ADAPTIVE FEATURE MAPPING FOR CUSTOMIZING DEEP LEARNING BASED FACIAL EXPRESSION RECOGNITION MODEL

Seminar Report

Submitted in partial fulfillment of the requirements for the award of degree of

BACHELOR OF TECHNOLOGY

In

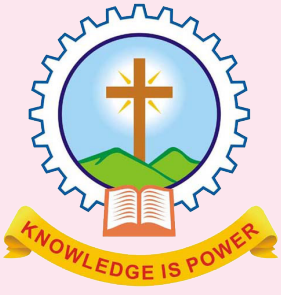
COMPUTER SCIENCE AND ENGINEERING

of

APJ ABDUL KALAM TECHNOLOGICAL UNIVERSITY

Submitted By

PRANAV PK



Department of Computer Science & Engineering
Mar Athanasius College Of Engineering Kothamangalam

ADAPTIVE FEATURE MAPPING FOR CUSTOMIZING DEEP LEARNING BASED FACIAL EXPRESSION RECOGNITION MODEL

Seminar Report

Submitted in partial fulfillment of the requirements for the award of degree of

BACHELOR OF TECHNOLOGY

In

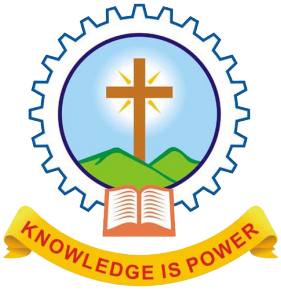
COMPUTER SCIENCE AND ENGINEERING

of

APJ ABDUL KALAM TECHNOLOGICAL UNIVERSITY

Submitted By

PRANAV PK



Department of Computer Science & Engineering
Mar Athanasius College Of Engineering Kothamangalam

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
MAR ATHANASIUS COLLEGE OF ENGINEERING
KOTHAMANGALAM

CERTIFICATE

This is to certify that the report entitled **Adaptive feature mapping for customizing deep learning based facial expression recognition model** *submitted by* **Mr. PRANAV PK, Reg. No.LMAC15CS064** *towards partial fulfillment of the requirement for the award of Degree of Bachelor of Technology in Computer science and Engineering from APJ Abdul Kalam Technological University for December 2018 is a bonafide record of the seminar carried out by him under our supervision and guidance.*

.....	.....	.....
Prof. Joby George	**Prof. Neethu Subash**	**Dr. Surekha Mariam Varghese**
Faculty Guide	***Faculty Guide***	***Head Of Department***

Date:

Dept. Seal

ACKNOWLEDGEMENT

First and foremost, I sincerely thank the 'God Almighty' for his grace for the successful and timely completion of the seminar.

I express my sincere gratitude and thanks to Dr. Solly George, Principal and Dr. Surekha Mariam Varghese, Head Of the Department for providing the necessary facilities and their encouragement and support.

I owe special thanks to the staff-in-charge Prof. Joby george, Prof. Neethu Subash and Prof. Joby Anu Mathew for their corrections, suggestions and sincere efforts to co-ordinate the seminar under a tight schedule.

I express my sincere thanks to staff members in the Department of Computer Science and Engineering who have taken sincere efforts in helping me to conduct this seminar.

Finally, I would like to acknowledge the heartfelt efforts, comments, criticisms, co-operation and tremendous support given to me by my dear friends during the preparation of the seminar and also during the presentation without whose support this work would have been all the more difficult to accomplish.

ABSTRACT

Facial expression recognition plays a vital role in the artificial intelligence era. Automated facial expression recognition can greatly improve the human-machine interface. The machine can provide better and more personalized services when it knows the human's emotion. This kind of improvement is an important progress in this artificial intelligence era. Many deep learning approaches have been applied in recent years due to their outstanding recognition accuracy after training with large amounts of data. The performance is limited, however, by the specific environmental conditions and variations in different persons involved.Weighted Center Regression Adaptive Feature Mapping (W-CR-AFM) is mainly proposed to transform the feature distribution of testing samples into that of trained samples. By means of minimizing the error between each feature of testing sample and the center of the most relevant category, W-CR-AFM can bring the features of testing samples around the decision boundary to the centers of expression categories. Therefore, their predicted labels can be corrected. Compared to the competing deep learning architectures with the same training data W-CR-AFM approach shows the better performance.

