上海交通大學

SHANGHAI JIAO TONG UNIVERSITY

学士学位论文

THESIS OF BACHELOR

Project Title Emotion Recognition based on Deep Learning

Name1	Shangquan Sun	ID1	515370910068	Major	ECE
Name2	Jie Gong	ID2	515370910078	Major	ECE
Name3	Jianshu Li	ID3	515370910024	Major	ECE
Name4	Xuanyu Wang	ID4	5142119039	Major	ECE

Supervisor Prof. Yong Long

School UM-SJTU Joint Institute

Semester 2018-Summer

上海交通大学

毕业设计（论文）学术诚信声明

本人郑重声明：所呈交的毕业设计（论文），是本人在导师的指导下，独立进行研究工作所取得的成果。除文中已经注明引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写过的作品成果。对本文的研究做出重要贡献的个人和集体，均已在文中以明确方式标明。本人完全意识到本声明的法律结果由本人承担。

作者签名：Shangquan Sun
Jie Gong
Jianshu Li
Xuanyu Wang

日期：2019 年 8 月 5 日

上海交通大学

毕业设计（论文）版权使用授权书

本毕业设计（论文）作者同意学校保留并向国家有关部门或机构送交论文的复印件和电子版，允许论文被查阅和借阅。本人授权上海交通大学可以将本毕业设计（论文）的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印或扫描等复制手段保存和汇编本毕业设计（论文）。

保密□，在___年解密后适用本授权书。

本论文属于

不保密☑。

（请在以上方框内打“√”）

作者签名： Shangquan Sun
Jie Gong
Jianshu Li
Xuanyu Wang　　　　指导教师签名：

日期： 2019 年 8 月 5 日　　　　日期： 2019 年 8 月 5 日

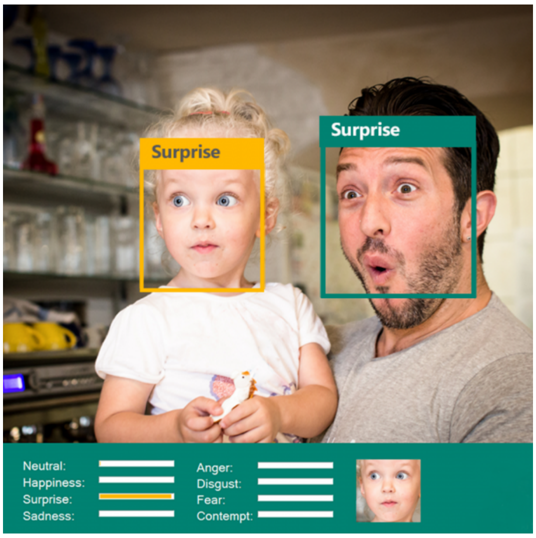
UM–SJTU Joint Institute

VE450 Major Design Experience

& VM450 Design and Manufacturing III
(VE450&VM450)

Design Review #3 Report

Pre-research of Emotion Recognition based on Deep Learning



Team #11

Shangquan Sun	Leader	515370910068
Jie Gong	Member	515370910078
Jianshu Li	Member	515370910024
Prof.Yong Long	Instructor	
Bo Xie	Co-Instructor	
HASCO VISION	Sponsor	
Date	July 24, 2019	

Team #11: Pre-research of Emotion Recognition based on Deep Learning

Shangquan Sun, Jie Gong, Jianshu Li

Sponsor: HASCO VISION, Instructor: Prof. Yong Long, Bo Xie

SJTU-Michigan Joint Institute, Shanghai Jiao Tong University

527112517@sjtu.edu.cn, kevinjiegong@gmail.com,

lijs007@qq.com

15021821189, 13162126869, 13606832334

July 24, 2019

摘要

面部表情识别 1(FER) 是深部学习最普遍的应用之一, 特别是在一些强大的深度学习方法出现后, 如卷积神经网络、生成对抗网络等, 面部表情识别变得尤为主要。但是, 在此领域中还有一些限制, 例如, 对于某些数据库最先进的方法的准确性仍然低于 80%。在本报告中, 模型将在多个数据库利用转移学习和深度学习进行训练, 使模型将能够识别人类的面部表情。更具体地说, 我们的赞助商,HASCO VISION 的愿景是旨在使用 FER 技术和深度学习模型来识别驾驶员的情绪以增加驾驶安全性, 因为 HASCO VISION 是汽车公司的子公司。在项目中, 包括 [LCK+10]、JAFFE[LAK+98] 等在内的多个数据库将会被使用。预训练的迁移学习的模型或特征值,[SZ14] 或初始 ResNetV2 [SIVA17], 等将被用于训练。总之, 我们的目的是产生一个高度准确的深度学习、专门安装在汽车内监控驾驶员情绪的 FER 模型。项目计划分为几个任务, 包括需求调查、文献研究、数据收集、数据处理、模型构建、模型改进和最终针对展示的准备工作。前三项任务在前四周内完成, 数据处理和模型构建任务在第 9 周之前完成。从第 9 周到第 12 周, 模型改进应该完成。在最后一周准备必须完成最后的演示工作。

关键字— 面部表情识别、情绪识别、深度学习、转移学习、人脸检测、数据增强、面部对齐

Abstract

Facial Expression Recognition[1] (FER) is one of the most prevalent applications of Deep Learning, especially after the appearance of some powerful deep learning methods, such as Convolutional Neural Network, Generative Adversarial Network, etc. However, there are still some limits in the territory. For instance, the accuracy of the most advanced method can still be lower than 80% for some databases. In this report, models will be trained on multiple databases by utilizing transfer learning and deep learning so that the model will be able to recognize human facial expressions. More specifically, our sponsor, HASCO VISION, aims to use FER techniques and deep learning models to recognize drivers' emotions in favor of safety, since HASCO VISION is a subsidiary of automotive company. In the project, databases including CK+ [LCK$^+$10], JAFFE [LAK$^+$98], etc. will be used. For pre-trained model or bottleneck for transfer learning, VGG16 [SZ14] or InceptionResNetV2 [SIVA17], etc. will be utilized In summary, Our purpose is to generate a highly accurate deep learning model for FER specifically installed inside a car monitoring drivers' emotions. The project plan is divided into several tasks, including requirement survey, literature research, data collection, data processing, model building, model improving and preparation for the final exposition. The first three tasks are finished within the first four weeks, while the tasks of data processing and model building is done before the 9th week. From the 9th week to the 12th week, model improving should be completed. In the last week, the preparation for the final demonstration has to be done.

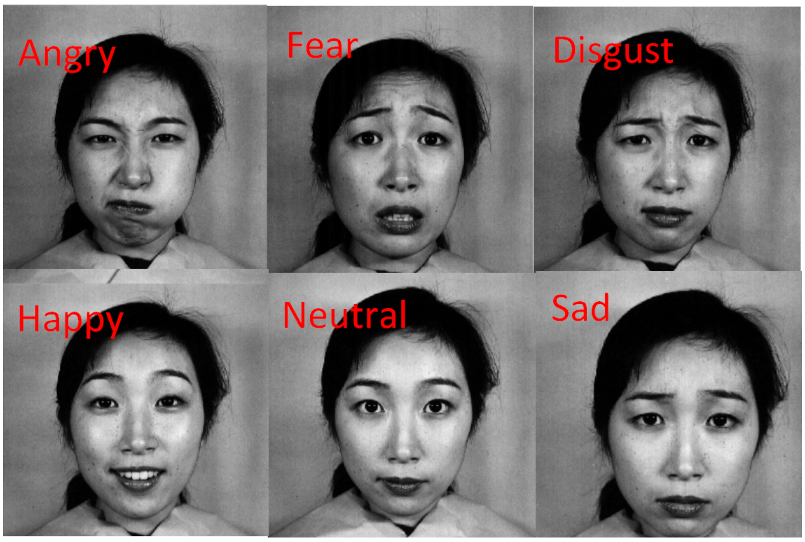


Figure 1: Facial Expression Recognition illustration from database JAFFE.

