	Neutral	Sad	Surprise
Forward	23%	14%	<b>20,4%</b>	13%	<b>25%</b>	18%	<b>24%</b>	16%
Left	14%	<b>28%</b>	<b>20,4%</b>	23%	<b>18%</b>	<b>23%</b>	15%	22%
Right	<b>28%</b>	<b>28%</b>	<b>20,4%</b>	18%	17%	20%	<b>24%</b>	<b>24%</b>
Up	14%	15%	18,4%	<b>28%</b>	<b>20%</b>	18%	21%	16%
Down	20%	15%	<b>20,4%</b>	18%	<b>20%</b>	21%	16%	22%
<b>Emotion accuracy</b>	<b>35%</b>	<b>46%</b>	<b>44%</b>	<b>39%</b>	<b>70%</b>	<b>56%</b>	<b>38%</b>	<b>37%</b>

Table 4.3: The highest percentage of direction per each emotion(by column)

Direction	Anger	Contempt	Disgust	Fear	Happy	Neutral	Sad	Surprise
Forward	23%	14%	20,4%	13%	25%	18%	<b>24%</b>	16%
Left	14%	<b>28%</b>	20,4%	23%	18%	23%	15%	22%
Right	<b>28%</b>	<b>28%</b>	20,4%	18%	17%	20%	24%	24%
Up	14%	15%	18,4%	<b>28%</b>	20%	18%	21%	16%
Down	20%	15%	20,4%	18%	20%	21%	16%	<b>22%</b>
<b>Emotion accuracy</b>	<b>35%</b>	<b>46%</b>	<b>44%</b>	<b>39%</b>	<b>70%</b>	<b>56%</b>	<b>38%</b>	<b>37%</b>

Table 4.4: The highest percentage of direction per emotion (by row)

## 4.2 Proposed model results

In the Proposed Model, in Table 4.5 MobileNetV3-small also took advantage, however, the GoogLeNet with the increase in the size of the dataset also showed 52%. In addition, ResNet-18 showed 49% in classification, which percentage has increased by 32% with the dataset size increase. However, other models such as AdaBoost, VGG-16, and AlexNet still showed the lowest results.

Pre-trained model	Accuracy	F1 score
MobileNetV3-small	52%	0.5177
GoogLeNet	52%	0.5160
ResNet-18	49%	0.4920
AdaBoost	16%	0.1658
VGG-16	12%	0.1277
Alexnet	12%	0.1276

Table 4.5: Overall Emotion Classification by each pre-trained model

Figure 4-4 and 4-5 show the Proposed model's Accuracy and Loss function Diagrams.



Figure 4-4: Proposed model Accuracy Diagram for MobileNetV3-small

The **Confusion Matrix** in Figure 4-6 shows true classified and false classified values. But, we will take into account only true positive values, that are located on

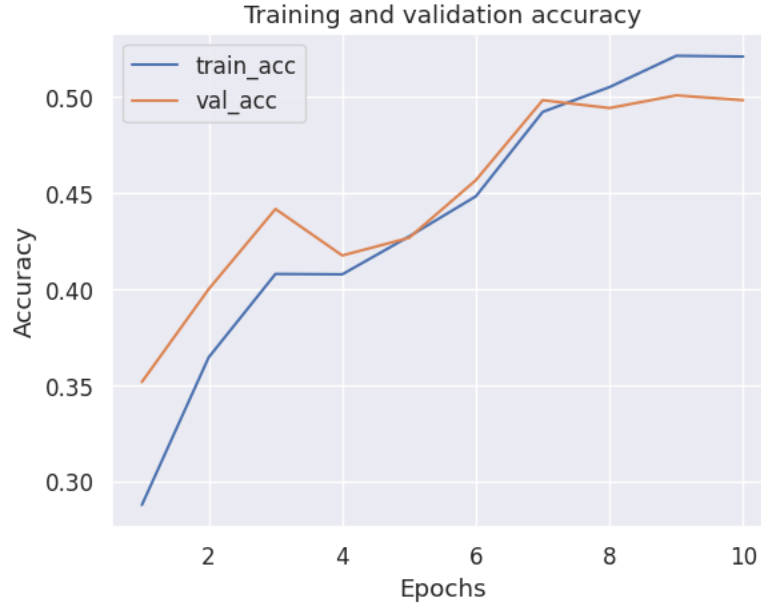


Figure 4-5: Proposed model Accuracy Diagram for MobileNetV3-small

the diagonal of the matrix, because we want to know which direction impacts most on true classifies emotion. Therefore, we do not need true negatives, false positives and false negatives of Confusion matrix.