Contents

List of Figures

List of Tables

List of abbreviation

AFM	Adaptive feature Mapping
WAFM	Weighted Adaptive feature Mapping
WCRAFM	Weighted Center Regression Adaptive feature Mapping
Rafd	Radboud Faces Database
CNN	Convolutional Neural network

Introduction

Facial expression recognition plays a vital role in the artificial intelligence era. According to the human's emotion information, machines can provide personalized services. Many applications, such as virtual reality, personalized recommendations, customer satisfaction, and so on, depend on an efficient and reliable way to recognize the facial expressions. This topic has attracted many researchers for years, but it is still a challenging topic since expression features vary greatly with the head poses, environments, and variations in the different persons involved.To mitigate these variations, some approaches modified the handcrafted features to gain the better performance, like and.Qirong Mao et al.make a Bayesian model by means of multiple head poses to conquer the feature variation caused by head poses. However, the handcrafted features have shown their limitations in practical applications, so deep learning methods are utilized to make the models learn to extract the complicated features from large amounts of facial expression data.Most of the standard database for facial expression recognition are not candid since they are built under the controlled environment with coached expressions. Therefore,apply data mining technique to search for the facial images on the internet to make the model more realistic. For deep learning neural networks, there is no clear rule to determine the architecture and learning parameters, so image pre-processing is often adopted to improve the neural network's performance. Andre et al. apply the spatial normalization, local intensity normalization, and facial image cropping to the Convolutional Neural Network (CNN). Mapped binary pattern method is utilized in[1]. They all have the betterresult after applying the preprocessing. In addition, some other approaches combine common machine learning models to gain the robustness and higher performance. Duc et al. take the second-to-last output layer as the encoded features, and utilize Support Vector Machine (SVM) to be the label predictor. Dennis et al. propose a 2-channel CNN, and the first convolutional layer in one of the channels is trained by Convolutional Auto-Encoder (CAE) to learn the better capability in order to extract better features. In order to mitigate the effect of head pose, a CNN learns the poserobust

features by regressing the features extracted from the Principal Component Analysis Network (PCANet) which has been trained by the frontal facial images with various expressions. Different from traditional learning algorithms in CNN, the model in learns the correlations among the training data. To mitigate the person-specific differences, Zibo Meng et al. and Chongsheng Zhang et al. propose a way to train an identity-aware structure to extract the person-specific features for recognizing the facial expressions. Rather than trying to recognize a single image, predict the expressions by passing a video, which seems more reasonable in practice; nevertheless, labeling video data is more labor intensive.

Facial expression database

This section addresses the usage of the facial expression database and the process of preparing the training and testing data. Only seven common facial expressions, anger, disgust, fear, happiness, sadness, surprise, and neutral, are considered in this paper. Other expressions are ignored even if they are collected in the public domain database.

Facial expression recognition is a challenging and interesting problem in computer vision and pattern recognition. Geometric variability in both emotion expression and neutral face is a fundamental challenge in facial expression recognition problem. This variability not only directly affects geometric facial expression recognition methods, but also is a critical problem in appearance methods. To overcome this problem, this paper presents an approach which eliminates geometric variability in emotion expression; thus, appearance features can be accurately used for facial expression recognition. Therefore, a fixed geometric model is used for geometric normalization of facial images. This model is defined as one of the emotional expressions. In addition Local Binary Patterns are utilized to represent facial appearance features. Experimental results show that the proposed method is more accurate than the existing works. Also for facial expression recognition, using geometric expression models of facial images where they have larger size in mouth/eyes regions, such as Surprise, gives better results indicating that mouth and eyes are important regions in emotion expression.

2.1 Extended cohn-kanade

Wilhelm Von Rosenberg[2] Proposed a smart helmet with embedded sensors for cycling and moto racing drivers. The helmet is fitted with special electrodes within the helmet padding to detect possible impact occured from any accidents. A respiration belt around the thorax area is used for continuous validation of heartbeat, referencing ECG from the chest. Also a multivariate R-peak detection algorithm tailored specially for optimising data acquisition and noise reduction in real time.

Figure 2.1: Extended cohn-kanade

2.2 Radboud faces database

The Radboud Faces Database (RaFD), [3]is a high quality database of faces.Which contains pictures of 8 emotional expressions, including Caucasian males and females,Caucasian children, both boys and girls, and Moroccan Dutch males. Head poses vary from left side to right, and each pose is shot with three eye gazing directions. Compared to CK+, RaFD is more challenging to the recognition model.

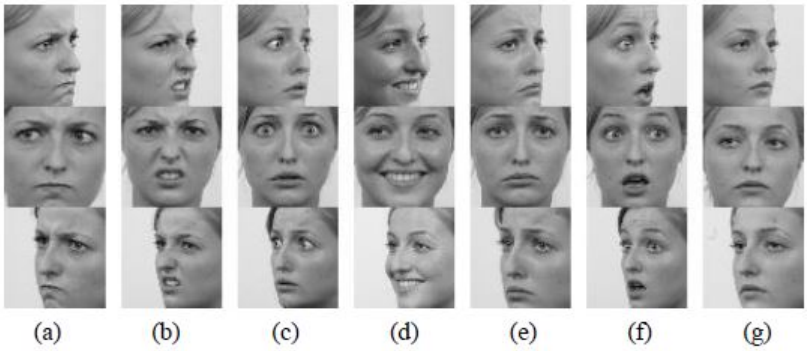


Figure 2.2: Radboud faces database

2.3 Amsterdam dynamic facial expression Set

Around 10 emotional expressions are collected in the, Amsterdam Dynamic Facial Expression Set (ADFES).[4]Most of them are videos with head pose variations, and the expression

intensity also changes from low to high, like in CK+. The facial images are captured with fixed time steps when the expressions start to become obvious.