Key Words— Facial Expression Recognition, Emotion Recognition, Deep Learning, Transfer Learning, Face Detection, Data Augmentation, Face Alignment

[1] The graph in the title page is a demonstration from Microsoft. [Lin15]

Contents

List of Figures

List of Tables

1 Executive Summary

The motivation of the project is quite meaningful, because driving safety is a significant problem in nowadays society. If the emotions of drivers can be monitored and well enlightened or reminded, some cases of traffic accidents will be avoided, such as road-rage, etc. The specifications are well designed. All the engineering characteristics can be corresponding to each customer requirements. Also the ranking of the engineering specifications is well done by selection matrices and quality function deployment. The concepts chosen are well organized. Each method for every part of data processing and model building is based on the most recent and best performing method or framework. The final design is a combination of the completion of all the sub-components. Manufacturing plan is prepared based on the analysis of the experiences of deep learning project the group members have done before. Cost analysis is simple, since almost all the databases and frameworks involved are free and the only necessary cost is for renting a GPU server.Test results are based on the indexes returned by the model as well as the data generated by our experiments. There is no critique on the project.

2 Introduction and Problem Description

Deep Learning is a type of algorithms or methods in the territory of Machine Learning, which uses neural networks when model does the task of classification or prediction. In the project, the input data of model are from databases containing the images of human faces of different emotions, poses and backgrounds. Based on the data set, a Convolution Neural Network, a Generative Adversarial Network, or other Deep Learning algorithms, can be built and trained to extract potential relationship between images' color pixels and the judgment or prediction of human emotions on images. As a result, a Deep Learning model can be generated, capable of classifying input images into seven labels of facial emotions1.

Driver's emotion is one of the major factors affecting driving safety. Our project monitors the driver's emotion in a visual way to improve driving safety and it's usually known as Facial Expression Recognition task. Most open-source databases contain 6 or 7 kinds of expression labels, which are angry, disgust, fear, happy, sad, surprise and neutral (optional). Some of them also include other emotions or different angles of view.

In recent years, deep learning develops rapidly. As a result, lots of state-of-art deep learning models of FER already exist and most of them can achieve around 75% accuracy on various famous databases like CK+ [LCK$^+$10], Oulu-Casia [LSWC14] and etc. HASCO VISION expects us to first implement a novel model from a recent paper and then to further modify and improve its performance. several benchmarks from recent papers have been investigated and compared before determining the final choice of benchmarking. Most of them were based on Convolutional Neural Networks(CNN) while some of them used the novel model Generative Adversarial Networks(GAN). In the next section, they will be discussed about in more details.

As deep learning as well as computer vision area develops rapidly, lots of state-of-art deep learning models of FER already exist and most of them can achieve around 75% accuracy on several famous databases like CK+ [LCK$^+$10], Oulu-Casia [LSWC14] and etc. However, those most recent papers with impressing performance haven't provided actual implementations. As a result, it's significant build our own method which can achieve competent performance with those state-of-art models.

HASCO VISION expects us to first implement a novel model from a recent paper and then to further modify the model and improve its performance. The baseline target is to implement a model with at least 75% accuracy. This model should be capable to adapt to conditions in the real world, i.e. the driving car, which means it should be able to recognize the emotion from different backgrounds, variant illumination situation with acceptable delay. Several benchmarks from recent papers have been investigated and compared before determining the final choice of benchmarking. Most of them were based on Convolutional Neural Networks (CNN) while some of them used the Generative Adversarial Networks (GAN).

The project design is shown in Figure 2, and summarised as follows. The original images' face regions are cropped by MTCNN also with 5 landmarks together. Then by OpenCV Homography, face regions can be stretched into a square and be aligned with each other based on the 5 landmarks. Then by Imgaug library, data set is expanded by 9 times. After that, the processed data are input in a model of transfer learning, InceptionResNetV2. Eventually, the final model is generated.

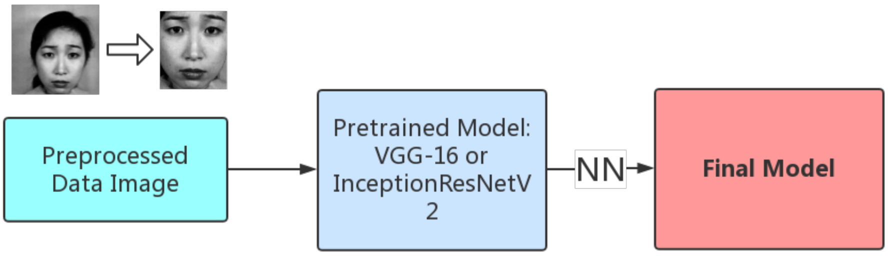


Figure 2: Flow chart for our project design.

3 Related Work and Benchmarks

Chieh-Ming Kuo et al [KLS18] applied a compact deep neural network. They did illumination normalization in order to better detect face in the wild. They also implemented a frame-to-sequence approach which takes a sequence of images as inputs instead and obtained an improvement in accuracy around 3%. After comparison, this benchmark has the best performance 83%.

Andre Teixeira Lopes et al's model [LdADSOS17] was also based CNN. They did preprocessing like rotation correction, down-sampling and intensity normalization before CNN. Since there are not enough public data for training, they synthetically generated new training images by translating, rotating and skewing the original images. The paper is helpful because its main focus is to process data before training for better performance, which is exactly the first step needed to do when constructing frameworks.

Heechul Jung el al [JLY$^+$15] combined two models together to obtain better performance. They trained Deep Temporal Appearance Network, which was composed of CNN, based on the extracted temporal appearance features. Also they trained Deep Temporal Geometric Network composed of Deep Neural Network (DNN) based on the extracted features from landmarks. The two models were combined together with joint fine-tuning method.

Huiyuan Yang et al's model [YCY18] is based on GAN. They first used GAN to exact no-expression images from the original ones and then used the residues in the network, which are supposed to contain only the information of expressions, to train and classify.

Finally Gaurav Sharma el al's model [KLS18] is chosen as the benchmark. Over 75% accuracy is expected on the combined dataset. They apply transfer learning and use the model structure and weights from VGG-16 [SZ14] or InceptionResNetV2 [SIVA17]. This method significantly eases the calculation burden on the hardware and reduces training time as well, which makes it possible for us to run the training locally at the first stage. Besides, there are still not enough datasets for FER task since most of the datasets only contain a few hundreds of images which are not enough for training and may cause overfitting. So VGG-16 would help us better extract geometric features from the images first.

Following instructions from HASCO VISION's engineer, the datasets should be further improved and processed. New data could be generated from the original ones using translating, rotating, skewing to enlarge the size of the training set. Besides, face alignment and rotation will also help improve the quality of the dataset. In the future, some pre-processing methods from other papers will be applied, like illumination normalization, intensity normalization and etc.

4 Information Sources and Databases

Up to July 24^{th}, 2019, five databases have been collected, which are CK+ [LCK+10], JAFFE [LAK+98], FacesDB [JPMC12], MMI [VP10], and FERG [ACF+16]. But FERG [ACF+16] is a huge database with only cartoon faces, which is obviously not fitful to our requirement. Also, MMI [VP10] is a database of videos rather than images. The facial expressions from NVIE [WLL+10] are poorly organized and expressed. So the three of them are not suitable in the project. Then the details of each database are illustrated as Table 1.

	CK+	JAFFE	FacesDB
Angry	45	30	36
Disgust	59	29	36
Fear	25	32	36
Happy	69	31	36
Neutral	327	30	36
Sad	28	31	36
Surprise	83	30	36
Total	636	213	252

Table 1: A statistics of three acquired and fitful databases, CK+, JAFFE, FacesDB (note that FacesDB also includes contempt, closed, open, leftside, etc.)

5 Customer Requirements and Engineering Specifications

5.1 Customer Requirements

After considerations, the probable customer requirements and also corresponding reasons are summarized as following:

1. Few mistakes:

   Since the model will be actually installed in a car to monitor drivers' emotions, it should make as few mistakes as possible. First and foremost, serious consequences might occur if dangerous emotion is not detected while driving. One the other hand, false alarm might also ruin the driver's driving experience and draw unnecessary cautions. So high accuracy should be the most important customer requirement.

2. Low-latency:

   In real-time and real-world driving, drivers are required to make decisions and actions in a few seconds frequently. As a result, our model should detect drivers' unstable and unsafe emotions as soon as possible. If it takes several minutes to generate a judgment, an accident might have occurred. So low latency should be another customer requirement.

3. Less power:

   Since the model will be installed in an automotive, it should consume as less energy as possible. Otherwise, the car loading it will spend much extra energy on it, which is not ideal. So less energy consuming should be one customer requirement.

