UNIVERSITY
OF WOLLONGONG
AUSTRALIA

Facial Expression Recognition of Archival Photographs

Afrida Tabassum
Student ID: 6015906
at057@uowmail.edu.au

November, 2019

This thesis is presented as part of the requirements for the conferral of the degree:

Master of Computer Science

Supervised by

Associate Professor Lei Wang
University of Wollongong
School of Computing and Information Technology

Abstract

As deep learning is relatively new attempt, past research on facial expressions have used datasets with modern photography which makes it unclear whether current methods can be employed in historical or archival photos. Archival photo expressions are very subtle, professional, pixel quality is different opposed to the existing datasets. In this paper, we present a novel dataset, National Archives Australia Expression Dataset(NAAE-DB), which contains about 5775 facial images from thousands of individuals. Each image has been individually labeled and unreliable labels were filtered out. To the best of my knowledge, the NAAE-DB is the first medium-scale dataset providing the labels of common expression perception of archival photos in unconstrained environment. At present, deep facial expression recognition faces two key problems: overfitting due to lack of sufficient training data and facial expression changes are subtle and changeable, the first-order information is insufficient to provide more discriminant information. To overcome these issues, this paper demonstrated that data augmentation and second-order pooling based deep convolutional neural networks for facial expression recognition outperforms conventional methods. The NAAE-DB benchmark experiments on the three class basic expressions (Happy, Concerned, Neutral), as well as the additional experiments on RAFDB and JAFFE databases, shows state-of-the-art performance (86.7%) compared to the handcrafted features and deep learning based methods for the expression recognition of archival photos in the wild.

Declaration

I, Afrida Tabassum, declare that this thesis is submitted in partial fulfilment of the requirements for the conferral of the degree Master of Computer Science, from the University of Wollongong, is completely my own work unless otherwise referenced or acknowledged. This document has not been submitted for qualifications at any other academic institution.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

Generally, archival photos crucially differ in resolution, lighting, pixels, poses etc. Past research on facial expressions haven't explored how the methods perform in historical or archival photographs which have very subtle, professional expressions, pixel quality is different opposed to the existing datasets. In this paper, our novel dataset, National Archives Australia Expression Dataset(NAAE-DB), a subset of NAA 300k dataset, has been leveraged individually labeled and unreliable labels were filtered out *National Archives of Australia* (n.d.). To the best of my knowledge, the NAAE-DB is the first medium-scale dataset providing the labels of common expression perception of archival photos in unconstrained environment.

Techniques to represent facial features for expression recognition are broadly divided into two types: Geometric based method and appearance based method. Geometric features hold location and shape information of those features such as eyes, mouth, nose etc., whereas, appearance features determine the appearance change of face like wrinkle, bulges and furrows. These together are called regional features. In this presented work, regional features are used to represent facial features for fundamental expressions (Happy, Sad, Disgust, Anger, Surprise, Fear) Moore & Bowden (2011). Automatic facial expression recognition techniques require two most important aspects: 1. Features representation 2. Modeling of appropriate classifier. Extraction of features those can represent the facial expressions effectively, are the key to have an accurate facial expression recognition system. These features need to provide more discriminant information, capturing most possible little variances for expression recognition. With first order pooling it is impossible to get these detail level features. Here, covariance matrix plays an important

role as traditional CNN can't get the second order features. Covariance matrix as a region descriptor can effectively combine numerous features and be methodically calculated via integral images Tuzel et al. (2006). Also, it is partially invariant to rotation and scale changes and robust against outliers. Covariance matrix is also used in generic feature representation for facial expression recognition. For this Riemannian manifold theory of symmetric positive-definite (SPD) matrices works as a driving force Huang & Van Gool (2017).

1.1 Motivation and Contribution

All the studies done previously on archival photos worked mainly with facial recognition. But this research is novel as this deals with facial expression recognition of archival photos. More specifically this work has particularly worked on NAAE-DB images to classify the facial expressions of people. Moreover, it has extensively done performance comparison between different models, parameter settings and classifiers which could be leveraged in related research such as building a robust Facial Expression Recognition System for NAAE-DB Archival. The NAAE-DB facial images are noticed to have very different characteristics from RAFDB benchmark dataset explained in Section 3.2. Such as NAAE-DB expressions are very subtle whereas RAFDB expressions are very explicit and clear which is easily distinguishable by a machine. As archival photos were captured with old model cameras the quality is not such high as modern ones such as RAFDB images. That makes the problem unexplored before.

NAA photos have metadata attached with it which contains information about the image. There isn't any information regarding the person's expression which this research aims to achieve. Hence, this research question guides the study: "How does covariance pooling perform on recognising the facial expressions from NAAE-DB archival photos?"

In summary, the contribution of this paper are following:

- State-of-art result on image-based facial expression recognition on archival photographs in unconstrained real-world environments
- Multiple data augmentation technologies to datasets has been added, and a group of experiments on different models has been performed to

illustrate the effectiveness of data augmentation. The result of experiments is shown in Table 6.4.

Chapter 2

Related Work

2.1 Deep Learning

Deep learning based machine learning models with various hidden layers trained on vast volumes of data could learn more useful features and thus enhance the accuracy of classification and prediction Wang et al. (2014). Convolutional Neural Networks, Deep Belief Networks and Deep Boltzmann Machines are some of them. Convolutional neural networks(CNN) have collections of small neurons in multiple layers that process the input image in parts that are the receptive fields. CNN consists of three layers: convolutional layers, max-pooling layers and fully-connected layer. Deep belief networks (DBNs) are multi-layer belief networks which are in several manners similar to convolutional neural networks. Deep Boltzmann Machines is a promising method as it can learn internal representations which is crucial for facial expression recognition Srivastava & Salakhutdinov (2012).