Figure 2.3: Amsterdam dynamic facial fxpression set

2.4 Proprietary database

To make the deep model more robust and general[5], a homegrown/ proprietary database is built to train the model. 372 videos are downloaded from YouTube, including movies, film reviews, variety shows, and some short videos.After that, a face detection method proposed by D. E. King, is employed to capture the face images with the time intervals set to 1, 2, or 3 second(s) to avoid repeating the images with similar expressions for one person. Then, 100,000 facial images are produced. Only the images that represent their corresponding categories are manually picked to be the training and testing samples. This database ended up with 17,655 images.

Figure 2.4: Propreitary database

2.4.1 Training and testing data rearrangement

Since CK+ is a well-known benchmark in facial expression recognition and the number of images is small, it will not be placed in the training set but instead in the testing set only in order to objectively show the performance of the proposed approach. Out of RaFD and ADFES, 10 and 4 persons images are chosen to be the testing data respectively, therefore, people in the training and the testing sets are definitely different. Altogether, the number of testing images is 630 from CK+, 616 from RaFD, and 562 from ADFES while the number of training data is 23,591 including 17,655 from the proprietary database, 3,377 from RaFD and 2,559 from ADFES.The fundamental goal of ML is to generalize beyond the data instances used to train models. We want to evaluate the model to estimate the quality of its pattern generalization for data the model has not been trained on. However, because future instances have unknown target values and we cannot check the accuracy of our predictions for future instances now, we need to use some of the data that we already know the answer for as a proxy for future data. Evaluating the model with the same data that was used for training is not useful, because it rewards models that can "remember" the training data, as opposed to generalizing from it.

To balance the image count throughout all categories, the category with less images is

supplemented by copying the randomly selected images before combining the training data. In this way, the total number of images in every category will be the same.

Researches show that if the data is augmented in a reasonable way[6], the model can perform much better. Thus, the training set is mirrored and also augmented by two Gamma transformation, three Gaussian blur, and three sharpening filter, so one image is extended to 42 images. As a result, the total number of training data is increased to 2,315,544, and the resolution is set to 6 4 * 6 4 pixels in grayscales.

Proposed system

Image processing is a method to perform some operations on an image, in order to get an enhanced image or to extract some useful information from it. It is a type of signal processing in which input is an image and output may be image or characteristics/features associated with that image. Nowadays, image processing is among rapidly growing technologies. It forms core research area within engineering and computer science disciplines too.Previous researches, and, have shown that if the image is pre-processed appropriately, the recognition performance can be improved.There are two types of methods used for image processing namely, analogue and digital image processing. Analogue image processing can be used for the hard copies like printouts and photographs. Image analysts use various fundamentals of interpretation while using these visual techniques. Digital image processing techniques help in manipulation of the digital images by using computers. The three general phases that all types of data have to undergo while using digital technique are pre-processing, enhancement, and display, information extraction. Here a proposed pre-processing method which contains spatial normalization and feature enhancement is introduced.

3.1 Spatial normalization

The purpose of spatial normalization is to adjust the alignment of the position and rotation angle of the detected facial images. An example is shown in Fig. 3.1, A face alignment algorithm is utilized to detect some landmarks on the face. The tip of the nose will be shifted to the center of the image so that the placement offset can be mitigated.

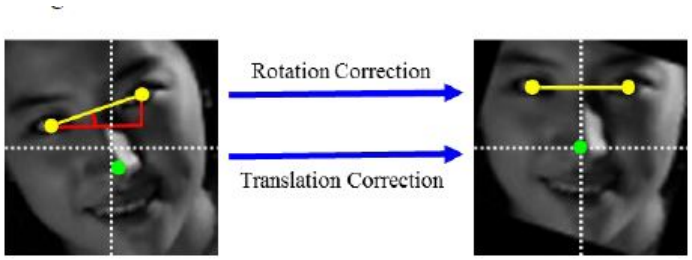


Figure 3.1: Spatial Normalization

3.1.1 Face alignment algorithm

A novel boundary-aware face alignment algorithm by utilising boundary lines as the geometric structure of a human face to help facial landmark localisation. Unlike the conventional heatmap based method and regression based method, this approach derives face landmarks from boundary lines which remove the ambiguities in the landmark definition. Three questions are explored and answered by this work: 1. Why using boundary? 2. How to use boundary? 3. What is the relationship between boundary estimation and landmarks localisation? Here boundary-aware face alignment algorithm achieves 3.49percentage mean error on 300-W Fullset, which outperforms state-of-the-art methods by a large margin.This method can also easily integrate information from other datasets. By utilising boundary information of 300-W dataset, our method achieves 3.92percentage mean error with 0.39percentage failure rate on COFW dataset, and 1.25percentage mean error on AFLW-Full dataset. Moreover, we propose a new dataset WFLW to unify training and testing across different factors, including poses, expressions, illuminations, makeups, occlusions, and blurriness.

