IDENTIFYING FACIAL LANDMARKS, ACTION UNITS AND EMOTIONS USING DEEP NETWORKS

BY

NAMRATA PRABHU

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Master of Science in Computer Science
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2016

Urbana, Illinois

Adviser:

Associate Professor Derek Hoiem

Abstract

The goal of this thesis it to use deep neural networks, specifically Convolutional Neural Networks (CNNs) to predict facial landmarks, facial action units and emotions and to study the results of intermediate experiments while doing so. Learning the different features of facial images has always been a difficult task and primarily involves using hand-crafted features which would almost definitely ignore some information related to the different dynamics of facial features. We train our network model using the raw facial images and study its effectiveness in predicting facial landmarks, action units and emotions. In this thesis we learnt that CNNs are highly effective in predicting facial landmarks and AUs, mainly because of their ability to learn features from raw images. We also established that feature sets which can effectively outline the different properties of a face are more useful in classifying facial emotions than either images or facial landmarks.

Acknowledgments

This project could not have been completed without the support of many people. I would like to thank my adviser Derek Hoiem for his invaluable support, guidance and patience while helping me through this project. I also could not have completed my project without constant help and advice from Kevin J Shih, who helped me navigate through all the different experiments with terrific patience and understanding. Finally I would like to thank the UIUC Vision group for their equipment.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Overview

Facial expressions convey non-verbal cues and and play a prominent role in inter-personal relations. Recognizing them can be used to determine the users emotional and mental state, and automatically obtaining this information can be incredibly useful in the fields of psychology, mental health, human-computer interaction, etc. The research in [9],[6] and [12], show several advances in the terms of facial detection and recognition, facial feature extraction mechanisms and facial expression classification, but developing an automated system which accomplishes these tasks has been difficult. In this thesis, we explore the use of deep learning for localizing facial landmarks, predicting facial action units and classifying emotions. We also investigate whether intermediate representations learned for AU prediction are useful to learning to classify emotions in a limited data-set.

Facial muscle movements are an important precursor to determining facial expressions or emotions. Edman and Friesman developed the Facial Action Coding System and proposed that different emotions are in fact a combination of different facial muscle movements. In [3] , they describe how every facial expression is in fact a combination of different facial muscle movements (facial action units or AUs). Each AU is characterized by a visible facial movement or its resultant deformation. Currently, FACS is the best known psychological framework for describing nearly the entirety of facial movements. Although humans can recognize different facial expressions and movements quite effortlessly, reliable detection movements in faces and understanding their meaning is still a challenging problem for machines primarily due to the high degree of variation in the appearance of human faces, low intensity action of specific units in particular expression, co-occuring AUs and the scarcity of training data.

More recently deep learning algorithms have shown significant improvement for object detection tasks mainly because of the availability of high performance hardware such as GPUs which allow for more efficiency and faster computation, and the availability of large datasets which allow for a large training and testing data [13], [7]. However there has not been much work using deep learning techniques in facial expression

detection and facial action unit recognition. Usually facial expression and AU recognition involve using appearance features such as HoG, Gabor [20], etc. as well as geometric features which can be computed from the locations of different facial landmarks. Since these features are not tuned to a particular task, they limit the performance of the classifier which has been trained on these features. Making use of deep learning techniques involves learning features directly from the pixel values in the pre-processed form of the raw image. Thus these methods provide an algorithm which can be trained directly from the pixels in the image to their corresponding label values, while also providing feature sets in the intermediate stages which can be then used to improve the performance of the learned model.

An efficient and automatic facial AU detector involves the analysis of the following facial features: shape of the face, appearance and the location of facial landmarks (eyes, eyebrows, nose, mouth, cheeks, head and jawline). A network which can learn these facial features in unison can produce a highly accurate model for detecting AUs and classifying emotions.

Convolutional Neural Networks(CNNs) are a type of deep neural net which are particularly effective in learning representations of images from raw input images without any extraction of handcrafted features. This form of learning is referred to as "representation learning" and can be applied to unseen data effectively. It was this property of CNNs which motivated us to use them to explore predicting facial landmarks positions and classify AUs in raw facial images.

The main contributions of this paper are summarized below:

- We use CNNs to predict facial landmark positions namely the positions of different points on the eyes, eyebrows, nose, mouth and jawline by training the network on raw facial images.
- We compared the performances of different models to classify AUs from raw facial images processed using a CNN, ground truth landmarks and predicted landmarks from the previous experiment.
- We explored different feature sets to classify facial expressions. Specifically we made use of facial landmarks as well as the information learned from training the neural net to detect AUs, to determine which feature set is able to most accurately understand variances in facial features in terms of emotions.

