Facial Expression Classification Using R-CNN Based Methods

Pengfei Sun

A thesis submitted to the Auckland University of Technology
in partial fulfillment of the requirements for the degree of
Master of Computer and Information Sciences (MCIS)

2018

School of Engineering, Computer and Mathematical Sciences

Abstract

With the rise of artificial intelligence technology and the development of human vision research, most of researchers are gradually putting more and more attention on the machine recognition of face images. In this thesis, we will study facial expression recognition based on deep learning algorithm. Facial expression is the display of one or more movements or states of facial muscles. Facial expression recognition technology is primarily used in human-computer interaction, intelligent control, security, medical, communication and other fields.

In this research, we utilize two deep learning algorithms to implement facial expression recognition system. The first recognition algorithm based on Faster R-CNN, which consists of a fully convolutional network and detector over a region of interest. The second algorithm is Mask R-CNN, which is an extension of Faster R-CNN algorithm that performs image segmentation. Facial expressions are divided into seven categories: anger, contempt, disgust, fear, happy, sadness and surprise. They have been used for object detection and recognition but have not been applied to facial expression classification before. Our experiments show that, compared with the conventional methods, these methods avoided the tedious manual feature extraction, reduced the number of parameters and significantly improved the recognition rate. Moreover, the performance of the trained model in the more realistic settings where the position and angle of the face, lighting, background, etc. are varied are reported in this work.

Table of Contents

List of Figures

List of Tables

Attestation of Authorship

I hereby declare that this submission is my own work and that, to the best of my knowledge and belief, it contains no material previously published or written by another person (except where explicitly defined in the acknowledgments), nor material which to a substantial extent has been submitted for the award of any other degree or diploma of a university or other institution of higher learning.

Signature: Date: 17 September 2018

Acknowledgment

I would like to deeply thank my parents for their financial support during my entire time of academic study in Auckland.

My deepest thanks are to my primary supervisor Professor Edmund Lai who has provided me with much technical guidance and support. I believe that I could not have been able to achieve my Master Degree without his invaluable help and supervision. In addition, I would like to appreciate my secondary supervisor Dr. Wei Qi Yan and administrators in our School for their support and guidance through the MCIS in the past two years.

Pengfei Sun

Auckland, New Zealand

September 2018

Chapter 1 Introduction

1.1 Background and Motivation

Facial expression recognition is a cross-field between computer vision and pattern recognition. As one of the most important external discriminant features of human beings, facial expressions can be used in human-computer interaction, intelligent control, security and medical fields. For example, in Mitchell Gray's study, a facial recognition surveillance mechanism was proposed to respond to urban spatial insecurity perception [90]. The recognition system implemented in Milligan can accurately identify the face with glasses or try to disguise [91]. Moreover, facial expressions, as one of the most important manifestations in social life and human communication, contain a large amount of information.

Facial expression is expressed through facial muscle activities to convey emotions. Facial expression is a form of nonverbal communication. The six common facial expressions of human beings are happy, surprise, anger, disgust, and fear. Besides the above six expressions, the primary research expression of this thesis also has contempt expression. As one of the most important external discriminant features, human face has the uniqueness as human fingerprint and can be used to identify a person. Therefore, facial expression recognition technology is widely used in human-machine interaction, intelligent surveillance, security, medical communication and other fields.

Two examples of real facial expression applications are given below. The first is a customer sentiment analysis of a shopping mall or physical store. The merchant uses a camera to capture a customer's image of a shopping mall or a store. By analyzing the facial expression, then further interpreting the consumer's emotional information, and

finally analyzing the consumer's experience in shopping malls or stores. The second example is about human-machine interaction. Facial expression recognition system is applied to a project of the educational assistant robot. One of the functions of the educational assistant robot is to judge the emotion and psychology of the user in the front of the robot through facial expression analysis.

Facial expression recognition is a cross-field between computer vision and pattern recognition. Computer vision is the use of various imaging systems instead of visual wonders as input means of visual information. Specifically, computers replace the brain to complete processing and interpretation and make corresponding decisions based on the interpretation results. Pattern recognition is the process of matching computer-generated images with known classes. In the process of recognition, objects represented in computer vision are mapped from feature space to model space. In the following content, we are concerned with some literature on these two theories.

Next, we need to clarify the main problems of facial expression recognition. In the process of facial expression recognition, different illumination, whether there is background and whether the image has rotation will affect the quality of the image and then affect the recognition results. In the following studies, we focus on the next five aspects.

The influence of background on facial expression recognition. Theoretically, the background will have an adverse effect on the recognition system. It is a recognition task to distinguish the background and the target object from an image. Therefore, the complexity of the background will affect the complexity of the recognition calculation. The effect of illumination on facial expression recognition. Illumination can change the structure of the target image, and the object's texture and contour will deviate. Even for the same face, the face images obtained under different illumination will be different. The

influence of rotation of targets on facial expression recognition. In the process of face image acquisition, the human face will not always be positive. The face will rotate, including deep rotation and plane rotation. The impact of facial sizes on facial expression recognition. In the process of facial emotion recognition, the face size in the image can successfully detect in the effective range; the face size is too extreme small or extreme large will cause difficulties in the recognition process. The influence of face position on facial emotion recognition. For facial emotion recognition, face position whether or not create change of recognition results. This question is a crucial factor in this thesis.

Facial expression recognition is trying to determine the class attribute of facial expression. Specifically, a facial expression is assigned to a particular type. Basically, facial expression recognition involves image acquisition, image preprocessing, feature extraction and classification. Image acquisition is to obtain static images or dynamic image sequences by image capturing tools such as cameras. Image preprocessing is to extract images from a specific environment without being influenced factors. As far as feature extraction is concerned, image data are dimensionality reduction under the premise of ensuring stability and recognition rate as far as possible. Classification divides the image features into image types and determines the unknown attributes of the image as a certain type of image. The main classification methods for expression recognition include linear classification, neural network classification, support vector machine and hidden Markov model. The following content we have introduced some methods of facial expression recognition preliminarily. The related research of facial expression recognition methods will be introduced explicitly in the following sections.

The geometry feature analysis is one of the earlier methods in face recognition research. This method combines the feature values of the target into a vector. Then, the feature vectors of the target to be identified are compared with the feature vectors of the samples in the database by the distance formula to get the recognition results. The

principal component analysis is widely used in facial expression recognition because of its simplicity and fast recognition speed. In the recognition process, the principal component features of the image are compared with the principal component features of each face image in the face image library. And then, the recognition result of the image to be identified can be obtained. In addition to the above algorithms, the artificial neural network is the most popular recognition algorithm for most of face recognition research. Meanwhile, the regions with CNN algorithm can not only realize the recognition of the object that to be identified but also achieve accurate positioning of the target object in the image. Therefore, the algorithm has been widely used. The task of facial expression classification from a single image is not trivial because of the richness and variety of facial expressions, the complexity of facial expression changes and the dynamic characteristics of facial expressions. Facial expressions can change continuously and a single image can only capture one instant of the expression. Furthermore, there are also gender and cultural differences in expression. There is a significant difference between males and females in expressing the same emotion. In addition, people from different cultural backgrounds express emotions differently. These factors make it difficult to analyze and recognize facial expression features.

In the field of facial expression classification, there are many kinds of recognition techniques. The Eigenface [85] is a classical facial recognition technique. It is widely used in studies of facial expression recognition. The method projects the image into different spaces and uses the distance to judge the expression. Sequences of images are used by Maninderjit Singh et al. to create Bayesian networks [86] to tackle the task while Abhishek Kumar and Anupam Agarwal made use of anatomical information [87]. The latter method has the characteristics of simple operation and fast classification speed.

The above methods rely on features that are selected manually by the researchers. More recently, deep neural networks which do not require manual feature are being favored. Frank Wallhoff et al. [88] used a hybrid of connectionist and hidden Markov model approach to identify face contours in databases. Hyun-chul Choi et al. [89] used

multilayer perceptron to achieve real-time facial expression classification. Better classification accuracies have been reported in these works compared with manual feature selection-based methods.

In the deep learning method, R-CNN introduced CNN into the field of object detection, which greatly improved the effect of object detection. Therefore, more and more researches are devoted to the realization of R-CNN related algorithms in object detection. R-CNN algorithm is composed of four modules, which is candidate region generation, feature extraction, category judgment and position correction [92]. For example, in the research of Ren Shaoqing et al, region proposal network is introduced on the basis of traditional convolution neural network. It shares the full image convolution features with detection network, thus realized high efficiency object detection [93]. For R-CNN related researches, almost all the R-CNN research is aimed at real-time object detection, and the related R-CNN algorithm is not widely used in the research of facial expression recognition. Moreover, some traditional recognition methods and deep learning methods have poor expression recognition rate. Therefore, according to the comparison of traditional recognition methods and deep learning methods in the section of literature review, we proposed our recognition network model Faster R-CNN and Mask R-CNN.

In this thesis, we realized facial expression recognition based on Faster R-CNN and Mask R-CNN respectively. The results show that the designed two methods have excellent recognition ability. Under different facial sizes and illumination, the Mask R-CNN is less affected that the Faster R-CNN model. Moreover, the Mask R-CNN model also has better facial localization ability than the Faster R-CNN model.

1.2 Objectives

The primary objective of this thesis is to investigate the effectiveness of R-CNN methods for the classification of facial expressions based on single images. The performance of two variants of R-CNN, namely Faster R-CNN and Mask R-CNN, will be studied. The primary research question is therefore:

Q1: Which of the deep learning architectures – Faster R-CNN or Mask R-CNN, is able achieve better facial expression classification results based on single images?

Datasets with different facial sizes, positions, postures, illumination and backgrounds will be used for this study in order to simulate images taken in real life situations. Hence a secondary research question is:

Q2: What are the effects of different facial sizes, positions, illumination and backgrounds on facial expression classification?

Some preliminary research will also be conducted to address cases where the facial expression does not definitely belong to one class or another. Hence a third research question is:

Q3: How can ambiguous recognition results be resolved?

1.3 Structure of This Thesis

The thesis is structured as follows:

In Chapter 2, literature review will be described, such as the previous studies in traditional recognition method and deep learning-based method for face recognition and expression recognition. Facial expression recognition has many kinds of research and methods. Thus, Chapter 2 described the basic of feature extraction and various recognition and classification methods which will be applied to facial expression

recognition.

In Chapter 3, the explanation of research methodology of this thesis will be introduced. Part of solutions and answers of research questions are illustrated. In addition, the design of the experiment and algorithm, as well as datasets and realization process with the evaluation methods, will be presented.

In Chapter 4, our proposed methologies and algorithms are implemented. In addition, experimental results and outcomes will be detailed with support of confusion matrixes and demonstrations.

In Chapter 5, precision, recall, F-score and accuracy are applied to evaluate the classification of different facial expressions in two trained models. Based on the data, we carried out detailed analysis and discussion. Finally, the conclusion and future work will be presented in Chapter 6.

Chapter 2 Review of Facial Expression Classification

Published research in facial expression classification generally use one of two approaches. They are feature-selection based methods and deep learning methods. Some of the representative works using these methods are reviewed in this chapter.

2.1 Feature-Selection Based Methods

In the process of facial expression recognition, the establishment of facial expression model is an important part. In this part, facial expression feature extraction is also a key link. With the development of face recognition and expression recognition, various and efficient feature extraction methods have been developed. In this section, we are introduced to some typical facial expression feature extraction methods.

2.1.1 Geometric Feature Analysis

Geometric feature method is one of earliest feature extraction methods in face recognition research. The method is that combines the feature value of the object into a vector and then use distance formula to compare the feature value of the object that to be identified and the feature vectors of the known objects.

R. J. Baron has proposed a method that called template matching [1]. This method is based on geometric facial features to establish the face models. In addition, Brunelli and Poggio have proposed another template matching method [2] that similar to Baron's research work. In the work of Brunelli and Poggio, they built geometric features method

on 47 individual databases and achieved better recognition efficiency. In the implementation of the template matching method, the facial image used is based on constant scales, orientations, and illumination. For real environment, it is difficult to achieve all conditions in the continuous state. Therefore, it is impossible to conclude that the universality of the study according to their research work.

Compared with the research results of R. J. Baron and Brunelli, the research work of Turk and Pentland has been further improved. Principal component analysis method [3] is used in the study of facial feature extraction. In the experiment, they map each facial image to a single point in extreme high dimensional space. In the high dimension space, a pixel matrix of size L*L dimension is formed. By a weighted combination of base vectors arbitrarily in high-dimensional space, the intensity of each pixel is obtained. They use principal component analysis to find the best image data that is low-dimensional projection. Finally, the collection data is classified by the nearest neighbor classification [4]. The experiment eventually reached 95% recognition rate based on 3000 different face images of the database. Although Turk and Pentland's research has achieved a high level of recognition rate, all the facial images in the experiment still maintain the constant scales, orientation, and illumination.

In an early study, Goldstein et al. reported a “face-feature questionnaire” [6]. They manually extracted 22 features from the facial images, and then used them to identify faces in a database with a range of 64 to 255. The only drawback is that the study achieved only 50% accuracy. Kanade describes a method to extract facial features automatically. The realization of this method is that the recognition rate reaches 45% to 75%. Ingemar J. Cox proposed a mixture-distance [8] technology that has been dramatically improved regarding method reliability and recognition rate. By using the inverse Euclidean distance weighting [9] of variance, the Ingemar J. Cox’s method optimizes the early workload of Kanade’s research. This study has successfully made up for the lack of previous research

work. Their research results ensure that the method is efficient based on varying scales, rotations, or illuminations, with high recognition rates. In their approach, the researchers extracted 30 features from each face by hand, while the experiment was to identify 95 face images from 685 samples. Experimental results show that the use of mixture-distance method can improve facial recognition performance from 84% to 95%.

In 2007, a method for recognition of six basic facial expressions was proposed by Irene kotsia and Ioannis Pitas [10]. For the feature extraction of facial expressions, the investigators used the Candide grid nodes and using support vector machines [11] [12] [13] [14] to identify dynamic facial expressions. This method maps and tracks facial model Candide to each frame of video. Therefore, the researchers manually placed the Candide grid nodes [15] on the face image. As time goes by, the tracking system detects facial expression until the expression reaches the highest value. The results show that the recognition rate of facial expression reached 95.1%. The influence of illumination or orientation on feature extraction also decreased.

2.1.2 Principal Component Analysis

The implementation of the principal component analysis method is based on the average value of the face which is calculated by the expected values of all facial image matrices in the database. Also, the principal eigenvector and the principal eigenvalue are computed according to the standard deviation of each face image. In the process of recognition, the recognition result is obtained by comparison of the principal component features of the image that to be identified and principal component features of each face image in the database. Also, PCA is used for feature extractors in many face recognition studies [16] [17] [18].

Different from the traditional PCA method, the 2-dimensional PCA method can

estimate the covariance matrix more accurately and reduce the computational complexity of the feature vector. In addition, the 2-dimensional PCA method has fewer correlation coefficients than the traditional PCA method [19] [20] [21]. Jian Yang et al. have proposed a 2-dimensional PCA method [22] for processing images. The proposed method does not require to transform the original image matrix into one-dimensional vector. Instead, the original image matrix is directly converted into the image covariance matrix [23]. In the research work of Daoqiang Zhang et al., they proposed a (2 Dimension)2PCA [24] method for facial expression and recognition. The 2-dimensional PCA method is more concerned with the lateral processing of the image. Differently, the (2Dimension)2PCA method processes data both horizontally and vertically. Experimental results show that both (2Dimension)2PCA and 2-dimensional PCA method have higher recognition rate. However, (2Dimension)2PCA needs more coefficients than the 2-dimensional PCA method.

In order to reduce the number of correlation coefficients [25], in 2011, Luiz Oliveira et al. proposed a feature extractor based on the 2D principal component analysis method [26] for facial expression recognition. Due to the existence of a large number of coefficients, mixture algorithms that is a combination of multi-objective genetic algorithm and feature selection algorithm are used to analyze or select coefficients. K-nearest neighbor (KNN) and support vector machine (SVM) is used for the classification of facial expression classes. The experimental results show that the combination of the 2DPCA and feature selection algorithm can significantly reduce the number of correlation coefficients. Comparing with the simple 2DPCA, the recognition rate of this combination method has been dramatically improved.

In the research work of Ajit P. Gosavi et al., principal component analysis with the singular value decomposition method [27] is applied to facial feature extraction. Unlike traditional PCA method, the singular value decomposition method gives more efficient

help after obtaining the average of the database images [28]. Also, the use of the singular value decomposition method minimized image storage space. Euclidean distance classifier is used to classify facial expression images. In the end, the researchers compared the PCA method and the new proposed method. The average recognition rate of the test results is 67.14% based on the PCA method. However, the average recognition rate of the new method is 78.57%. Although the use of the singular value decomposition method has improved the recognition rate, the average recognition rate of 78.57% is not a satisfactory result in theory.

In 2016, based on feature extraction of Log Gabor filter method, PCA dimensionality reduction and Euclidean distance classifier [29] make the research of facial expression recognition reach a higher level. Five scales and eight different directions of the Log Gabor filter method are applied to feature extraction. In order to reduce image dimensions, the PCA method is applied to the Log Gabor filter method. Experimental results show that the application of the new method achieves a better average expression recognition rate. Moreover, the limitation of the whole experiment is the prediction of disgust, sadness and surprise emotions.

In 2014, Sukanya et al. 's research have explored the performance of PCA algorithm [30] in face recognition and facial expression recognition. The application of the PCA method reduced the storage space of all images. In other words, the form of face images that used for training is not the raw images [31]. Experimental results show that the PCA algorithm has a low error rate in face recognition and emotion recognition. Moreover, the recognition performance is maintained at a high level under diverse illumination.

2.1.3 Local Binary Patterns Analysis

The local binary pattern is an algorithm that used to describe the texture features of images. Directly speaking, the computational flow of the LBP algorithm is to obtain the local texture features by comparing the gray scale of the central pixel as the threshold value to the other pixel values. The method was proposed by T. Ojala in 1994 and used to extract local texture features of images [32]. Experimental results show that the algorithm has significant advantages such as rotation invariance and gray invariance. However, with the wide application of LBP method, it has two factors that affect recognition performance, that is, the change of illumination and a large number of dimensions [33].

Although LBP method has strong ability of texture recognition, it cannot accurately describe the facial muscle texture, wrinkles and other local deformation. Therefore, the single LBP method does not perform well in the research of facial expression recognition. In 2008, a facial expression recognition algorithm based on orthogonal locality preserving projection [34] and feature fusion was proposed by Quanyou Zhao et al. this study combines LBP and Gabor wavelets method [35] to overcome some shortcomings of LBP algorithm in facial expression recognition. The Gabor wavelets method is a linear filter for edge extraction. Due to insensitivity to illumination changes, this method is one of the best ways to describe the local appearance and texture of human faces [36]. In the process of extracting feature vectors, the average and standard deviation of each block are calculated, which is sub-block of each Gabor wavelets in the image. According to the experimental results of the CED-WYU database [37] and the TFEID database [38], the application of the proposed algorithm has significantly improved for the recognition rate. Moreover, the mixture algorithm has a substantial floating effect on the test results of different databases and poor recognition rate for disgust, surprise and fear. Since there is no definite definition of any kind of expression, the recognition system may have a larger recognition result for the same expression of different people.

Theoretically, the change of capturing angle will lead to the change of face contour. Due to the change of face angle, some features of human face cannot be accurately extracted, which further leads to the error recognition of the face. In 2014, a face data normalization method based on mutual mapping between face image depth values and texture information was proposed by Xiaoli Li et al. Moreover, polytypic local binary patterns (p-LBP) [39] are used to construct 3D facial expression features. The method can identify three-dimensional coordinate information of every point in the field. And combining texture information to ensure the integrity of local faces. Experimental results show that the proposed method maintains high efficiency in facial expression recognition. Also, the new method achieves hugely high recognition rate for each expression.

3D face recognition is realized by 3D camera stereo imaging. As the information increase gained by the system, the accuracy of recognition will also increase. In 2016, Shu An et al. proposed local binary patterns-based expression recognition method [40] that used for increasing more local texture information. Different from the traditional LBP method, the binary threshold of the algorithm is calculated by the difference between the gray values of adjacent pixels and central pixels. Then, the HOG [41] is applied to the LTBP. The new proposed algorithm achieves high recognition rate in BU-3DFE database.

In 2015, Shiwen LV et al. proposed a 3D face recognition method based on region extended local binary patterns (eLBP) [42] [43]. After normalizing the depth image, the LBP feature extraction for different facial expression regions is expressed as an extended LBP uniform pattern. Weight-sparse representation classifier is applied to facial expression image classification. The experimental results obtained very significant expression recognition rate.

2.2 Deep Learning Methods

The difference between deep learning and traditional recognition algorithms is that deep learning can automatically learn features from a large number of data and not rely on manually extracted features. On the contrary, traditional recognition algorithms extremely rely on designer's prior knowledge. Therefore, it cannot use big data to complete learning. Based on the above-related literature analysis, we can know that all experiments require the designer to adjust the parameters such as the related research of PCA method. However, deep learning can contain thousands of parameters. It can be automatically adjusted to achieve optimal performance. Thus, deep learning is a methodology to simplify the design of recognition system. The following content, we are concerned with the application of deep learning in facial expression recognition.