3.2 Feature enhancement

Local Binary Pattern (LBP) may be an efficient way to extract the features from images. Nonetheless, it may lose a lot of intrinsic information. Jiwen et al.[7]Try to fix this problem by finding a mapping from the Neighbor-Center Difference Vector (NCDV) into the binary space so that the patterns can better represent the images in the original database, but it needs more computing effort. Neighbor-Center Difference Image (NCDI) is presented to enhance the edges

efficiently and retain the original information. The concept is the same as NCDV. NCDIs are extracted by subtracting the center pixel from the neighboring pixels, so the pixel values fall in the range from 255 to 255 . An NCDI collects the subtraction results of the selected channel from all patches to reconstruct the image. Thus, eight images which have been sharpened in eight different directions are produced if the 8-channel NCDI is applied.

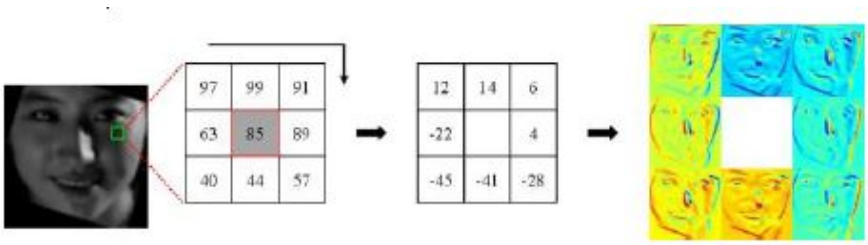


Figure 3.2: feature Enhancement

After enhancing the edges, the facial contour and background become sharper, but they have nothing to do with facial expressions. Hence, facial image cropping should be applied. Since the facial contour is often confused with the background, the detected landmarks may drift between the facial contour and background.

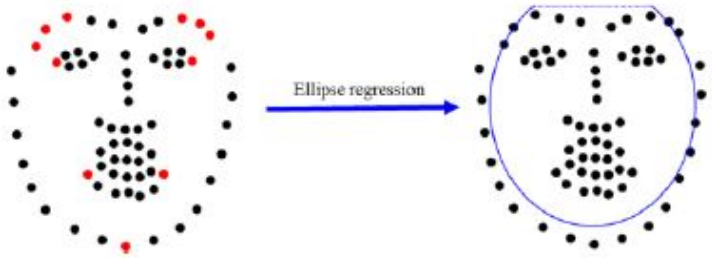


Figure 3.3: Landmarking

It is not recommended to crop the facial image by connecting the landmarks, which is considered as polygon cropping. Except for the contour, other landmarks are more stable. An elliptical region which regresses the suitable landmarks, as shown in Fig. 3.3, is the better way to crop the facial image effectively. The ellipse function is

$$f(x, y|\mathbf{a}) = a_1 x^2 + a_2 xy + a_3 y^2 + a_4 x + a_5 y + a_6 ,$$

(Equ: 1)

where a = a1,a2,a3,a5,a6.

x and y are the selected positions of the landmarks, N is the sample number, and alpha is the hyper parameter used to regularize the optimization. By setting the gradient of the cost function to zero, the equation becomes to

$$(\mathbf{D}-\mathbf{\Psi})\,\mathbf{a}=\mathbf{\Lambda},$$

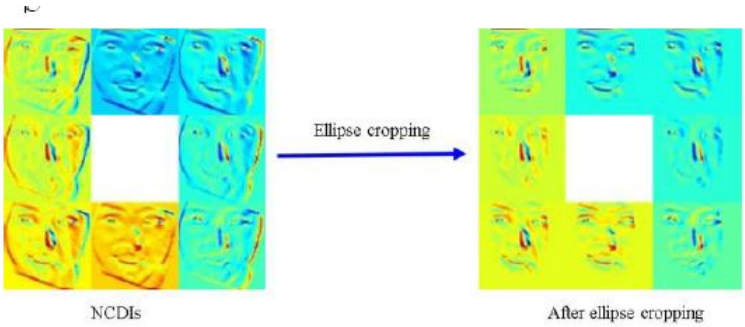
$$\mathbf{D}=\sum_{i=1}^{N}\begin{bmatrix} x_i^4 & x_i^3y_i & x_i^2y_i^2 & x_i^3 & x_i^2y_i & x_i^2 \\ x_i^3y_i & x_i^2y_i^2 & x_iy_i^3 & x_i^2y_i & x_iy_i^2 & x_iy_i \\ x_i^2y_i^2 & x_iy_i^3 & y_i^4 & x_iy_i^2 & y_i^3 & y_i^2 \\ x_i^3 & x_i^2y_i & x_iy_i^2 & x_i^2 & x_iy_i & x_i \\ x_i^2y_i & x_iy_i^2 & y_i^3 & x_iy_i & y_i^2 & y_i \\ x_i^2 & x_iy_i & y_i^2 & x_i & y_i & 1 \end{bmatrix},$$