A huge research interest is going on since the last decade on the formulation of methods of facial expression recognition. A detailed overview of approaches can be found in Sariyanidi et al. (2014), Căleanu (2013)

2.2 Facial Expression Recognition

In twentieth century, Ekman and Friesen identified six basic emotions based on cross-culture study indicating that humans comprehend certain basic emotions in the same way regardless of culture Ekman & Friesen (1971). These common facial expressions are anger, disgust, fear, happiness, sadness

and surprise.
Covariance matrix used in covariance pooling can capture second-order statistics. It is used as region descriptor which has shown promising performance in object detection, recognition and tracking [Tuzel et al. (2006),Tuzel et al. (2007)]. The past several years have seen an expansion of covariance matrix in vision implementation. Furthermore it has now been used as a generic feature representation and used for various tasks including pedestrian detection Tuzel et al. (2007), face recognition Pang et al. (2008), action recognition Yuan et al. (2009). Few early works employed covariance pooling for feature extraction and used it as regional descriptor Carreira et al. (2012), Tuzel et al. (2006). Yu & Salzmann (2017) et. al. proposed numerous architectures based on VGG network to apply covariance pooling. Acharya et al. (2018) used covariance pooling after final convolution layers and combined with Riemannian manifold network inspired by Huang & Van Gool (2017)'s work to enhance facial expression recognition. This research is an extended work of that version. Data augmented images are distinct and independent images for untrained neural networks. In order to get more data, I made minor changes to the main NAA-EDB dataset by using salt-pepper noise, horizontal flipping, rotating and random crop.

2.3 Data Augmentation

CNN(convolutional neural networks) is invariant to directions, scales or illuminations, that is, CNN can classify objects robustly arranged in various directions, scales, or illuminations He et al. (2016). Data augmentation is based on this premise. Data augmentation is mainly to better the robustness and generalization ability of the model, and decrease the over-fitting phenomenon of the network. By transforming the training data, the network with stronger generalization ability can be obtained, which can better adapt to the application scenarios. Over-fitting is mainly caused by two reasons: too little data and too complex model. Hence, data augmentation is the most direct and effective method to avoid over-fitting. Common data augmentation methods are as follows:

- *Rotation* : Random rotates the image at a certain angle.

- *Reflection* : Change the orientation of the image content.

- *Flip* : Flip the image horizontally or vertically.
- *Scale* : The image can be scaled outward or inward as per the desired scale factor. Outward scaling makes image larger and inward makes it smaller than the original size.
- *Crop* : Random select the region of the image for clipping and scaling to a certain scale.
- *Shift* : Shifts the image in a certain way on the image plane.
- *Noise* : Random disruption of each pixel of the image. Salt-pepper noise and Gaussian noise are the common noise modes.
- *Color jitter* : Randomly vary the brightness, contrast and saturation of the image.
- *Random Erasing* : Random selection of an area on a picture and random erasure of image information.

Salt-and-pepper noise

Since the number of images for each class in NAAE-DB dataset was less, more images were fed to fine tune the model by adding noise components in them. Salt and Pepper noise is added to an image by addition of both random bright (with 255 pixel value) and random dark (with 0 pixel value) all over the image.

Figure 2.1: After adding noise(left) and before adding noise(right).

2.4 Classifiers

After extracting the features in above mentioned way I used different classifiers to observe performance variation. The classifiers used in this research are discussed below:

Suppport Vector Machines (SVM)

A Support Vector Machine (SVM) is a discriminative classifier based on the concept of a separating hyperplane. Another definition is, for supervised learning using labeled training data the algorithm finds an optimal hyperplane that can categorise unseen data points Michel & El Kaliouby (2003). This hyperplane in a two dimensional space would be a line dividing a plane in two parts where each class lies in either side. In linear SVM the hyperplane learning is done by transforming the problem using some linear algebra. Kernel SVM uses the kernel trick where linear classifier is used to solve a non-linear problem. The kernels used for experimentation in this research are Linear and RBF.

K-Nearest Neighbours (KNN)

K-nearest neighbors algorithm is based on the idea that similar things exist in close proximity thus classifies new cases based on a similarity measurement usually, Euclidean distance. Other approaches are Manhattan, Minkowski and Hamming distance methods.

Random Forest Classifier

This algorithm is constructed by a number of decision trees which can solve the multi-data classification problem and each decision tree is a classifier having one vote, and the final result of random forest classification is the average of all the decision trees' voting results.

Naive Bayes

A Naive Bayes classifier hypothesize that the existence of a particular feature in a class is not related to the existence of any other feature. This is a set of supervised learning algorithms Rish et al. (2001).

Chapter 3

Datasets

In this paper, facial expression dataset NAA-EDB built from from National Archives Australia (NAA) dataset has been built to better approximate the real-world scenarios and challenging prediction situations. This paper presented the evaluation results based on NAAE-DB and supported with RAFDB and JAFFE datasets. Due to the small amount of facial expression data, this paper uses different data augmentation methods to increase the amount of training data and improve the generalization ability of the model.

3.1 NAAE-DB

The NAA dataset is a collection of 344,894 photos. It will be called NAA 300K dataset in this report. Photos are either in forms of digital or digitized (from negative films) photographs, either grayscale or color. They are collected and archived by the National Archives of Australia (NAA). The photos are taken from the year 1900 to the year 2008. Figure 3.1a shows some image samples from the dataset. Each image has an associated metadata file and the metadata is stored as plain humanreadable XML tags. Each metadata file contains a range of additional information such as the year the image was captured, the person or people's name in that image, the event or location of the image, the photographer who took the image, copyright information etc. This research works with a small subset of whole NAA 300k that is called NAA-EDB which has been created after doing all the preprocessing steps as discussed in Chapter 4. All the basic seven expressions were hard to find in NAA-EDB as NAA photos don't have many fear, angry,

sad, surprise, disgust faces. So these five expressions are merged together and a new class has been formed as "Concerned". Thus NAA-EDB has three basic expressions(Happy, Concerned, Neutral). NAA-EDB statistics for this research is given below:

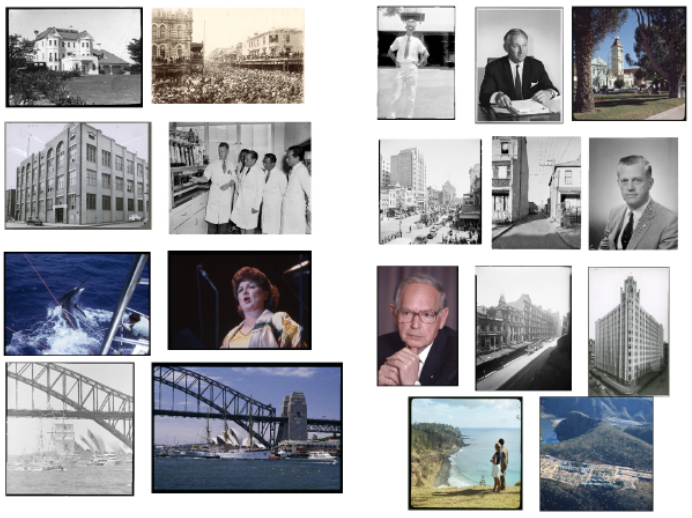
Number of train images: 6244

Number of test images: 2059

Total images: 8303

Train and test splitted into 4:1 proportion.

Train	**Test**
Happy: 1696	Happy: 504
Neutral: 2540	Neutral: 885
Concerned: 2010	Concerned: 670



(a) Samples.

Figure 3.1: NAAE-DB samples.

3.2 RAFDB

Real-world Affective Faces Database (RAFDB) Li et al. (2017) is an extensive facial expression database containing around 30K great-diverse facial images downloaded from the Internet. This dataset contains 15331 images labeled for seven basic expressions. Training set has 12271 samples and test set has 3068 samples. The RAF-DB has RGB images of size $100X100$ as shown in Figure 3.2a. Seven rows representing a category with 7 images per category. RAFDB provides already aligned faces.

(a) Samples.

Figure 3.2: RAFDB samples.

3.3 JAFFE

JAFFE database is made up of 213 images containing seven basic facial expressions posed by 10 Japanese female models Lyons et al. (1998). Every image has been semantic rated on six emotion adjectives by 60 Japanese subjects. The database was prepared by Michael Lyons, Miyuki Kamachi

and Jiro Gyoba. Images are $256X256$ gray level in .tiff format with no compression.

Chapter 4

Face Detection

Face Detection is classified under computer vision and is a very crucial aspect in facial recognition system. It is a method of developing and training algorithms to properly locate faces or objects in images, either in real time from a video camera or from photographs.

4.1 Collecting training data

Different sets of portrait images are manually sampled from the NAAE-DB 29k dataset containing approximately 28,912 images and from a small subset of the NAAE-DB 300k dataset. The Histogram of Oriented Gradients (HOG) descriptors Dalal & Triggs (2005) are used to extract the features of training data. Images are resized for efficient extraction of HOG descriptive features so that they all have same pixels. The HOG features will be reasonably larger if the image size is too large. However, a problem which can occur when resizing images is that, faces in some images will be distorted. A common aspect ratio is determined to avoid this issue and the common size to scale all the images down is based on this ratio.

If the size of faces are too small, the facial recognition algorithm may not work efficiently. Thus, a threshold value can be set to locate faces above the threshold. In this research, if the detected face is with pixels $32X32$ or more they will be selected for experimentation.

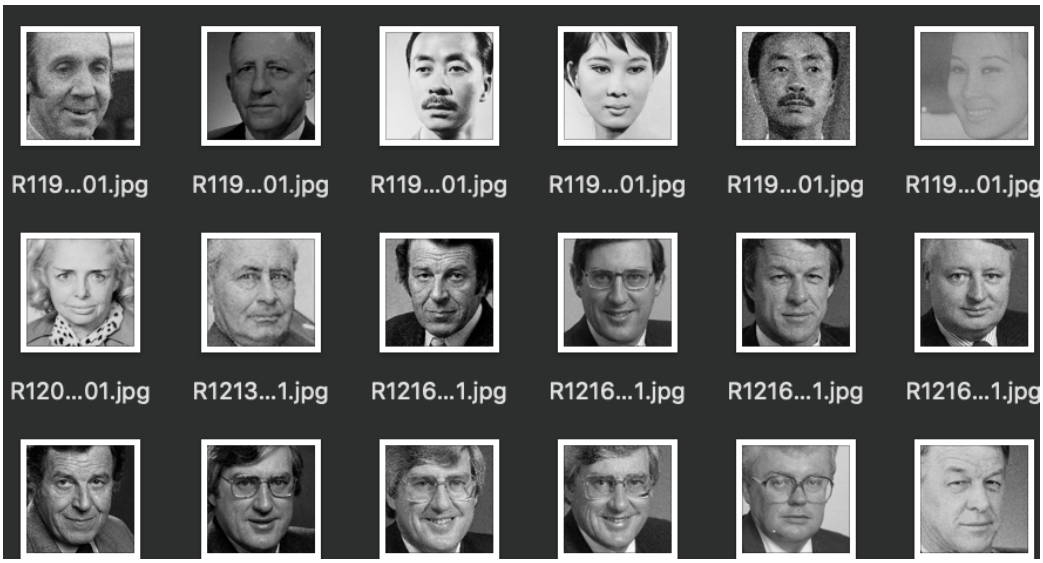


Figure 4.1: After all preprocessing of images

4.2 Extracting faces

For extracting faces from the raw hand-collected NAAE-DB dataset Multi-Task Cascaded Convolutional Neural Network (MTCNN) has been used. Using MTCNN the bounding boxes of faces in an image along with their five point face landmarks are detected at first. The five facial landmarks include left eye, right eye, nose, left mouth corner, and right mouth corner. Then gradually in each stage the detection results has been improved by passing it's inputs through a CNN, which returns candidate bounding boxes with their scores, followed by non max suppression(NMS). The three step processing (Zhang et al. 2016) of images through MTCNN has been discussed below :

1. In this step, for each scaled image, a 12x12 kernel runs through the image, searching for a face. It predicts bounding boxes using the regression technique. Then it merges the candidates which are highly overlapped.*Proposal Network (P-Net)*

2. Next Refine Network(R-Net) CNN, which has more layers, again rejects a large number of false candidates. It takes the P-Net bounding boxes as its inputs and predicts more roubust bounding boxes.