1.2 DataSets

In the course of this experiment, we made use of two different datasets: Japanese Female Facial Expression Dataset (JAFFE) [18] and the MMI Facial Expression Database [19]. We specifically chose the MMI dataset since it is primarily a video dataset, which provided a wide range of data across 31 different AUs which we

aimed to detect in this experiment. While the MMI dataset was used exclusively for AU detection, we used the JAFFE dataset to classify facial expressions as one of the following: Happy, Disgust, Surprise, Anger, Fear,Sadness and Neutral.

We also initially considered using the Cohn-Kanade Extended (CK+) dataset [16] for classifying different facial expressions. The CK+ dataset is very popular for algorithm detection in facial expression detection. Our initial method was to use the facial landmark data provided for each image to classify different facial expressions but we surprisingly received an unusually high accuracy of 98 per cent on our first trial. After investigating the data, we realized that the labels in the dataset were from the output of a classifier rather than the ground truth label, which ultimately made our classification meaningless. Although CK+ also provides facial images as well as their corresponding AUs, we decided to rely solely on the MMI Facial Expression Dataset to ensure uniformity.

1.2.1 JAFFE

This dataset comprises of 213 images of 7 facial expressions (6 basic facial expressions plus 1 neutral) portrayed by 10 Japanese female models. Each image has been rated on 6 emotion adjectives by 60 Japanese subjects.

1.2.2 MMI Facial Expression Database

The MMI Facial Expression Database is a large dataset which is a valuable resource for building and evaluating facial expression recognition algorithms. It contains video recordings of the full temporal patterns of facial expressions, from neutral, through a series of onset,apex and offset phrases and back again to a neutral face.

The database addresses a number of key omissions in other databases of facial expressions. In particular, it contains recordings of the full temporal pattern of a facial expressions, from Neutral, through a series of onset, apex, and offset phases and back again to a neutral face.The entire database consists of over 2900 videos and high-resolution still images of 75 subjects. It is fully annotated for the presence of AUs in videos (event coding), and partially coded on frame-level, indicating for each frame whether an AU is in either the neutral, onset, apex or offset phase. For the purpose of this experiment, we considered 329 video files involving 21 different subjects. We considered only those videos which were annotated with a frame-by-frame description of the onset-apex-offset of the action unit being represented in the video. Individual frames were extracted from videos and were labeled according to the presence or absence of the AUs being considered. We considered 31 different action units for this experiment.

Chapter 2

Background

There have been some prominent approaches in using deep learning for detecting facial landmarks, action units and emotions. For our experiments to predict facial landmarks and AUs, we make use of AlexNet [13] which is a pre-trained CNN model.The authors trained a large, deep convolutional neural network to classify the 1.2 million high-resolution images in the ImageNet LSVRC-2010 contest into the 1000 different classes. On the test data, they achieved top-1 and top-5 error rates which were considerably better than the previous state-of-the-art [13].

Predicting facial landmarks in images has also seen some considerable work. While we use the pre-trained AlexNet to predict landmark positions, we were primarily motivated by the efforts in [24].Yi Sun, Xiaogang Wang and Xiaoou Tang [24] make use of three-level carefully designed Convolutional Neural Networks(CNNs) to detect the positions of five facial keypoints. At each level, the outputs of multiple networks are fused for robust and accurate estimation. Their experiment outperforms state-of-the-art methods in both detection accuracy and reliability. We used similar measures to evaluate the performance of our classifier and found comparable results. However this is not completely conclusive since the datasets used are different. [27] presents a novel CNN design for facial landmark regression. Intermediate features of a standard CNN which was trained for detecting landmarks was analyzed and features from later and more specialized layers were able to capture rough landmarks locations. Thus the authors introduced a Tweaked CNN model which takes advantage of the robustness of CNNs to analyze appearances.

Our primary approach to detect AUs in facial images involves using a CNN to process raw facial images. Several works in the past (eg. [[25], [15],[8], [5]]) primarily make use of handcrafted features as inputs to their network model and utilise all the key facial features i.e. face shape, appearance and dynamics. Tian et al [25] constructed a three-layer neural network with one hidden layer for identifying lower face AUs. However instead of processing raw facial images, he made use of handcrafted facial features such as lip height, width and corner as well as left nasio-labal furrow angle, and presence of nose wrinkles, as different inputs to the neural network. In [15], the authors trained a multi-layer boosted deep belief network and achieved state-of-the-art accuracy on the CK+ [16] and JAFFE [18] datasets. Gudi et al [8] used a deep