$$\mathbf{\Psi}=\begin{bmatrix} 0 & 0 & \frac{2\delta}{N} & 0 & 0 & 0 \\ 0 & -\frac{\delta}{N} & 0 & 0 & 0 & 0 \\ \frac{2\delta}{N} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix},$$

$$\mathbf{\Lambda}=\begin{bmatrix} \frac{N\delta}{2} & 0 & \frac{N\delta}{2} & 0 & 0 & 0 \end{bmatrix}^T,$$

(Equ: 2)

Only the pixels in the ellipse and those lower than the highest landmarks of the eyebrows are kept, as Fig. 3.3 shows. There are some differences between ellipse cropping and polygon cropping.



The pre-processing procedure follows the steps below. Detect the face and find the bounding box. Resize the facial image into 6 4 * 6 4 pixels. Then, extract the landmarks on the face, and perform the spatial normalization and feature enhancement last.

3.3 Deep convolutional neural network

Based on Caffe framework, a CNN model is designed from the concept of and. There are parallel structures in the network to extract the features using different sizes of windows. The model consists of nine convolutional layers, two max pooling layers, one mean pooling layer, [8]three fully connected layers, and a Local Response Normalization (LRN) layer. The activation functions are all set to rectified linear functions.

The output of Full connection layer is regarded as the encoded features. The combination of Full connection layer and Softmax output layer is regarded as a classifier. Except for the classifier, the whole structure is a feature extractor for the input image. The Convolutional Feature Extractor (CFE) is defined as from Convolution layer to Mean pooling layer while the Fully Connected Feature Extractor (FCFE) is defined as from Full connection layer to Full connection layer.

To learn about thousands of objects from millions of images, we need a model with a large learning capacity. However, the immense complexity of the object recognition task means that this problem cannot be specified even by a dataset as large as ImageNet, so our model should also have lots of prior knowledge to compensate for all the data we don't have. Convolutional neural networks (CNNs) constitute one such class of models. Their capacity can

be controlled by varying their depth and breadth, and they also make strong and mostly correct assumptions about the nature of images (namely, stationarity of statistics and locality of pixel dependencies). Thus, compared to standard feedforward neural networks with similarly-sized layers, CNNs have much fewer connections and parameters and so they are easier to train, while their theoretically-best performance is likely to be only slightly worse. ep convolutional neural network is capable of achieving recordbreaking results on a highly challenging dataset using purely supervised learning. It is notable that our network's performance degrades if a single convolutional layer is removed. For example, removing any of the middle layers results in a loss of about 2 percentage for the top-1 performance of the network. So the depth really is important for achieving our results.

3.4 Adaptive feature mapping

This section addresses the design principle and the mechanism of AFM. In the following description, the trained and testing data sets.

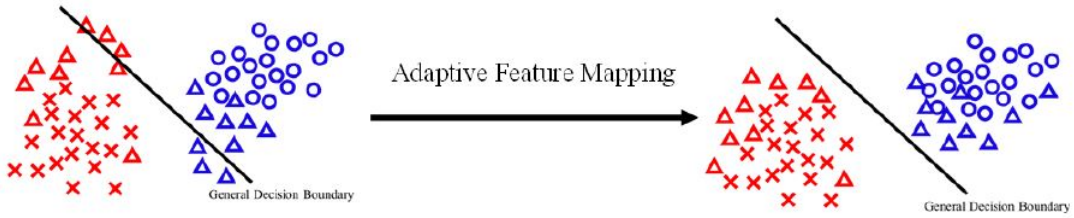


Figure 3.4: The purpose of adaptive feature mapping.

The trained and testing sets are denoted as Xs and Xt.and s Ns is the number of trained samples while t Nt is the batch size of the testing samples. The feature extractor consists of CFE and FCFE, and it is denoted as h(x—W). W is the parameter set of the whole feature extractor while x is the input sample. Here, x is the 8-channel NCDIs.