For example, for emotion "Anger", there are 54 true classified images as "Anger". For emotion "Contempt"- 78, for emotion "Disgust"-68, for emotion "Fear"-77, for emotion "Happy"-113, for emotion "Neutral"-75, for emotion "Sad"-75, for emotion "Surprise"-84 true classified images.

As you can see in Confusion Matrix in Figure 4-6, the emotion "Happy" still shows the highest number of true classified images, while the emotion "Anger" is still the hardest task for the model to classify correctly.

In Table 4.6, the true labels from the Figure 4-6 are taken to see the distribution of true labels for each direction. Also, mean, variance, and standard deviation are calculated for each emotion and for each direction.

For Emotions:

"Happy" showed the highest variance and standard deviation, while "Disgust" presented the lowest variance and the lowest standard deviation. For Directions:

"Forward" showed the highest variance and standard deviation, while "Left" showed

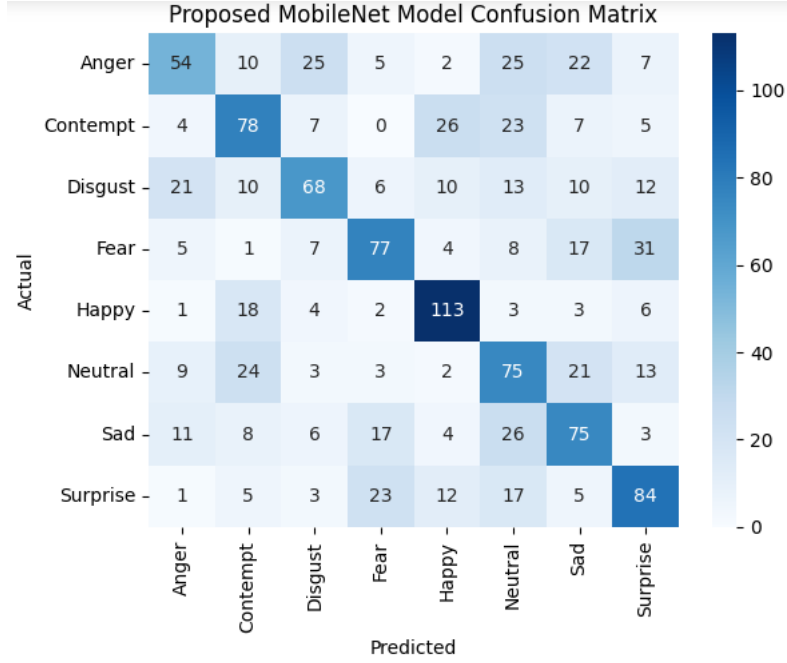


Figure 4-6: Proposed model Accuracy Diagram for MobileNetV3-small

both the lowest variance and standard deviation results.

Overall, in Tables 4.7 and 4.8 that most images, which were taken in "Forward" orientation could well identify "Anger", "Contempt", and "Surprise" emotions. While, images where the head position is in the "Up" direction, are mostly "Happy" images, and in the "Down" direction are "Sad" images. The majority of "Neutral" images are taken in the "Left" orientation, while "Disgust" and "Fear" are taken in the "Right" orientation. The results are visualized in Figure 4-7.