CNN consisting of 3 convolutional layers, sub-sampling layer and 1 fully connected layer to predict the occurrence and intensity of Facial AUs. Fasel [5] used CNNs for the task of facial expression recognition. He used a 6 layer CNN architecture (2 convolutional layers, 2 sub-sampling layers and 2 fully connected layers) for classifying 7 facial expressions (6 basic emotions + 1 neutral). In [9] Shashank Jaiswal and Michel Valstar from the University of Nottingham use combination of Convolutional and Bi-directional Long Short-Term Memory Neural Networks (CNN-BLSTM), which jointly learns shape, appearance and dynamics in a deep learning manner. They also use the locations of facial landmarks to compute binary image masks, to encode facial features. [11] combined face models learnt using a deep CNN architecture with various other models for facial expression classification. Their method involves using a CNN model of a face with a bag of words model, a deep belief network for audio information and deep auto-encoder to model spatio-temporal information. They then used a weighted average of the results from all their constructed models to classify the facial expression. Our approach to detect AUs using a CNN is most closely aligned with [6]. Sayan Ghosh, Eugene Laksana, Stefan Scherer and Louis-Philippe Morency's experiment which highlights a multi-label convolutional neural network approach to learn a shared representation between multiple AUs directly from the input image. Thus their CNN processed the raw facial image without making use of any handcrafted features. They perform their experiments on three AU datasets - CK+, DISFA and BP4D and obtain competitive results on all datasets. We used the same evaluation parameters such as accuracy and AUC score (which is comparable to their 2AFC score) to analyze our results.

Our experiment to classify facial expressions involves using several features constructed from facial landmarks and the CNN used to predict AUs, on a limited dataset. Previous work (eg. [[17]]) involves using Gabor filters to produce image representations. Jung et al [10] used a deep CNN to learn temporal appearance features for recognising facial expressions. Additionally, they also employed a deep Neural network to learn temporal geometric features from detected facial landmarks. Both the networks were learned independently to predict facial expression. The output decision values from each of the networks were combined (linear combination) to compute the final score for any example face image. In [14], feature sets are constructed by extracting parametric descriptions of the eyes, lips, eyebrows, furrows are extracted. Using these parameters, a group of AUs are recognized on whether they occur alone or in combinations. While this paper attempts to explicitly define facial expressions as a combination of different AUs, in our approach we try to understand if the feature set extracted from the network used to predict AUs is effective in classifying facial expressions.

Chapter 3

Learning and Detecting Facial Landmarks

Facial Landmarks can be a helpful way to determine different facial expressions and AUs. Slight movements in the positions of basic keypoints such as the eyes, eyebrows, nose, mouth and jawline would cause different AUs and facial expressions.

Convolutional Neural Networks(CNNs) are particularly effective in learning representations of images from raw input images without any extraction of handcrafted features. In this chapter, we describe our experiment to train a CNN to predict the positions of facial landmarks from processed facial images.

3.1 Facial Landmarks

The facial landmarks we try to predict are the positions of eyebrows, around the eyes, the nose, mouth and the jawline. The CSIRO Face Analysis SDK [2] is a software tool, which when given a facial image, first detects the face and then lists the positions of the facial landmarks. Figure 3.1 shows a facial image with the 66 ground truth facial landmarks plotted in red. We use this tool to obtain the ground-truth landmarks for all the facial images. The software does have a few limitations, the most important being that it relies on a frontal face to actually detect a face. Even then, it sometimes fails to detect certain facial images which show large variations from the average facial image. Thus those images in which a face could not be detected are discarded.

3.2 CNN Architecture

We use the open-source MatConvNet [26] to perform this experiment. Our network is primarily based on AlexNet, but we modified the loss layer to reflect the euclidean loss function. AlexNet [13] is an architecture with 5 convolutional layers and 3 fully connected layers. We initialized the network with the pre-trained AlexNet network which was trained on the ImageNet data-set. This network is then trained on the pre-processed images from the data-set which are each re-sized down to 227x227 pixels before they are fed into the network.

3.3 Pre-Processing the Dataset

We first processed the given video files to extract the individual frames. We then used the CSIRO Face Analysis SDK to detect a face in each of the images and obtained the position of a set of landmarks. These landmarks comprise of the position of the eyes, eyebrows, nose, mouth and jawline giving us a total of 66 x-y coordinates which is represented in the form of a vector of size 132.

Using these landmarks we obtained the bounding box surrounding the face, and cropped the images accordingly and re-sized down to a size of 227x227.The facial landmarks were then normalized to fit the bounding box in the following manner:

$$x' = x - min(x)/w * 227, y' = y - min(y)/h * 227 \tag{3.1}$$

where w and h are the width height of the bounding box respectively.