3.4.1 Cost function

The main purpose of AFM is to tune the parameters of the feature extractor for the testing samples so that the tuned feature extractor can make the feature distribution of the testing samples similar to that of the trained samples.That is,p(h(Xt—W‘))=p(h(Xs—W)), where W is the generic parameter set, and W‘ is the new parameter set for the testing samples. To accomplish this, the discrepancy between the means of the trained and testing samples must be minimized.Then the cost function can be written as

$$E\left(\tilde{\mathbf{W}}\left|\mathbf{X}^{t},\mathbf{X}^{s}\right.\right)=\left\|\frac{1}{N_{t}}\sum_{i=1}^{N_{t}}\varphi\left(\mathbf{h}\left(\mathbf{x}_{i}^{t}\left|\tilde{\mathbf{W}}\right.\right)\right)-\frac{1}{N_{s}}\sum_{j=1}^{N_{s}}\varphi\left(\mathbf{h}\left(\mathbf{x}_{j}^{s}\left|\mathbf{W}\right.\right)\right)\right\|_{\mathrm{H}}^{2},$$

(Equ: 3)

Xs=Testing Set,Xt=Trained Set

Where H stands for RKHS which can be defined by a kernel.Since we are only concerned with the cross relations between the features of trained and testing samples, other terms can be eliminated. Therefore,][9] the cost function can be modified as

$$E\left(\tilde{\mathbf{W}}\left|\mathbf{X}^{t},\mathbf{X}^{s}\right.\right)=\sum_{i=1}^{N_{t}}\sum_{j=1}^{N_{s}}\frac{\alpha_{i,j}}{N_{s}N_{t}}\left\|\mathbf{h}\left(\mathbf{x}_{i}^{t}\left|\tilde{\mathbf{W}}\right.\right)-\mathbf{h}\left(\mathbf{x}_{j}^{s}\left|\mathbf{W}\right.\right)\right\|^{2},$$

(Equ: 4)

Xs=Testing Set,Xt=Trained Set

Where alpha are the weight to represent the relevance between h(Xt—W‘) and h(Xs—W).A large alpha value is required when the features of trained and testing samples are relevant, i.e. error becomes small. On the contrary, a smaller alpha value is required when the features of trained and testing samples are less relevant, i.e. error becomes larger. If alpha is not

appropriate, the cost function will oscillate drastically and may not converge during the training process. Moreover, the computational complexity will increase as the number of trained samples increases.

Since some bad samples may limit the performance of AFM, the probability distribution on the output of the classifier can be taken into consideration to regularize the cost function. By simply multiplying the prediction confidence, the cost function can be written as

$$E\left(\tilde{\mathbf{W}} \middle| \mathbf{X}^t, \mathbf{X}^s, \mathbf{y}^s\right) = \frac{1}{2N_t}\sum_{i=1}^{N_t} f_{y_r^s}\left(\mathbf{h}\left(\mathbf{x}_r^s \middle| \mathbf{W}\right)\right) \cdot \left\| \mathbf{h}\left(\mathbf{x}_i^t \middle| \tilde{\mathbf{W}}\right) - \mathbf{h}\left(\mathbf{x}_r^s \middle| \mathbf{W}\right) \right\|^2 ,$$

(Equ: 5)

Xs=Testing Set,Xt=Trained Set

where the entries of the Ys are the trained samples,f() is the classifier,and Nk is the number of categories.This form of AFM is defined as Weighted Adaptive Feature Mapping(W-AFM).

The classifier in CNN is a linear transformation. After a CNN model is well trained, the features extracted by the model must be linearly separable. Therefore, there must be a unique center in each category. If the centers of the categories are considered, the person-specific bias can be mitigated further. The cost function, then, can be modified as

$$E\left(\tilde{\mathbf{W}} \middle| \mathbf{X}^t, \mathbf{X}^s, \mathbf{y}^s\right) = \frac{1}{2N_t}\sum_{i=1}^{N_t} f_{y_r^s}\left(\mathbf{h}\left(\mathbf{x}_r^s \middle| \mathbf{W}\right)\right) \cdot \left\| \mathbf{h}\left(\mathbf{x}_i^t \middle| \tilde{\mathbf{W}}\right) - \mathbf{h}_{y_r^s}^c \right\|^2 ,$$

(Equ: 6)

Xs=Testing Set,Xt=Trained Set

such form of AFM is regarded as Weighted Center Regression Adaptive Feature Mapping (W-CR-AFM).

3.4.2 System operation

After training the CNN model, the extracted features of training samples shall be stored as the feature database. In the testing phase, AFM can tune the weights based on the relationship between features of testing samples and the feature database in order to transform the features of testing samples into a new space so that its distribution can be similar to that of the feature database. Most of the parameters are distributed in the fully connected layers, so AFM is only applied to tune FCFE for higher efficiency.The premise of AFM is that the feature distribution of the testing samples is assumed to be similar to that of the training samples. The features around the decision boundary, therefore, shall be moved to the centers of categories. This way, the misclassified labels can be corrected. In addition, misclassified trained samples must be removed in advance so that the newly mapped features can be better. To make it more reliable, the testing samples with lower confidence of prediction can be ignored.