4. Less heat:

   Since the model will be installed on a board, a hardware, in a car, it should generate as less heat as possible. Otherwise, the car loading it will generate much heat and the board loading may be damaged duet to heat, which is not ideal. So less heat consuming should be one customer requirement.

5. Applicability to all races:

   Since HASCO VISION is a global company, whose customers are not limited to Asian people, drivers of different races should be recognized by the model. Otherwise, an European driver living in Shanghai and owning the car will not be protected by the system. So applicability to all races should be one customer requirement.

6. Applicability to all poses:

   Drivers may appear on camera with slightly different angles of view. So our model should detect all poses of faces. So applicability to all poses should be one customer requirement. But in the most recent meeting, the sponsor has said we only need to do experiments on frontal view of faces.

7. Applicability with glasses or masks, etc.:

   Drivers may appear on camera with different ornaments, e.g. glasses, masks, etc. So our model might be able detect all faces with those ornaments. So applicability with glasses or masks, etc. should be one customer requirement. But in all databases, there is no such kind of data images. Our sponsor also says we only need to focus on the available datasets without those ornaments.

8. Functionality in different backgrounds:

   In a car, there may be different kinds of backgrounds of human faces. So human faces from all kinds of backgrounds must be perfectly cropped. Otherwise, a fashionable seat decoration behind driver's face might influence the performance of the model, which is not ideal. So functionality in different backgrounds should be one customer requirement.

5.2 Engineering Specifications

Then several engineering specifications corresponding to customer requirements are listed. They are listed as Table 2.

No.	Engineering Specification	Target Value	Current Value
1	F1 score	>75%	93.90%
1	Area Under Curve (AUC)	>75%	92.20%
3	Intersection Over Union (IOU)	High	High
4	Size of model	<2048MB	750MB
5	Precision matrix	>75% for each diagonal entry	≈90%
5	Recall matrix	>76% for each diagonal entry	≈90%
7	Delay	<2000ms	500-1800ms

Table 2: List of engineering specifications

For the engineering specifications, how those customer requirements are related to them are also showed as following.

1. F1 score

   F1 score is one direct index that indicates the performance of a model. Its mathematical expression is shown in equation 1

   $$F_\beta = \frac{(\beta^2 + 1)P \cdot R}{\beta^2 P + R} \tag{1}$$

   where $P$and R denotes the Precision and Recall. When $\beta = 1$, F1$=2\frac{P \cdot R}{P+R}$. In a figure of Precision Recall Curve, optimal F1 score is the cross point of $P = R$ and the curve, which denotes the model's performance. Besides, if the curve is smoother, the model performs better. By the value of F1 score, the model performance can be generally indicated. And our basic target is to reach 0.75, a reasonable baseline for a 7-class classification task.

2. Area Under Curve (AUC)

   Area Under Curve is specifically the area under Receiver Operating Characteristic Curve, with false positive rate as X axis and true positive rate as Y axis. If the curve is less smoother, it is more likely that the model is overfitting. The larger the AUC of a model is, the better its performance is. Our target is also to make AUC 75% at least.

3. Intersection Over Union (IOU)

   An illustration of IOU is showed in Figure 3. IOU equals to the cropped region area over the perfect cropped region area. If IOU approaches to 1, it means sufficient and necessary

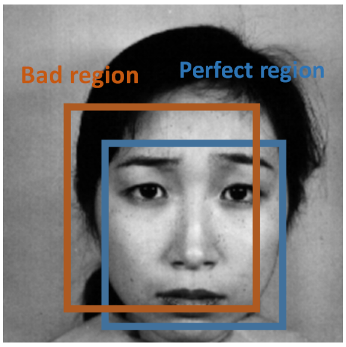


Figure 3: Illustration of Intersection Over Union, where blue region is a perfect cropping and the orange one is a bad one (the photo is from JAFFE).

region is cropped. If it is small, many irrelevant or many necessary information has been excluded or included, which is not ideal. The value is highly related to applicability to poses, ornaments, backgrounds, since it functions to crop face region from disregarding to those interfering factors. The target IOU fraction should be larger than 0.9, which means that the cropped face should be nearly perfect in human eyes in most cases. The target IOU fraction should be larger than 0.9, which means that the cropped face should be nearly perfect in human eyes in most cases.

4. Size of model

   Since size of model influences significantly the computing speed, heat generation, and energy consuming, it should be taken into consideration as an engineering characteristics. Its unit is byte or megabyte. The upper bound of target size is 2 GB or 2048 MB.

5. Precision matrix Recall matrix

   Since our task is multi-class classification, a precision value and a recall value exist for each pair of classes. Therefore, a precision matrix and a recall matrix are required as two engineering characteristics. The precision equals to the fraction of the true positive set over the whole retrieved set, while the recall equals to the fraction of true positive set over the whole golden set. They're two aspects of F1 score and AUC. They're both related to model performance and reliability, i.e. accuracy, applicability to races, poses, ornaments, backgrounds, but not significantly since the F1 already takes precisions and recalls into consideration. In order to accurately recognize each of the seven emotions, every precision and recall should be higher than 0.75.

6. Delay

   Delay is the engineering specification exactly responds to the customer requirement of low latency. Its unit is second or microsecond. The ideal delay should be no more than 2 seconds or 2000 milliseconds in order to give feedback to the driver in time.

Please note that since the task in the project is multi-class classification, F1 socre and AUC must be calculated between each pair of classes, e.g. happy & fear, etc. Then two average values of F1 scores and AUC between all pairs needed to be generated and they are the final indexes indicating performance of the multi-class model, which are called macro F1 and micro F1.

Since F1 score and AUC are related the performance, including its accuracy, applicability to all races, poses, ornaments, backgrounds. They are equally the most important two engineering characteristics.

Following them, the next most important characteristic is IOU, since it indicates how the faces in the image data are detected, cropped and processed. Then size of model is the fourth important one, since it is related to time latency, heat generation and energy consuming. Precision matrix and Recall matrix are less significant since they only reflect one aspect of model's performance and F1 score already covers the precision and recall somehow. Delay is the least important aspects since it's not difficult to limit the processing time in a few seconds. And it seems that there's no necessity to further reduce the delay into like a few milliseconds currently.

5.3 Quality Function Deployment (QFD)

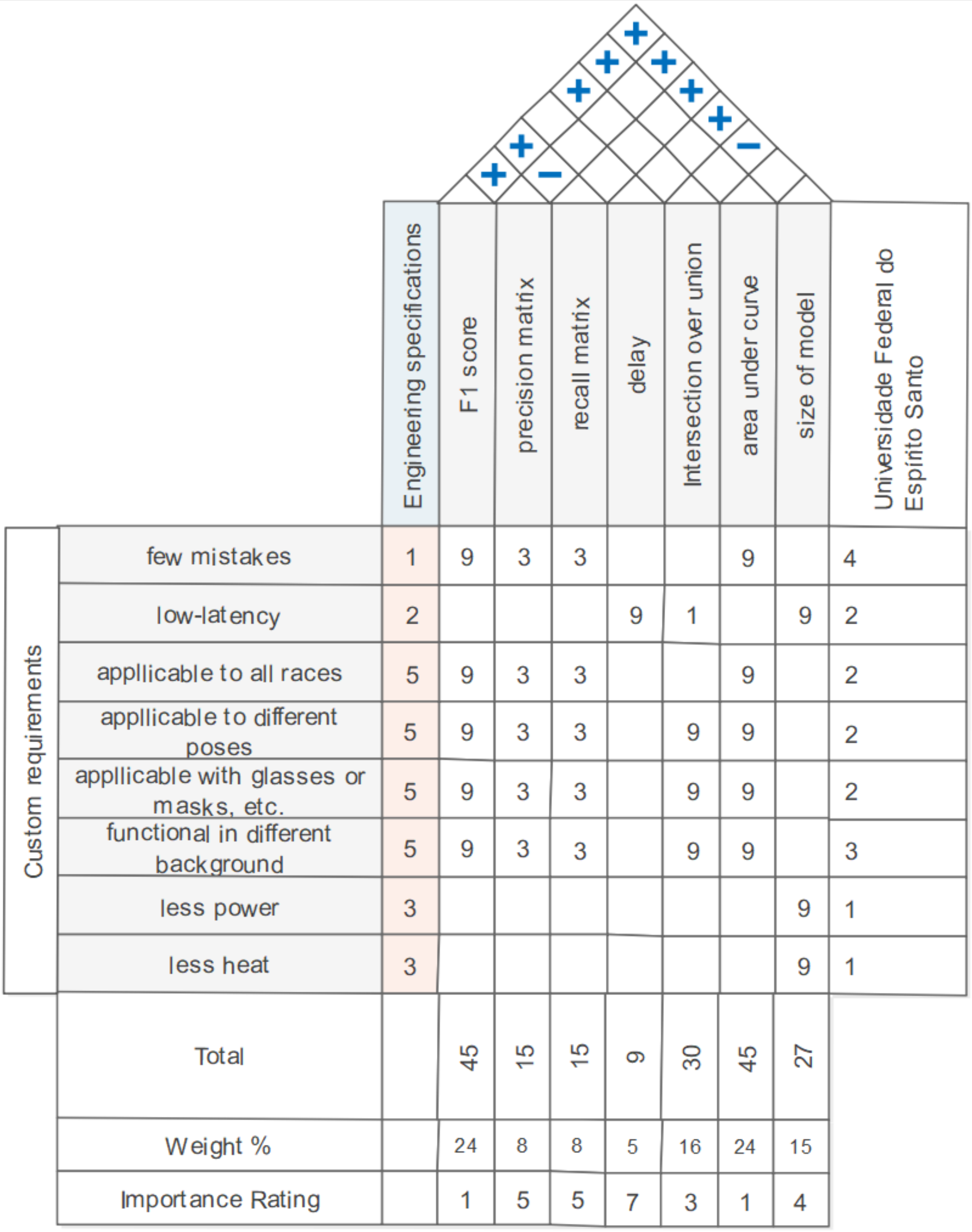
After all these discussions, a Quality Function Deployment is created as Figure 4.