Deep learning is a method of machine learning based on data representation learning. Deep learning is a set of algorithms that use machine learning algorithm on multilayer neural network to solve problems of image, text and so on. It can use various ways to represent input value, such as vectors of each pixel's intensity values. The core of deep learning is feature learning which includes supervised and unsupervised feature learning. For example, convolutional neural network is a deep network model based on supervised learning. Deep belief nets are an unsupervised network model [44]. DBN model can be used for handwritten numeral recognition and expression recognition [45] [46]. In addition, deep learning can obtain hierarchical feature information through hierarchical network to solve the problem of manual design features. Deep learning includes many essential algorithms, including convolutional neural network, recurrent neural network and so on. For different research objects (image, language, text), the correct choice of network model can achieve the best results. In this thesis, the convolutional neural network is a fundamental key algorithm. Convolutional neural network is specially designed to process data with net structure. It is a multilayer neural network and is good at handling machine learning of images [47].

2.2.1 Application of deep learning in FER

Compared with geometric structure and artificial design features, facial expression recognition based on deep learning can learn quickly and consume less time. This method can obtain a better recognition rate by learning the feature rules of the face image in detail. Therefore, a large number of studies on recognition have shifted from traditional learning methods to deep learning.

The pooling strategy in conventional CNN acts a disadvantageous behavior [48], so the method is modified. In 2016, Tong Zhang et al. proposed a new feature learning method based on deep neural network [49] and implement it in the research of multi-view facial expression recognition. In the study, they proposed the new methodology that a series of landmark points that corresponds to the scale invariant feature transform (SIFT) was extracted from each face images. The eigenvector matrix which composed by SIFT eigenvectors is used as the input of DNN model. Specifically, the significant low-level features extracted from facial landmark points can be used as DNN input to decrease the impact of misalignment. In the proposed neural network, the protection layer is used to learn the facial features of different facial landmark points. Unlike traditional CNN, projection layer and convolution layer reduce the complexity of network model.

According to this research, we can know that the network model is only suitable for small simple study. In the research of facial expression recognition, most facial expression databases have thousands of face images. For some typical expression databases such as CK, FER datasets, the application of this model may cause over-fitting problems.

In 2015, Heechul Jung et al. utilized deep learning technology to recognize human

facial expressions. Depp neural network and convolutional neural network [50] are used to detect facial expressions. Then, two different network models are compared. In the past few decades, a large number of studies have concerned on training methods with human design such as HOG, LBP and so on [51] [52]. Therefore, a lot of experiments and efforts are often needed to achieve high recognition performance. In this study, the complicated experimental process was solved. The network model can automatically classify and extract the features required. Experimental results show that convolutional neural network has better performance. However, the average of expression recognition needs to be further improved.

Micro-expression is a rapid, faint facial movement. Nowadays, more and more researchers turn their attention to the recognition of micro-expressions. Patel et al. have proposed a method for identifying micro-expressions [53]. Features of facial micro-expressions are extracted from ImageNet and then recognizing micro-expressions by using a designed CNN. Investigators found that feature extraction by designed CNN is extremely better than conventional methods. In 2016, Devangini Patel et al. explored deep learning based on micro-expression recognition. Due to the lack of micro-facial data, the investigators did not use the CNN model for training. They used transfer learning and facial expression based on CNN model. In order to delete the unrelated deep features in the experiment through feature selection. Experimental results show that the proposed method has good performance in micro-expression recognition.

In the study of deep learning, more and more studies are devoted to the application of transfer learning. Because most of data or tasks are related, the trained model parameters can be transferred to the new model to help the new model training. Through transfer learning [54], we can share the parameters [55] of the trained model to the new model accelerate and optimize the learning efficiency [56]. Obviously, this method has some disadvantages. For example, negative transfer may occur if the source task is not

related to the target task. Conversely, if source task and target task are related, transfer learning is a good way to improve the training efficiency of the model.

In 2018, Min Peng et al. [57] trained two micro-expression databases using transfer learning. The ResNet10 training model is pre-trained for the ImageNet dataset. Then the model parameters are provided to train another micro-expression data set. Experiment results show that higher WAR and UAR [58] [59] can be achieved by using transfer learning.

In 2018, a new transudative transfer subspace learning method was proposed by Wenming Zheng et al. [60]. The combination of the marked image in the source domain and the unmarked image in the target domain is used for training. Class labels are used to predict unmarked image sets. After testing two different expression image databases, the performance of the new proposed method in this study is better than some traditional methods.

In 2015, Zhiding Yu et al. proposed a method that used for image-based static facial expressions recognition [61]. They used face recognition modules and multiple convolutional neural networks in the proposed method. In the facial expression recognition challenge 2013 [62], the CNN has completed the pre-training of facial expression images. Then the trained model is fine-tuned on the training set of SFEW 2.0 [63]. The minimum the logarithmic likelihood loss and minimum hinge loss are added in the process of combining several CNN models. The experiment results show that the accuracy of validation set, and test set is 55.96% and 61.29%.

According to these latest deep learning-based transfer learning studies, we can conclude that the result of the experiments can only be slightly improved. Due to the large

error recognition, the proposed technologies are also difficult to implement in the real environment.

In 2016, SE Kahou et al. proposed a facial expression recognition method based on multi-model deep learning [64]. Convolutional neural networks, deep belief nets represented by audio streams and “bag-of-mouths” model based on K-means [65] are used in recognition system. Investigators used these models to solve the problems of facial information capture, visual feature extraction and video space-time. Moreover, the illumination, posture and other factors in the data set have great changes. Thus, the multi-model feature combination methods are used to classify labels. Similar to previous studies, the experiment results show that the recognition accuracy is still not satisfactory. In the dataset provided by EmotiW challenge [66], the accuracy of the test set is achieved to 47.67%.

In the research of facial expression recognition, it is more difficult to identify random video or real environment face images than some facial expression datasets. In order to adapt to the recognition problem under complex conditions, the multi-model method is widely applied. Jingjie Yan et al. proposed a method combining convolution neural network with feature fusion [67] for facial expression recognition. Convolution neural network, Gabor method and openSMILE tool [68] are used to extract face features in video. Principal component analysis (PCA), sparse kernel reduced-rank regression (SKRRR) [69] and kernel cross-modal factor analysis (KCFA) [70] is used in the process of feature fusion. According to the test results on the AFEW 6.0 dataset, the proposed fusion method obtained 53.46% and 50.93% respectively.

In 2016, Xinyu Wang et al. have proposed a dense trajectory algorithm [71] for identifying human behavior. In their research, the effective detection of human behavior in complex background is realized by online weighted multi-instance learning to tracking

objects and calculating dense trajectories in human body boundary boxes. Moreover, 89% accuracy was obtained in the background dataset affected by illumination, body pose and other variables.

The successful application of dense trajectory algorithm in human behavior detection has attracted the attention of researchers in facial expression recognition. Some investigators hope to transplant dense trajectory feature algorithm into facial expression recognition. In 2016, Sadaf Afshar proposed an improved dense trajectory recognition [72] system for facial expression recognition. In the proposed system, LGBP-TOP method, geometric feature processing and improved dense trajectory algorithm are used to extract the features of face images from variety of videos. By testing the proposed method on Cohn-Kanade (CK+) [73] and EmotiW 2015 [74], the dense trajectory algorithm achieves high accuracy. Moreover, compared with other conventional methods, it has shorter training time.

According to the above reports about the challenges of facial expression recognition, we can conclude that the deep learning method exhibits many advantages compared with conventional algorithms in application of facial expression recognition. For example, automatically facial features extraction, automatically network parameters adjustment, and so on. Moreover, the convolutional neural networks, transfer learning and fusion method et are applied in the above studies. Based on the results of the report, we can conclude the following issues that are low recognition accuracy, low training efficiency and not suitable for big dataset training. In our research, these problems have been effectively promoted.

2.2.2 Classification

In the research of facial expression recognition, it is a multi-classification problem. The classifiers commonly used by researchers include Naïve Bayes, logistic regression, support vector machines and so on. Convolutional neural network and support vector machine (SVM) are widely used classifiers for deep learning-based emotion recognition. CNN can predict the class labels of the classification tasks or the real values of the regression tasks. Basically, the classification task is achieved through the combination of the full-connected layer and the SoftMax loss function. Support vector machine is one of the most commonly used classification methods in supervised learning. It mainly realizes the classification of input data by constructing segmentation planes.

In 2016, Deepjoy Das et al. proposed a deep neural network [75] for emotion recognition. The researchers input the extracted facial features into restricted Boltzmann machine and deep belief networks. Moreover, stacked autoencoder with SoftMax function is used to convert class score generated by second level automatic encoder into probability value. The experiment results obtained rational recognition accuracy. The researchers found that the performance of the model increased with the increase in the number of hidden layer nodes.

In [76], a stacked automatic encoder based deep learning classification network is proposed. In this deep learning model, the SoftMax function is used to classify three types of emotions and the output of the corresponding accuracy. The SoftMax function that used in experiment outputs the K-dimension vector, which equals to the probability of K-class. The class that is correctly predicted is the maximum absolute value in the output vector. The experiment results obtained a high recognition rate.

Face expression recognition that has illumination, angle or other invariant factors

requires combination of feature extraction and classifier. In part of the research, convolutional neural network is chosen as a model of learning and classification. However, some studies repute that support vector machines can generate decision segmentation surfaces from well-learned feature vectors [77].

In 2013, Yichuan Tang et al. proposed an L2-support vector machine-based expression recognition method. Most of the deep learning models use cross-entropy loss function to predict and classify. In their research, the learning minimizes a margin-based loss function instead SoftMax function. The SVM target generated the reverse second vector. The use of L2-SVM's target to train deep neural network. The gradient of top-level linear SVM helps to learn the lower weight by backpropagation. The experiment results show that the use of the L2-SVM method can minimize the loss in the prediction process. However, this method is different from the SoftMax function. It can only output classes but not output probabilities.

In Ibrahim's report, they tested the performance of four different SVM kernels in facial expression recognition [79]. Radial basis function, linear function, quadratic function and polynomial function are used as kernel functions of the support vector machine. Experiment results show that SVM can provide better performance in emotion recognition. In addition, the quadratic function-based SVM method has achieved the best performance in emotion recognition.

2.2.3 Deep Learning for Multi-Variable Face Recognition

In addition to insufficient dataset, various facial pose, size, position, background and illumination will also affect the accuracy of recognition. For example. Simple background will make face feature extraction easier. On the contrary, complex background may lead to difficulty in extracting face features and furtherly cause error recognition. Generally

speaking, the intensity of illumination will have a great impact on human visual judgment. Similarly, illumination has great influence on the object structure, texture and contour of the face image in the process of expression recognition. Although these factors will have a severe impact on expression recognition, these factors are ubiquitous in real environment. Therefore, more effective and reasonable new technologies to overcome these problems are necessary. The following descriptions are the relevant research reports about deep learning on different face pose, illumination and so on.

In unconstrained environment, various face pose, illumination and occlusion increase the difficulty of facial recognition and alignment. The research based on deep learning method has achieved a breakthrough in solving this problem. In 2016, Kaipeng Zhang et al. proposed a deep cascade multitask framework [80] to predict facial and landmark positions. The accuracy of recognition is improved by utilizing the correlation between detection and alignment. Moreover, a new online hard, simple mining strategy is applied to the recognition system. The proposed framework is achieved through the combination of a cascaded architecture and convolution neural networks. According to the experiment results, the test results of the face recognition dataset get higher accuracy. For the non-frontal face, it can achieve better localization performance.

Related research can enhance facial recognition accuracy under the influence of external factors. In the report of Dong Chen and other researchers, an advanced face detection method [81] is applied to detect different facial states. They explore the intrinsic relationship between face alignment and detection to observe the aligned facial features further. Therefore, the learning of facial features is more efficient. Also, face detection and alignment are synchronized in the same cascading framework. Experiment results show that the proposed method achieved the best accuracy in face detection dataset. Moreover, this method has improved the performance of cascade detection.

For the research of facial expression recognition, extreme facial pose, illumination and rotation increase the difficulty of recognition. In 2017, Rajeev Ranjan et al. proposed a deep convolutional neural network [82] for face recognition and landmark position estimation. The proposed method is a multi-task learning algorithm for fusion feature processing. Specifically, the core of the algorithm is independent CNN model merging the middle layer of deep convolutional neural network. The ResNet-101-based algorithm variants and Fast-HyperFace method are used to improve the recognition performance and the speed of the algorithm. The test demonstration shows that the proposed method can accurately locate faces with different facial pose. However, some partially occluded faces cannot be detected. The proposed model can capture the global and local information from faces.

Further research on face recognition have focused on the recognition based on different facial poses and illumination. In 2013, Zhengyao Zhu proposed a method of facial identity-preserving [83] to improve the recognition problem based on extreme pose and illumination. The proposed method can extract face features in a state of facial pose and illumination, and then reconstruct the facial structure in the canonical view. Moreover, this method can naturally reduce the variation of internal identity and distinguish the difference between identities. Also, a deep convolution neural network is proposed to combine feature extraction layer and refactoring layer. The extracted features are put into the CNN structure as input. Experiment results show that the proposed method can recognize and locate different facial pose accurately.

In 2008, Hsiuao-Ying Chen et al. proposed an advanced multi-class hybrid-boost learning algorithm [84] for different face poses, sizes, occlusion and different facial expression recognition. The classifier of the mixture-boost algorithm consists of Haar-like method and Gabor-like method. Moreover, the learning algorithm selects the mixed feature of the weak classifier to minimize the loss function. The experiment results

obtained high recognition performance and accurate face localization.

2.2.4 R-CNN

In 2014, Girshick et al. [94] proposed the R-CNN method for the object detection problem. This algorithm can be divided into four parts, which is candidate region selection, CNN feature extraction, object classification and boundary regression. However, the images corresponding to multiple candidate regions need to be extracted in advance, so they will occupy a large disk space. Moreover, the traditional CNN requires a fixed size of the input image, normalization processing produced object truncation or stretching will lead to the loss of input CNN information. In addition, each proposal region needs to enter the CNN network for computing, thousands of regions overlap resulting in huge waste of computing.

After the development of R-CNN and Fast R-CNN, Ross B. Girshick proposed a new Faster R-CNN [95] in 2016. In structure, Faster R-CNN has integrated feature extraction, candidate box generation, bounding box regression and classification into a network. The experimental results show that the use of Faster R-CNN makes the comprehensive performance of object detection greatly improved, especially in terms of detection speed. With the successful application of Faster R-CNN in object detection, some researches use Faster R-CNN to achieve face detection. In the research of Sun et al, they proposed a new face detection method using deep learning [96]. They combine various strategies to improve the Faster R-CNN framework, including feature concatenation, hard negative mining, multi-scale training, model pre-training, and proper calibration of key parameters. Finally, the experimental results obtained a higher accuracy of face detection.

Mask R-CNN is a new convolutional network proposed by Kaiming He et al based on Faster R-CNN architecture. This method achieves high-quality semantic segmentation while accurately detecting objects [97]. Mask R-CNN mainly extends the original Faster

R-CNN and adds a branch to predict the object in parallel. Based on the experimental results, the network structure is easier to implement and train than Faster R-CNN, and the recognition speed is faster.

2.3 Summary

In this section, we introduced the traditional facial expression recognition methods and the application of deep learning in facial expression recognition. Compared with the conventional method of complicated manual feature extraction, deep learning method has shown absolute advantages in image recognition research. Moreover, we introduced in detail the deep learning-based classification problems and the performance under the different face poses, illumination and other factors. We can conclude that most of the deep learning research is based on the original convolutional neural networks. Also, the research on face pose and illumination is not comprehensive enough. Moreover, the relevant R-CNN research is based on real-time object detection to achieve, there is little R-CNN research on facial expression recognition. There is not relevant research to conduct a comprehensive test of the proposed model. According to these research gaps, we propose two face expression recognition methods based on deep learning technology in the thesis. The region of interest (ROI) method, region proposal network and CNN combined multi-task learning method is applied with the first learning model. For the second learning method, ROI Align layer is used to replace ROI Polling layer by the first model and a new loss function is added to the classification layer.

Chapter 3 Faster R-CNN and Mask R-CNN

This chapter provides more details of Faster R-CNN and Mask R-CNN which have been briefly reviewed in Section 2.2.4. The applications of these two algorithms to facial expression classification is the focus of this research.

3.1 Faster R-CNN

The Faster R-CNN model is composed of two modules. The first module is the made up of convolutional neural network that is used in region proposal. The second module is the Fast R-CNN recognizer of region proposal. For the proposed model, feature extraction, proposal extraction, bounding box regression and classification are integrated into a network. Experimental results show that the recognition speed and accuracy of the model reach a high level.

3.1.1 Detection Box Generation

Traditional detection box generation methods, such as the edge box or the selective search method used by R-CNN. These methods are very time-consuming in generating boxes. The Faster R-CNN algorithm directly used Region Proposal Network to generate detection boxes. Compared with the traditional detection box generation method, the detection box generation speed of Faster R-CNN model has been greatly improved.

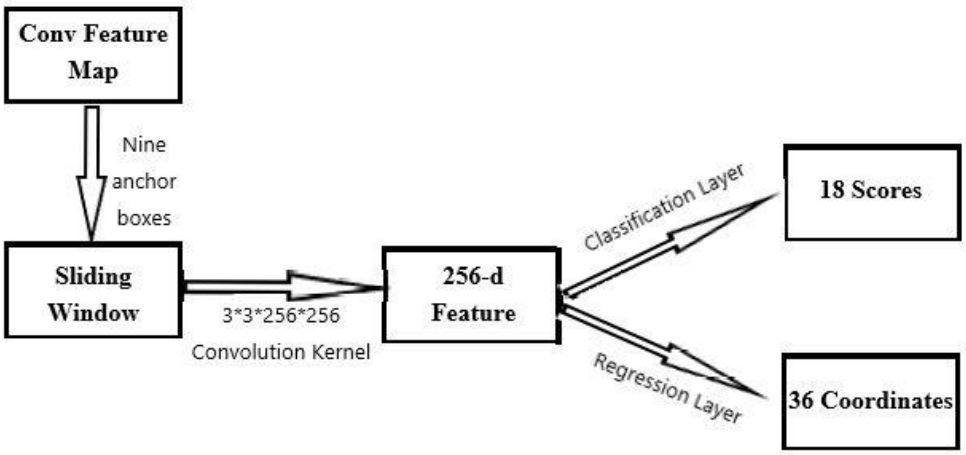

Figure 3.1 Simple flow chart of Region Proposal Network

Figure 3.1 shows the concrete structure of region proposal network(RPN). The RPN is used to extract region proposal from input images. The network structure is based on convolutional neural network, and its output contains the multitask model of classification and regression. According to the Figure 3.1, the dimension of convolution feature map is three dimensional and the size of sliding window is 3×3. We used a 4-dimensional convolution kernel of $3 \times 3 \times 256 \times 256$ to map each 3×3 sliding window to a 256-dimensional feature. In the real implementation process, there is multiple sliding windows, so the final multidimensional 256-d vector is obtained. The low dimensional feature vectors that are mapped are sent to the next two full connection layers, including classification layer and regression layer. In the experiment, we set the number of anchor to nine. The working principle of anchors is described later. By using a $1 \times 1 \times 256 \times 18$ convolution kernel between 256-d and classification layer. We can obtain 10 output nodes in the classification layer. At the same time, by using a convolution kernel of $1 \times 1 \times 256 \times 36$, 36 outputs of the regression layer can be obtained. Then, classification layer and regression layer connect to their own loss function. The value of the loss function is given, and the results of backward propagation is derived based on derivative. Figure 3.2 shown

some examples of facial expression recognition using RPN proposals on CK+ dataset.

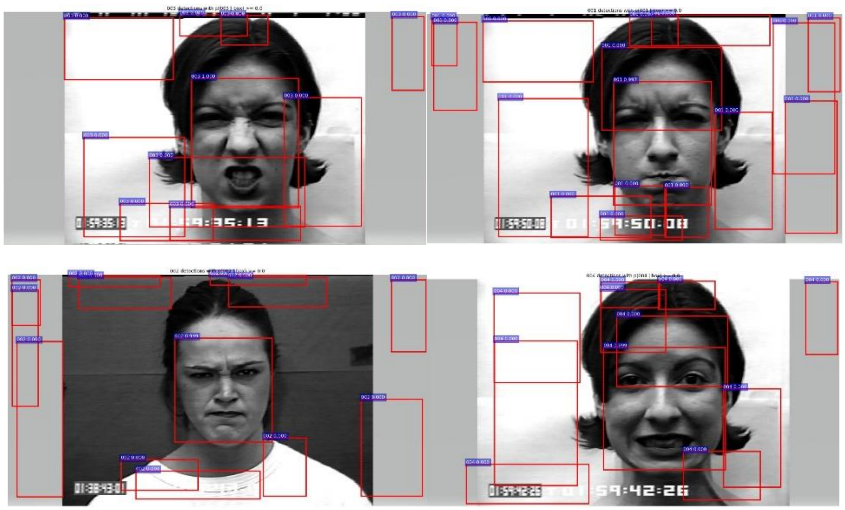

Figure 3.2 Examples of RPN proposals on CK+ dataset

Anchor is the core of region proposal network. Because the object size and length-width ratio are different, we need to set up multiple scale windows. Anchor is the size of a reference window. In the experiment, we set three ratios of length and width of 0.5, 1 and 2 respectively. The windows of different sizes can be further obtained by changing the ratio of length to width. Therefore, when we slide the windows of M×N convolution feature map, we can get M×N×9 anchors in total.