3. *Output Network (O-Net)*: The *O-Net*'s inputs are the R-Net bounding boxes and marks down the coordinates of facial landmarks.

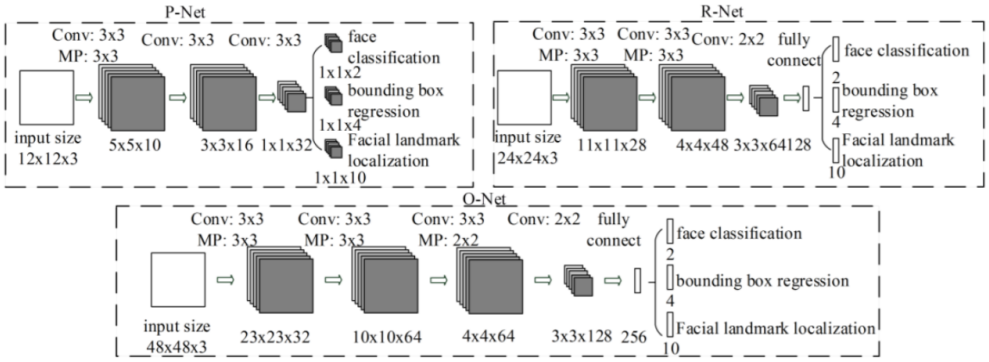


Figure 4.2: MTCNN Structure

Through MTCNN, face detection and face alignment has been done jointly in a multi-task training fashion. This allowed the model to better detect faces that are initially not aligned.

4.3 Face Alignment

To increase the face recognition accuracy **face alignment** I did face alignment. It normalized technique and makes the face-centered to the image. Image is rotated such that line joining the center of two eyes is parallel to the horizontal line and it resizes the faces to an identical scale.

4.3.1 Image Rotation

There are two key parameters which are required for rotating of a face:

1. *The anchor point*: The middle point between the eyes and face is rotated around this point.

2. *The angle of rotation - α*: The angle in which face is rotated to ensure it has zero-tilt. It is the angle between the line joining the center points of the left and right eye and the horizontal line.

$$\alpha = \arctan\left(\left|\frac{x_1 - x_2}{y_1 - y_2}\right|\right) \tag{4.1}$$

where (x_1, y_1) and (x_2, y_2) are the coordinates of the centroids of the left and right eye.

4.3.2 Image cropping, translating and scaling

The extracted face region is translated and scaled such that the distances from the edges of the image to the centroids of left and right eye are equal. At last after all these processing, NAAE-DB imageset looks like Figure 4.1.

Figure 4.3: Detected faces getting aligned and cropped

Chapter 5

Facial Expression Recognition Methodology

5.1 Overview

Facial expression is restricted in the facial region and images in the wild contain large irrelevant information. For this, face detection is performed first and then aligned depending on facial landmark locations. Next, the normalized faces are fed into a deep CNN. I used covariance pooling for pooling the feature maps spatially from the CNN and then the manifold network is used to deeply learn the second-order statistics. In Figure 5.1 the leveraged model's pipeline for facial expression recognition is shown. The manifold network described in Huang & Van Gool (2017) has also been used for learning the second-order features deeply, dimensionality reduction and introducing non-linearity on covariance matrices. The main methods of the two models are spatial covariance pooling and the manifold network. These two techniques will be introduced in the following:

5.2 Covariance Pooling

As discussed earlier, traditional CNNs that consist of fully connected layers, max or average pooling and convolutional layers only capture first-order information Yu & Salzmann (2017). ReLU introduces non-linearity

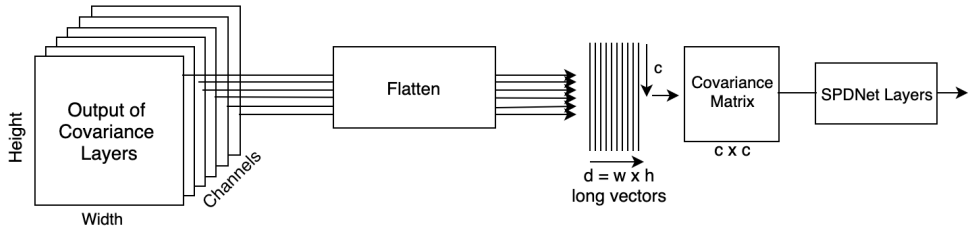


Figure 5.1: The output of convolutional layer is flattened as illustrated.

only at individual pixel level. Covariance matrices calculated from features better able to gain regional features than first-order statistics Tuzel et al. (2006). For a set of features covariance matrix can be used to integrately sum up the second-order details in the set. If $x_1, x_2, ..., x_n \, \epsilon \, \mathbb{R}^d$ be the set of features, the covariance matrix can be found as:

$$C = \frac{1}{n-1} \sum_{i=1}^{n} (x - \overline{x})(x - \overline{x})^T$$

$\overline{x}$ is the sample mean. Given an image region, a feature vector, x, is extracted from each pixel to describe its location, colour, gradient, filter response, etc. Covariance matrix characterises this region with these feature vectors. As a region descriptor it has a natural way to combine different features; being stable against illumination change and outliers; allowing two regions of different sizes to be compared; is rotation invariance when rotation-independent features are used; and fast computation via integral images. A covariance matrix is said singular when sample size is less than the number of variables Cai et al. (2011).