As we can see, F1 score and AUC are the two most important engineering specifications, because they are the direct indexes showing how good the model's performance is. Behind them, it is IOU, since it indicates how the image data are processed and cropped. Then size of model is the forth important one, since it is related to time latency, heat generation and energy consuming. Then Precision matrix and Recall matrix are not so important since they only reflect one aspect of model's performance. Delay is the least important, since it may not be so significant if it is limited in several seconds, which is quite easy to realize.

6 Concept Description and Selection

6.1 Concept Generation

A deep learning project can be divided into two large sub-components, data processing and model training. More specifically, data processing includes face detection, face alignment,



		Engineering specifications	F1 score	precision matrix	recall matrix	delay	Intersection over union	area under curve	size of model	Universidade Federal do Espírito Santo
Custom requirements	few mistakes	1	9	3	3			9		4
	low-latency	2				9	1		9	2
	appllicable to all races	5	9	3	3			9		2
	appllicable to different poses	5	9	3	3		9	9		2
	appllicable with glasses or masks, etc.	5	9	3	3		9	9		2
	functional in different background	5	9	3	3		9	9		3
	less power	3							9	1
	less heat	3							9	1
	Total		45	15	15	9	30	45	27	
	Weight %		24	8	8	5	16	24	15	
	Importance Rating		1	5	5	7	3	1	4	

Figure 4: Quality Function Deployment for our project.

and data augmentation.

	Data Processing			Model Training
Method	Face Detection	Face Alignment	Data Augmentation	Model Building
1	HAAR	Homography	Keras Img Generator	SVM
2	OpenCV DNN		ImgAug Library	VGG16
3	MTCNN			InceptionResNetV2

Table 3: Morphological chart of the generated concepts and methods

6.1.1 Face Detection

The original image data contain large part of background, while only the face region is the desired region. Therefore, it is necessary to detect and crop the face region. There are three ways and their demonstrations are shown in Figure 5.

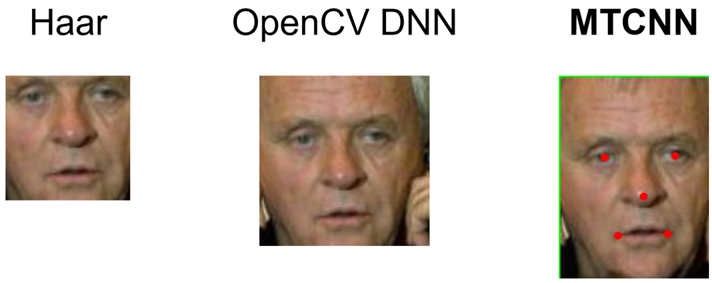


Figure 5: Demonstration of face detection.

6.1.2 Face Alignment

A deep learning can learn from the values of pixels in a image. Therefore, there should be an increase of accuracy if the facial pixels are aligned as landmarks located in assigned positions. The method used in OpenCV Homography. The experiment has been done to test the necessity of the step of face alignment. As shown in Table 4, after adding the step of face alignment, the accuracy can increase from 67.8% to 82.3%.

Accuracy before face alignment	Accuracy after face alignment
67.80%	82.30%

Table 4: Overall test accuracy trained on dataset of CK+

6.1.3 Data Augmentation

Up to now, the basic data processing has been done. But the size of data set is still small. There are only 1101 images. A large data set can increase the accuracy as well as decrease the possibility of overfitting. Therefore, the step of data augmentation is included. An experiment is conducted to test the necessity of the step. As shown in Table 5, after adding the step of data augmentation, the accuracy can increase from 82.3% to 93.9%.

Accuracy before data augmentation	Accuracy after data augmentation
82.30%	93.90%

Table 5: Overall test accuracy trained on dataset of CK+.

6.1.4 Model Building

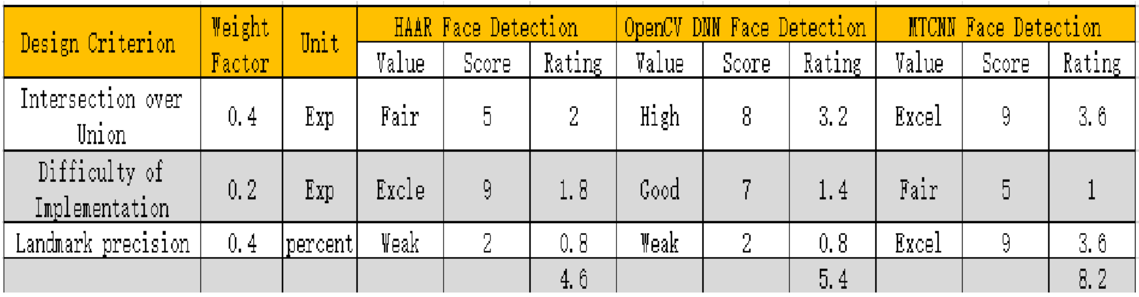
The deep learning model can be further trained based on a pre-trained model provided by Keras framework. Since it is much better in performance and faster in speed. There are three methods, SVM, VGG16, and InceptionResNetV2.

6.2 Concept Selection Process

There are three components that require a selection of concepts. They are face detection, data augmentation and model building. For the sub-component of face alignment, OpenCV Homography's performance is good enough. Therefore, no other methods are taken into considered.

6.2.1 Face Detection

For face alignment, three specifications are related, including IOU, difficulty of implementation, and landmark precision. Among the three methods of HAAR, OpenCV DNN, and MTCNN, MTCNN has the best performance as shown in Figure 6. Although its implementation is slightly more difficult that other two, it performs best in the other two specifications of totally 0.8 weight factor. The weight factor of each specifications is discussed and decided by all team members and instructors. In addition, MTCNN is the only one that can return facial landmarks of eyes, nose, mouth, etc.Therefore, MTCNN is selected.