3.4 Training and Testing Validation Split

Similar to [6], we employed a leave-one-subject out testing scheme, and split the MMI Facial Action Unit Dataset into training, testing and validation sets. All images related to subject 1 were categorized as testing images. From the remaining images, 75% of the subjects were used for training, while the remaining 25% of the subjects was used for testing. Thus this experiment was performed in a subject independent manner. We had a total of 17,554 training images, 5851 validation images and 4,186 testing images.

3.5 Evaluation

We evaluate the results by measuring the average detection error and the failure rate of each facial landmark, similar to [24]. These 2 measures determine the accuracy as well as the reliability of the algorithm. The average detection error for each landmark is measured as

$$error = sqrt((x - x')^2 + (y - y')^2)/d, \tag{3.2}$$

where (x',y') is the predicted position for a landmark, while (x,y) is the ground truth position of the landmark, and d is the diagonal of the bounding box surrounding the face. Usually the bi-ocular distance is used as the detection error normalizer, but this metric utilises the positions of the pupils which unfortunately is not one of the landmarks determined by the CSIRO Face Analysis SDK. Using the bi-ocular distance also

assumes that both eyes have been properly detected in a face, which is not always true [22]. Sometimes the software is slightly off with its keypoint estimation, which is why it was safer to normalize the error using the diagonal distance of the facial image. Thus we would calculate this measure for each landmark and determine the failure rate by checking if the error is great than 0.05, which would then by counted as a failure.

3.6 Result and Conclusions

In order to tabulate our results, we decided to group certain landmarks together and report their average detection error as well as the failure rate. We hypothesized that extracting the error rates for landmarks positions clustered around the jawline, eyes, eyebrows, nose and mouth separately would help us determine which areas this method was the most accurate in detecting. Table 3.1, displays the average detection error as well as the failure rate for the 5 groups of landmarks. From these results, we see that the average detection error is highest for landmarks 0-16 which reflect the points around the jawline, while it is lowest for landmarks 37-48 which represent the points around the eyes. This led us to believe that the network is more adept at detecting the eyes in facial images, while it has some difficulty in determining the positions of the other landmarks. This could be due to the fact that several of these facial images portray different facial expressions most of which involve movements around the eyebrows, cheeks, nose and mouth, while there are limited facial expressions which significantly alter the eyes. Movements around these areas would make it difficult to predict the landmarks clustered around these areas. Our results might have been significantly different if the facial images all portrayed neutral faces.

Figure 3.2 displays one of the test images with the ground truth landmarks plotted in green, while the predicted landmarks are plotted in red. From the figure it is apparent that the landmarks related to the eyes align more closely with its ground truth counterparts than the other landmarks which are slightly deviated and show a much higher average error.

We also considered 2 different landmarks and plotted the number of cumulative images for which this particular landmark which fall within a certain error range. One of the landmarks is 36 (represented in Figure 3.3), which corresponds to the corner of the right eye, while we also plotted a similar graph for landmark 6(represented in Figure 3.4) which is located along the jawline. From these two plots we notice that landmark 36 is detected within the 0.05 threshold much more than landmark 6.

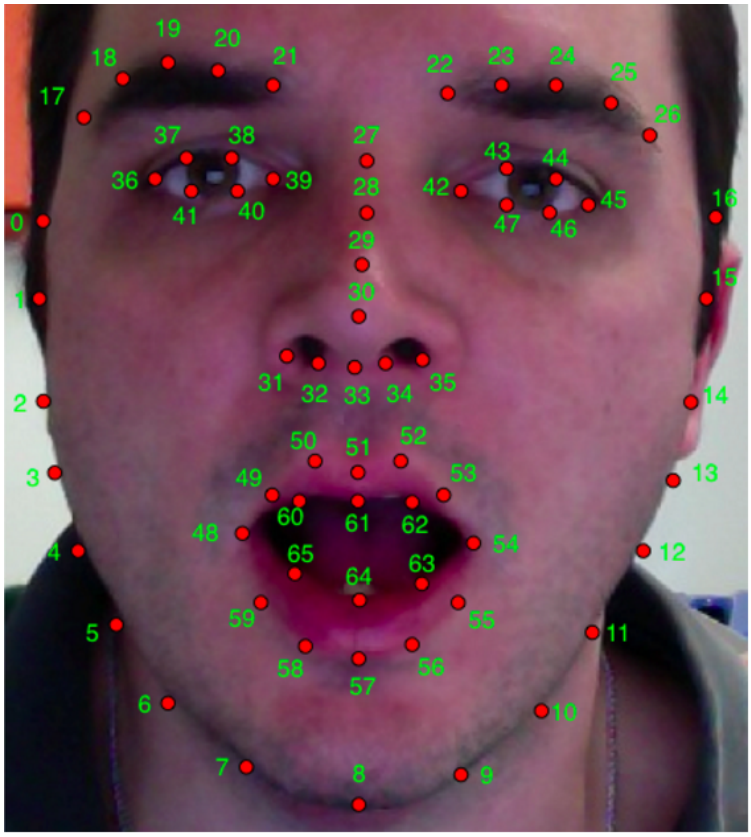


Figure 3.1: Facial Landmark Positions for a facial image as determined by the csiro-face-sdk-analysis