3.1.2 ROI Pooling

For the traditional convolutional neural network, the input of the recognition system is a fixed size image. In the process of facial expression recognition test, the datasets used in preliminary test and proper test are all facial expression images of different sizes. If the recognition model is unable to recognize input images of different sizes, it will increase the complexity of the experiment. The Regions of Interest(ROI) pooling layer is used to solve this problem in the Faster R-CNN model. Is does not require fixed size input feature maps, but it can output fixed size feature maps. The following content is a description of

the execution process of the ROI Pooling layer.

First, we mapped the ROI coordinates to the feature map by the coordinates of ROI dividing the ratio of the input image and the feature size. After obtaining the box coordinates on the feature map, proposal with different sizes is mapped to rectangular boxes with fixed sizes by pooling. From the fixed rectangle box, we get the features of ROI and make further classification and regression.

3.1.3 Training

In model training, we divided the training process into four stages. First, we used the pre-training model to train Region Proposal Network separately. Then, the output region proposal of the first stage RPN is used as input to train the Fast R-CNN network individually. Generally, the original image is intercepted by the region proposal of RPN output and the intercepted image is repeatedly convolution-pooling and passing ROI Pooling and full connection network. The final result outputs two branches that are object classification based on SoftMax function and bounding box regression. Up to now, there are not shared parameters between the two networks. Figure 3.3 shows the test flow chart after the end of the network training.

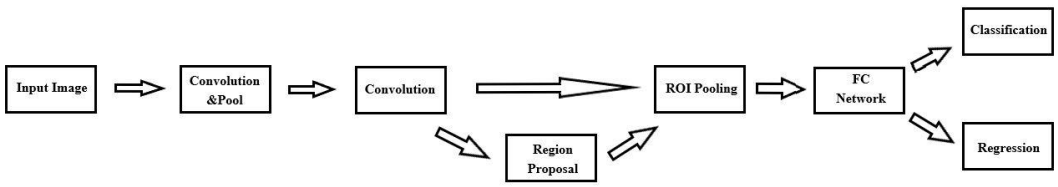

Figure 3.3 Flow chart of test phase

During the third training stage, Region Proposal Network was retrained. In this process. The parameters of the common part of the fixed network are retained and the parameters of the RPN unique part are updated. For the fourth training stage, the Fast R-CNN network is fine-tuning again through the results of RPN. The parameters of the common part of the network are remained and the parameters of the unique part of the Fast R-CNN network are updated.

In the process of training RPN, we assign a binary label to each anchor. Use 0 and 1 to indicate whether there is an object. We have adopted the following rules to achieve the object determination in anchor. First, if the largest IoU of an anchor and any object region, then the anchor is determined to be an object. Secondly, of the IoU of an anchor and any object region is greater than 0.7, then it is determined to have an object. Third, if the IoU of an anchor and any object region is less than 0.3, then the background is determined. The IoU(Intersection-over-Union) mentioned above is that the coverage of predicted box and ground-truth box equals to the intersection of two box divided by the union of two box. With these definitions, the loss function is defined as Eq. (3.1).

$$\mathrm{L}(\{p_i\},\{t_i\}) = \frac{1}{N_{cls}}\sum_i L_{cls}\,(p_i, p_i^*) + \lambda\frac{1}{N_{reg}}\sum_i p_i^*\,L_{reg}(t_i, t_i^*) \tag{3.1}$$

Here, i is the index of anchor and p_i is the prediction probability of an object. If p_i is positive, the value of p_i^* is 1. If p_i is negative, the p_i^* value is 0. t_i represents the four coordinate vectors of the predicted bounding box. t_i^* indicates that the vector of the ground-truth box is associated with positive anchor. $\{p_i\}$ and $\{t_i\}$ represents the output of the classification layer and the regression layer. The output is achieved through normalization of Ncls and Nreg and balance parameter λ weighting. For the classification error of object and the regression error of bounding box, we use the smooth

L_1 loss function. Compared with L_2, sooth L_1 function is easier to adjust the learning rate. Bounding box regression only calculate the error that anchor containing the object through multiplication of Leg and p_i. For the ground-truth of bounding box, the system value is calculated with the object anchor and the coordinates marked as ground-truth. The way of coordinate calculation is represented as Eq. (3.2).

$$t_x = \frac{(x - x_a)}{w_a}, \quad t_y = \frac{y - y_a}{h_a}, \quad t_w = \log\left(\frac{w}{w_a}\right), \quad t_h = \log\left(\frac{h}{h_a}\right),$$
$$t_x^* = \frac{x^* - x_a}{w_a}, \quad t_y^* = \frac{y^* - y_a}{h_a}, \quad t_w^* = \log\left(\frac{w^*}{w_a}\right), \quad t_h^* = \log\left(\frac{h^*}{h_a}\right), \qquad (3.2)$$

Here, x, y, w and h represent the center coordinates, width and height of the box. x, x_a and x^* represent predicted box, anchor box and ground-truth box respectively.

3.2 Mask R-CNN

The Mask R-CNN model is built on the basis of Faster R-CNN model. The Faster R-CNN model provides two outputs for each candidate object that are class label and bounding box offset respectively. The algorithm is added third mask prediction branch based on two branches. Moreover, ROI Align is replaced by ROI pooling. Our method can detect objects in images effectively.

3.2.1 ROI Segmentation

Unlike Faster R-CNN model, Mask R-CNN model needs to correctly detect tall the objects in the image and precisely segment each instance. We use this algorithm to classify each expression and use the box to locate each object. The target of semantic segmentation is to classify each pixel as a fixed category without distinguishing the object.

In the Mask R-CNN model, we added a branch to each ROI region to predict the segmentation masks. This branch is parallel to the existing classification and bounding box regression. Figure 3.4 shows the framework of Mask R-CNN. The mask branch is a fully convolutional neural network acting on each ROI, which is implemented as a pixel to pixel segmentation mask. The Mask R-CNN model is based on the Faster R-CNN model. Through our experiments, we find it easier to implement and train.

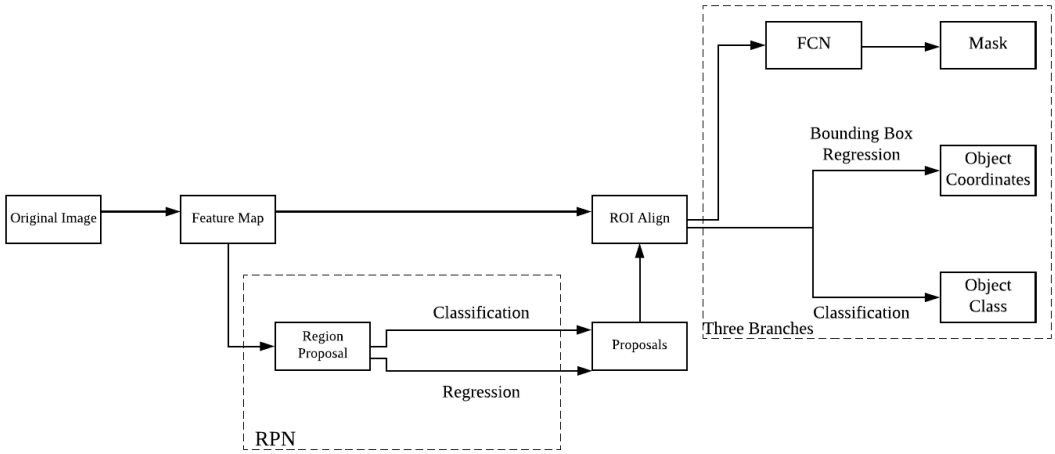

Figure 3.4 Framework of Mask R-CNN

FCN (Fully Convolutional Network) is a classic semantic segmentation algorithm, which can accurately segment object in images. The algorithm is an end-to-end network, and the main modules include convolution and deconvolution. In this experiment, convolution and pooling of images are first aimed at reducing the size of feature map. Then the deconvolution operation makes the size of feature map increase. Finally, each pixel value is classified. Since the experimental dataset is not masked, the experiment results only focus on the bounding box of the object, not the object mask.

3.2.2 ROI Align

For extracting corresponding feature maps from ROI, the Faster R-CNN model is implemented by using ROI Pooling layer. ROI divided the reduced feature map into N blocks, and then selects the maximum value in each block and puts it into the corresponding region of N*N. the experiment of Faster R-CNN shows that this method has no obvious effect on ROI classification but has a significant influence on pixel-by-pixel object prediction. The reason is that each ROI feature does not align with ROI. Therefore, we replaced ROI pooling layer with ROI Align in the Mask R-CNN experiment.

In order to obtain a fixed size of feature map, we do not use quantization operation in the process of extracting feature map from ROI. The aim is to reduce the error caused by quantization. However, we use bilinear interpolation algorithm to solve the floating-point problem caused by pooling and convolution. It is an image scaling algorithm that uses for real pixel values around the virtual point in the original image (all pixel values are integers without floating-point numbers) to determine a pixel in the object image. We found that the number and location of these sampling points not have a significant effect on performance through the experiment. Using this algorithm, the whole process can avoid quantization operation, and the pixels in the original image are completely aligned with the pixels in feature map. This is also the reason why the recognition rate of Mask R-CNN is higher than of Faster R-CNN.

3.2.3 Loss Function

For the Mask R-CNN model, the execution process calculates the bounding box of the candidate region through Region Proposal Network, and then performs the classification of the candidate boxes and bounding box regression. Finally, the system output binary

mask to each ROI. Therefore, the loss function of Mask R-CNN is defined as Eq. (3.3).

$$L = L_{cls} + L_{box} + L_{mask} \tag{3.3}$$

For mask branch and other classification branches, convolution networks are used for computation and processing. In the experiment, the mask used the sigmoid function. The binary mask is finally output by comparing with threshold value. Also, we adjusted the threshold to 0.7. This way avoids the competition between classes and the classification task is assigned to the classification branch. Moreover, L_{mask} used the average binary cross-entropy loss function to train the mask branch.

3.3 Ambiguous Problem Handling

In the processing of the experimental test, there is an ambiguous problem that is arduously obtain the only answer. Next, we analyze the real test results of the two groups. First, the results of the one group that are 0.80, 0.20 and 0.11 respectively correspond to anger, happy and sadness. In addition, the second test results that is 0.72, 0.68 and 0.19 respectively corresponding to angry, happy and sadness. Based on this situation, we found that the recognition rate of anger is much higher than the recognition rate of other classes according to the test results of first group. Obviously, we can completely use anger class as our prediction result. However, the recognition rates of angry and happy are excessively close for the second group test results. Thus, we are unable to further analyze the real predictions by only classes scores. In order to solve this ambiguity problem, we convey a function that can be used to convert the class score into a probability value. Thus, the final results are further obtained by simple comparison of the probability values. The following content is the implementation process and related explanation about the SoftMax function.

The output of the algorithm is the classes score, which is unbaled to distinguish intuitively between classes. Thus, we introduce the SoftMax function that used to convert class scores into probability value. In this way, we can reasonably carry out multi-class classification. It maps the output of multiple neurons to 0 to 1 intervals. Therefore, the final output of the function can be viewed as a probability for multiple classifications. The following is the expression of the SoftMax function.

$$f_j(z) = \frac{e^{f_{y_i}}}{\sum_k e^{z_k}} \tag{3.4}$$

For this function expression, tis input value is a vector whose element is the score of any real number. After completing the calculation, a vector is output with each element value between 0 and 1. Also, the all of element sums being 1. The SoftMax function contains two core parts, which are exponential and normalization respectively. Through exponential operations, a small class scores become slightly larger, large scores change greatly and negative scores are converted to small decimals. Next, the scores amplified by exponential operation is converted to probability by normalization.

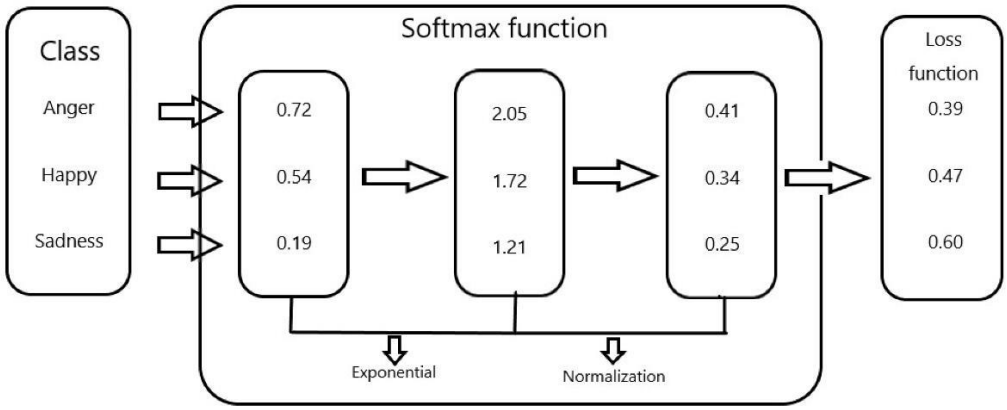

Figure 3.5 Example of SoftMax function operation flow

Figure 3.5 is about the calculation process of SoftMax function. The scores of classes as input is exponentially calculated and normalized, and finally 3 probability values are obtained. When we finally select the output node, we choose the node with greatest probability as our prediction target. Therefore, anger expression is identified as the ultimate prediction goal.

Chapter 4 Facial Expression Recognition Experiments

4.1 Research Design

In this section, the specific design process and implementation process of the research. Based on the description of objectives and research questions, our study focuses on solving the research problems. Here, we have made a detail project plan as the following steps:

1. Choosing research topics, we will contribute to the study of facial expression recognition in this project.
2. Defining the objective and problems of the research through relevant literature study. Throughout the survey, we primarily explored the performance of Faster R-CNN and Mask R-CNN algorithms for different facial states and improved the accuracy of recognition.
3. Selection and marking of datasets. In order to provide the training data of two models, we choose the CK+ dataset as the training dataset of the two algorithms. Moreover, we made my own facial dataset for testing.
4. Specific experimental contents are designed to solve the research problems. In this study, we proposed two facial expression recognition models based n deep learning algorithm. The first method is Faster R-CNN algorithm. The second method is the Mask R-CNN algorithm. In this section, we should complete the implementation and training of two models.
5. Design the testing content of trained model. After completing the model training, we divided the test process into two parts, preliminary test and proper test respectively. In addition, the CK+ and my own datasets are adjusted based on different facial sizes, position, poses, illumination and background.

6. Performing comparative analysis of the experimental results of the two algorithms. Then, conducting the concrete conclusion.

4.1.1 Evaluation Methods

In this thesis, we used variety of evaluation items to assess the performance of classifier. Based on the results of classifier with different facial expressions, we produced a large number of confusion matrices. According to the confusion matrix, we calculate precision, recall, F-score and accuracy to judge the classification performance of different classes. Moreover, we also collected the test time of classifiers in different facial states.

In the field of object detection, the confusion matrix is a display tool for evaluating the quality of the classification model. In detail, each column of the matrix represents the prediction result of the model, and the real result of the sample represented by each row of the matrix. Table 3.1 illustrates a simple confusion matrix. There are several standard terms defined in the confusion matrix involving precision, Recall, F-score and accuracy.

The meaning of letters in the Table 3.1 is:

- True positives (TP) is the number of true predictions based on positive instance.
- True negatives (TN) is the number of true predictions, but the instance is negative.
- False positives (FP) is the number of false predictions, but the instance is positive.
- False negatives (FN) is the number of false predictions based on negative instance.

Table 3.1 The confusion matrix

Predict Class / Actual Class	Negative	Positive
Negative	TN	FP
Positive	FN	TP

The Precision expresses the proportion of positive cases in the example which is predicted to be positive, the equation is shown as Eq. (3.5).

$$\text{Precision} = \frac{FN}{FP + TP} \tag{3.5}$$

The Recall expresses the ability to detect the all real object in a dataset. The way to calculate Recall in following equation:

$$\text{Recall} = \frac{TP}{FN + TP} \tag{3.6}$$

F-score is a good metric when data is imbalanced. The F-score is the harmonic average of the recall and precision. It is calculated by the Eq. (3.7).

$$\text{F} - \text{score} = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{3.7}$$

Accuracy is calculated as the total number of true predictions divided by the total number of instance. The equation is presented as Eq. (3.8).

$$\text{Accuracy} = \frac{(TN + TP)}{(TN + FP + FN + TP)} \tag{3.8}$$

4.1.2 Data Collection and Experimental Environment

In this thesis, we focus on the concrete expression of faster R-CNN and mask R-CNN algorithm in facial expression recognition research. We explored the performance of Faster R-CNN and Mask R-CNN in facial expression classification. The performance of the proposed methods is explored by testing different facial pose, position, size, illumination and background. Therefore, the collection of training datasets and testing datasets is one of the most important tasks. In the whole research process, datasets are divided into seven classes, which are anger, contempt, disgust, fear, happy, sadness and surprise respectively.

Table 4.1 Datasets class labels and corresponding expressions

Class	Expression	Class	Expression
001	Anger	005	Happy
002	Contempt	006	Sadness
003	Disgust	007	Surprise
004	Fear		

The Extended Cohn-Kanade Dataset (CK+) is a popular dataset in facial expression recognition. Most relevant studies use this dataset to train and testing. Therefore, the CK+

dataset is used in training and testing of the proposed models. Table 4.1 shows seven output class names of the CK+ dataset and the corresponding expressions. The Cohn-Kanade dataset expands this dataset. Moreover, it is much larger than some general expression datasets, such as JAFFE dataset.

The dataset used two synchronous Panasonic AG-7500 cameras to record facial expressions of 210 adults, which including 123 subjects and 593 image sequences. The last frame of each image sequence had the label of the action units, and 327 sequences had the label of the expression in the 593 image sequences. The age range of the person recorded is between 18 to 50 years old. Moreover, the image sequence contains 8 bit of gray-scale or 24-bit color values of 640 × 490 or 640 × 480 pixel arrays. Figure 4.1 shows some examples of the dataset which is adopted in this research. A total of 7 examples presented all the types of expressions. Each image represents an expression of a person.

Figure 4.1 Examples of the CK+ dataset. Examples of expression and the corresponding class name: (a) Anger - 001, (b)Contempt – 002, (c)Disgust – 003, (d)Fear – 004, (e)Happy – 005, (f)Sadness – 006, (g)Surprise – 007.

For testing the performance of the models, the test experiments are divided into two parts which are preliminary test and proper test respectively. The execution of preliminary

testing is used to preliminarily discover the performance of the model so as to make relevant assumptions. What needs to be mentioned is that my facial images generate the data used in this part experiment. My facial video is based on different facial size, location, illumination and background. Utilizing the MATLAB2016, the image frames are extracted from video and used in preliminary test experiments. In the proper test, we used the CK+ dataset to verify further the hypotheses put forward in the preliminary test. Moreover, the dataset of proper test also made a similar adjustment with preliminary test.

The all of experiments is processed on a desktop that installed Ubuntu 16.04 Operating System using the AMD Ryzen 5 1500X CPU 3.50 GHz. The training and testing of the models are accomplished by Nvidia GeForce GTX 1050 Ti. The all facial expression recognition methods are developed and implemented in Python and complete training and testing in Caffe.

4.2 Results with Faster R-CNN

In this thesis, we proposed two advanced deep learning algorithms for facial expression recognition. In this part, the tests of faster R-CNN and mask R-CNN are divided into two tasks. The execution of preliminary test that based on my facial dataset is used to put forward some assumptions. The assumptions are further verified through the execution of proper test. The following content shows all the test results of Faster R-CNN and Mask R-CNN. In this section, we conducted preliminary tests to explore factors that affect the performance of the faster R-CNN. First, we explore the effect of face size, position, poses, illumination and background on this algorithm. Also, the trend of algorithm performance and the best recognition rate will be found through tests of different face sizes, position, poses, illumination and background.

4.2.1 Preliminary Tests

Test results with different face sizes

In this section, the facial sizes are divided into seven levels from 10%-20% to 70%-80%. Videos of seven different facial sizes were recorded with the same background, illumination, poses and position. Then, 25 images are extracted from each video to test the faster R-CNN algorithm. Table 4.2 shows the confusion matrix of the dataset with face size from 60% to 70%.

Table 4.2 The confusion matrix for dataset with 60-70% face size

Predict Class / Actual Class	Anger	Contempt	Disgust	Fear	Happy	Sadness	Surprise
Anger	24	1	0	0	0	0	0
Contempt	4	21	0	0	0	0	0
Disgust	2	0	23	0	0	0	0
Fear	0	2	0	23	0	0	0
Happy	0	1	0	0	24	0	0
Sadness	3	0	0	0	0	22	0
Surpise	4	1	0	1	0	1	18

According to the results of Table 4.2, the recognition system maintains a high classification ability for most of the expressions. The classifier recognizes 25 Contempt expression images into 4 Anger expressions. The 25 Sadness expression images are judged to be 3 Anger expression. Surprise expression has poor classification results. There are 4 Anger expressions, one Contempt expression, one Fear expression and one Sadness expression are classified incorrectly.

Figure 4.2 shows the trends in the recognition rate of different emotion classes with different facial sizes. The line chart is the average recognition rate by which calculated the results of successful recognition. According to this chart, we can conclude that the recognition model at 60% - 70% achieves the beat recognition rate for most classes. Except for Sadness and Surprise, the results reached the highest value at 40% - 50% and

50% - 60% respectively. Also, the model has the best classification results for Anger, Happy and Fear. At 50% - 60% and 60% - 70%, their recognition rate is close to 100%.