5.3 SPD Network Layers

A nonsingular covariance matrix refer to the set of symmetric positive-definite (SPD) matrices. SPD matrices form a connected Riemannian manifold. SPD manifold network is called "Covariance pool". This is used here for my research combining manifold networks with traditional convolutional networks in an end-to-end depth learning approach. The obtained covariance matrices are usually large and their dimension needs to be decreased without the lose of geometric structure. In Huang & Van Gool (2017), authors

proposed special layers for dimensionality reduction of SPD matrices and to flatten the Riemannian manifold to be able to apply standard loss functions. Logarithm operation is applied to flatten and to be able to apply standard loss functions of Euclidean space Tuzel et al. (2006). So, covariance features are flattened through Bilinear Mapping Layer and then through Eigenvalue Rectification non-linearity is introduced. The layers introduced in Huang & Van Gool (2017) for learning on Riemannian Manifold have been briefly discussed in the following subsections.

Bilinear Mapping Layer (BiMap)

As obtained covariance matrices can be big, it is not feasible to directly apply fully connected layers after flattening them. Furthermore, it is also important to preserve geometric structure while reducing dimension. The BiMap layer overcomes both of these and works as the traditional fully connected layers by performing element-wise square root normalization. It does not consider the manifold of covariance matrices.

Eigenvalue Rectification (ReEig)

To introduce non-linearity like the Rectified Linear Unit (ReLU) layers in traditional neural networks, ReEig layer is used here.

Log Eigenvalue Layer (LogEig)

The final layer is LogEig which enables elements in Riemannian manifold with a certain structure so that matrices can be flattened and standard euclidean operations can be applied.

This work mainly deals with three classes for NAAE-DB dataset as I merged five classes (angry,fear,surprise,sad,disgust) and renamed it into "Concerned". The code for this research would be uploaded in Github.

Chapter 6

Experiments

The purpose of this chapter is to evaluate the performance of the facial expression recognition method on NAAE-DB dataset under different algorithmic settings. Since the NAAE-DB dataset has a variety of images in both good and bad clarity, the dataset was further confined and picked only the images that can properly evaluate or recognise the facial expressions. The portrait images filtered from the NAAE-DB dataset were aligned as discussed in 4.3. These aligned facial images were noticed to have very different characteristics from RAFDB benchmark dataset explained in Section 3.2. As the RAFDB benchmark dataset has great-diverse facial images, so the model trained on RAFDB dataset was used to predict on the NAAE-DB dataset. The prediction accuracies were recorded and later the trained model with RAFDB dataset was finetuned with NAAE-DB dataset expecting to achieve an improvement in accuracy. The parameters and kernels were also varied to enrich the findings. While implementing covariance pooling, we experimented with various models and classifiers for which the details are summarized in Section 6.2 below.

6.1 Computational issues

This research for classifying expression as shown in Figure 5.1 was experimented with and without the Manifold Network Layer. It was expected to get a better performance without the Manifold Network layer, but because of higher feature dimensionality, the SVM was being trained for days. Thus due to lack of computation power it failed at a point.

As mentioned earlier the model was initially trained on RAFDB benchmark dataset which had a total of 12771 images. Different pretrained model like VGG-16, Facenet and Inception-Resnet was used to train the dataset and features dimension of long vector was around 131328 after covariance pooling as explained in Section 5.2. Experimental procedures were run in 4GB Memory of GPU. Number of images is 12271 (n observations) and each of them a p-long vector, where p is 131328, i.e.,$(512X513)/2$.The authors in Cai et al. (2011) stated that the sample covariance matrix calculated with p>n is singular. The covariance matrix has to be non singular for this research purpose.

The rank of covariance matrix obeys the rule: $rank(C) \leq min(d, n-1)$. When C is used as a region descriptor, the number of feature vectors $(14X14X512)$ extracted from an image region, n, is usually much larger than the dimensions, d. This ensures C to be nonsingular and allows it to be reliably estimated. However, this situation is not the situation in this application, and singularity occurred. In that case, in order to utilize Riemannian metrics, a small scaled identity matrix has been appended.

The covariance matrices thus obtained are often large and processing them was causing computational issue. This was mainly due to the limitations with the memory in the environment I was working on. It was then necessary to adopt SPDNet mentioned in Section 5.3 which can reduce the dimension without losing geometric structure.

6.2 Evaluation

The model trained on the benchmark dataset RAFDB is tested with a variety of datasets including JAFFE and NAAE-DB. The predicted accuracies are recorded with different settings.

Comparison of Standard Architectures

In Table 6.1 the comparison of accuracies of training or finetuning various standard network architectures are presented. For experimental purpose I reproduced the result of Acharya et al. (2018) to adopt this methodology for this research. The scores reported on RAFDB Bechmark Dataset for Acharya et al. (2018)'s network is more compared to Fan et al. (2018) and Li et al. (2018)'s model. Thus the network architecture of Acharya et al. (2018)

is considered for this research. So the networks are not trained again here. It is worth pointing out that all these work recognised seven basic expressions (Angry, Happy, Disgust, Neutral, Surprise, Fear).

Table 6.1: The comparison of the accuracy in RAFDB.

Methods	No of classes	Accuracy
Fan et al. (2018)	7	77%
Li et al. (2018)	7	85%
Acharya et al. (2018)	7	87%

However, in using the fully supervised deep learning models, the chance of overfitting is high due to insufficient training samples for training the model. Thus pretrained models are used to work with facial expression recognition. The pretrained models like VGG, Alexnet were initially popular for facial recognition, which fall short of discrimination ability of expression characteristic. So in this research the model trained from scratch with RAFDB by Acharya et al. (2018) is used which is based on Li et al. (2017)'s architecture, Deep Locality-preserving CNN (DLP-CNN), as shown in figure 6.1.