Direction	Anger	Contempt	Disgust	Fear	Happy	Neutral	Sad	Surprise	mean	var	std
Forward	14	19	12	13	24	11	14	20	15.87	18.35	4.28
(240)											
Left(240)	12	15	14	17	16	18	17	18	15.87	3.85	1.96
Right(240)	8	16	16	18	24	16	14	15	15.87	17.10	4.13
Up(240)	11	15	14	14	25	14	12	14	14.87	16.10	4.01
Down(240)	9	13	12	15	24	16	18	17	15.5	17.75	4.21
mean	10.8	15.6	13.6	15.4	22.6	15.0	15.0	16.8			
variance	4.56	3.84	2.23	3.44	11.039	5.6	4.8	4.56			
st deviation	2.13	1.95	1.49	1.85	3.32	2.36	2.19	2.13			
Emotion accuracy	54/150	78/150	68/150	77/150	113/150	75/150	75/150	84/150			

Table 4.6: Direction by Emotion true labeled table for Proposed Model

Direction	Anger	Contempt	Disgust	Fear	Happy	Neutral	Sad	Surprise
Forward	<b>26%</b>	<b>24%</b>	18%	17%	21%	15%	19%	<b>24%</b>
Left	22%	19%	20%	22%	14%	<b>24%</b>	23%	21%
Right	15%	21%	<b>24%</b>	<b>23%</b>	21%	21%	19%	18%
Up	20%	19%	20%	18%	<b>23%</b>	19%	15%	17%
Down	17%	17%	18%	20%	21%	21%	<b>24%</b>	20%
Emotion accuracy	36%	52%	45%	51%	75%	50%	50%	56%

Table 4.7: The highest percentage of direction per each emotion(by column)

Direction	Anger	Contempt	Disgust	Fear	Happy	Neutral	Sad	Surprise
Forward	<b>26%</b>	24%	18%	17%	21%	15%	19%	24%
Left	22%	19%	20%	22%	14%	<b>24%</b>	23%	21%
Right	15%	21%	<b>24%</b>	23%	21%	21%	19%	18%
Up	20%	19%	20%	18%	<b>23%</b>	19%	15%	17%
Down	17%	17%	18%	20%	21%	21%	<b>24%</b>	20%
Emotion accuracy	36%	52%	45%	51%	75%	50%	50%	56%

Table 4.8: The highest percentage of direction per emotion (by row)

### 4.3 Baseline model vs Proposed model

In Table 4.1 and Table 4.5, it can be seen that, in both Baseline and Proposed models pre-trained model called "MobileNetV3-small" showed the highest result in Accuracy and F1-score.

Also, take into consideration, that the balanced dataset size was increased 2 times in proposed model. If in Baseline model "MobileNetV3-small" showed 46%, then in Proposed model "MobileNetV3-small" showed 52%. Which means, that the increase in dataset size, also impacts on the model better performance.

In Table 4.2 and in Table 4.6 both lowest variance and the lowest standard deviation for Baseline Model and Proposed Model are seen in emotion "Disgust". It means that the data in the images ("Disgust") is very similar to each other, indicating that there is little variation in the values across the images.

While in Baseline Model the highest variance and the highest standard deviation are seen in emotion "Contempt" and in Proposed Model they are seen in emotion "Happy", respectively. It means that a set of images with different objects or scenes captured from different angles and lighting conditions. So, there are more versions of "Contempt" or "Happy" images to classify them as "Contempt" or "Happy".

In addition, In Table 4.2 and in Table 4.6 both highest variance and the highest standard deviation for Baseline Model and Proposed Model are seen in direction


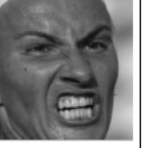

<b>Anger-Forward</b>	<b>Contempt-Forward</b>	<b>Surprise-Forward</b>	<b>Neutral-Left</b>	<b>Disgust-Right</b>	<b>Fear-Right</b>	<b>Happy-Up</b>	<b>Sad-Down</b>
							

Figure 4-7: Visualisation of results of mostly classified directions by emotions

"Forward". It means that, there are more different versions of images, which are classified as "Forward".

While, the lowest variance and standard deviation for Baseline Model is "Right" direction, the lowest variance and standard deviation for Proposed Model is "Left" direction. The reason could be because of different direction classification approaches, that were applied, therefore, they are showing different directions.





# Chapter 5

## Conclusion

To conclude, it was found out that, Head Pose Orientation has impact on Facial Emotion Classification. In this research, 3 datasets such as AffectNet, the Pointing'04 and the AFLW2000-3D were used. Two models such as Baseline Model and Proposed Model were introduced.