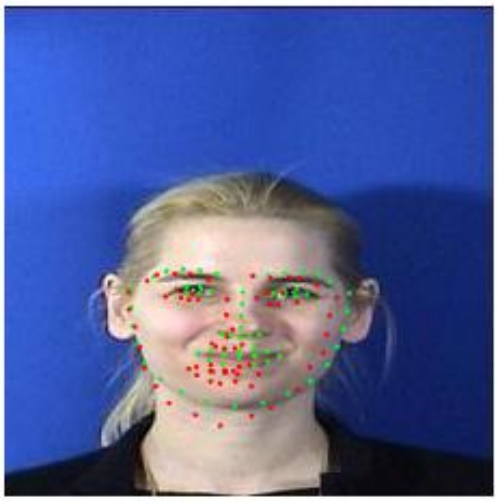


Figure 3.2: Difference In Ground Truth Facial Landmark Positions and Predicted Landmark Positions. The Ground Truth Landmarks positions are marked in red, while the predicted landmarks positions are marked in green.

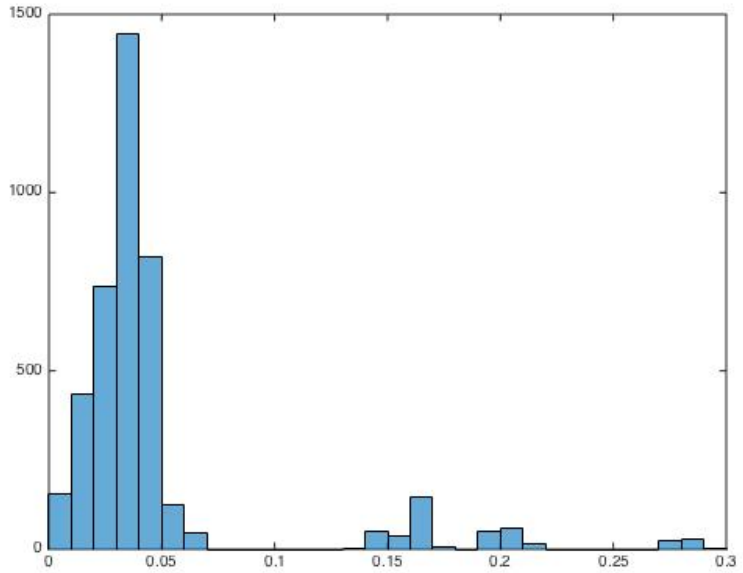


Figure 3.3: Histogram plot for Average Detection Error for Landmarks 33. The x-axis is the diagonal normalized distance from the ground truth and the y axis is the cumulative number of images that fall within this error radius

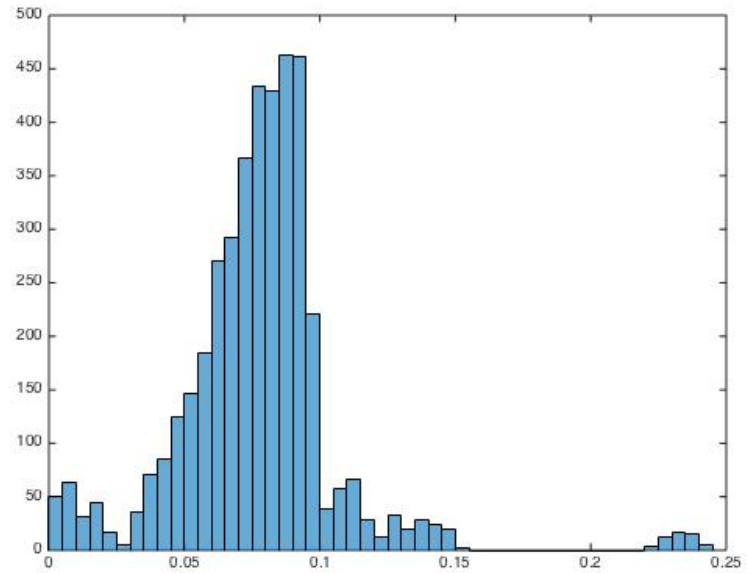
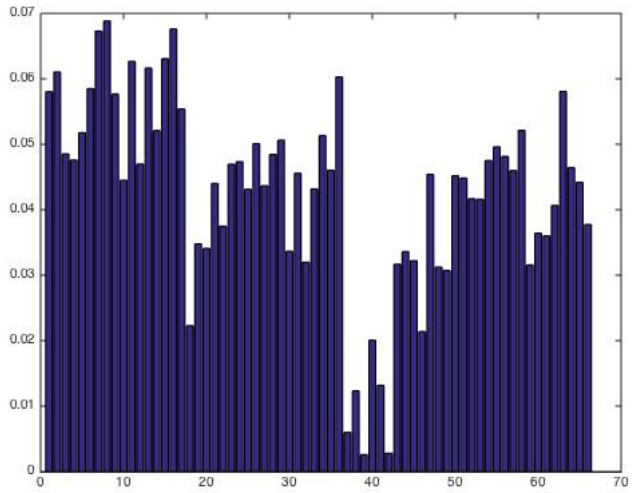