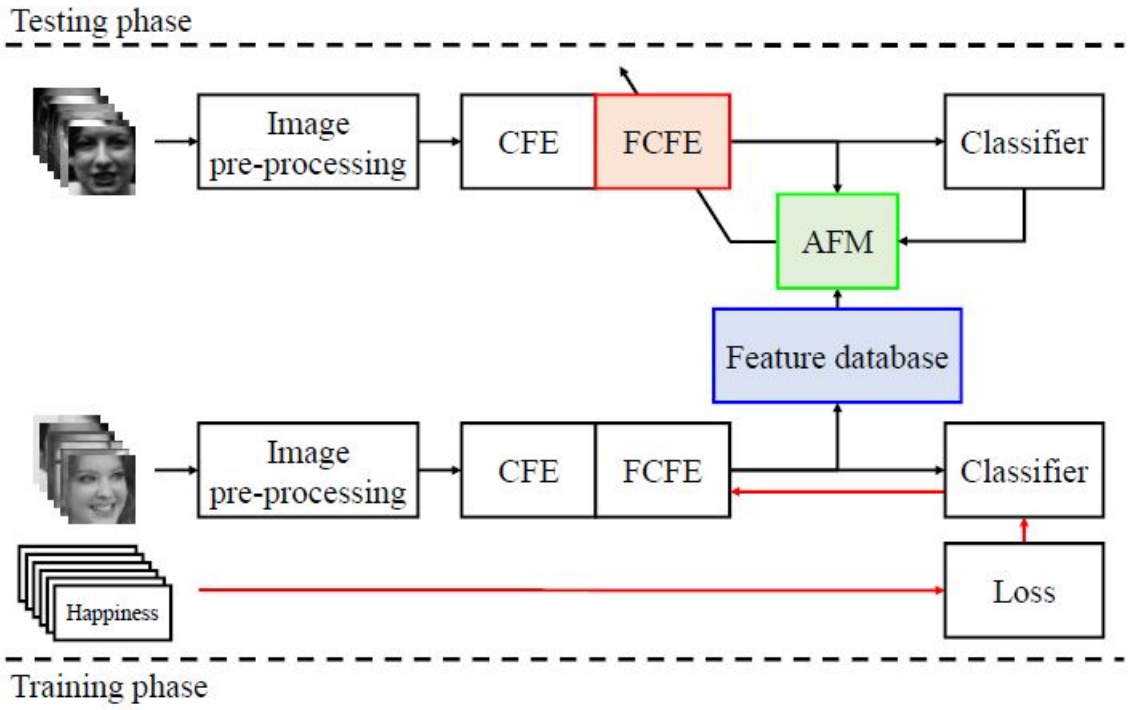


Figure 3.5: System architecture.

Performance analysis

4.1 Image pre-processing comparison

TABLE 4.1 shows the results of the proposed model with different pre-processing methods. As can be seen in TABLE 4.1, the spatial normalization does not always seem to help the recognition accuracy since the edges of bounding box may appear and become the main feature of the image after spatial normalization is applied, which impairs the recognition function, causing the accuracy to be lower. Also, the model has been trained with many candid images from YouTube,so it can extract some features that are not affected by rotation. Thus, the recognition accuracy can be higher than when spatial normalization is applied. These may be the reasons why the spatial normalization appears ineffective in Table.

IMAGE PRE-PROCESSING COMPARISON

Pre-processing	Accuracy (%)		
	CK+	RaFD	ADFES
None	82.22	90.26	85.59
SN	81.75	94.16	83.10
SN + FE	**86.83**	**95.78**	**87.37**

SN stands for spatial normalization while FE stands for feature enhancement.

The feature enhancement operation not only makes the facial edges more distinct but also removes the areas that are irrelevant to facial expressions, so the accuracy can be increased by about 4.61percentage in CK+, 5.52percentage in RaFD, and 1.78percentage in ADFES. These

Table 4.1: Process Comparison

results demonstrate that the proposed pre-processing method is really effective.

4.2 The effect on adaptive feature mapping

The proposed CNN model with spatial normalization and feature enhancement is regarded as our Generic Model (GM). As for AFM, the learning rate is set to 0.001 while the regularizing factor is set to 0.0005. The training iteration is set to 1000. The batch size ranges from 16 to 512. The trained samples and testing samples are mirrored, and the trained samples which are misclassified should be removed in advance while the testing samples whose prediction confidence is lower than 90percentage are not taken into consideration when tuning the model with AFM.In most cases, WAFM performs better than AFM, and W-CR-AFM is the best of all. The performance will be more stable when the batch size is large enough, otherwise the effect may be limited.