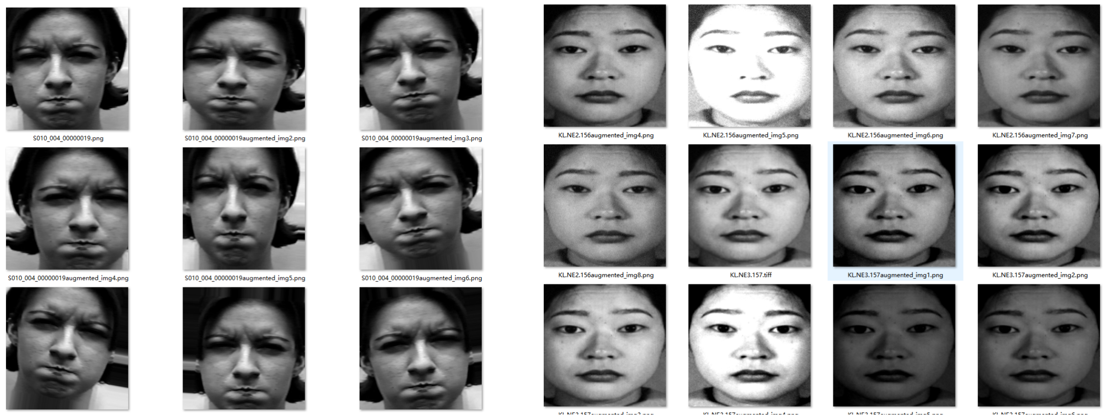


Design Criterion	Weight Factor	Unit	HAAR Face Detection			OpenCV DNN Face Detection			MTCNN Face Detection		
			Value	Score	Rating	Value	Score	Rating	Value	Score	Rating
Intersection over Union	0.4	Exp	Fair	5	2	High	8	3.2	Excel	9	3.6
Difficulty of Implementation	0.2	Exp	Excle	9	1.8	Good	7	1.4	Fair	5	1
Landmark precision	0.4	percent	Weak	2	0.8	Weak	2	0.8	Excel	9	3.6
					4.6			5.4			8.2

Figure 6: Scoring matrix of face detection concepts, including the methods of HAAR, OpenCV DNN, and MTCNN.

6.2.2 Data Augmentation

There are two tools that are used in data augmentation, Kears Image Data Generator and Imgaug Library. But the function and effect of two tools are completely different. The former one, as shown in Figure 7(a), is to change the positions and coordinates of an image, by rotation, shifting and so on. The latter one, as shown in Figure 7(b), is to change the color, lightness, contrast and so on. Since the images need to be aligned up, modification of coordinates are unnecessary. Therefore, Imgaug library is selected. Typically, data augmentation is done by horizontal flip, Gaussian flur, contrast normalization, additive Gaussian noise, and lightness modification.


S010_004_00000019.png
S010_004_00000019augmented_img2.png
S010_004_00000019augmented_img3.png
S010_004_00000019augmented_img4.png
S010_004_00000019augmented_img5.png
S010_004_00000019augmented_img6.png
KL.NE2.156augmented_img4.png
KL.NE2.156augmented_img5.png
KL.NE2.156augmented_img6.png
KL.NE2.156augmented_img7.png
KL.NE2.156augmented_img8.png
KL.NE3.157.tiff
KL.NE3.157augmented_img1.png
KL.NE3.157augmented_img2.png

(a) Demonstration of the effect of Keras Image Data Generator.

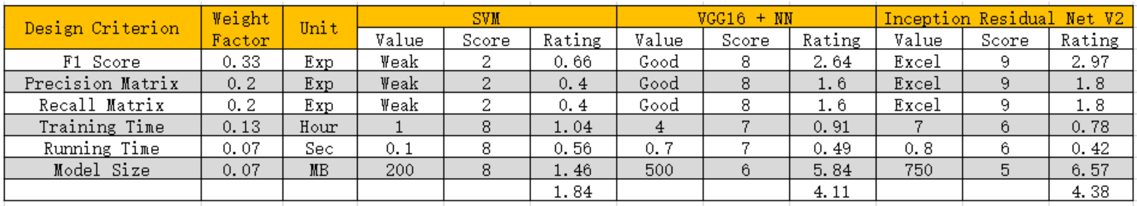
(b) Demonstration of the effect of Imgaug.

Figure 7: Demonstration of the effect of two data augmentation methods.

6.2.3 Model Building

For model building, there are three methods are considered, including Support Vector Machine (SVM), transfer learning model VGG16, and transfer learning Inception Residual

Network V2 (InceptionResNetV2). As shown in Figure 8. There are six specifications related to the task. They are F1 score, precision matrix, recall matrix, training time, running time and model size. Among them, SVM performs worst. InceptionResNetV2 performs slightly better than VGG16, because it has a higher accuracy.



Design Criterion	Weight Factor	Unit	SVM			VGG16 + NN			Inception Residual Net V2		
			Value	Score	Rating	Value	Score	Rating	Value	Score	Rating
F1 Score	0.33	Exp	Weak	2	0.66	Good	8	2.64	Excel	9	2.97
Precision Matrix	0.2	Exp	Weak	2	0.4	Good	8	1.6	Excel	9	1.8
Recall Matrix	0.2	Exp	Weak	2	0.4	Good	8	1.6	Excel	9	1.8
Training Time	0.13	Hour	1	8	1.04	4	7	0.91	7	6	0.78
Running Time	0.07	Sec	0.1	8	0.56	0.7	7	0.49	0.8	6	0.42
Model Size	0.07	MB	200	8	1.46	500	6	5.84	750	5	6.57
					1.84			4.11			4.38

Figure 8: Scoring matrix of model building concepts, including the methods of SVM, VGG16 transfer learning model, and InceptionResNetV2 transfer learning model.

6.3 Selected Concept Description

In Figure 9, a face image will be detected and cropped out. Then it can be stretched and put in a 299-by-299 square with five landmark points located in the five assigned points. After that, the data set can be expanded by ImgAug, containing the images modified from the original one.

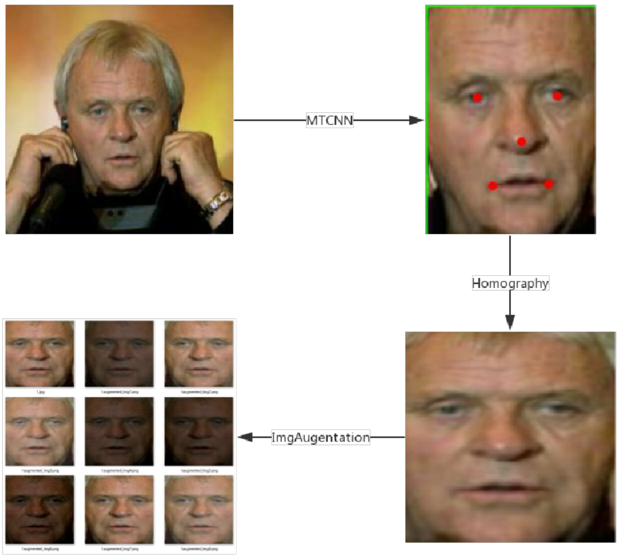


Figure 9: Selected concepts and corresponding demonstration in data processing part, including MTCNN for face detection, OpenCV Homography for face alignment, and ImgAug for data augmentation .

For model building and training part, InceptionResNetV2 is used and its structure is shown in Figure 10

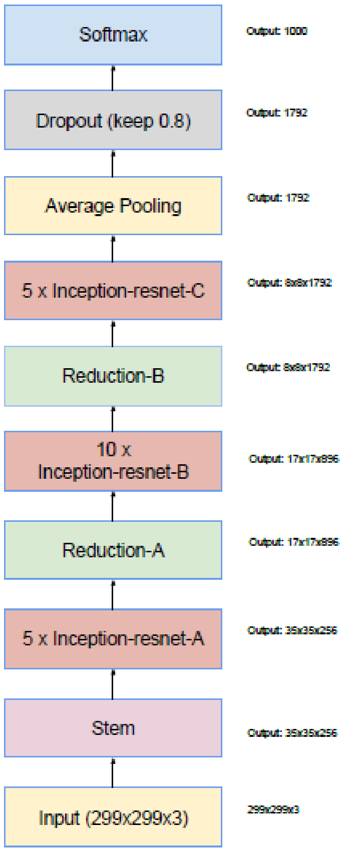


Figure 10: Selected concept and demonstration in model building part, InceptionResNetV2 for transfer learning. [SIVA17]

6.3.1 Engineering Design Analysis

In the project, the engineering fundamentals are how well the human face is cropped, how well the face is aligned, how well the data set is classified and expanded, as well as how well the training model performs. In particular, the image data processing is related to the Computer Vision and deep learning territory. On the other hand, the model building of emotion recognition is related to deep learning and it is essentially a classification problem. In general, both of them can be seen as deep learning problem.