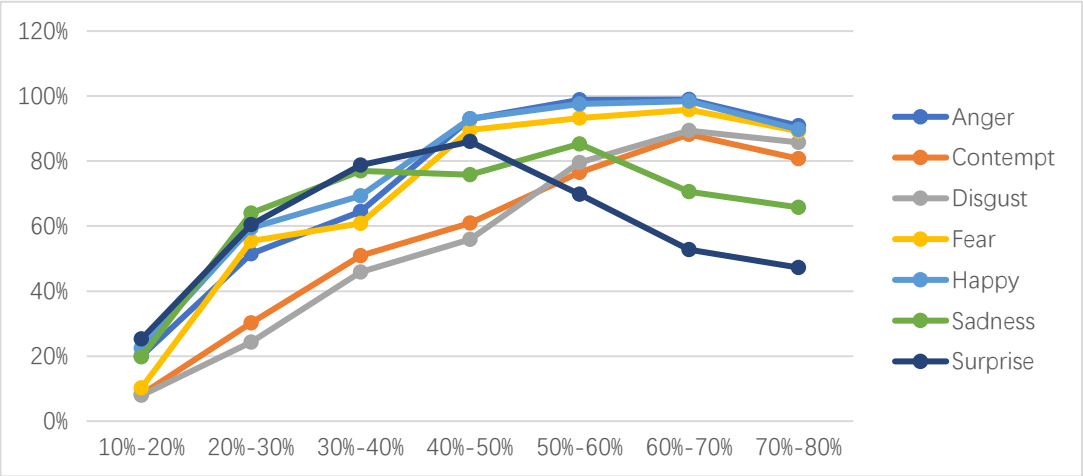

Figure 4.2 Average recognition rate with different facial sizes

Based on the above results, we can make the following assumptions. The number of successful recognition and recognition rate will increase with the increase of facial sizes. The recognition rate will decrease slightly from 60% to 80%. The faster R-CNN only maintains the best classification performance for Anger, Fear and Happy. These assumptions will be further verified in the proper test.

Figure 4.3 shows some examples of facial expression recognition for different facial sizes. The recognition system can accurately capture the location of the face. But as the face shrinks, the recognition rate will decrease.

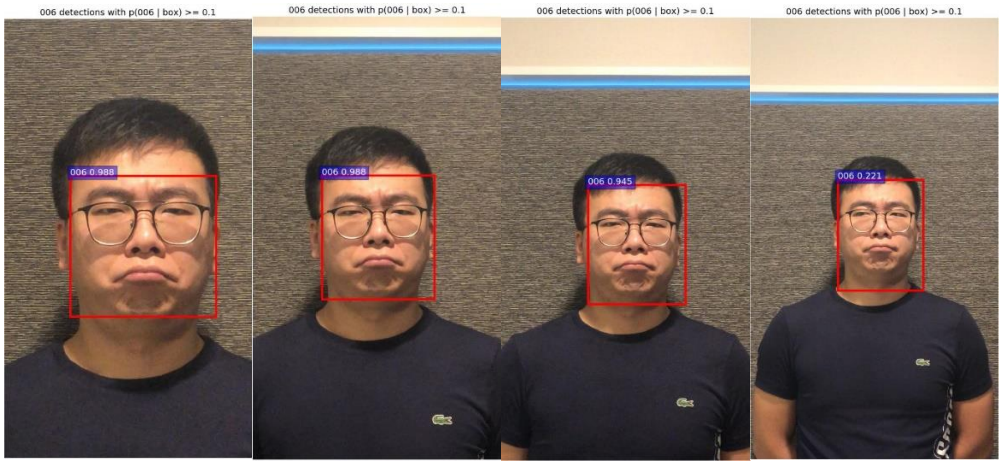

Figure 4.3 Examples of images of different face sizes

Test results with different face positions

In this part, we divided the facial position into nine categories, which are the way of distribution from top to bottom and from left to right. It is like a matrix of 3 × 3 distribution. The categories of the facial position are labeled by way of (1,1), (1,2) et… 25 images for each facial position are used to test the performance of the faster R-CNN. Table 4.3 shows the confusion matrix with face position dataset at (3, 1).

Table 4.3 Confusion matrix for dataset with face position at (3, 1)

Predict Class / Actual Class	Anger	Contempt	Disgust	Fear	Happy	Sadness	Surprise
Anger	22	0	0	0	0	0	3
Contempt	6	17	2	0	0	0	0
Disgust	5	2	18	0	0	0	0
Fear	4	0	0	19	0	0	2
Happy	0	3	0	0	21	0	1
Sadness	6	0	0	0	0	19	0
Surpise	7	4	0	0	0	0	14

Overall, the recognition system has a large number of misclassifications for all expressions. Six expressions of Anger and two Disgust expressions in 25 Contempt expressions were misclassified. In the Surprise expression test, seven Anger and 4

Contempt were mistaken. Only Anger expression can be easily classified from other expressions.

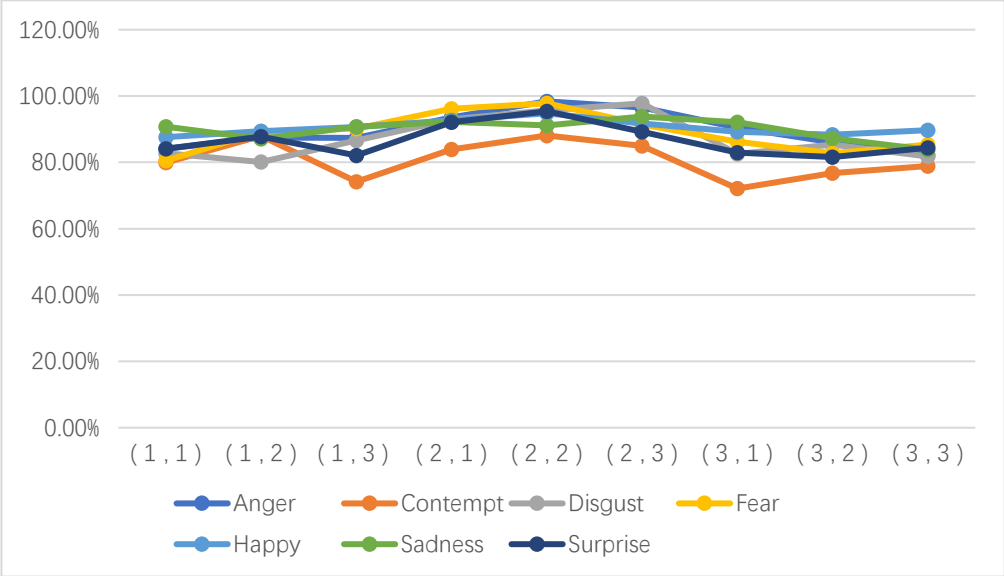

Figure 4.4 Average recognition rate with different facial positions

Figure 4.4 shows the trends in the recognition rate of different expression classes based on different facial position. The mean values of the successful recognition results are used to make the line chart. For this chart, the recognition rate of the seven expressions is basically maintained at the reasonable level. The nodes of each expression line can be near 90%. Except for the line of Contempt, its whole trend nodes are lower than the other lines corresponding to the nodes. Apparently, the recognition system performs poorly on the classification of Contempt.

Based on the above results, we can make the following assumptions. Different facial positions will not affect the faster R-CNN model. Sadness and Surprise are difficult to be detected successfully and Contempt has a low recognition rate.

Figure 4.5 shows the examples of emotion recognition results with the different facial

position. The recognition system can accurately locate the faces of different positions. In addition, the recognition rate can also be maintained in a stable state.

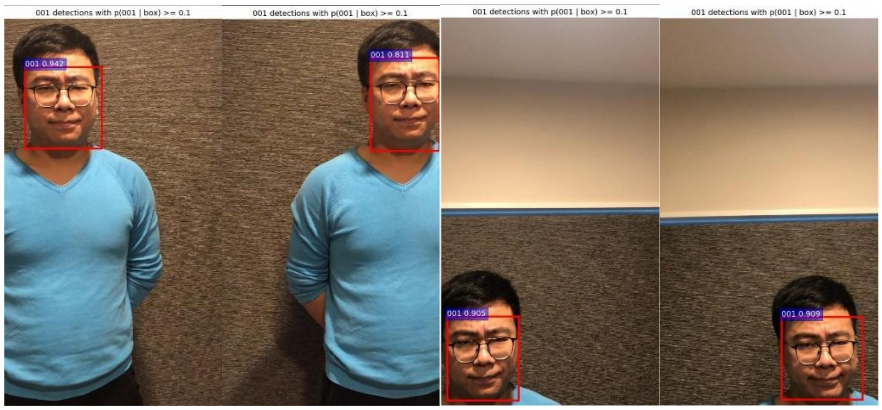

Figure 4.5 Examples of images with different facial positions

Test results with different face poses

In this section, the facial poses are divided into six categories, which the gap both of them is 30°. In the process of recording facial pose images, the face images of different angles are obtained by way of deep rotation. Then, all face images are integrated into six types of the facial pose. Table 4.4 shows the confusion matrix with face poses dataset from 30° to 60°.

Table 4.4 Confusion matrix for dataset with 30°-60° face pose

Predict Class / Actual Class	Anger	Contempt	Disgust	Fear	Happy	Sadness	Surprise
Anger	20	1	2	0	0	2	0
Contempt	4	18	2	0	0	0	1
Disgust	2	1	19	0	0	3	0
Fear	0	0	0	21	1	0	3
Happy	2	0	0	4	16	0	3
Sadness	1	4	3	0	0	17	0
Surpise	0	0	2	4	0	0	19

According to the Table 4.4, we can see that the recognition system can recognize

most of the expressions. In the test of Contempt expression, two Anger expressions, three Fear expressions, three Sadness expressions and two Surprise expressions were mistaken. Changes in facial angles lead to deformation of facial features. The recognition system cannot accurately determine the expression furtherly.

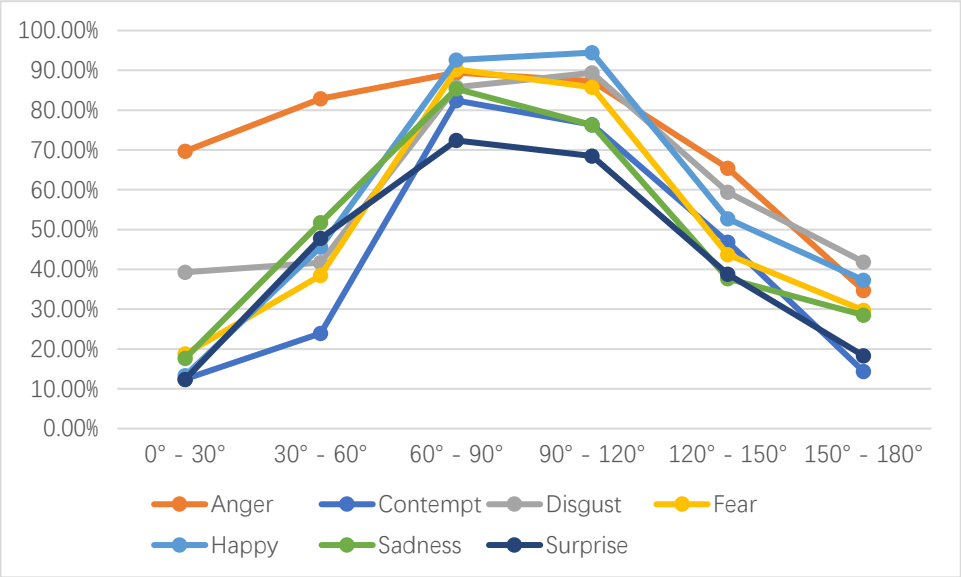

Figure 4.6 Average recognition rate for different facial poses

Based on Figure 4.6, we can see that the change line of all facial expressions is presenting a positive cone. In the angle range of 60° to 120°, the recognition rate of most expressions can be close to or exceed 90%. For this part of the test, the image of the original dataset cannot be deeply rotated, failing to verify the assumptions. So, we do not make any assumptions in this section. In Figure 4.7, we present some examples of emotion recognition results based on different facial poses. Based on the output of the recognition system, we can find that the facial localization ability remains stable as the facial angle changes.

Figure 4.7 Examples of images with different facial poses

Test results with different facial illumination

In the process of testing different facial illumination, facial illumination variables are divided into eight levels. The 3D painting software of Windows 10 system is used to adjust the light intensity of face images based on normal illumination. From the experimental results, the above operations are very rational. Figure 4.5 shows the confusion matrix with face illumination dataset under -2 level.

Table 4.5 The confusion matrix for dataset with -2 level facial illumination

Predict Class / Actual Class	Anger	Contempt	Disgust	Fear	Happy	Sadness	Surprise
Anger	17	3	2	0	1	1	1
Contempt	4	8	0	5	1	6	1
Disgust	5	3	4	2	2	7	2
Fear	2	9	0	1	2	10	1
Happy	1	1	2	6	5	3	7
Sadness	4	2	0	0	0	19	0
Surpise	4	1	2	7	7	1	3

The diagonal number of the confusion matrix illustrates the correct number of classifications in all classes. According to the results shown in Table 4.5, we can explore that the recognition system cannot make correct classification for most weak illumination facial expression images. Only Anger and Sadness expression can be easily distinguished

from other expressions. Similar to changes in facial angles, changes in facial illumination can cause distortion and extrusion of facial features.

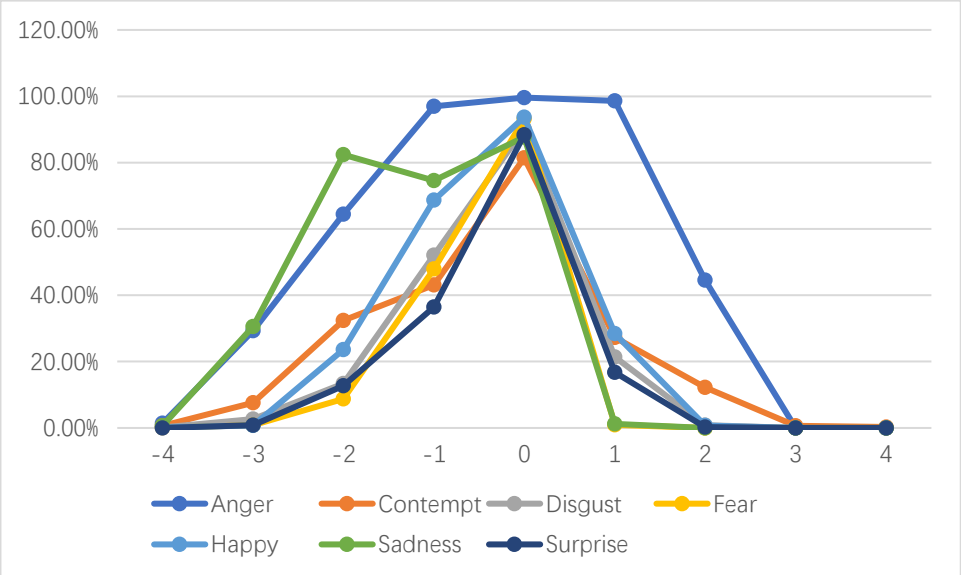

Figure 4.8 Average recognition rate with different facial illumination

According to the recognition trends shown in Figure4.8, we can find that the illumination intensity has a great influence on the recognition system. From the chart, we can see that only under standard illumination, facial expression recognition can reach 100%. On the contrary, any change in illumination intensity will make the recognition performance drop dramatically.

Based on the above results, we can make the following assumptions. Any change in illumination will have a massive impact on the faster R-CNN model. In the process of recognizing performance changes, the influence of strong illumination on the system is much higher than that of weak illumination.

Figure 4.9 shows some examples of expression recognition results based on the different facial illumination. With the change of facial image illumination, the expression

recognition rate has been greatly affected. However, the localization of the facial region is still very accurate by the recognition system.

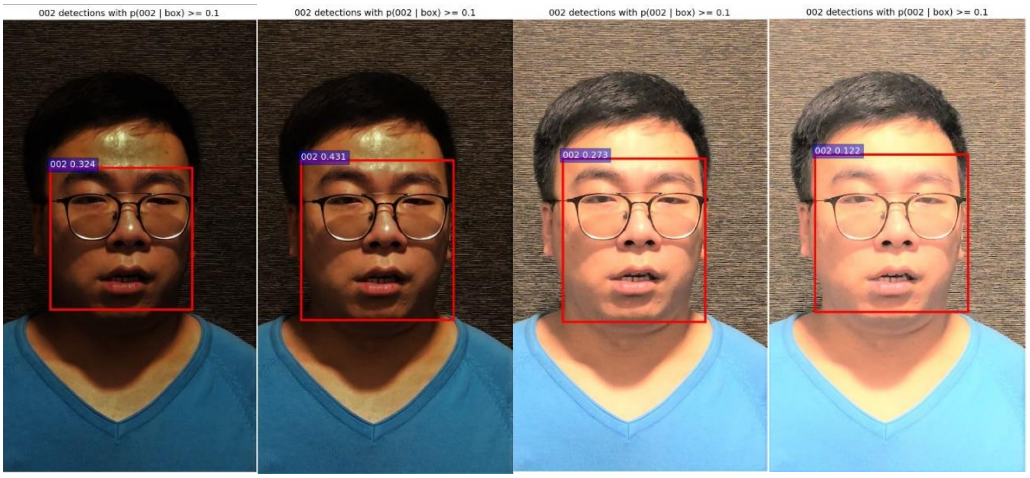

Figure 4.9 Examples of images for different face illumination

Test results with different image backgrounds

In different facial background of test, facial backgrounds are classified into seven categories. From the background 1 to background 6, the background color changes from simple to complex. All the background images are from some colorful landscape photos on the Internet. We intercepted the face area of the original images and spliced it into the background image. Table 4.6 shows the confusion matrix of face background dataset with background 3.

Table 4.6 The confusion matrix for dataset with the third facial background

Predict Class / Actual Class	Anger	Contempt	Disgust	Fear	Happy	Sadness	Surprise
Anger	22	0	2	0	0	1	0
Contempt	1	22	0	0	0	2	0
Disgust	1	1	23	0	0	0	0
Fear	1	0	0	22	0	0	2
Happy	0	3	0	0	21	0	1
Sadness	1	1	1	0	0	22	0
Surpise	0	0	0	3	0	0	22

From Table 4.6, we can see that the recognition system can classify all the face expressions correctly. In addition, Anger, Contempt and Disgust were all mistaken once in the Sadness expression test. In the Surprise expression test, Fear was mistaken three times. Thus, we can conclude that some expressions are easy to be confused with single expression, while some expressions are easy to be misidentified with multiple expressions. It shows that there are similar features between facial expressions.

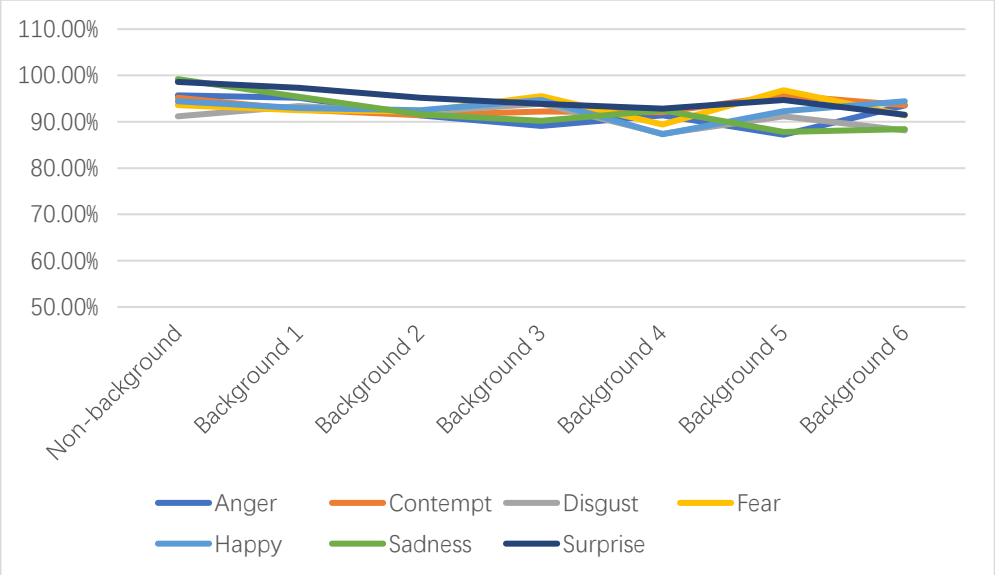

Figure 4.10 Trends of average recognition rate for different emotion classes with different facial backgrounds

Based on the information shown in Figure 4.10, we can conclude that all the expression recognition results can be maintained roughly between 90% and 100%. In other words, the change of facial background does not have much impact on the recognition system. Based on the above recognition results, we can make the following assumptions. Changes in the facial background do not affect the accuracy of recognition. Moreover, the number of background colors does not affect the performance of the system.

Figure 4.11 shows some examples of expression recognition results based on the different facial background. The location of the face area will not be affected by the change of facial background. In addition, the following recognition results maintain high recognition rate and classification results.