All datasets went through Preprocessing stage, where bad samples( large intra-class variances) were removed from the dataset and images were converted into grayscale. Then image names and their Euler angles(pitch, yaw, roll values) after calculation using HopeNet were written in separate CSV file. At the end, into CSV file, direction labels and emotion labels for each image were written.

Balanced datasets for Baseline Model and Proposed Model were made and divided into train, validation and test subdatasets. The size of Proposed Model Dataset is twice larger than the Baseline Model Dataset.

There were two approaches in Direction Classification, where one used Neural Network by training on Pointing'04 dataset(Proposed Model), showing 85% accuracy, and other one used Manual Algorithm by finding thresholds for each direction without training on any dataset(Baseline Model). Then the images were went through Emotion Classification, where pre-trained models were applied.

Overall, 7 pre-trained models (MobileNetV3-small, GoogleNet, ResNet-18, VGG-16, AlexNet, AdaBoost, HopeNet) were used.

The best pre-trained model for Emotion Classification is MobileNetV3-small, which

showed 46% in Baseline Model and 52% in Proposed Model, respectively. With the increase in the size of the dataset, the accuracy of pre-trained models also increased. At the end, Confusion Matrix was found for each Model and true classified images(true positives) from Confusion Matrix were selected to identify which of 5 directions(Forward, Left, Right, Up, Down) impacts most on each of 8 emotions(Anger, Contempt, Disgust, Fear, Happy, Neutral, Sad, Surprise).

So, "Forward" direction could classify emotions such as "Anger", "Contempt", "Surprise", "Left" direction could classify "Neutral" emotion, "Right" direction could classify "Disgust", "Fear", "Up" direction could classify "Happy" emotion and "Down" direction could identify "Sad" emotion.

Even though the percentage distribution of directions is almost the same, it can be seen that some directions suppress others, and could predict emotion.

In the future, the dataset size should be increased and other machine learning or deep learning models for direction and emotion classification should be tried.

In conclusion, Facial Emotion Classification based on Head Pose Orientation is a promising area of research that has the potential to improve the accuracy and reliability of Emotion Detection Systems.



# Bibliography

- [1] Konstantina VEMOU, Anna HORVATH, "Facial Emotion Recognition", the Technology and Privacy Unit of the European Data Protection Supervisor (EDPS), ISBN 978-92-9242-472-5QT-AD-21-001-EN-N, doi:10.2804/014217
- [2] James, William (1 April 2007). *The Principles of Psychology*. Cosimo, Inc. ISBN 9781602063136. Retrieved 20 October 2017 – via Google Books.
- [3] Cowen, Alan S.; Keltner, Dacher (2017). "Self-report captures 27 distinct categories of emotion bridged by continuous gradients". *Proceedings of the National Academy of Sciences of the United States of America*. 114 (38): E7900–E7909. doi:10.1073/pnas.1702247114. PMC 5617253. PMID 28874542.
- [4] S. Shojaeilangari, W. -Y. Yau, K. Nandakumar, J. Li and E. K. Teoh, "Robust Representation and Recognition of Facial Emotions Using Extreme Sparse Learning," in *IEEE Transactions on Image Processing*, vol. 24, no. 7, pp. 2140-2152, July 2015, doi: 10.1109/TIP.2015.2416634.
- [5] Shi Yi-Bin, Zhang Jian-Ming, Tian Jian-Hua and Zhou Geng-Tao, "An improved facial feature localization method based on ASM", *Computer-Aided Industrial design and Conceptual design*, 2006.
- [6] Jiamin Liu and Jayaram K. Udupa, "Oriented Active Shape Models", *IEEE Transactions on medical Imaging*, vol. 28, no. 4, pp. 571-584, April 2009.