Figure 3.4: Histogram plot for Average Detection Error for Landmark 6. The x-axis is the diagonal normalized distance from the ground truth and the y axis is the cumulative number of images that fall within this error radius



=

Figure 3.5: Average Detection Error for each facial landmark. Landmarks in the range 37-48 (36-47 with zero-indexing) which represent the landmarks around the eyes show the lowest average error.

Landmarks	Average Detection Error	Failure Rate
0-16 (Jawline)	0.058	0.759
17-26 (Eyebrows)	0.040	0.453
27-35 (Nose)	0.0463	0.489
36-47 (Eyes)	0.0251	0.232
48-65 (Mouth)	0.043	0.565

Table 3.1: Average Detection Error and Failure Rate for each Facial Landmark

Chapter 4

Learning and Detecting Action Units

For this experiment, we attempt to train different models to detect the presence/absence of 31 action units(AUs) in Table 4.1 and determine the model that performs best. We consider 3 different feature sets and models for this experiment:

- A CNN trained using the pre-processed facial images
- Ground truth landmarks positions trained using a linear logistic regression model
- Predicted landmark positions from the network used to predict landmark positions, trained using a linear logistic regression model

We hypothesize that the CNN will be able to classify AUs more effectively than the other models.

For all the experiments, we used the same training, testing and validation split. Similar to [6], we employed a leave-one-subject out testing scheme, and split the MMI Facial Action Unit Data-set [19] into training, testing and validation sets. All images related to subject 1 were categorized as testing images. From the remaining images, 75% of the subjects were used for training, while the remaining 25% of the subjects was used for testing. Thus these experiments were performed in a subject independent manner. We had a total of 17,554 training images, 5,851 validation images and 4,186 testing images. For the experiment involving detecting AUs using a linear logistic regression model, the validation images were discarded.

The sparsity in certain AUs proved to be a problem. Previous work showed a slight improvement in performance after balancing the data-set prior to training [23]. Following our mechanism for splitting the input image set into training, testing and validation sets, we noticed an imbalance in the number of images for certain AUs in each set. We solved this issue by balancing the fraction of positive/negative labels for those AUs as much as possible.

To measure the performance of these models on the task of AU detection, we used the following measures %positive examples, accuracy and the Area under the ROC curve (AUC). The accuracy is equal to the percentage of testing examples correctly classified, while the AUC score is defined as the area under the curve arising from the plot of the true positive rate over the false positive rate measures how well the

classifier is performing. Knowing the number of positive examples, will help us determine when the accuracy is trivially high, specifically in cases where the number of negative examples is much higher than the number of positive examples.

One important characteristic of the AUC score is that it is independent of the fraction of the test population in different classes. This makes the AUC score useful for evaluating the performance of classifiers on unbalanced data sets. If the classifier is very good, the true positive rate will increase quickly and the area under the curve will be close to 1. If the classifier is no better than random guessing, the true positive rate will increase linearly with the false positive rate and the area under the curve will be around 0.5.

4.1 From Pre-processed Facial Images

For this experiment we use Convolutional Neural Networks(CNNs) to detect AUs in raw facial images. Since each facial image can have more than one AU associated with it, we use a multi-label logistic loss function and train the CNN using this loss function.

4.1.1 Multi-Label Logistic Log Loss

The problem of detecting multiple AUs in a facial image is a multi-label binary classification task, where the presence of an AU is denoted by +1, while its absence is denoted by -1. The label vector for each image would be represented as a vector binary attributes in (+1,-1).The CNN is trained using the logistic log loss where the prediction scores for a particular label are normalized into a probability. The loss is calculated by

$$L(x, c) = log(1 + exp(-cx)) \tag{4.1}$$

where class c is a assigned a prediction score of x if is present i.e $c = +1$, while class $c = -1$ is a assigned a score of 0.