Figure 4.11 Examples of images with different backgrounds

4.2.2 Proper Tests

In the proper tests, we will focus on using the original dataset to test the performance of the faster R-CNN. Also, we will further verify the assumptions that put forward in the preliminary test.

Test results with different face size

If only the pixels of the dataset are sized, the face target can be fully localization and the recognition rate can reach 100%. In this way, we can change face size by adjusting the image pixels. However, if the fill-in size is adjusted, the detection rate will change. For this solution, we have adjusted the face size based on a constant image pixel. Specifically, the constant pixel value is 1656 × 912. The following experiment result is that the change of recognition rate based on image of fill-in adjustment. Table 4.7 shows the confusion matrix with face size dataset at 60% - 70% in proper test.

Table 4.7 The confusion matrix for dataset with 60% - 70% face size

Actual Class \ Predict Class	Anger	Contempt	Disgust	Fear	Happy	Sadness	Surprise
Anger	21	1	3	0	0	0	0
Contempt	0	23	1	0	0	1	0
Disgust	1	0	24	0	0	0	0
Fear	0	0	0	22	0	0	3
Happy	0	1	0	0	24	0	0
Sadness	0	0	2	0	0	23	0
Surpise	0	0	0	1	0	0	24

From the results shown in Table 4.7, we can see that the classifier can accurately classify most facial expressions. Most of the expressions can be accurately identified by 100%. In the Anger expression test, one Contempt and three Disgust were mistaken.

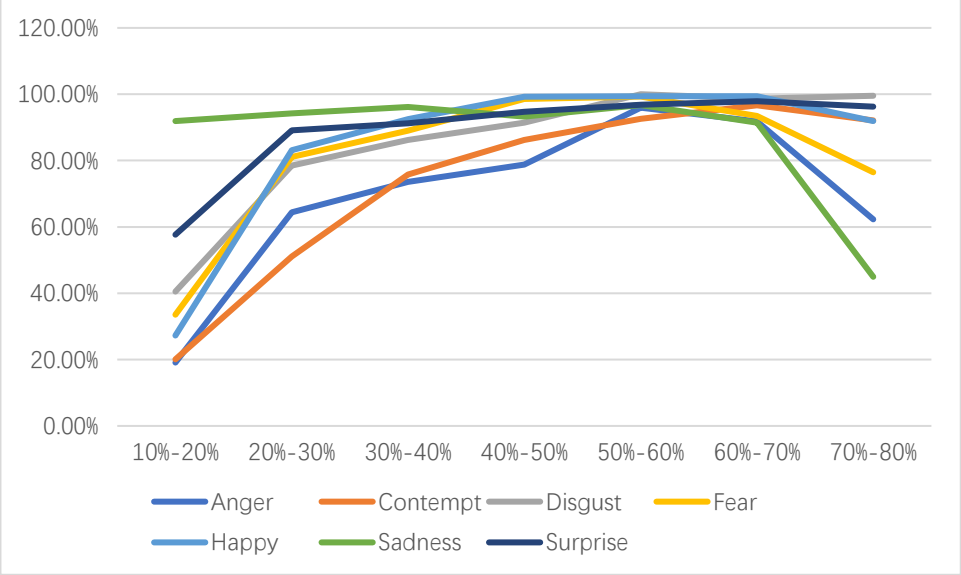

Figure 4.12 Average recognition rate with different face sizes

Figure 4.12 shows the change in average recognition rate for different emotion classes with different face size. Based on the results, we can conclude that the recognition rate of all expressions reaches the best level from 50% to 70%. Only the maximum or minimum face size caused the recognition rate decreasing.

So, we verified the previous assumptions. The recognition performance of the system will increase with the increase of the face size, but recognition system has problems of recognition or unrecognizable for some large face size images. Moreover, the recognition model has a better classification effect for all expressions.

Figure 4.13 shown examples of expression recognition results based on different face size. Based on the test of the original dataset, we can conclude that the recondition system has a good facial positioning ability. Also, the recognition of different expressions is very accurate under the appropriate face size.

Figure 4.13 Examples of images with the different face sizes

Test results with different face position

In this step, we have tested the original datasets with different facial positions. Table 4.8 shows the confusion matrix with face position dataset at (3, 1).

Table 4.8 The confusion matrix with face position dataset at (3, 1)

Actual Class \ Predict Class	Anger	Contempt	Disgust	Fear	Happy	Sadness	Surprise
Anger	24	1	0	0	0	0	0
Contempt	4	16	1	0	3	1	0
Disgust	1	0	24	0	0	0	0
Fear	0	1	0	23	1	0	0
Happy	0	0	0	1	24	0	0
Sadness	0	0	0	0	0	25	0
Surpise	0	0	0	0	0	0	25

Based on the results shown in Table 4.8, we can find that the recognition system can easily classify facial expressions accurately. Moreover, in the test of Contempt expression, four Anger, one Disgust, three Happy and one Sadness were recognized incorrectly.

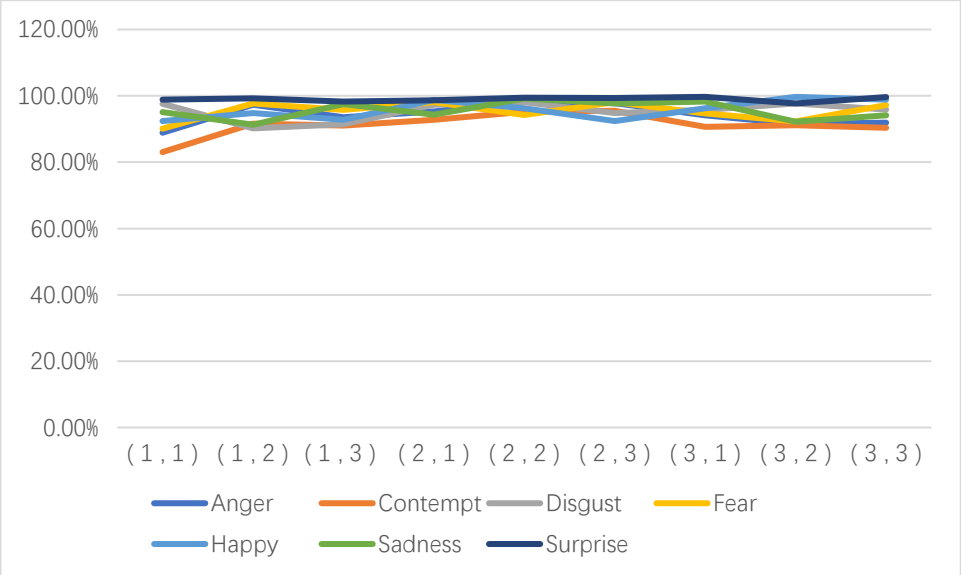

Figure 4.14 Average recognition rates with different face positions

Figure 4.14 shows the recognition results for different emotion recognition based on different face position. The trend of expression curve is very stable. Basically, the expression recognition results in all different positions are very high. Now, we can answer the previous assumptions. The recognition performance of the faster R-CNN model is not

affected by facial position. Although part of the expression recognition results in the preliminary test are not very reasonable, but all the expression categories have excellent recognition results in this test.

Figure 4.15 shows some examples of expression recognition results with different face position. Based on different facial positions, the recognition system can still maintain stable facial positioning ability and high recognition rate.

Figure 4.15 Examples of images with different face positions

Test results with different illumination

In this section, we tested the different facial illumination of the original datasets. Table 4.9 shows the confusion matrix with face illumination dataset based on -2 illumination level.

Table 4.9 The confusion matrix with face illumination dataset under -2 level

Actual Class \ Predict Class	Anger	Contempt	Disgust	Fear	Happy	Sadness	Surprise
Anger	22	0	3	0	0	0	0
Contempt	4	17	0	0	0	3	1
Disgust	3	0	21	0	0	1	0
Fear	0	0	2	23	0	0	0
Happy	0	0	2	0	22	0	1
Sadness	1	0	2	0	0	22	0
Surpise	1	0	0	0	0	2	22

From the information shown in Table 4.9, we can conclude that the recognition system can basically make an accurate classification. Only the Contempt expression test results present the poor performance compared with other facial expressions test results.

Figure 4.16 shows the trends in the average recognition rate for different expressions recognition based on different facial illumination. According to the following chart, the illumination change has significant influence on the recognition rate. Under extremely weak illumination, the recognition system almost lose recognition ability to some expressions. In addition, the recognition of all expressions can achieve very significant results between -1 and 1 levels.

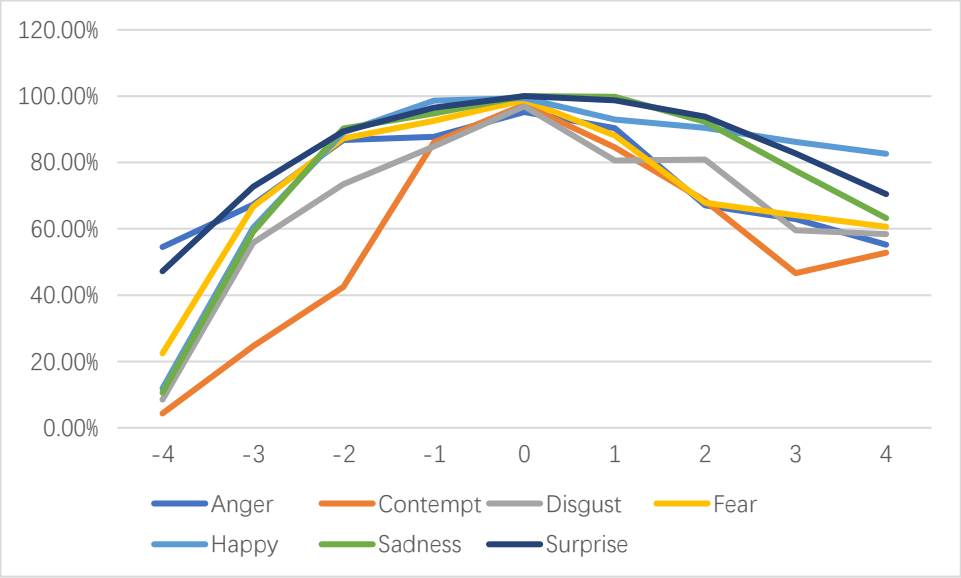

Figure 4.16 Average recognition rate with different face illumination

Based on the above test results, we can prove that the assumptions of the previous test are very reasonable. Figure 4.17 shows some examples of one expression with different face illumination.

Figure 4.17 Examples of images with different face illumination

Test results with different facial background

In this section, we tested the original datasets under different facial backgrounds. Table 4.10 presents the confusion matrix with face position dataset based on the third face background.

Table 4.10 The confusion matrix with face position dataset under the third background

Predict Class / Actual Class	Anger	Contempt	Disgust	Fear	Happy	Sadness	Surprise
Anger	25	0	0	0	0	0	0
Contempt	1	21	1	0	0	2	0
Disgust	1	0	24	0	0	0	0
Fear	0	0	0	25	0	0	0
Happy	0	0	0	0	24	0	1
Sadness	0	1	3	0	0	21	0
Surpise	0	0	0	1	0	0	24

After changing the background of the original training data, the test results did not have a very significant impact. Basically, the classifier can accurately classify all facial expressions.

Figure 4.18 shows the trend of different facial expression recognition rates based on different facial backgrounds. All facial expression recognition rates can be maintained between 90% and 100%.

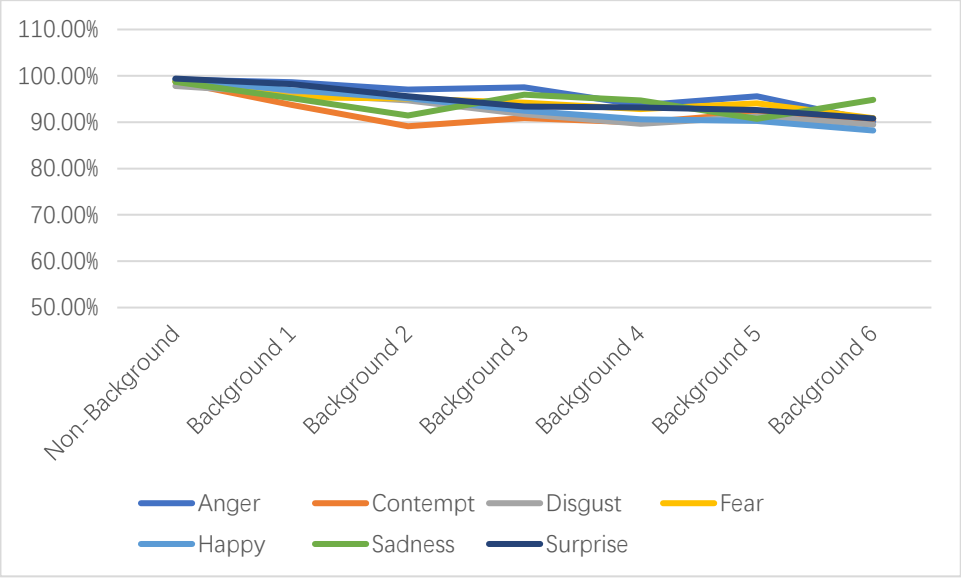

Figure 4.18 Average recognition rate with different face backgrounds

Based on the above test results, we can draw further conclusions on the previous assumptions. Changes in the facial background (include changes of color) will not affect the faster R-CNN model performance. Figure 4.19 shows some examples of one expression recognition with different facial backgrounds.

Figure 4.19 Examples of images with different backgrounds

4.3 Results with Mask R-CNN

In the test of the second method, we explored the effect of related facial image variables on the performance of Mask R-CNN model. In the preliminary test, we modified my facial dataset in different sizes, positions, poses, illumination and backgrounds. Then, the modified dataset is used for testing the Mask R-CNN model. In the proper test, we verified the original dataset by face size, position, illumination and background.

What needs to be raised is that there is not much difference of results between the preliminary test and proper test, and only some small part of the details is different. Therefore, we no longer make any assumptions in this section.

4.3.1 Preliminary Tests

Test results with the different face size

In this section, we used my facial dataset of different facial sizes to test the performance of the Mask R-CNN model. Table 4.11 presents the confusion matrix with face size dataset of 60% - 70%.

Table 4.11 The confusion matrix for dataset with 60-70% face size

Predict Class / Actual Class	Anger	Contempt	Disgust	Fear	Happy	Sadness	Surprise
Anger	24	0	1	0	0	0	0
Contempt	0	23	1	0	0	1	0
Disgust	1	1	23	0	0	0	0
Fear	0	0	0	24	0	0	1
Happy	0	1	0	0	24	0	0
Sadness	0	2	0	0	0	23	0
Surpise	0	0	0	1	0	0	24

According to Figure 4.20, we can conclude that the classifier showed excellent classification performance in 60% to 70% face size. Compared with the faster R-CNN model, the Mask R-CNN model can distinguish all facial expression classes. The above results further illustrate that the training model has an excellent expression recognition ability.

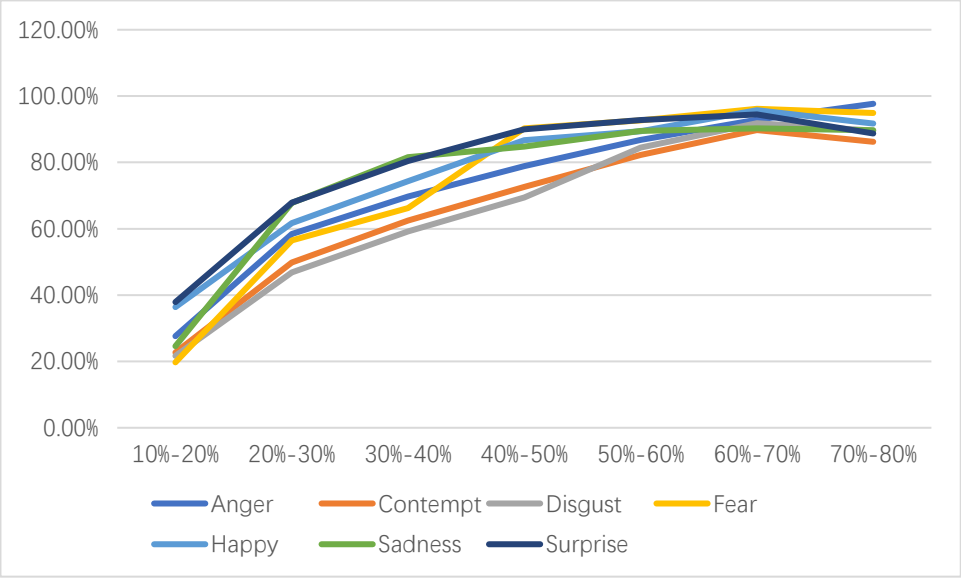

Figure 4.20 Average recognition rate with the different face sizes

Figure 4.21 shows some examples of expressions recognition based on the different face size. From the following images, we can see that the Mask R-CNN model has an excellent performance on the face localization. Moreover, the recognition rate of facial

size changes with this range remained above 95%.

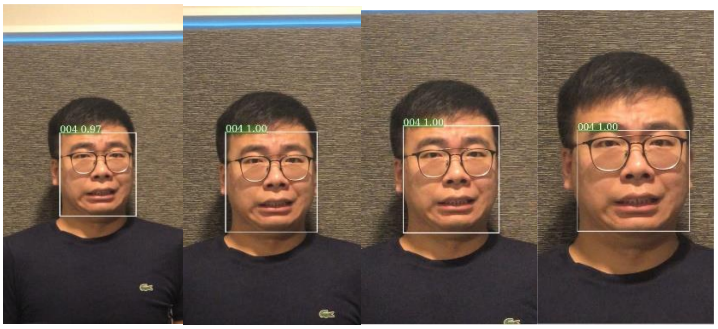

Figure 4.21 Examples of images with the different face sizes

Test results with different face positions

In this section, we tested the performance of the Mask R-CNN model at different face positions. Table 4.12 presents the confusion matrix with face position dataset under (3, 1).

Table 4.12 The confusion matrix with face position dataset under (3, 1)

Predict Class / Actual Class	Anger	Contempt	Disgust	Fear	Happy	Sadness	Surprise
Anger	22	0	2	0	0	1	0
Contempt	1	22	0	0	0	2	0
Disgust	1	0	24	0	0	0	0
Fear	0	0	0	25	0	0	0
Happy	0	2	0	0	23	0	0
Sadness	0	0	0	0	0	25	0
Surpise	0	0	0	0	0	0	25

For the confusion matrix of facial expression in (3, 1) position, all facial expression can be easily distinguished from other facial expressions. Compared with the test results of Faster R-CNN model, the Mask R-CNN model can accurately distinguish the features of different expressions.

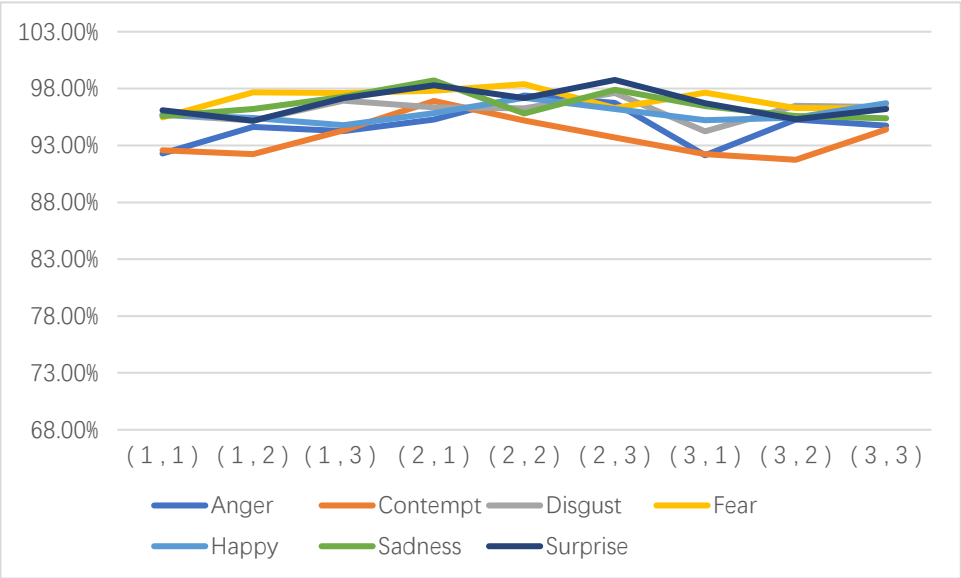

Figure 4.22 Average recognition rate with different face positions

Figure 4.23 presents some examples of one expression recognition based on different face positions. As shown in the images, the Mask R-CNN model has excellent performance in facial localization and recognition.

Figure 4.23 Examples of images with different face positions

Test results with different face poses

For this part, we tested the recognition system performance with different face poses. Table 4.13 shows the confusion matrix of face poses from 30° to 60° angles.

Table 4.13 The confusion matrix with face pose dataset under 30° to 60° angles

Actual Class \ Predict Class	Anger	Contempt	Disgust	Fear	Happy	Sadness	Surprise
Anger	21	2	0	0	0	2	0
Contempt	2	18	0	3	0	3	2
Disgust	3	0	20	0	0	2	0
Fear	0	0	0	21	0	2	2
Happy	0	0	0	2	22	1	0
Sadness	2	3	0	0	0	20	0
Surpise	2	0	0	3	0	0	20

The test of different facial pose is limited to preliminary test. The original data image cannot be processed in depth rotation, so we cannot further verify in the proper test. From Table 4.13, we can see that the Mask R-CNN model cannot make more accurate classification for most facial expressions compared with the faster R-CNN model results. Figure 4.24 shows the trend of all expression recognition rate based on different face poses.