In the project, there are two parameters that can be tuned. One is the threshold of confidence for face detection. The number chosen is 80%, which is tuned after many trials. The value can make MTCNN recognize all wanted faces. The other one is the positions in a 299-by-299 square assigned for landmarks to be located, The length of the side is set to be fitful to the transfer learning model, InceptionResNetV2, while the positions are obtained from an average value of multiple squared faces.

Threshold	Result
70%	1 wrong faces detected & 10 failures in detection (Total 1101)
80%	3 wrong faces detected & 6 failures in detection (Total 1101)
90%	13 wrong faces detected & 4 failures in detection (Total 1101)

Table 6: The comparison between different thresholds of confidence for face detection in MTCNN

As shown in Table 6, when the threshold of confidence for face detection is either 70% or 90%, there are many wrong faces detected or failures in detection. But when the value is 80%, the two mistakes are minimized to single digit.

	Positions of Squared Face									
	Left Eye		Right Eye		Nose		Left Mouth		Right Mouth	
	X	Y	X	Y	X	Y	X	Y	X	Y
1	69.3	111.5	213.2	111.5	139.7	170.2	76.0	222.5	196.7	222.5
2	68.4	111.7	213.8	111.7	139.8	169.8	77.3	222.1	196.2	222.3
3	68.4	111.2	212.8	111.2	140.2	171.3	76.7	221.6	197.1	222.6
4	69.5	111.0	214.3	111.0	140.5	171.4	76.9	222.6	197.1	222.0
5	69.0	112.2	213.3	112.2	140.7	170.5	76.1	222.7	197.1	222.5
Avg	69	112	213	112	140	171	77	222	197	222

Table 7: Experiments on deciding assigned positions of landmarks in face alignment

When deciding the parameter of positions of landmarks in face alignment, five samples are manually made into squared and the positions of five landmarks are measured. In Table 7, they are listed and the average values are set as the assigned positions of landmarks.

7 Final Design Description

In Figure 11, the flow chart of the layout and demonstrations of all sub-components is shown. The original images' face regions are cropped by MTCNN also with 5 landmarks together. Then by OpenCV Homography, face regions can be stretched into a square and be aligned with each other based on the 5 landmarks. Then by Imgaug library, data set is expanded by 9 times. After that, the processed data are input in a model of transfer learning, InceptionResNetV2. Eventually, the final model is generated.

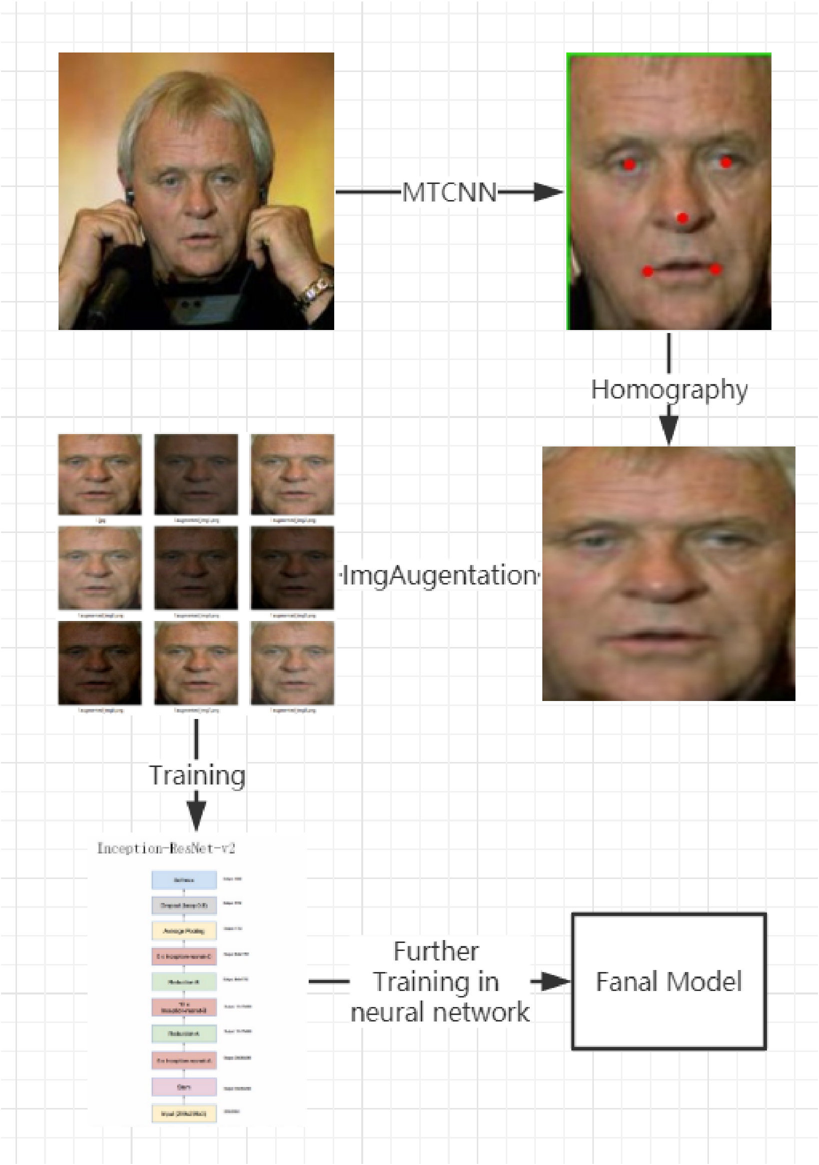


Figure 11: Flow chart of the layout of all sub-components in the project.

8 Manufacturing Plan

In the project, all the required and used languages, tools, and softwares are listed in Table 8. The detailed explanation of the algorithms are as follows.

For MTCNN, there are three nets, P-Net, R-Net and O-Net. P-Net is a shallow CNN, whose function is to produce candidate windows. R-Net is a complex CNN, whose function is to refine the windows by rejecting non-face windows. R-Net is a complex CNN, whose function is to refine the windows & outputting 5 landmark positions, including 2 eyes, nose and 2 corners of mouth. The detailed network structures are shown in Figure 12. [ZZLQ16]

	Item
Language	Python Shell Script
Tools & softwares	Tensorflow OpenCV Homography OpenCV HAAR OpenCV DNN MTCNN Framework Keras Img Generator Imgaug Library SVM Keras VGG16 Keras InceptionResNetV2

Table 8: Tools, languages and softwares used in the project.

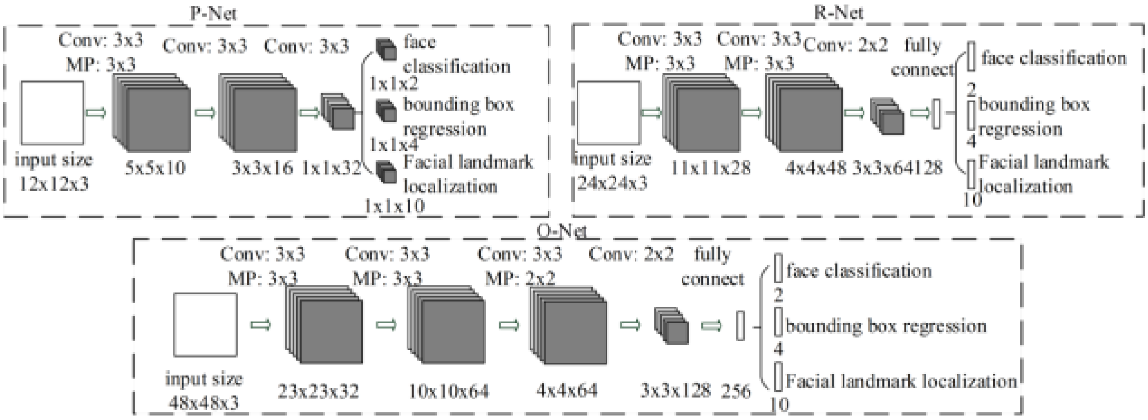


Figure 12: The structures of MTCNN's P-Net, R-Net and O-Net. [ZZLQ16]

For Homography, it is a transformation between two planes in 3 dimension space, which is literally a 3×3 matrix. The Homography matrix (noted as H) can be calculated by using 4 corresponding points between the two planes with more points, more accurate. Since H is a 3×3 matrix, it can be written as Eq. (2).