4.1.2 Preprocessing the Input Datasets

The data-sets were preprocessed by first extracting individual frames from the provided video files. Each video files had a complimentary file detailing the presence of each AU by providing its onset, apex and offset frame numbers. Thus AU labels was assigned as follows:

- For all frames from 0 to (but not including) the onset, the AU label was as assigned as -1.
- For all frames between the onset and offset (including both), the AU was assigned to be +1.

- For all frames after the offset frame, the label was set as -1.

Thus each frame would represent a facial image which would map onto a binary vector of the 31 possible AUs. Additionally, if a particular image had no AUs, its first label (AU0) which determined if a face was neutral, was set to 1. This step was done after assigning the rest of the AUs, since each video could represent more than 1 AU in the same frame.

We also used the CSIRO Face Anaylsis SDK [2] to detect a face in every image and obtain the position of certain facial landmarks. The facial images were then normalized in a manner similar to the experiment for predicting facial landmark positions, so that each facial image was just a bounding box of the face (refer to equation 3.1).

4.1.3 Results

Table 4.2 shows the results of this experiment. In terms of accuracy, AU0 which represents a neutral face showed the lowest accuracy with 66.77%, while the model showed significantly higher accuracies for the remaining AUs. However it must be noted that the AUs which showed high accuracies also showed a low number of positive examples which doesn't necessarily indicate that this model was able to effectively learn these AUs.

However in terms of the AUC score, AU12 (lip corner puller), AU20(lip stretcher) and AU5 (upper lid raiser) showed some of the highest scores with 0.9883, 0.9752 and 0.9479 indicating that this classifier is particularly efficient in detecting these units. The average AUC score for this model across all AUs was 0.7899.

4.2 From Ground Truth Landmarks

Another experiment to detect AUs in facial images is through the position of facial landmarks. Most facial muscle movements involve tiny movements around the eyebrows, eyes, nose, mouth and jawline. Thus we hypothesized that knowing the positions of ground truth facial landmarks might be helpful in detecting the presence of AUs.

4.2.1 Logistic Regression Model

For the purpose of this experiment, the feature set comprised of the 66 normalized ground truth landmarks which are in the form of x-y coordinates. The label vector is represented in the form of binary vector of size

31 (representing the 31 AUs being detected) where +1 indicates the presence of an AU while +1 indicates its absence.

We made use of LIBLINEAR [4] which is an open source library for large-scale linear classification. It supports logistic regression and linear support vector machines. In order to detect AUs in a facial image, we decided to use a logistic regression (LR) model as our primary binary linear classification model. Since we're trying to detect multiple labels (AUs) for each feature vector, we trained a logistic regression model for each AU independently and used said model to detect its presence.

4.2.2 Pre-Processing Data

The ground truth landmarks as obtained by the csiro-face-analysis-sdk [2] indicate the raw positions of the eyebrows, eyes, nose, mouth and jawline from a facial image. The software works by first detecting a face within the image and then marking the locations of the mentioned points. One drawback of this software is that it relies on a fully frontal face to detect facial features. It it possible that the algorithm will fail to detect facial features which includes a slight head tilt. Thus all those images in which a face could not be detected were ignored.

Once we obtained ground truth landmarks positions for all the training and testing images, they had to be normalized across all subjects. This was done in the same manner in which we normalized the landmarks to train our neural net to predict them.

Thus our normalized feature vector which comprised of 66 x-y coordinates (a total of 132 features) mapped either to a binary instance which represented the presence of an AU.

4.2.3 Results

Table 4.2 shows the results of this experiment. In terms of accuracy, AU0 which represents a neutral face showed the lowest accuracy with 61.23%, while the model showed an average accuracy of 89.78% for the remaining AUs.

However in terms of the AUC score, AU12 (lip corner puller) and AU20 (lip stretcher) showed some of the highest scores indicating that the mentioned AUs have very good true positive rates, emphasising that this classifier is particularly efficient in detecting these units. Both these AUs would indicate a very obvious shift in position for the landmarks centered around the mouth, which would make it easy for the classifier to correctly identify them. The average AUC score for this model across all AUs was 0.7543.

4.3 From Predicted Landmarks

For this experiment, we aim to use the landmarks predicted by training the CNN over the pre-processed facial images. Our main aim is to identify how well the predicted landmarks classify action units, in comparison to the ground truth landmarks and network obtained by training the raw facial images.

4.3.1 Logistic Regression Model

We used the same logistic regression model as the previous experiment to classify different AUs. Using the same model with a different feature set but the same label vectors helped us determine if the model obtained by training the predicted landmarks performed better than the model obtained by training the ground truth landmarks.

Similar to the experiment with the ground truth landmarks, the feature set comprised of the 66 normalized predicted landmarks which are in the form of x-y coordinates. The label vector is represented in the form of binary vector of size 31 (representing the 31 AUs being detected) where +1 indicates the presence of an AU while +1 indicates its absence. These landmarks are obtained by running all the images (training and testing) through the network.