According to the experiments shown in, the best result of each AFM is listed. The category of happiness can always be predicted correctly in these three databases because its feature is obvious and its training data is ample. After applying AFM, most predicted labels are corrected. Compared to other categories, the number of images in anger, disgust and fear is less, so the ability of extracting features for these expressions is poor; hence, most features of testing samples do not fall into the center of the category, and will be drawn to other categories. Besides, the expression of anger is usually not explicit, so it is sometimes confused with neutral expression even if AFM or W-AFM is utilized. The main feature of surprise is the exaggerated mouth while the minor feature is the eyes, but the feature of eyes is difficult to extract properly because it varies greatly with the person. Hamid et al. have proven that the mouth is the primary feature for facial expressions. However, some surprised faces do not clearly express the feature on the mouth in ADFES, so they are misclassified into neutral expression if AFM or W-AFM is applied.

For W-CR-AFM, since it minimizes the distance between the feature of the testing sample and the center of the most relevant category rather than the most relevant feature of the trained sample, the person-specific bias can be mitigated much more. Moreover, the feature

distribution of neutral expression contains the largest area in the feature space, so the features of neutral expression that are far away from the center of neutral expression category will be brought to other categories. That is why the recognition accuracy of neutral expression is reduced after applying WCR- AFM while accuracy of other expressions are raised. According to the experiments, all three types of AFM can assist in improving the performance of a model in specific cases. For the overall recognition accuracy, W-CRAFM works the best.

4.3 Benchmark comparison

Some other deep learning approaches in facial expression recognition are introduced and compared with ours. They are trained with our training data for fair comparison.To make GoogLeNet and AlexNet perform better, they have been trained with ImageNet previously. The second-to-last layer of the trained AlexNet is utilized to train a SVM. To present the original performance of the competing models, the architectures and training parameters are set based on the original works. Since CK+ is not included in the training data, it is reasonable that the recognition accuracy in TABLE V is lower than the results of state-of-the-arts. If the models are trained by CK+ [2], the recognition accuracy is expected to be much higher. The results in TABLE V show that our approach performs better than others. The parameter quantities of GoogLeNet, AlexNet and the CNN designed by Heechul et al. [5] are around 40MB, 222MB, and 5MB respectively. Although our parameter quantity, around 3.5MB, which is much lower than others, the performance can be comparable to these state-of-the-art architectures through the use of proposed pre-processing method. Besides, AFMs can adapt the testing samples so that the model can perform better than other approaches.

Conclusion

Feature selection is an important preprocessing method in machine learning and data mining.Traditional feature selection methods evaluate the dependency and redundancy of features separately, which leads to a lack of measurement of their combined effect. . This procedure overcomes the hard constraint on the number of features, enables the combined evaluation of each subset as a whole, and improves the search ability of conventional binary particle swarm optimization.The technology that is described has two main contributions are, One contribution is that the proposed pre-processing method can assist the CNN model to gain the higher accuracy rate in the applications of facial image processing. The other contribution is that three types of AFMs can reformulate the features of new samples which do not have label information so that some misclassified samples can be corrected, which means it can tune a generic model to adapt to a specific condition. Moreover, AFMs can be deployed to real-time systems since it learns batch by batch rather than calculating all of the training and testing data in one batch. The concept drift problem is restrained because AFMs map the features of the testing samples to a static feature distribution. With the pre-processing and AFMs, a light CNN can outperform the state-of-the-art architectures.

References

[1] R.E.Jack,O.G.B.Garrod, H. Yu, R. Caldara, and P. G. Schyns, "Facial expressions of emotion are not culturally universal," Proc. National Academy of Sci. of the United States of America, vol. 109, no. 19, pp. 7241–7244, 2012.

[2] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended cohn-kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression,", 2010.

[3] O. Langner, R. Dotsch, S. T. Hawk, and A. van Knippenberg, "Presentation and validation of the Radboud Faces Database," Cognition and Emotion, vol. 24, no. 8, pp. 1377–1388, 2010.

[4] J. van der Schalk, S. T. Hawk, A. H. Fischer, and B. J. Doosje, "Moving faces, looking places: The Amsterdam Dynamic Facial Expressions Set (ADFES)," Emotion, vol. 11,pp. 907–920, 2011.

[5] W. Li, M. Li, and Z. Su, "A deep-learning approach to facial expression recognition with candid images," MVA2015 IAPR Int. Conf. Mach. Vision Appl., May 18–22, 2015.

[6] A. Mollahosseini, D. Chan, and M. H. Mahoor, "Going deeper in facial expression recognition using deep neural networks," IEEE Winter Conf. Appl. of Comput. Vision (WACV), pp. 1–10, 2016.

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

[8] A. Mollahosseini, B. Hassani, M. J. Salvador, "Facial expression recognition from world wild web," IEEE Conf. on Comput. Vision and Pattern Recognition (CVPR) Workshops, pp. 59–65, 2016.