$$H = \begin{bmatrix} h_{00} & h_{01} & h_{02} \\ h_{10} & h_{11} & h_{12} \\ h_{20} & h_{21} & h_{22} \end{bmatrix} \tag{2}$$

Then for a group of corresponding points, (x_1, y_1) in the first plane and (x_2, y_2) in the second plane, their relation can be represented as Eq. (3).

$$\begin{bmatrix} x_2 \\ y_2 \\ 1 \end{bmatrix} = H \begin{bmatrix} x_1 \\ y_1 \\ 1 \end{bmatrix} = \begin{bmatrix} h_{00} & h_{01} & h_{02} \\ h_{10} & h_{11} & h_{12} \\ h_{20} & h_{21} & h_{22} \end{bmatrix} \begin{bmatrix} x_1 \\ y_1 \\ 1 \end{bmatrix} \tag{3}$$

So yy using Homography matrix, all faces can be aligned into same position. [Mal16]

For InceptionResNetV2, as shown in Figure 10, it is mainly a deep convolutional neural network, which contains three residual blocks. The output is the classification results of the images. An Adam optimization is applied to find the weights in the network. Residual network is the piling up of the residual blocks, which can make the network deeper. For a usual deep neural network, the loss will usually decrease in the beginning and then increase later. But for a residual network, the loss will keep decreasing as the network gets deeper. On the other hand, the design of inception can enable the model self-adaptively choose proper parameters.

Our budget is mainly used for renting GPU server from Shanghai Jiao Tong University as shown in Table 9. The databases and tools we use are all free. Our budget will only come from JI and HASCO VISION. The original budget is enough for our project.

Item	Cost
Tensorflow & Keras	0
Databases, e.g. CK+, FacesDB, JAFFE, etc.	0
OpenCV Library	0
Imgaug Library	0
MTCNN Framework	0
GPU Server	≈100
Total	≈100

Table 9: The budget of the project [Unit: CNY]

9 Validation Plan

For precision matrix and recall matrix, the test set is input into model and the output result is compared with corresponding labels. Then precision and recall matrices can be obtained. For delay, the model is run on real-time camera. The average time of generating a result can be measured. For intersection of union, an ideal face region is between lower jaw and forehead vertically, and between left and right ears horizontally. The cropped face area over the ideal face area is high. The performance of IOU can only be judged manually. The experiments are done on cross-validation (CV) set and test set. Therefore, the data set is divided into training set, cv set, and test set, with proportion of 0.67, 0.13 and 0.2 respectively.

No.	Engineering Specification	Target Value	Current Value
1	F1 score	>75%	93.90%
1	Area Under Curve (AUC)	>75%	92.20%
3	Intersection Over Union (IOU)	High	High
4	Size of model	<2048MB	750MB
5	Precision matrix	>75% for each diagonal entry	≈90%
5	Recall matrix	>76% for each diagonal entry	≈90%
7	Delay	<2000ms	500-1800ms

Table 10: Validation results with respect to engineering specifications

10 Test and Validation

We validate our model basically on the F1 Score, which denotes the general performance, while other parameters mentioned in the specification part, including AUC, IOU, size of model, precisions, recalls and delay, are also taken into consideration. The combined dataset is split randomly into three parts, the training, validation and test sets, by the ration of 8:1:1. Training set is utilized to train the model, the validation set is utilized to model and method selection while the test set gives the final performance. The images are augmented after the sets division to make the results more convincing. The F1 score of our final model is some value , much higher than the baseline mentioned 75%. All the parameters related are shown in the following table and we can see that all the test results meet the baseline and some of them largely exceeds the target values.

Our model's performance is also competent to those models in recent papers. Since most of them are validated on the accuracy on CK+ and OuluCasia separately, we do the same test as well.

11 Discussions and Recommendations

11.1 Discussions

Since the method involved in the project is transfer learning, a completely new Convolutional Neural Network is not able to be trained and tuned due to time limitation. Also, the main target of the project is to on data processing. The tuning and training of the model is not paid enough attention on. But the test accuracy and training result have been already good enough. Therefore, the tuning and self-designed deep learning model is unnecessary. If time permits, training and tuning a new model might be helpful in increasing the performance of the model.

11.2 Recommendations

First and foremost, we would recommend to use CNN as the base model since it performs generally the best according to the papers we investigated and compared. Data augmentation is also strongly recommended since there're only limited number of free open-source databases. In our project, the average accuracy increases around 20% after augmenting.

There're still some possible improvements they could try to implement in the future. Currently, our model doesn't perform as well on faces with glasses. They could try to utilize GAN to remove the glasses before training and testing. Although this would add to the processing time, the adaptability of the model would further improve since not a few drivers would wear sunglasses or glasses while driving.

The representative expressions in those databases are quite exaggerated actually, in order to make the emotions distinguishable enough. However, in the real world, those unintentional expressions are more difficult to detect, which will significantly influence the performance. There do exist some databases with micro expressions, but it won't be a good idea to simply combine them with the original images. We recommend them to do the micro expression recognition as a split task or they could try to do a 14-classes (or less if they remove some not so important classes) classification where, for example, the anger and mirco-anger are considered two different classes.

12 Project Plan & Design

The project plan is organized as Figure 13. Requirement survey and literature research have been done within the beginning three weeks, except the first week for forming our team. Then from week 3 to week 4, necessary datasets have been done. From 5 to Week 6, raw data are processed by the procedure of face cropping, face alignment and data augmentation. Before week 9, we have finished building our framework and models. After that to week 12, the model and framework can be improved to be even better. In the last week, the demonstration and prepare for exposition should be done.

Among all the milestones, data processing is the most important because with well processed data, performance will be improved significantly. Then model building and improving are less important, but they are still very vital. Requirement survey should also be vital, since it is a direction of our heading.

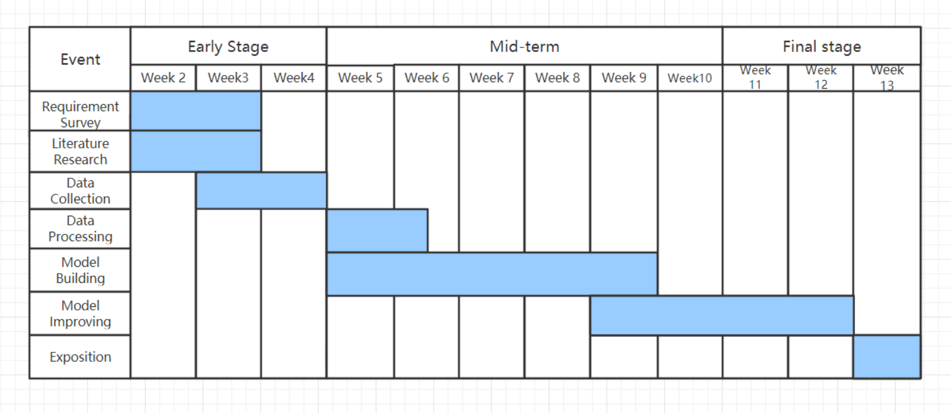


Figure 13: Gantt chart for our project plan.

In the project, Shangquan Sun is responsible for almost everything, including collecting data, processing data, writing and organizing reports, preparing presentations, training models, improving models and so on. And Jie Gong mainly does the work of collecting data, writing and organizing reports, doing presentations, training and building models, improving models and so on. Jianshu Li does the work of writing reports, doing presentations, processing data, improving models and so on. Xuanyu Wang does the work of doing presentations, and so on.

13 Analysis of Potential Problems

At first, we didn't fix the training, test and CV sets but randomly generated them by a fix ratio (0.67:0.13:0.2) which makes the comparison among the performance of different methods not very convincing. We first fix the division of the sets and compare the methods and models again.

We did the data augmentation before splitting the data sets. As a result, images generated from the same original image may fall into both the test and training set. So the accuracy we obtained might be a little bit higher than the true value.

In the face cropping part, the MTCNN actually returns a series of faces with confidence over 0.8. Previously we choose the face with highest confidence as the driver's face since in most cases the closet face gives the highest confidence. However, after some experiments we found that this method may choose the further face as the target and fails. Then we modified it to detect the largest face with at least 0.8 confidence instead. Up till now, this method works well but it is not safe to say that 0.8 is the proper threshold value. More experiments are required to determine a more reasonable threshold value.