Layer	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
Type	Conv	ReLu	MPool	Conv	ReLu	MPool	Conv	ReLu	Conv	ReLu	MPool	Conv	ReLu	Conv	ReLu	FC	ReLu	FC
Kernel	3	-	2	3	-	2	3	-	3	-	2	3	-	3	-		-	
output	64	-	-	96	-	-	128	-	128	-	-	256	-	256	-	2000	-	7
Stride	1	1	2	1	1	2	1	1	1	1	2	1	1	1	1		1	
Pad	1	0	0	1	0	0	1	0	1	0	0	1	0	1	0		0	

Figure 6.1: The configuration parameters of DLP-CNN

Following up the idea mentioned by Li et al. (2017), each DCNNs for basic emotion recognition task are trained first, and then directly used the already trained DCNN models to extract deep features for expressions. 2000-dimensional deep features learnt from raw data were extracted from the penultimate fully connected layer of the DCNNs and then classified with different classifiers.

In this research, the model that has been used for experimenting with NAA-EDB has the basic architecture of Li et al. (2017)'s model which is trained on RAFDB. Then Acharya et al. (2018) has topped that with covariance pooling and SPDNet layers and finetuned with RAFDB. We are calling

this **DLP-CNN Finetuned**. This pre-trained model has been carried out and has been finetuned on NAA-EDB in this paper's research for classifying three facial expressions. We named it as **Finetuned NAA-EDB**. Effects of finetuning the already pretrained model with NAAE-DB is also recorded in Table 6.4. The table 6.2 represents the various models used in this research for evaluating their respective performance on NAA-EDB and JAFFE dataset when covariance pooling is used.

Table 6.2: Various models considered for covariance pooling

Baseline	DLP-CNN Finetuned	Finetuned NAA-EDB
Conv256	Conv256	Conv256
	Cov	Cov
	BiRe	BiRe
	LogEig	LogEig
FC2000	FC2000	FC2000
FC7	FC128	FC128
	FC7	FC7
Trained from scratch RAFB	Finetuned with RAFDB	Finetuned with RAFDB + Finetuned with NAAE-DB

JAFFE Dataset

JAFFE dataset were divided into with 185 train and 28 test images. Train image were extremely low. We experimented with a pretrained VGG-16 model and classified with SVM RBF which performed poorly. Even-though the JAFFE dataset is very limited with the number of images in training, we can see the covariance pooling method performed reasonably well from the below table 6.3. I also performed data augmentation (flipping, scaling,

adding noise) to increase the number of training data and the performance improved significantly with a increase of 6%.

Table 6.3: The comparison of the accuracy of each class in JAFFE.

Model	Classifier	Accuracy
DLP-CNN Finetuned	SVM - rbf + Data Augmentation	70 %
	SVM - linear	57%
	SVM - rbf	64%
VGG 16	SVM - rbf	40%

NAAE-DB Dataset

The performance of DLP-CNN Finetuned model was tested with and without finetuning with NAAE-DB. The achieved accuracies have been recorded in the last column of Table 6.4.

Table 6.4: The comparison of the accuracy of each class in NAAE-DB.

Model	Classifier	Concerned	Happy	Neutral	Average
DLP-CNN Finetuned	SVM (linear)	0.96	0.75	0.80	0.835
	SVM (rbf)	0.99	0.61	0.88	0.855
	KNN = 3	0.97	0.76	0.80	0.843
	KNN = 5	0.96	0.77	0.80	0.843
	KNN = 10	0.97	0.75	0.79	0.843
	Naive Bayes	0.82	0.61	0.51	0.63
	Random Forest	0.79	0.51	0.77	0.71
Finetuned NAAE-DB	SVM (rbf)	0.98	0.735	0.88	**0.869**
Basic CNN epoch=100	softmax	0.78	0.60	0.70	0.69

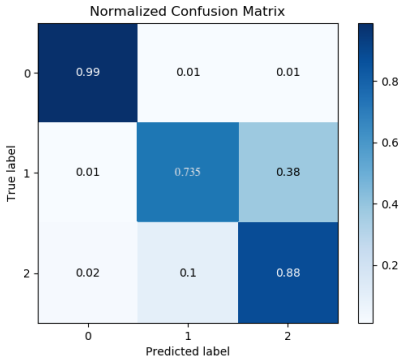
Table 6.5: Comparison of accuracies before data augmentation

Model	Classifier	Accuracy
DLP-CNN Finetuned	SVM - linear	75%
	SVM - rbf	75.12%

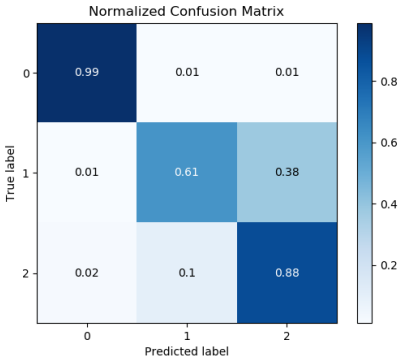
At first classification was attempted with a basic CNN, run for 100 epochs, which displayed only 69% classification accuracy. Then employing DLP-CNN Finetuned model without data augmentation showed 75.12% accuracy in Table 6.5. Evidently using covariance pooling raised the performance by 6.12%. Clearly data augmentation caused a significant improvement as well in the prediction results as can be seen from Table 6.4. The DLP-CNN finetuned model on data augmented NAA-EDB acquired 85.5% accuracy which is 10.38% more than the previous result. But it couldn't classify well for the Happy class, assuming because it has the lowest training data. The Finetuned NAAE-DB model achieves 1.4% higher accuracy coming to 86.9% with near perfect classification rate for Concerned class and improved accuracy for Happy class. Finetuning the top layer has extracted the specific NAAE-DB features which also grows individual class accuracy. My cross-dataset study shows that finetuned model with SVM-RBF or KNN performs better in all cases. I have divided classified images into two folders called "misclassified" and "rightclassified" which contain images of wrongly classified and correctly correctly classified along with the predicted and targeted class names. The folders would be provided on Google Drive upon reader's request.