- [7] K. -E. Ko and K. -B. Sim, "Development of a Facial Emotion Recognition Method Based on Combining AAM with DBN," 2010 International Conference on Cyberworlds, 2010, pp. 87-91, doi: 10.1109/CW.2010.65.
- [8] Mollahosseini, Ali et al. "AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild." IEEE Transactions on Affective Computing 10 (2017): 18-31.
- [9] Kollias, D., Cheng, S., Ververas, E. et al. Deep Neural Network Augmentation: Generating Faces for Affect Analysis. Int J Comput Vis 128, 1455–1484 (2020). <https://doi.org/10.1007/s11263-020-01304-3>
- [10] Kollias, Dimitrios and Stefanos Zafeiriou. "Expression, Affect, Action Unit Recognition: Aff-Wild2, Multi-Task Learning and ArcFace." ArXiv abs/1910.04855 (2019): n. pag.
- [11] A. V. Savchenko, L. V. Savchenko and I. Makarov, "Classifying Emotions and Engagement in Online Learning Based on a Single Facial Expression Recognition Neural Network," in IEEE Transactions on Affective Computing, vol. 13, no. 4, pp. 2132-2143, 1 Oct.-Dec. 2022, doi: 10.1109/TAFFC.2022.3188390.
- [12] Nicolaou, A., Zafeiriou, S., Pantic, M. (2017). A robust 3D head pose estimation in-the-wild using hybrid CNN–local–global regression. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 6044-6052).
- [13] Nicolaou, A., Zafeiriou, S., Pantic, M. (2018). Hopenet: A convolutional neural network for head pose estimation in-the-wild. IEEE Transactions on Pattern Analysis and Machine Intelligence, 41(1), 85-98.
- [14] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications," *arXiv:1704.04861 [cs.CV]*, Apr. 2017.

- [15] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going Deeper with Convolutions," *arXiv:1409.4842 [cs.CV]*, Sep. 2014.
- [16] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," *arXiv:1512.03385 [cs.CV]*, Dec. 2015.
- [17] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," *arXiv:1409.1556 [cs.CV]*, Sep. 2015.
- [18] N. Gourier, D. Hall, J. L. Crowley Estimating Face Orientation from Robust Detection of Salient Facial Features Proceedings of Pointing 2004, ICPR, International Workshop on Visual Observation of Deictic Gestures, Cambridge, UK
- [19] Ewhaiep@gmail.com. Pointing04 DB. June 2017,doi: 10.6084/m9.figshare.5142466.v2.
- [20] Heisele, B. (2004). Pointing'04 - Database for Head Pose Estimation. Proceedings of the International Conference on Computer Vision Theory and Applications, 1, 67-74.
- [21] Krizhevsky, A., Sutskever, I., Hinton, G. E. (2012).ImageNet classification with deep convolutional neural networks. In Advances in neural information processing systems (pp. 1097-1105).
- [22] Freund, Y., Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. Journal of Computer and system sciences, 55(1), 119-139.
- [23] Tewari, A., Zollhoefer, M., Kim, H., Garrido, P., Bernard, F., Perez, P., Theobalt, C. (2017). Mofa: Model-based deep convolutional face autoencoder for unsupervised monocular reconstruction. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 6091-6100).

- [24] Felipe Zago Canal, Tobias Rossi Müller, Jhennifer Cristine Matias, Gustavo Gino Scotton, Antonio Reis de Sa Junior, Eliane Pozzebon, Antonio Carlos Sobieranski, A survey on facial emotion recognition techniques: A state-of-the-art literature review, *Information Sciences*, Volume 582, 2022, Pages 593-617, ISSN 0020-0255, <https://doi.org/10.1016/j.ins.2021.10.005>.
- [25] Basak, Shubhajit, Corcoran, Peter, Khan, Faisal, McDonnell, Rachel, Schukat, Michael. (2021). Learning 3D Head Pose From Synthetic Data: A Semi-Supervised Approach. *IEEE Access*. PP. 1-1. 10.1109/ACCESS.2021.3063884.
- [26] Nataniel Ruiz, Eunji Chong, and James M. Rehg, Fine-Grained Head Pose Estimation Without Keypoints, *arXiv:1710.00925*, 2018.
- [27] “blender.org - Home of the Blender project - Free and Open 3D Creation Software.” [Online]. Available: <https://www.blender.org/>. [Accessed: 10-Nov-2020]