14 Conclusion

Nowadays, Facial Expression Recognition is increasingly popular with the development of Deep Learning and Computer Vision. the project will be based on the preeminent works which have done by the predecessors. The experiments are conducted on three databases, i.e. CK+ [LCK$^+$10], JAFFE [LAK$^+$98], FacesDB [JPMC12]. After the process of data processing, including face detection, face alignment, and data augmentation, the facial images are detected, cropped, aligned and the data set is expanded and diversified. The task of face detection is done by MTCNN. The task of face alignment is done by OpenCV Homography. The task of data augmentation is done by Imgau Library. With the process, accuracy can be increased significantly. Then the processed database is divided into training set, cv set and test set. After inputting them into transfer learning model, InceptionResNetV2, and a successive neural network, a final model can be generated. The test accuracy of the model is about 98%, which means a success. In real-time test on camera, it is possible to have a wrong result if the facial expression is not inflated. This is in a territory of the classification and predition on micro-expression, which is still a tough problem to be figure out.

Reference

[ACF$^+$16] Deepali Aneja, Alex Colburn, Gary Faigin, Linda Shapiro, and Barbara Mones. Modeling stylized character expressions via deep learning. In *Asian Conference on Computer Vision*, pages 136–153. Springer, 2016.

[JLY$^+$15] Heechul Jung, Sihaeng Lee, Junho Yim, Sunjeong Park, and Junmo Kim. Joint fine-tuning in deep neural networks for facial expression recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 2983–2991, 2015.

[JPMC12] Luiz Velho Jesus P. Mena-Chalco, Roberto Marcondes Cesar-Jr. Impa-faces3d: 3d facial expression database (raw-data). In *FacesDB VISGRAF faces database*. VISGRAF, 2012.

[KLS18] Chieh-Ming Kuo, Shang-Hong Lai, and Michel Sarkis. A compact deep learning model for robust facial expression recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 2121–2129, 2018.

[LAK$^+$98] Michael J Lyons, Shigeru Akamatsu, Miyuki Kamachi, Jiro Gyoba, and Julien Budynek. The japanese female facial expression (jaffe) database. In *Proceedings of third international conference on automatic face and gesture recognition*, pages 14–16, 1998.

[LCK$^+$10] Patrick Lucey, Jeffrey F Cohn, Takeo Kanade, Jason Saragih, Zara Ambadar, and Iain 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*, pages 94–101. IEEE, 2010.

[LdADSOS17] André Teixeira Lopes, Edilson de Aguiar, Alberto F De Souza, and Thiago Oliveira-Santos. Facial expression recognition with convolutional neural networks: coping with few data and the training sample order. *Pattern Recognition*, 61:610–628, 2017.

[Lin15] Allison Linn. Happy? sad? angry? this microsoft tool recognizes emotions in pictures. `https://blogs.microsoft.com/ai/happy-sad-angry-thi` s-microsoft-tool-recognizes-emotions-in-pictures/, 2015. Accessed: 2019-06-18.

[LSWC14] Mengyi Liu, Shiguang Shan, Ruiping Wang, and Xilin Chen. Learning expressionlets on spatio-temporal manifold for dynamic facial expression recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1749–1756, 2014.

[Mal16] Satya Mallick. Homography examples using opencv. https://www.learnopencv.com/homography-examples-using-opencv-python-c/, Jan 2016.

[SIVA17] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[SZ14] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[VP10] Michel Valstar and Maja Pantic. Induced disgust, happiness and surprise: an addition to the mmi facial expression database. In *Proc. 3rd Intern. Workshop on EMOTION (satellite of LREC): Corpora for Research on Emotion and Affect*, page 65. Paris, France, 2010.

[WLL$^+$10] Shangfei Wang, Zhilei Liu, Siliang Lv, Yanpeng Lv, Guobing Wu, Peng Peng, Fei Chen, and Xufa Wang. A natural visible and infrared facial expression database for expression recognition and emotion inference. *IEEE Transactions on Multimedia*, 12(7):682–691, 2010.

[YCY18] Huiyuan Yang, Umur Ciftci, and Lijun Yin. Facial expression recognition by de-expression residue learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2168–2177, 2018.

[ZZLQ16] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, 2016.

Appendixes

Frameworks

The list of the frameworks and open-source projects are listed as following.

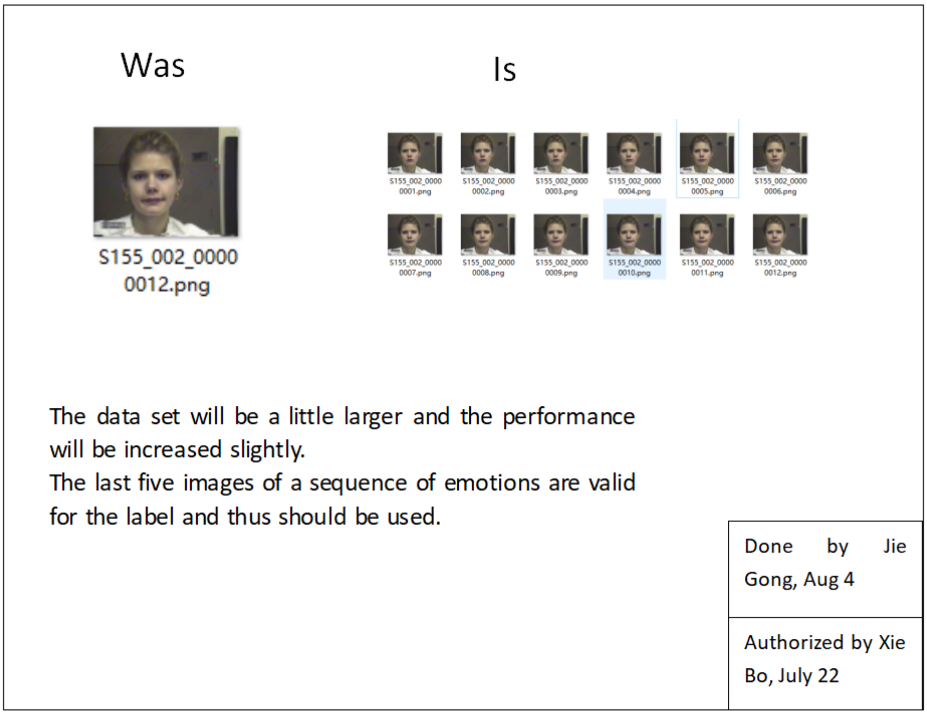
1. The framework for transfer learning is modified from `https://github.com/gauravtheP/Real-Time-Facial-Expression-Recognition`.

2. The codes of real-time-camera user interface are modified from the combination of `https://github.com/gauravtheP/Real-Time-Facial-Expression-Recognition` and `https://github.com/xionghc/Facial-Expression-Recognition`.

3. The codes for data augmentation are based on ImgAug Library. The official website is `https://imgaug.readthedocs.io/en/latest/source/augmenters.html`.

4. The codes of MTCNN for face cropping are from `https://github.com/AITTSMD/MTCNN-Tensorflow`.

Engineering Changes Notice



Was
S155_002_0000
0012.png
Is
S155_002_0000
0001.png
S155_002_0000
0002.png
S155_002_0000
0003.png
S155_002_0000
0004.png
S155_002_0000
0005.png
S155_002_0000
0006.png
S155_002_0000
0007.png
S155_002_0000
0008.png
S155_002_0000
0009.png
S155_002_0000
0010.png
S155_002_0000
0011.png
S155_002_0000
0012.png

The data set will be a little larger and the performance will be increased slightly.
The last five images of a sequence of emotions are valid for the label and thus should be used.

Done by Jie Gong, Aug 4

Authorized by Xie Bo, July 22

Bill of Materials

Item	Cost
Tensorflow & Keras	0
Databases, e.g. CK+, FacesDB, JAFFE, etc.	0
OpenCV Library	0
Imgaug Library	0
MTCNN Framework	0
Traffic cost	300
GPU Server	95
Total	395

Table 11: Bill of Materials

Acknowledgements

Thanks to Prof. Bo Xie, Yueping He and their colleagues for instructing and advising us.
Thanks to Prof. Yong Long for instructing and advising us.
Thanks to the Judge Panel for excellent lectures, answering our questions, listening our presentation and helping us.
Thanks to Allen Zhu from JI research office, for organizing the team and helping on requesting a GPU server.