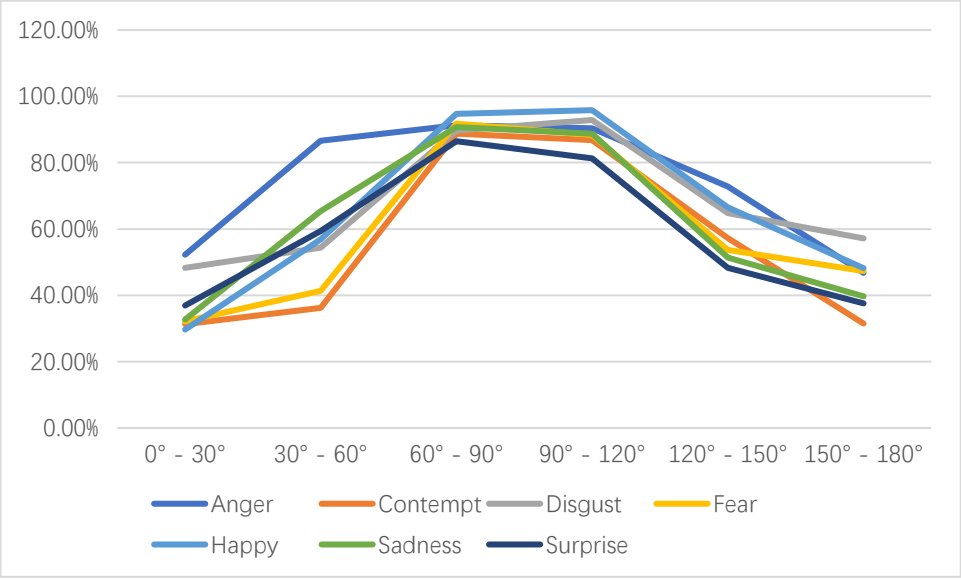

Figure 4.24 Average recognition rate with different face poses

Figure 4.25 presents some examples of one expression recognition based on different

face poses. The expression recognition system can accurately capture facial location with different face angles. But the recognition rate still fluctuated greatly with change of facial angle.

Figure 4.25 Examples of image with different face poses

Test results with different face illumination

In this part, we tested the Mask R-CNN model performance based on the different face illumination. Table 4.14 presents the confusion matrix with face illumination dataset under -2 illumination level.

Table 4.14 The confusion matrix with face illumination dataset in -2 level

Predict Class / Actual Class	Anger	Contempt	Disgust	Fear	Happy	Sadness	Surprise
Anger	23	0	2	0	0	0	0
Contempt	0	22	2	0	0	1	0
Disgust	2	0	22	0	0	1	0
Fear	0	0	1	22	0	0	2
Happy	0	3	0	0	22	0	0
Sadness	0	2	1	0	0	22	0
Surpise	0	0	0	0	0	0	25

Compared with the illumination test results of Faster R-CNN model, the facial expression images with weak illumination have no obvious influence on Mask R-CNN

model. Most facial expressions can be accurately identified. Figure 4.26 illustrates the recognition rate trends of different face illumination.

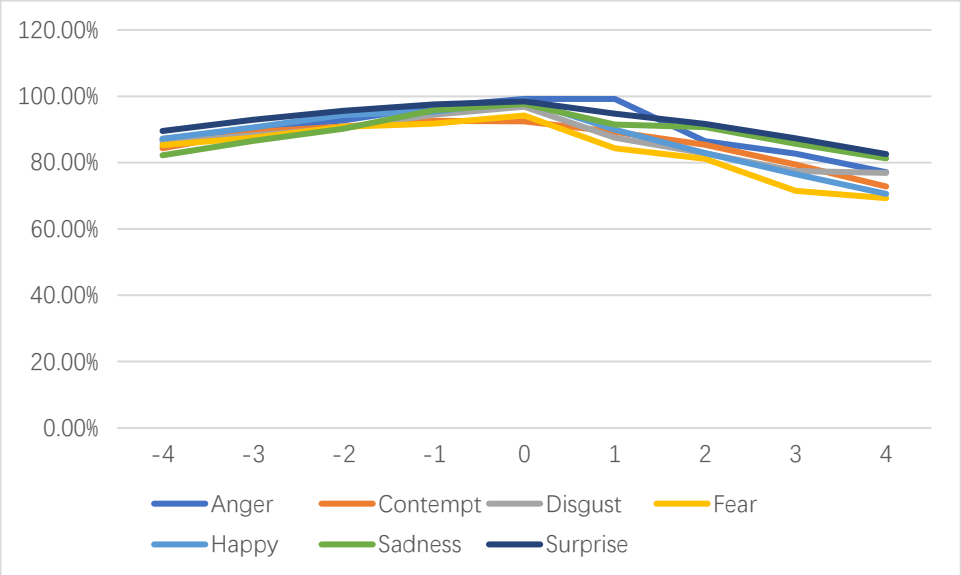

Figure 4.26 Average recognition rate with different face illumination

Figure 4.27 presents some examples for one expression recognition with different face illumination. According to the following images, we can observe that strong illumination has more significant impact on the recognition system than weak illumination. Besides, the face localization is not affected by the change of face illumination.

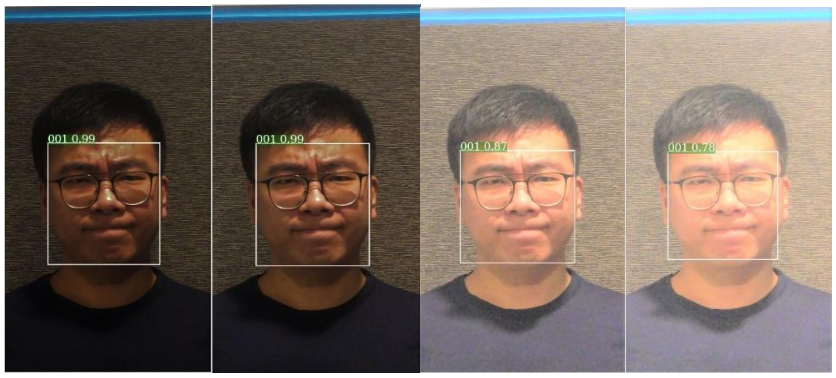

Figure 4.27 Examples of images with different face illumination

Test results with different face background

For this part, we tested the Mask R-CNN model performance by the different face background. Table 4.15 shows the confusion matrix with face background dataset in the third background.

Table 4.15 The confusion matrix with third face background dataset

Predict Class / Actual Class	Anger	Contempt	Disgust	Fear	Happy	Sadness	Surprise
Anger	23	0	1	0	0	1	0
Contempt	1	24	0	0	0	0	0
Disgust	1	0	24	0	0	0	0
Fear	0	0	0	25	0	0	0
Happy	0	0	0	0	24	0	1
Sadness	1	0	1	0	0	23	0
Surpise	0	0	0	1	0	0	24

From the information shown in Table 4.15, we can see that the recognition system can accurately classify all facial expressions. Compared with background test results of Faster R-CNN model, all test results did not deviate greatly. Figure 4.28 presents the trend of expressions recognition rate with different face backgrounds. From the line chart, we can see that the recognition rate of all expressions has reached a very high level. With the increase of background complexity, the recognition rate is stable and infinitely close to 100%.

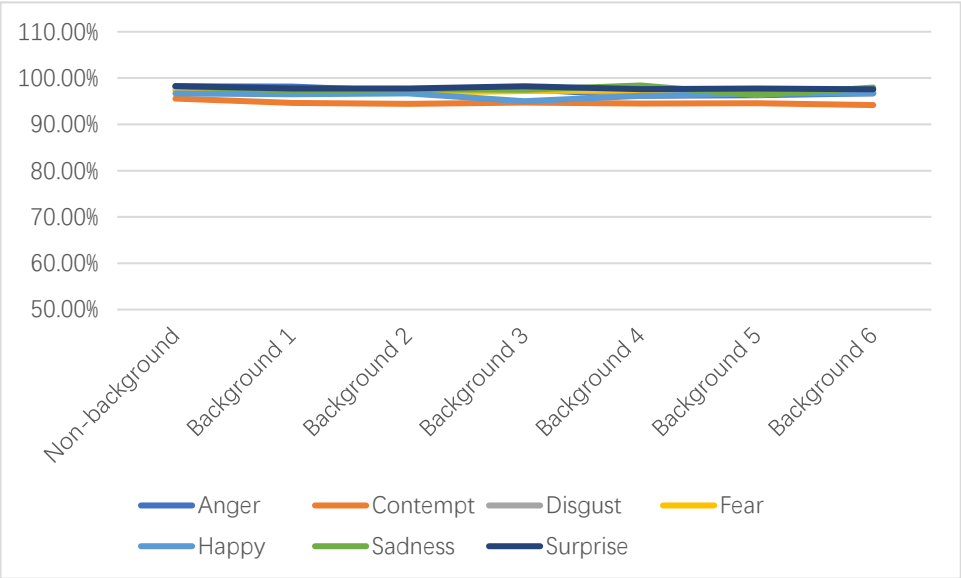

Figure 4.28 Average recognition rates different image backgrounds

Figure 4.29 presents some examples of one expression recognition based on different face backgrounds. From the following recognition images, we can find that the recognition system can accurately capture the face position and make accurate expression recognition. The complexity of the face background will not affect the Mask R-CNN model.

Figure 4.29 Examples of images with different face background

4.3.2 Proper Test

Test results with the different face size

In this section, we proper tested the recognition performance of the Mask R-CNN model on different facial sizes. Table 4.16 shows the confusion matrix with face size dataset under 60% to 70%.

Table 4.16 The confusion matrix for dataset with 60-70% face size

Predict Class / Actual Class	Anger	Contempt	Disgust	Fear	Happy	Sadness	Surprise
Anger	23	0	1	0	0	1	0
Contempt	1	24	0	0	0	0	0
Disgust	1	0	24	0	0	0	0
Fear	0	0	0	25	0	0	0
Happy	0	0	0	0	24	0	1
Sadness	1	0	1	0	0	23	0
Surpise	0	0	0	1	0	0	24

From Table 4.16, we can see that most of the test results are acceptable. Compared with the face size test results of the faster R-CNN model, all the rest results obtained a higher level. Figure 4.30 presents the trend of the average recognition rate of all expression with different face size. The recognition rate of all expressions increases with the increase of facial size. Moreover, in the range of 60% to 80%, the recognition rate of most expressions is close to 100%.

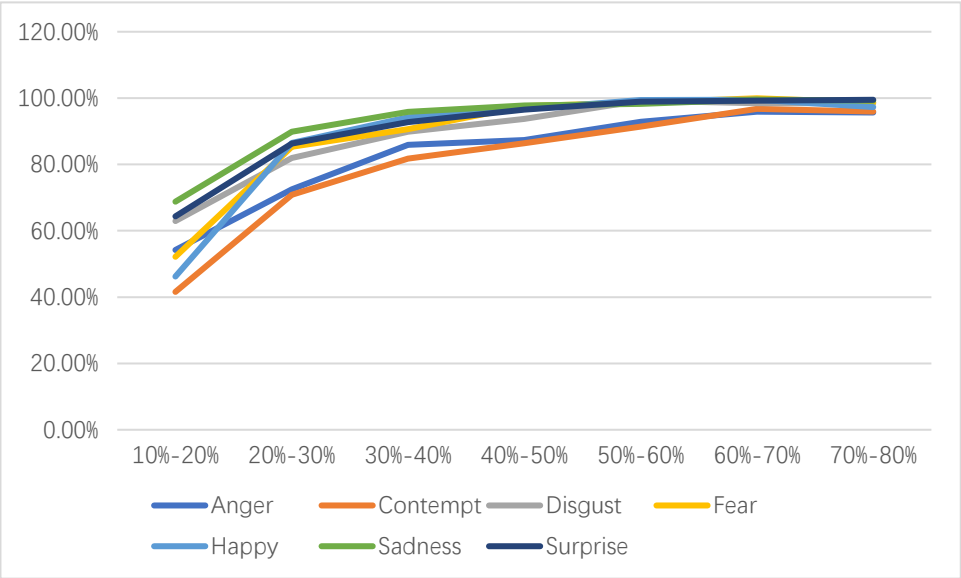

Figure 4.30 Average recognition rates for different face sizes

Figure 4.31 illustrates some examples of one expression recognition based on different face sizes. According to the following recognition images, we can observe that the recognition system can maintain a good recognition performance. But if the facial size is large, the localization of the face will be weakly affected.

Figure 4.31 Examples of images with the different face size

Test results with different face position

In this section, we proper tested the Mask R-CNN model performance on different face positions. Table 4.17 shows the confusion matrix with face position dataset at (3, 1).

Table 4.17 The confusion matrix with face position dataset at (3, 1)

Predict Class / Actual Class	Anger	Contempt	Disgust	Fear	Happy	Sadness	Surprise
Anger	24	0	1	0	0	0	0
Contempt	0	21	2	0	0	2	0
Disgust	0	0	25	0	0	0	0
Fear	0	0	0	24	0	0	1
Happy	0	0	0	0	25	0	0
Sadness	0	0	0	0	0	25	0
Surpise	0	0	0	0	0	0	25

From the Table 4.17, we can see that the recognition system has excellent recognition ability for all expressions at (3, 1) positions. In particular, two Disgust expressions and two Sadness expression were erroneously recognized in the Contempt test results.

Figure 4.32 illustrates the trend of the average recognition rate of all expressions with different face positions. From the chart, the recognition rate of all facial expressions is maintained at 90% to 100%.

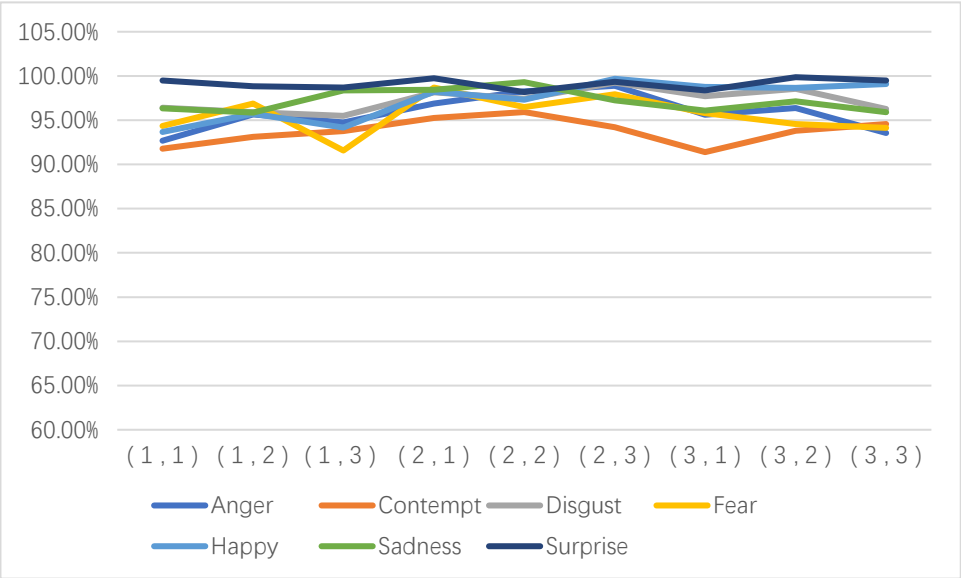

Figure 4.32 Average recognition rates for different face positions

Figure 4.33 presents some examples of one expression recognition results based on different face positions.

Figure 4.33 Examples of images with different face positions

Test results with different face illumination

For this section, we proper tested the Mask R-CNN model performance on different face illumination. Table 4.18 illustrate the confusion matrix with face illumination dataset in -

2 illumination level.

Table 4.18 The confusion matrix with face illumination dataset based on -2 level

Predict Class / Actual Class	Anger	Contempt	Disgust	Fear	Happy	Sadness	Surprise
Anger	25	0	0	0	0	0	0
Contempt	0	22	2	0	0	1	0
Disgust	0	0	25	0	0	0	0
Fear	0	0	0	25	0	0	0
Happy	0	0	0	0	25	0	0
Sadness	0	0	0	0	0	25	0
Surpise	0	0	0	0	0	0	25

From Table 4.18, we can conclude that the recognition system generally shows better recognition results. Compared with the illumination test of the Faster R-CNN model, all of the facial expressions were tested at a higher level. Figure 4.34 presents the trend of the average recognition rate of all expressions based on different face illumination. Illumination affects the performance of the recognition system to a certain extent. But the recognition rate of most expressions can be maintained at over 90%. Therefore, the Mask R-CNN model shows excellent performance under variable illumination environment.

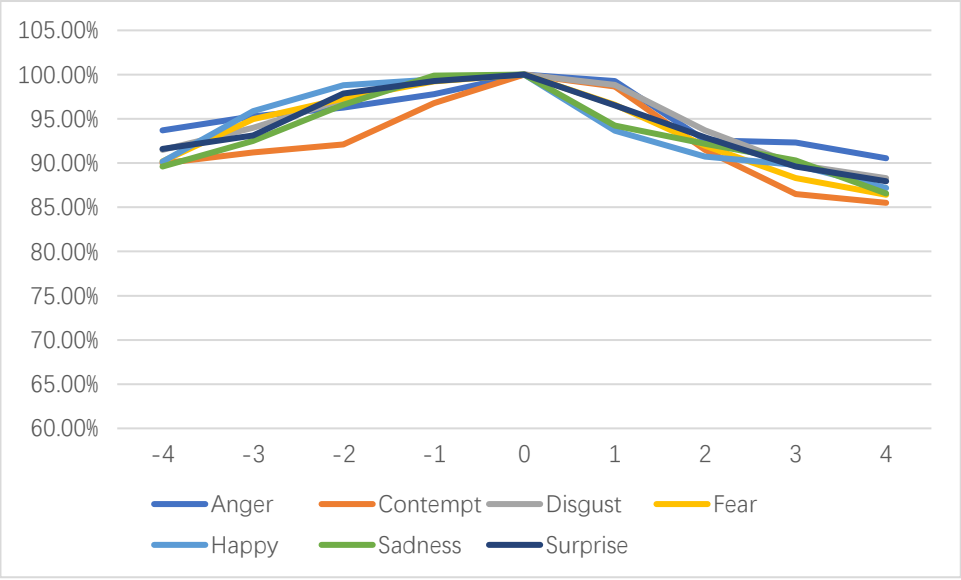

Figure 4.34 Average recognition rate with different face illumination

Figure 4.35 illustrates some examples of one expression recognition results based on different face illumination.

Figure 4.35 Examples of images with different face illumination

Test results with different background

In this section, we proper tested the Mask R-CNN model performance on different face backgrounds. Table 4.19 presents the confusion matrix with face background dataset based on the third background.

Table 4.19 The confusion matrix with the third background dataset

Predict Class / Actual Class	Anger	Contempt	Disgust	Fear	Happy	Sadness	Surprise
Anger	24	0	1	0	0	0	0
Contempt	0	23	1	0	0	1	0
Disgust	0	0	25	0	0	0	0
Fear	0	0	0	23	0	0	2
Happy	0	0	0	0	24	0	1
Sadness	0	0	1	0	0	24	0
Surpise	0	0	0	1	0	0	24

Compared with results of the preliminary test, Mask R-CNN model has better recognition results for the original datasets at the third background. Figure 4.36 illustrates

the trend of the average recognition rate for all expressions with different face backgrounds. According to the line chart, all facial expression recognition rates are close to 100%. The Mask R-CNN model is not affected by background changes.

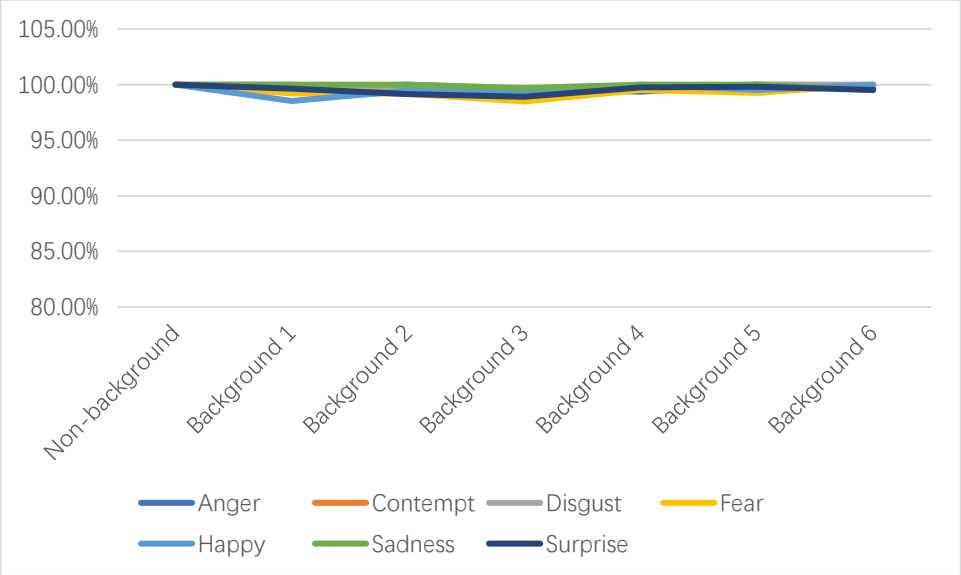

Figure 4.36 Average recognition rate with different image backgrounds

Figure 4.37 presents some examples of one expression recognition results with different face backgrounds. Face localization and recognition results have reached a very high level.