6.3 Discussion

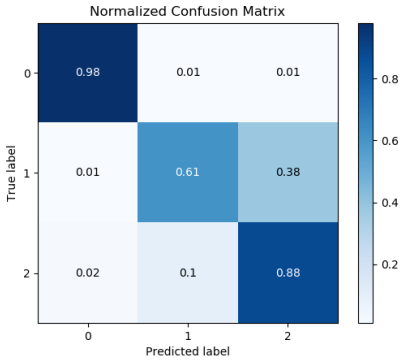
From the confusion matrices given in Figure 6.2 we see that the accuracy of "Concerned" is significantly higher than "Neutral", but the accuracy of "Happy" is the lowest. All the classifiers failed to recognise the happy images. The reason behind this consistent classification error might be the number of Happy image is significantly lower than the other two classes, particularly 844 images lesser than Neutral. This imbalance should have made the SVM biased to make the classification error. Additionally, the Concerned class was made of surprise, disgust, fear, angry and sad as we didn't have enough of



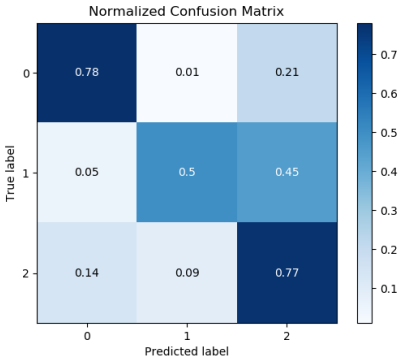
(a) SVMrbftuned 86.7%.



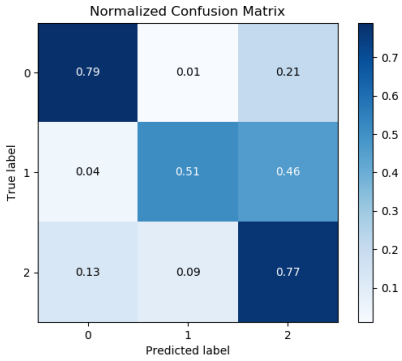
(b) SVMrbf 85%.



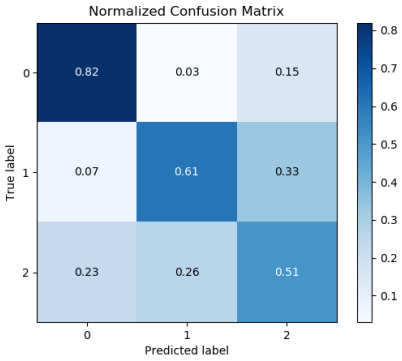
(c) SVMlinear 84.4%.



(d) Randomforest100 70.6%.



(e) Randomforest70 71.2%.



(f) Naive Bayes 63.4%.

Figure 6.2: 1.Concerned 2.Happy 3.Neutral

those expressions in NAAE-DB. Also, there are certain similarities between surprise,disgust,angry. In real life, humans also find it difficult to distinguish these three kinds of expressions, especially when they are not acquainted with each other.

After analysing the results, and misclassified and correctly classified images, I noticed following few properties of the images for which the the algorithm was failing to perform expression classification properly used in this research and were having difficulties recognising them:

- Faces are occluded by other people, objects or side posture, thus only half or part of faces can be seen (Figure 6.3f).
- Face images that are extremely blurred, in very bright or dark conditions, the details on faces such as eyes, mouth, nose, ears and other parts cannot be seen clearly (Figure 6.3c, 6.3e).
- Faces that are left-rotated, right-rotated or have extreme poses, expressions, angles, orientations (Figure 6.3a).
- Distinct expressions but they have the same distortion looking nearly similar. The facial expression recognition algorithm cannot distinguish between these expressions. Suppose in Figure 6.3d the happy expression is so subtle that it is almost near to neutral.
- Although the ratio is very low, some expressions were wrongly hand classified while preparing the dataset. When expression is actually Happy it has been annotated as Neutral during hand annotation (Figure 6.3g). This problem was detected after careful analysis.

(a) Overbrightness.

(b) Extreme Angle.

(c) Overblurred.

(d) Too near to Neutral.

(e) Darkness hid details .

(f) Side Posture.

(g) Handmisclassification.

Figure 6.3: Misclassified Expression Scenarios.

Chapter 7

Future Work

7.1 Expressions

The subset that we used as NAA-EDB in here has approximately 85% male, 15% female, and 5% children. In future a more diverse subset would increase the generality of the research. Facial expression is a series of action, and the difference of expression mainly reflects the dynamic changes in information Fanelli et al. (2013). For example, on the static images anger, disgust and sadness have more similarities. The cause of humans accurately distinguishing between a static image expression is that human can think of a series of facial expression action based on past knowledge and experience. It is thus believed that if one can take advantage of multi-frame image sequences to extract dynamic information for the expression classification, they would have a better recognition effect Chen et al. (2013). It is feasible to distinguish more expressions, such as smiling and laughing in happy etc.

7.2 Faces

The number of faces which has been identified, only occupies a small part of the NAA300k dataset, while the large portion of images in the dataset with faces cannot be identified by the algorithm used here. However, I believe that because the face detection and recognition algorithms are not fully optimised for the NAA 300K dataset, they couldn't detect and recognize many faces. Future scope can explore to make face detection more robust.

Chapter 8

Conclusion

In this work, we exploit the use of SPDNet on facial expression recognition for archival photographs. In this research, I have addressed the lack of metadata containing facial expression of a person and presented the key components of facial expression recognition for archival photos using advanced deep learning techniques collaborated with covariance matrix and SPDNet. I finetuned an existing algorithm to find promising performance of NAAE-DB dataset. This would help to provide more functionalities in image search and retrieval. This method can be used to further build a system for the NAA staff to manage the archival photos as well as the public to access these visual information effectively. As seen above, SPDNet applied to covariance of convolutional features perform facial expressions classification more efficiently. We study that second-order networks are better able to capture facial landmark distortions. In facial expression recognition, we obtain results comparable to state-of-the-art results.