Figure 4.37 Examples of images with different face backgrounds

4.4 Analyses

We have introduced the direct results of each training algorithm in the previous chapter. In this chapter, the analysis of experimental results will be analyzed explicitly with our training methods. Moreover, the performance of comparative analysis of the two methods will be presented in this chapter.

4.4.1 Analysis for Faster R-CNN

Faster R-CNN model is a deep learning target detection algorithm based on region proposal. For region proposal, the possible location of the object in the image can be found by using the texture, edge, color and other information in the image. The faster R-CNN model used the region proposal method to replace the sliding window method in the conventional deep learning method. Experiments show that the use of region proposal

method can obtain higher quality target region compared with the traditional convolutional neural network.

According to the previous chapter, the influence of facial size changes on the Faster R-CNN model has a linear increase. With the rise of the facial size, the recognition rate has also increased significantly. Because of the single face size of the training dataset, the recognition model is easily influenced by objective factors. The change of facial size will result in the change of facial contour, edge and so on, which makes the recognition system more difficult to extract facial features. In the facial position test, we found that the change of facial position would not cause a significant change in recognition rate. Under the premise of ensuring the integrity of facial features, region proposal method can capture thousands of regions proposal for the input image by sliding window. Wherever the face position in the image, the recognition system can get the exact position of the face. In the test of face illumination, the change of illumination has great impact on the faster R-CNN model. In particular, the influence of strong illumination on the accuracy of the recognition is far greater than that of weak illumination. The illumination affects the image structure of the face, which is deviated from face contour and texture. The region proposal method uses the edge, texture and color of the image to extract the face region proposal. The inaccuracy of facial region proposal extraction further leads to a decrease in recognition ability. The change of facial background did not significantly affect the faster R-CNN model. Theoretically, changes in the facial background do not affect the texture, contour and color of the facial image. Therefore, in the process of facial region proposal extraction, the target can still be accurately captured.

Table 4.20 shows the precision, recall, F-score and accuracy values of the faster R-CNN model in the preliminary test and proper test. We use these values to evaluate the performance of the Faster R-CNN model in facial expression recognition. Because there are different test sets in Faster R-CNN experiment, we choose representative datasets to

generate a confusion matrix under preliminary and proper test. In constant facial illumination, angle, position, and background, 50% to 60% of the face size is used to generate confusion matrixes. The data shown in Table 4.20 are calculated from the generated confusion matrixes.

Table 4.20 Evaluation of Faster R-CNN

Evaluation / Expression	Precision		Recall		F-score		Accuracy	
	Preliminary	Proper	Preliminary	Proper	Preliminary	Proper	Preliminary	Proper
Anger	0.8400	1.0000	0.8400	0.9600	0.8400	0.9796	98.87%	95.94%
Contempt	0.7917	1.0000	0.7600	1.0000	0.7755	1.0000	76.50%	92.60%
Disgust	0.6897	1.0000	0.8000	1.0000	0.7408	1.0000	79.56%	98.21%
Fear	0.8333	1.0000	0.8000	1.0000	0.8163	1.0000	93.20%	99.30%
Happy	1.0000	1.0000	0.8800	1.0000	0.9362	1.0000	97.60%	99.40%
Sadness	0.8077	1.0000	0.8400	1.0000	0.8936	1.0000	85.30%	96.73%
Surprise	0.8400	1.0000	0.8400	0.9600	0.8400	0.9796	69.83%	96.82%
Average value	0.8289	1.0000	0.8229	0.9886	0.8346	0.9941	85.84%	97.00%

In the preliminary column, the precision and recall values indicate that most of the expression classes are mistaken in the confusion matrix. Only the Happy expression reached 100% of the ratio, which the instance of true positive account for instance that positively predicted. Accuracy can intuitively see the recognition ability of classifier and F-score is the metric to quantify its performance. According to the F-score and accuracy results, the faster R-CNN model predicts well on multiple class classification on imbalanced given data in proper test. Because the Faster R-CNN used my facial dataset, F-score and accuracy value can illustrate the difference between my facial dataset and the original dataset. Due to no accurate measurement of facial expressions, the inaccuracy of my own facial expressions dataset is the primary cause of low F-score and accuracy value.

4.4.2 Analysis for Mask R-CNN

The Mask R-CNN model has two stages processes. In the first stage, the region proposal

network is used to generate a large number of candidate regions. In the second stage, the Mask R-CNN model outputs binary masks for each region of interest under the premise of predicting class and box offset. The same operation as Faster R-CNN model is used in bounding box classification and regression. Mask R-CNN model adds a branch of the third output object mask to each candidate by providing class labels and boundary box offsets.

From the experimental results of Mask R-CNN model, we find that the model has excellent recognition performance under different facial conditions. From the test result of facial size, we can explore that facial size has little effect on Mask R-CNN model compared with the experimental results of the Faster R-CNN model. More smooth curves are presented in the figure of trend recognition rate. For different facial positions, the application of the region proposal method makes the Mask R-CNN model less susceptible to the facial position. Moreover, the competition between classes is reduced by using a binary mask of each class. As we are known in the previous chapter, the recognition rate for the different facial position can be achieved at very high levels. For face illumination experiments, we can find that the performance of the Mask R-CNN model is much better than that of the Faster R-CNN model. In the process of building the model, we replace the ROI pooling layer with the ROI Align layer. The ROI layer corrects the alignment problem using direct sampling, for pixel level recognition, there will be large errors. For illumination change images, the application of ROI Align layer can significantly improve the recognition rate under different illumination intensities. For the test of facial background, the test results of Faster R-CNN model and Mask R-CNN model did not deviate greatly.

Table 4.21 shows the precision, recall, F-score and accuracy results of Mask R-CNN model in the preliminary and proper test. Mask R-CNN is an optimization deep learning algorithm based on Faster R-CNN model. On the basis of the class label and boundary

box offset of the candidate objects, the third branch is adopted into the Mask R-CNN model to generate the binary mask of each candidate object. This model is proposed in this thesis to explore the recognition performance of the model based on different facial conditions. The following experimental results are obtained from the confusion matrix generated by preliminary and proper tests. The dataset used in the test is 50% to 60% face size images.

Table 4.21 Evaluation of Mask R-CNN

Evaluation / Expression	Precision		Recall		F-score		Accuracy	
	Preliminary	Proper	Preliminary	Proper	Preliminary	Proper	Preliminary	Proper
Anger	0.8400	0.9600	0.8400	0.9600	0.8400	0.9600	86.84%	92.85%
Contempt	0.8333	1.0000	0.8000	0.9200	0.8163	0.9583	82.27%	91.36%
Disgust	0.7407	0.9615	0.8000	1.0000	0.7692	0.9804	84.47%	99.23%
Fear	0.7857	1.0000	0.8800	1.0000	0.8302	1.0000	92.72%	98.62%
Happy	1.0000	1.0000	0.8800	1.0000	0.9362	1.0000	89.39%	99.40%
Sadness	0.9565	0.9615	0.8800	1.0000	0.9167	0.9804	89.50%	98.24%
Surprise	0.8846	1.0000	0.9200	1.0000	0.9020	1.0000	92.76%	98.93%
Average value	0.8630	0.9833	0.8571	0.9829	0.8587	0.9827	88.28%	96.95%

From the Mask R-CNN model evaluation information, we can conclude that Mask R-CNN model shows acceptable recognition results in both preliminary and proper tests. According to the previous chapter, the Mask R-CNN model illustrates an effective ability to locate faces. Based on the average of F-score and accuracy value of all expressions, Mask R-CNN model has better recognition ability in the proper test than preliminary test. Theoretically, the ROI Align layer constructed in the model solves the problem that the feature map and original image are difficult to align. This problem is particularly difficult to achieve in the Faster R-CNN. Moreover, the model removes competition between classes by predicting a binary mask independently for each class. The classification of each binary mask is based on the classification given by the branch of the region of interest network. The model successfully dismantled mask prediction and classification prediction. Therefore, a high recognition rate and accurate facial localization are realized based on the above methods.

4.4.3 Comparison and Discussion both of Two Methods

In this section, the comparison between Faster R-CNN model and Mask R-CNN model will be further discussed. The following contents present the contrast results with two deep learning algorithms. Because each experiment was tested under different facial conditions, we only selected one of the facial states to complete the test. Based on the confusion matrix generated in the test, precision, recall, F-score and accuracy are computed to measure the specific performance of the two methods. Moreover, the testing time of the two methods is also collected.

We first discuss the different ways of constructing the two methods. Faster R-CNN and Mask R-CNN model are built on the basis of the region of convolution neural network(R-CNN). The core of R-CNN is to add region proposal network into the CNN. The general process of the algorithm is to generate a large number of region proposal at first. Then, we use the trained CNN to extract every region proposal feature. Finally, each region proposal is graded by the trained support vector machine. The structure of Faster R-CNN and Mask R-CNN follows the above recognition process. The primary purpose of the Faster R-CNN algorithm is to achieve a faster training and recognition speed compared to R-CNN and Fast R-CNN. Therefore, the core promotion of Faster R-CNN is the application of region proposal network(RPN). RPN outputs a predicted object bounding box and objects score at each image position. Also, the RPN and Fast R-CNN share the feature map. From the following experimental results, we can see that the test results of Faster R-CNN have faster test time than the other one. In the Mask R-CNN model, avoiding any quantization of region of interest boundaries or bins solves the problem that the feature map and the original image cannot be aligned in Faster R-CNN. Moreover, a parallel FCN layer is added to the Mask R-CNN model. The FCN layer can classify each pixel in multiple classes. The use of a sigmoid function avoids competition

among classes. Next, we turn to the performance of the two methods in practical applications based on our test results.

Table 4.22 shows the evaluation results of two methods for original dataset with 60-70% face size. During the evaluation test, the original dataset with 60-70% face size was used to test Faster R-CNN and Mask R-CNN respectively. Each expression contains 25 images as input to the models. According to the test results, we generate the confusion matrix of Faster R-CNN and Mask R-CNN for the dataset with 60-70% face size. Based on the confusion matrix, we further calculated each evaluation values.

Table 4.22 Evaluation of two methods for the test of facial size

Evaluation / Expression	Precision		Recall		F-score		Accuracy		Test time	
	F-RCNN	M-RCNN	F-RCNN	M-RCNN	F-RCNN	M-RCNN	F-RCNN	M-RCNN	F-RCNN	M-RCNN
Anger	0.9545	1.0000	0.8400	0.9600	0.8936	0.9796	91.84%	95.90%	0.120s	0.156s
Contempt	0.9200	1.0000	0.9200	0.9200	0.9200	0.9583	96.66%	96.75%	0.114s	0.159s
Disgust	0.8000	0.8929	0.9600	1.0000	0.8727	0.9434	98.70%	98.49%	0.112s	0.155s
Fear	0.9565	0.9583	0.8800	0.9200	0.9167	0.9388	93.44%	99.94%	0.113s	0.160s
Happy	1.0000	1.0000	0.9600	0.9600	0.9796	0.9796	99.40%	99.50%	0.115s	0.155s
Sadness	0.9583	0.9600	0.9200	0.9600	0.9600	0.9600	91.42%	99.56%	0.113s	0.155s
Surprise	0.8889	0.8889	0.9600	0.9600	0.9231	0.9231	97.90%	99.20%	0.115s	0.154s
Average value	0.9255	0.9572	0.9200	0.9543	0.9237	0.9547	96.62%	98.48%	0.114s	0.156s

According to the values of Faster R-CNN and Mask R-CNN, we can find that the results of Mask R-CNN model are generally higher than the result of Faster R-CNN model, which means, the Mask R-CNN model has fewer error recognition in each expression test process. The test results of some expressions reached 1, indicating that there was no error recognition in the test. From the average results of Faster R-CNN and Mask R-CNN, it can be concluded that the Mask R-CNN model has better classification ability for multi-class classification that the other method. However, the Faster R-CNN model has less recognition time that the test time of the Mask R-CNN model.

Table 4.23 shows the evaluation results of two methods for original dataset with (3, 1) face position. In this test, the confusion matrixes of Faster R-CNN and Mask R-CNN

is generated based on the different facial expressions under the same facial position. The confusion matrixes further calculate the values of precision, recall, F-score and accuracy of the two methods. Moreover, the average test time was collected by repeated experiments.

From the results shown in the Table 4.23, we can conclude that Mask R-CNN model is better than Faster R-CNN model in the classification of each expression class. From the results of F-score and accuracy, we can see that the average result of the Mask R-CNN model is higher than that of the Faster R-CNN model, which means, Mask R-CNN model has better classification performance in multi-expression classification. The test time of the Mask R-CNN model is slightly longer than that of the Faster R-CNN model. On the whole, the test results of the two methods did not cause much deviation. From previous experiments, the Mask R-CNN model is more dependent on the Faster R-CNN model in facial localization.

Table 4.23 Evaluation of two methods for the test of facial position

Evaluation / Expression	Precision		Recall		F-score		Accuracy		Test time	
	F-RCNN	M-RCNN	F-RCNN	M-RCNN	F-RCNN	M-RCNN	F-RCNN	M-RCNN	F-RCNN	M-RCNN
Anger	0.8276	1.0000	0.9600	0.9600	0.8889	0.9796	94.12%	95.60%	0.113s	0.158s
Contempt	0.8889	1.0000	0.6400	0.8400	0.7442	0.9130	90.62%	91.39%	0.115s	0.164s
Disgust	0.9600	0.8929	0.9600	1.0000	0.9600	0.9434	95.80%	97.71%	0.114s	0.159s
Fear	0.9583	1.0000	0.9200	0.9600	0.9434	0.9796	94.70%	95.81%	0.115s	0.160s
Happy	0.8571	1.0000	0.9600	1.0000	0.9056	1.0000	96.28%	98.77%	0.116s	0.163s
Sadness	0.9615	0.9259	1.0000	1.0000	0.9804	0.9615	98.30%	96.10%	0.114s	0.157s
Surprise	1.0000	0.9615	1.0000	1.0000	1.0000	0.9803	99.70%	98.37%	0.115s	0.158s
Average value	0.9219	0.9686	0.92	0.9657	0.9175	0.9653	95.65%	96.25%	0.114s	0.159s

Table 4.24 shows the evaluation results of two methods for my dataset with 30°-60° face pose. For facial pose testing, we only tested our facial dataset without testing the testing the original dataset. The two methods do not have reasonable experimental results in facial expression recognition. But we still operate to compare the performance of the two methods in 30°-60° facial angles.

Table 4.24 Evaluation of two methods for the test of facial pose

Evaluation / Expression	Precision		Recall		F-score		Accuracy		Test time	
	F-RCNN	M-RCNN	F-RCNN	M-RCNN	F-RCNN	M-RCNN	F-RCNN	M-RCNN	F-RCNN	M-RCNN
Anger	0.6897	0.7000	0.8000	0.8400	0.7408	0.7636	82.86%	86.61%	0.115s	0.158s
Contempt	0.7500	0.7826	0.7200	0.7200	0.7347	0.7500	23.90%	36.19%	0.119s	0.155s
Disgust	0.6786	1.0000	0.7600	0.8000	0.7170	0.8889	41.68%	54.37%	0.118s	0.153s
Fear	0.7241	0.7241	0.8400	0.8400	0.7778	0.7778	38.49%	41.30%	0.116s	0.155s
Happy	0.9412	1.0000	0.6400	0.8800	0.7619	0.9361	45.62%	56.91%	0.119s	0.157s
Sadness	0.7727	0.6667	0.6800	0.8000	0.7234	0.7273	51.73%	65.35%	0.115s	0.158s
Surprise	0.7308	0.8334	0.7600	0.8000	0.7451	0.8164	47.83%	59.43%	0.111s	0.151s
Average value	0.7553	0.8153	0.7429	0.8114	0.7430	0.8086	47.44%	57.17%	0.116s	0.155s

According to the results of precision and recall, we can intuitively see that Faster R-CNN and Mask R-CNN model have a large number of error recognitions for the test with 30°-60° facial angle. The accuracy of all facial expression recognition is at a low level. Due to the lack of multi-angle facial instances in the training set, it is complicated to align the feature map with the original image when the non-positive facial image is input into the system. Apparently, this is the primary reason why Faster R-CNN and Mask R-CNN model fail to obtain acceptable results.

Table 4.25 shows the evaluation results for original dataset with -2 facial illumination. The confusion matrixes are generated through images of -2 level facial illumination. Based on the confusion matrixes, we further calculated the test results of precision, recall, F-score, and accuracy. During the experiment, the test time is also collected to measure the performance of each model.

According to the results of precision and recall, we can find that Mask R-CNN model can almost achieve all the accurate recognition of -2 level facial illumination images. Based on the average values of F-score and accuracy, the Mask R-CNN model test for the -2-facial illumination dataset shows better recognition performance that Faster R-CNN model. Moreover, there is a big gap between the two methods, which means, Mask R-

CNN has the better adaptability to illumination change. Changes in facial illumination lead to changes in facial texture, contour, and color. For Faster R-CNN model, such changes make it difficult to align between the feature map and the original image. For the Mask R-CNN model, ROI Align layer is used instead of ROI pooling layer. The original image with illumination changes can still be accurately aligned with the feature map. As shown in the Table 4.25, the Mask R-CNN model still maintains a high recognition accuracy under -2 face illumination conditions.

Moreover, we can find that the test time of the Mask R-CNN model under -2 facial illuminations is slightly increased. The increased time is used for the alignment process between the feature map and the original image. However, the test time of Faster R-CNN has not changed significantly.

Table 4.25 Evaluation of two methods for facial illumination

Evaluation / Expression	Precision		Recall		F-score		Accuracy		Test time	
	F-RCNN	M-RCNN	F-RCNN	M-RCNN	F-RCNN	M-RCNN	F-RCNN	M-RCNN	F-RCNN	M-RCNN
Anger	0.7097	1.0000	0.8800	1.0000	0.7857	1.0000	86.80%	96.25%	0.115s	0.163s
Contempt	1.0000	1.0000	0.6800	0.8800	0.8095	0.9362	42.42%	92.12%	0.117s	0.162s
Disgust	0.7000	1.0000	0.8400	1.0000	0.7636	1.0000	73.40%	97.16%	0.120s	0.160s
Fear	1.0000	1.0000	0.9200	1.0000	0.9583	1.0000	87.22%	97.28%	0.116s	0.164s
Happy	1.0000	1.0000	0.8800	1.0000	0.9361	1.0000	89.10%	98.79%	0.113s	0.162s
Sadness	0.7857	1.0000	0.8800	1.0000	0.8302	1.0000	90.27%	96.58%	0.114s	0.160s
Surprise	0.9167	1.0000	0.8800	1.0000	0.8980	1.0000	89.40%	97.84%	0.112s	0.164s
Average value	0.8732	1.0000	0.8514	0.9829	0.8545	0.9909	79.80%	96.57%	0.115s	0.162s

Table 4.26 shows the evaluation results of two methods for original dataset with the third facial background. We use the confusion matrixes generated in the experiment to calculate further the data shown in Table 4.26. According to the results of F-score and accuracy, we can conclude that there is no significant deviation between Faster R-CNN and Mask R-CNN model. Changes in facial images will not have a significant impact on the two methods.

Table 4.26 Evaluation of two methods for facial background

Evaluation / Expression	Precision		Recall		F-score		Accuracy		Test time	
	F-RCNN	M-RCNN	F-RCNN	M-RCNN	F-RCNN	M-RCNN	F-RCNN	M-RCNN	F-RCNN	M-RCNN
Anger	0.9259	1.0000	1.0000	0.9600	0.9615	0.9796	97.50%	99.62%	0.102s	0.157s
Contempt	0.9545	1.0000	0.8400	0.9200	0.8936	0.9583	90.92%	99.47%	0.090s	0.139s
Disgust	0.8571	0.8929	0.9600	1.0000	0.9056	0.9434	91.70%	98.63%	0.095s	0.156s
Fear	0.9615	0.9583	1.0000	0.9200	0.9803	0.9388	94.20%	98.49%	0.105s	0.154s
Happy	1.0000	1.0000	0.9600	0.9600	0.9796	0.9796	92.50%	99.29%	0.097s	0.153s
Sadness	0.9130	0.9600	0.8400	0.9600	0.8750	0.9600	95.94%	99.71%	0.092s	0.158s
Surprise	0.9600	0.8889	0.9600	0.9600	0.9600	0.9231	93.40%	98.91%	0.093s	0.155s
Average value	0.9389	0.9572	0.9371	0.9543	0.9365	0.9547	93.74%	99.16%	0.096s	0.153s

4.5 Summary

In this section, we compared the recognition performance of Faster R-CNN model and Mask R-CNN model in one facial state. We found that Faster R-CNN and Mask R-CNN had relatively close experimental results in the tests of 60-70% face size, (3, 1) face position, 30°-60° face pose and third facial background. But the Mask R-CNN model still has a weak lead-in F-score and accuracy. In the facial illumination test, we found that the adaptability of the Mask R-CNN model to facial illumination was significantly stronger than Faster R-CNN.

Generally, the Faster R-CNN model improves the recognition accuracy and speed and realizes the end to end recognition framework. However, the model still cannot achieve real-time object detection and for each region proposal classification computation is still very large. For the Mask R-CNN model, the application of the ROI Align layer makes the feature map and the original image more accurately aligned. Even for images with varying illumination, it can maintain high recognition performance. But it can be achieved at the expense of recognition time.