Bibliography

Acharya, D., Huang, Z., Pani Paudel, D. & Van Gool, L. (2018), Covariance pooling for facial expression recognition, *in* 'Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops', pp. 367–374.

Cai, T., Liu, W. & Luo, X. (2011), 'A constrained l-1 minimization approach to sparse precision matrix estimation', *Journal of the American Statistical Association* **106**(494), 594–607.

Căleanu, C.-D. (2013), Face expression recognition: A brief overview of the last decade, *in* '2013 IEEE 8th International Symposium on Applied Computational Intelligence and Informatics (SACI)', IEEE, pp. 157–161.

Carreira, J., Caseiro, R., Batista, J. & Sminchisescu, C. (2012), Semantic segmentation with second-order pooling, *in* 'European Conference on Computer Vision', Springer, pp. 430–443.

Chen, J., Chen, D., Li, X. & Zhang, K. (2013), 'Towards improving social communication skills with multimodal sensory information', *IEEE Transactions on Industrial Informatics* **10**(1), 323–330.

Dalal, N. & Triggs, B. (2005), Histograms of oriented gradients for human detection, *in* 'Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 1 - Volume 01', CVPR '05, IEEE Computer Society, Washington, DC, USA, pp. 886–893.
URL: *https://doi.org/10.1109/CVPR.2005.177*

Ekman, P. & Friesen, W. V. (1971), 'Constants across cultures in the face and emotion.', *Journal of personality and social psychology* **17**(2), 124.

Fan, Y., Lam, J. C. & Li, V. O. (2018), Multi-region ensemble convolutional neural network for facial expression recognition, *in* 'International Conference on Artificial Neural Networks', Springer, pp. 84–94.

Fanelli, G., Dantone, M., Gall, J., Fossati, A. & Van Gool, L. (2013), 'Random forests for real time 3d face analysis', *International Journal of Computer Vision* **101**(3), 437–458.

He, K., Zhang, X., Ren, S. & Sun, J. (2016), Deep residual learning for image recognition, *in* 'Proceedings of the IEEE conference on computer vision and pattern recognition', pp. 770–778.

Huang, Z. & Van Gool, L. (2017), A riemannian network for spd matrix learning, *in* 'Thirty-First AAAI Conference on Artificial Intelligence'.

Li, S., Deng, W. & Du, J. (2017), Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild, *in* 'Proceedings of the IEEE conference on computer vision and pattern recognition', pp. 2852–2861.

Li, Y., Zeng, J., Shan, S. & Chen, X. (2018), 'Occlusion aware facial expression recognition using cnn with attention mechanism', *IEEE Transactions on Image Processing* **28**(5), 2439–2450.

Lyons, M., Akamatsu, S., Kamachi, M. & Gyoba, J. (1998), Coding facial expressions with gabor wavelets, *in* 'Proceedings Third IEEE international conference on automatic face and gesture recognition', IEEE, pp. 200–205.

Michel, P. & El Kaliouby, R. (2003), Real time facial expression recognition in video using support vector machines, *in* 'Proceedings of the 5th international conference on Multimodal interfaces', ACM, pp. 258–264.

Moore, S. & Bowden, R. (2011), 'Local binary patterns for multi-view facial expression recognition', *Computer vision and image understanding* **115**(4), 541–558.

National Archives of Australia (n.d.).
URL: *http://www.naa.gov.au/*

Pang, Y., Yuan, Y. & Li, X. (2008), 'Gabor-based region covariance matrices for face recognition', *IEEE Transactions on Circuits and Systems for Video Technology* **18**(7), 989–993.

Rish, I. et al. (2001), An empirical study of the naive bayes classifier, *in* 'IJCAI 2001 workshop on empirical methods in artificial intelligence', Vol. 3, pp. 41–46.

Sariyanidi, E., Gunes, H. & Cavallaro, A. (2014), 'Automatic analysis of facial affect: A survey of registration, representation, and recognition', *IEEE transactions on pattern analysis and machine intelligence* **37**(6), 1113–1133.

Srivastava, N. & Salakhutdinov, R. R. (2012), Multimodal learning with deep boltzmann machines, *in* 'Advances in neural information processing systems', pp. 2222–2230.

Tuzel, O., Porikli, F. & Meer, P. (2006), Region covariance: A fast descriptor for detection and classification, *in* 'European conference on computer vision', Springer, pp. 589–600.

Tuzel, O., Porikli, F., Meer, P. et al. (2007), Human detection via classification on riemannian manifolds., *in* 'CVPR', Vol. 1, p. 4.

Wang, W., Yang, J., Xiao, J., Li, S. & Zhou, D. (2014), Face recognition based on deep learning, *in* 'International Conference on Human Centered Computing', Springer, pp. 812–820.

Yu, K. & Salzmann, M. (2017), 'Second-order convolutional neural networks', *arXiv preprint arXiv:1703.06817* .

Yuan, C., Hu, W., Li, X., Maybank, S. & Luo, G. (2009), Human action recognition under log-euclidean riemannian metric, *in* 'Asian Conference on Computer Vision', Springer, pp. 343–353.

Zhang, K., Zhang, Z., Li, Z. & Qiao, Y. (2016), 'Joint face detection and alignment using multi-task cascaded convolutional networks', *CoRR* **abs/1604.02878**.
URL: *http://arxiv.org/abs/1604.02878*

Appendix A

Appendix Title

RBF: Radial Basis Function
CNN: Convlution Neural Network
NAAE-DB: National Archives of Australia Expression Dataset
DLP: Locatily Preserving
MTCNN: Multi-Task Cascaded Convolutional Neural Network
NAA300k: National Archives of Australia Dataset with 30000 images