Chapter 5 Conclusions and Future Work

5.1 Conclusions

The objective of this thesis is to study facial expression recognition from expression images. We proposed two facial expression recognition algorithms, namely Faster R-CNN and Mask R-CNN. In the thesis, we divided the testing process into two parts: preliminary test and proper test. Preliminary and proper test correspond to my facial dataset and CK+ dataset respectively to test the algorithm models. In order to verify the performance of algorithm models for different datasets. Experimental results show that the two algorithm models have excellent stability for test different datasets, and part of accuracy achieved satisfactory results. Moreover, we verified that both algorithms achieved high accuracy by adjusting the face size, position, background and other factors. After comparing the results of the two methods, the main contributions are summarized as follows:

For the Faster R-CNN model experiments, we used the RPN method to extract the candidate box of an image. Experimental results show that the Faster R-CNN model has a faster recognition speed. For example, the model has an average time of 0.114s for 60-70% of the face size. In the experiment of different facial states, facial size and illumination have obvious influence on the test results of Faster R-CNN model. The results of face position and background test reached satisfactory accuracy.

In the Mask R-CNN model experiment, we obtained better expression recognition rate and facial localization. The application of ROI Aligned played a positive role in the face localization operation. However, the addition of new mask branch increases the expression recognition time of the Mask R-CNN model. The average recognition time of the Mask R-CNN model is 0.156s for the 60-70% face size.

Overall, Faster R-CNN and Mask R-CNN model exhibit excellent recognition and facial localization abilities for seven facial expressions. The Faster R-CNN has a faster recognition speed, while the Mask R-CNN model has more stable recognition rate and accurate localization.

5.2 Future Works

This research could be extended in several different directions. Four of them are listed below.

1) In this study, we focus on 7 categories of facial expression recognition. In addition, more complex facial expressions should be added to the project.
2) In the process of data training, the facial expression images of CK+ dataset keep the same facial size, illumination and so on. Facial images of different sizes should not be used for training in this project. Therefore, the recognition rate of Faster R-CNN and Mask R-CNN models with 10-60% face size should be improved greatly.
3) During the test, only the preliminary test performed facial expression recognition with different facial expressions. The proper test failed to perform the test because the original image could not be rotated deeply. The lack of this section should be one of the main tasks in the future works.
4) Due to the lack of mask expression dataset, the experimental results do not achieve mask processing. In future, we should create a mask dataset to further mask the image object.

Reference

[1] Baron, R. J. (1981). Mechanisms of human facial recognition. *International Journal of Man-Machine Studies, 15*(2), 137-178.

[2] Brunelli, R., & Poggio, T. (1993). Face recognition: Features versus templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 15*(10), 1042-1052.

[3] Turk, M. A., & Pentland, A. P. (1991). Face recognition using eigenfaces. *IEEE Computer Society Conference,* pp. 586-591.

[4] Hall, P., Park, B. U., & Samworth, R. J. (2008). Choice of neighbor order in nearest-neighbor classification. *The Annals of Statistics, 36*(5), 2135-2152.

[6] Goldstein, A. J., Harmon, L. D., & Lesk, A. B. (1971). Identification of human faces. *Proceedings of the IEEE, 59*(5), 748-760.

[8] Cox, I. J., Ghosn, J., & Yianilos, P. N. (1996, 18-20 June 1996). Feature-based face recognition using mixture-distance. *1996 IEEE Computer Society Conference on Computer Vision and Pattern Recognition,* pp. 209-216.

[9] Zimmerman, D., Pavlik, C., Ruggles, A., & Armstrong, M. P. (1999). An experimental comparison of ordinary and universal kriging and inverse distance weighting. *Mathematical Geology, 31*(4), 375-390.

[10] Kotsia, I., & Pitas, I. (2007). Facial expression recognition in image sequences using geometric deformation features and support vector machines. *IEEE Transactions on Image Processing, 16*(1), 172-187.

[11] Drucker, H., Wu, D., & Vapnik, V. N. (1999). Support vector machines for spam categorization. *IEEE Transactions on Neural networks, 10*(5), 1048-1054.

[12] Ganapathiraju, A., Hamaker, J. E., & Picone, J. (2004). Applications of support vector machines to speech recognition. *IEEE Transactions on Signal Processing, 52*(8), 2348-2355.

[13] Pontil, M., & Verri, A. (1998). Support vector machines for 3D object recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 20*(6), 637-646.

[14] Tefas, A., Kotropoulos, C., & Pitas, I. (2001). Using support vector machines to

enhance the performance of elastic graph matching for frontal face authentication. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 23*(7), 735-746.

[15] Kotsia, I., & Pitas, I. (2007). Facial expression recognition in image sequences using geometric deformation features and support vector machines. *IEEE transactions on image processing*, *16*(1), 172-187.

[16] Chen, X., Flynn, P. J., & Bowyer, K. W. (2003). PCA-based face recognition in infrared imagery: Baseline and comparative studies. *IEEE International Workshop on Analysis and Modeling of Faces and Gestures, 2003. AMFG2003,* pp. 127-134.

[17] Moon, H., & Phillips, P. J. (2001). Computational and performance aspects of PCA-based face-recognition algorithms. *Perception, 30*(3), 303-321.

[18] Zheng, W.-S., Lai, J.-H., & Yuen, P. C. (2005). GA-fisher: a new LDA-based face recognition algorithm with selection of principal components. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 35*(5), 1065-1078.

[19] Dandpat, S. K., & Meher, S. (2013). Performance improvement for face recognition using PCA and two-dimensional PCA. *2013 Intnernational Conference on Computer Communication and Informatics (ICCCI),* pp. 1-5.

[20] Kiby, M., & Sirovich, L. (2010). Application of the kl procedure for the characterization of human faces. *IEEE Trans. Pattern Anal. Mach. Intell, 98*(6), 1031-1044.

[21] Sirovich, L., & Kirby, M. (1987). Low-dimensional procedure for the characterization of human faces. *Josa a, 4*(3), 519-524.

[22] Yang, J., Zhang, D., Frangi, A. F., & Yang, J.-y. (2004). Two-dimensional PCA: a new approach to appearance-based face representation and recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 26*(1), 131-137.

[23] Velicer, W. F. (1976). Determining the number of components from the matrix of partial correlations. *Psychometrika, 41*(3), 321-327.

[24] Zhang, D., & Zhou, Z.-H. (2005). (2D) 2PCA: Two-directional two-dimensional PCA for efficient face representation and recognition. *Neurocomputing, 69*(1-3),

224-231.

[25] McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological methods, 1(*1*), 30.*

[26] Oliveira, L., Mansano, M., Koerich, A., & de Souza Britto, J. A. (2011). 2D principal component analysis for face and facial-expression recognition. *Computing in Science & Engineering, 13*(3), 9-13.

[27] Gosavi, A. P., & Khot, S. R. (2014). Emotion recognition using principal component analysis with singular value decomposition. *2014 International Conference on Electronics and Communication Systems (ICECS),* pp. 1-5.

[28] Face recognition with singular value decomposition [29] Mehta, N., & Jadhav, S. (2016, 12-13 Aug. 2016). Facial Emotion recognition using Log Gabor filter and PCA. *2016 International Conference on Computing Communication Control and automation (ICCUBEA),* pp. 1-5.

[30] Meher, S. S., & Maben, P. (2014). Face recognition and facial expression identification using PCA. *Advance Computing Conference (IACC), 2014 IEEE International,* pp. 1093-1098.

[31] Garg, A., & Choudhary, V. (2012). Facial expression recognition using principal component analysis. *Int. J. Sci. Eng. Res. Technol*, pp. 39-42.

[32] Ojala, T., Pietikainen, M., & Harwood, D.(1994). Performance evluation of texture measures with classification based on Kullback discrimination of distributions. *Proceedings of the 12th IAPR Internation Conference on Pattern Reocgnition, 1994. Vol. 1-Conference A: Computer Vision & Image Processing., ,* pp. 582-585.

[33] Ding, Y., Zhao, Q., Li, B., & Yuan, X. (2017). Facial expression recognition from image sequence based on lbp and taylor expansion. *IEEE Access, 5*, 19409-19419.

[34] Cai, D., He, X., Han, J., & Zhang, H. J. (2006). Orthogonal laplacianfaces for face recognition. *IEEE Transactions on Image Processing, 15*(11), 3608-3614.

[35] Zhao, Q.-Y., Pan, B.-C., Pan, J.-J., & Tang, Y.-Y. (2008). Facial expression recognition based on fusion of Gabor and LBP features. *International Conference*

on Wavelet Analysis and Pattern Recognition, 2008. ICWPR'08, pp. 362-367.

[36] Lee, T. S. (1996). Image representation using 2D Gabor wavelets. *IEEE Transactions on Pattern Analysis & Machine Intelligence*(10), 959-971.

[37] Nan, Z., & Junmei, G. (2011). The inducement analysis of local facial expression recognition. *2011 International Conference on System Science, Engineering Design and Manufacturing Informatization (ICSEM),* pp. 306-309.

[38] Mohseni, S., Kordy, H. M., & Ahmadi, R. (2013). Facial expression recognition using DCT features and neural network based decision tree. *ELMAR, 2013 55th International Symposium IEEE,* pp. 361-364.

[39] Li, X., Ruan, Q., An, G., & Jin, Y. (2014). Automatic 3D facial expression recognition based on polytypic Local Binary Pattern. *2014 12th International Conference on Signal Processing (ICSP),* pp. 1030-1035.

[40] An, S., & Ruan, Q. (2016). 3D facial expression recognition algorithm using local threshold binary pattern and histogram of oriented gradient. *2016 13th International Conference on Singal Processing (ICSP),* pp. 265-270.

[41] Dalal, N., & Triggs, B. (2005). Histograms of oriented gradients for human detection. *IEEE Conputer Society Conference on Computer Vision and Pattern Recogntion, 2005,* pp. 886-893.

[42] Huang, D., Shan, C., Ardabilian, M., Wang, Y., & Chen, L. (2011). Local binary patterns and its application to facial image analysis: a survey. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 41*(6), 765-781.

[43] Lv, S., Da, F., & Deng, X. (2015). A 3D face recognition method using region-based extended local binary pattern. *2015 IEEE Internation Conference on Image Processing (ICIP),* pp. 3635-3639.

[44] Lee, H., Grosse, R., Ranganath, R., & Ng, A. Y. (2009). Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. *Proceedings of the 26th annual international conference on machine learning,* pp. 609-616.

[45] Susskind, J., Mnih, V., & Hinton, G. (2011). On deep generative models with applications to recognition. *2011 International Conference on Computer Vision and Pattern Recognition (CVPR),* pp. 2857-2864.

[46] Susskind, J. M., Hinton, G. E., Movellan, J. R., & Anderson, A. K. (2008). Generating facial expressions with deep belief nets. In *Affective Computing*: InTech, pp. 421-440.

[47] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *nature, 521*(7553), 436.

[48] Chen, H., Ni, D., Qin, J., Li, S., Yang, X., Wang, T., & Heng, P. A. (2015). Standard plane localization in fetal ultrasound via domain transferred deep neural networks. *IEEE journal of biomedical and health informatics, 19*(5), 1627-1636.

[49] Zhang, T., Zheng, W., Cui, Z., Zong, Y., Yan, J., & Yan, K. (2016). A deep neural network-driven feature learning method for multi-view facial expression recognition. *IEEE Transactions on Multimedia, 18*(12), 2528-2536.

[50] Jung, H., Lee, S., Park, S., Kim, B., Kim, J., Lee, I., & Ahn, C. (2015). Development of deep learning-based facial expression recognition system. *Frontiers of Computer Vision (FCV), 2015 21st Korea-Japan Joint Workshop on IEEE,* pp. 1-4.

[51] Fang, Y., Luo, J., & Lou, C. (2009). Fusion of multi-directional rotation invariant uniform LBP features for face recognition. *Intelligent Information Technology Application,* pp. 332-335.

[52] Qin, J., Zhang, Z., & Wang, Y. (2013). Cross-view action recognition via transductive transfer learning. *Image Proceeding from IEEE International Conference,* pp. 3582-3586.

[53] Patel, D., Hong, X., & Zhao, G. (2016). Selective deep features for micro-expression recognition. *IEEE International Conference on Pattern Recognition,* pp. 2258-2263.

[54] Tong, R., Wang, L., & Ma, B. (2017). Transfer learning for children's speech recognition. *IEEE International Conference on Asian Language Processing,* pp.

36-39.

[55] Pan, S. J., & Yang, Q. (2010). A survey on transfer learning. *IEEE Transactions on knowledge and data engineering, 22*(10), 1345-1359.

[56] Wei, L., Runge, L., & Xiaolei, L. (2018). Traffic sign detection and recognition via transfer learning. *2018 Chinese Control and Decision Conference (CCDC) IEEE,* pp. 5584-5587.

[57] Peng, M., Wu, Z., Zhang, Z., & Chen, T. (2018). From Macro to Micro Expression Recognition: Deep Learning on Small Datasets Using Transfer Learning. *Automatic Face & Gesture Recognition from IEEE International Conference,* pp. 657-661.

[58] Ali, M., Dong, L., Liang, Y., Xu, Z., He, L., & Feng, N. (2014). A color image retrieval system based on weighted average. *IEEE International Conference on Signal Processing, Communications and Computing (ICSPCC),* pp. 184-189.

[59] Schuller, B., Steidl, S., & Batliner, A. (2009). The interspeech 2009 emotion challenge. *Tenth Annual Conference of the International Speech Communication Association,* pp. 312-315.

[60] Zheng, W., Zong, Y., Zhou, X., & Xin, M. (2016). Cross-domain color facial expression recognition using transductive transfer subspace learning. *IEEE Transactions on Affective Computing*, pp. 21-37.

[61] Yu, Z., & Zhang, C. (2015). Image based static facial expression recognition with multiple deep network learning. *Proceeding of the 2015 ACM on International Conference on Multimodal Interaction,* pp. 435-442.

[62] Dhall, A., Goecke, R., Joshi, J., Wagner, M., & Gedeon, T. (2013). Emotion recognition in the wild challenge 2013. *Proceeding of the 15th ACM on International Conference on Multimodal Interaction,* pp. 509-516.

[63] Yao, A., Shao, J., Ma, N., & Chen, Y. (2015). Capturing au-aware facial features and their latent relations for emotion recognition in the wild. *Proceeding of the 2015 ACM on International Conference Multimodal Interaction,* pp. 451-458.

[64] Kahou, S. E., Bouthillier, X., Lamblin, P., Gulcehre, C., Michalski, V., Konda,

K., . . . Boulanger-Lewandowski, N. (2016). Emonets: Multimodal deep learning approaches for emotion recognition in video. *Journal on Multimodal User Interfaces, 10*(2), 99-111.

[65] Ashraf, A. B., Lucey, S., Cohn, J. F., Chen, T., Ambadar, Z., Prkachin, K. M., & Solomon, P. E. (2009). The painful face–pain expression recognition using active appearance models. *Image and Vision Computing, 27*(12), 1788-1796.

[66] Dhall, A., Ramana Murthy, O. V., Goecke, R., Joshi, J., & Gedeon, T. (2015). Video and image based emotion recognition challenges in the wild: Emotiw 2015. *Proceeding of the 15th ACM on International Conference on Multimodal Interaction,* pp. 423-426.

[67] Yan, J., Yan, B., Lu, G., Xu, Q., Li, H., Cheng, X., & Cai, X. (2017). Convolutional neural networks and feature fusion for bimodal emotion recognition on the emotiW 2016 challenge. *IEEE International Conference on Image and Signal Processing, Biomedical Engineering and Informatics (CISP-BMEI),* pp. 1-5.

[68] Eyben, F., Wöllmer, M., & Schuller, B. (2010). Opensmile: the munich versatile and fast open-source audio feature extractor. *Proceeding of the 18th ACM International Conference on Multimedia,* pp. 1459-1462.

[69] Yan, J., Zheng, W., Xu, Q., Lu, G., Li, H., & Wang, B. (2016). Sparse Kernel Reduced-Rank Regression for Bimodal Emotion Recognition From Facial Expression and Speech. *IEEE Trans. Multimedia, 18*(7), 1319-1329.

[70] Wang, Y., Guan, L., & Venetsanopoulos, A. N. (2012). Kernel cross-modal factor analysis for information fusion with application to bimodal emotion recognition. *IEEE Transactions on Multimedia, 14*(3), 597-607.

[71] Wang, X., Chen, D., Yang, T., Hu, B., & Zhang, J. (2016). Action recognition based on object tracking and dense trajectories. *IEEE International Conference on Automatica (ICA-ACCA),* pp. 1-5.

[72] Afshar, S., & Ali Salah, A. (2016). Facial expression recognition in the wild using improved dense trajectories and fisher vector encoding. *Proceeding of the IEEE Conference on Computer Vision and Pattern Recognition Workshops,* pp. 66-74.

[73] Lucey, P., Cohn, J. F., Kanade, T., Saragih, J., Ambadar, Z., & Matthews, I. (2010). The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. *IEEE Computer Conference on Computer Vision and Pattern Recognition Workshops (CVPRW),* pp. 94-101.

[74] Levi, G., & Hassner, T. (2015). Emotion recognition in the wild via convolutional neural networks and mapped binary patterns. *Proceedings of 2015 ACM on international conference on multimodal interaction,* pp. 503-510.

[75] Das, D., & Chakrabarty, A. (2016). Emotion recognition from face dataset using deep neural nets. *Inovations in Intellegent Systems and Applications (INISTA),* pp. 1-6.

[76] Yang, B., Han, X., & Tang, J. (2017). Three class emotions recognition based on deep learning using staked autoencoder. *IEEE International Congress on Image and Signal Processing, BioMedical Engeering and Informatics (CISP-BMEI),* pp. 1-5.

[77] Huang, F. J., & LeCun, Y. (2006). Large-scale learning with svm and convolutional netw for generic object recognition. *IEEE Computer Society Conference on Computer Vision and Pattern Recogntion.*

[79] Adeyanju, I. A., Omidiora, E. O., & Oyedokun, O. F. (2015). Performance evaluation of different support vector machine kernels for face emotion recognition. *SAI Intellgent Systems Conference IEEE,* pp. 804-806.

[80] Zhang, K., Zhang, Z., Li, Z., & Qiao, Y. (2016). Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters, 23*(10), 1499-1503.

[81] Chen, D., Ren, S., Wei, Y., Cao, X., & Sun, J. (2014). Joint cascade face detection and alignment. *European Congerence on Computer Vision,* pp. 109-122.

[82] Ranjan, R., Patel, V. M., & Chellappa, R. (2017). Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and genfer recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* pp. 1-1.

[83] Zhu, Z., Luo, P., Wang, X., & Tang, X. (2013). Deep learning identity-preserving face space. *Proceedings of the IEEE International Conference on Computer Vision,* pp. 113-120.

[84] Chen, H.-Y., Huang, C.-L., & Fu, C.-M. (2008). Hybrid-boost learning for multi-pose face detection and facial expression recognition. *Pattern Recognition, 41(3),* 1173-1185.

[85] Shakyawar, P., Choure, P., & Singh, U. (2017, April). Eigenface method through through facial expression recognition. *IEEE International Conference on Electronics, Communication and Aerospace Technology (ICECA),* pp. 500-505.

[86] Singh, M., Majumder, A., & Behera, L. (2014, July). Facial expressions recognition system using Bayesian inference. *Neural Networks (IJCNN), 2014 International Joint Conference on IEEE,* pp. 1502-1509.

[87] Kumar, A., & Agarwal, A. (2014, December). Emotion recognition using anatomical information in facial expressions. *IEEE International Conference on Industrial and Information Systems (ICIIS),* pp. 1-6.

[88] Wallhoff, F., Muller, S., & Rigoll, G. (2001). Recognition of face profiles from the MUGSHOT database using a hybrid connectionist/HMM approach. *IEEE International Conference on Acoustics, Speech, and Signal Processing,* pp. 1489-1492.

[89] Choi, H. C., & Oh, S. Y. (2006, October). Realtime facial expression recognition using active appearance model and multilayer perceptron. *SICE-ICASE, 2006. International Joint Conference*, pp. 5924-5927.

[90] Gray, M. (2003). Urban Surveillance and Panopticism: will we recognize the facial recognition society?. *Surveillance & Society*, *1*(3), 314-330.

[91] Milligan, C. S. (1999). Facial Recognition Technology, Video Surveillance, and Privacy. *S. Cal. Interdisc. LJ*, *9*, 295.

[92] Jiang, H., & Learned-Miller, E. (2017, May). Face detection with the faster R-CNN. In *2017 12th IEEE International Conference on Automatic Face &*

Gesture Recognition, pp. 650-657.

[93] Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems* pp. 91-99.

[94] Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* pp. 580-587.

[95] Ren, S., He, K., Girshick, R., & Sun, J. (2017). Faster R-CNN: towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (6), 1137-1149.

[96] Sun, X., Wu, P., & Hoi, S. C. (2018). Face detection using deep learning: An improved faster RCNN approach. *Neurocomputing*, *299*, 42-50.

[97] He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017, October). Mask r-cnn. In *Computer Vision (ICCV), 2017 IEEE International Conference on* (pp. 2980-2988). IEEE.