4.3.2 Pre-Processing Data

The predicted landmarks for the facial images were determined by running all the facial images (23405 images) through the network used for predicting the positions of facial landmarks. Similar to the ground truth landmarks, the final predicted landmarks comprise of 66 x-y coordinates (a total of 132 features), which are already normalized to fit the bounding box surrounding the face (since the raw facial images as well as their corresponding landmarks were normalized before being fed into the network). Each of these feature vectors then mapped to a binary instance which represented the presence of a particular AU.

4.3.3 Results

Table 4.2 shows the results of this experiment. In terms of accuracy, the results were slightly lower than the accuracies calculated for the CNN and the ground-truth-landmarks model.

However in terms of the AUC score,the average across all AUs was 0.5864 which indicates that this classifier wasn't able to efficiently characterize any of the AUs, and functioned just a little better than a random classifier.

4.4 Discussion

For each of the experiments described above to detect AUs in facial images, we have recorded the number of positive examples, accuracy and AUC score, as well as their corresponding averages across all AUs. Comparing the overall average accuracies for each model indicates that the CNN was the best at identifying the true positive and true negative examples, while the model trained using the predicted landmarks positions performed the worst.

However it is also worth noting that using the accuracy to determine the effectiveness of a model is not completely reliable. There are some imbalances in the data-set, since some AUs have a very high number of negative examples as compared to the number of positive examples, which explains why several AUs show very high accuracies for their respective models, while some AUs(AU0 - neutral face), are more balanced across the data-set. This explains why AU0 shows an average accuracy of 61.82% across the three different models, while for some of the other AUs, the average accuracy is above 90%.

Furthermore, accuracy is a more reliable metric for a constant threshold value. For this reason, we also consider the AUC score which evaluates the area under the graph of the True-Positive rate against the False-Positive rate across a range of thresholds.

Examining the AUC score for each AU across the different models, we notice that some models are particularly efficient in detecting certain AUs. For example, the CNN indicates high AUC scores for AUs 5,12 and 20 while the ground-truth landmark model indicates high AUC scores for AUs 12 and 20. AU12 and AU20 represent very obvious movements in the region around the mouth which could be why both models proved to be efficient in detecting them. However the ground truth landmarks model wasn't as efficient in predicting AU5, which indicates a movement around the upper eye lid. This could be because the ground-truth landmarks-model may be better suited to predicting very obvious changes in positions since it relies solely on landmarks as opposed the entire facial image. This explains why the ground-truth-landmarks-model is better at predicting AU12 and AU20, while being poor at detecting AU5. Additionally the CSIRO Face Analysis SDK is not completely accurate in predicting landmarks positions, which contributes to the lower AUC score.

The average AUC score for each model across the different AUs indicates that the CNN is able to classify AUs better than the ground-truth-landmarks-model and the predicted-landmarks-model. The ground-truth-landmarks model also had a higher average AUC score across all AUs than the predicted-landmarks model, indicating that it is a better model. This could be mainly attributed to the fact that the predicted landmarks differs from the ground truth landmarks with an average error of 0.0425%, and thus the predicted landmarks model couldn't learn the correlation between facial landmarks and AUs.

Our hypothesis that the CNN would perform better than both the landmarks model was true. We think this is because the network processed the entire facial image, without making use of any handcrafted features. It is also apparent from the positions of the landmarks from Figure 3.1, that while they primarily surround the more important features of the face, they fail to learn other parts of the face (like the cheeks and the forehead) and are thus are not robust enough to detect AUs. Thus the CNN can in fact learn features which are more discriminative between the presence/absence of AUs than the features learned by the ground-truth and predicted landmark models.

Action Unit Number	**FACS Name**
0	neutral face
1	Inner Brow Raiser
2	Outer Brow Raiser
4	Brow Lowerer
5	Upper Lid Raiser
6	Cheek Raiser
7	Lid Tightener
8	Lips Toward Each Other
9	Nose Wrinkler
10	Upper Lip Raiser
11	Nasolabial Deepener
12	Lip Corner Puller
13	Sharp Lip Puller
14	Dimpler
15	Lip Corner Depressor
16	Lower Lip Depressor
17	Chin Raiser
18	Lip Pucker
19	Tongue Show
20	Lip Stretcher
21	Neck Tightener
22	Lip Funneler
23	Lip Tightener
24	Lip Pressor
25	Lips Part
26	Jaw Drop
27	Mouth Stretch
28	Lip Suck
30	Jaw Sideways
38	Nostril Dialator
45	Blink

Table 4.1: Action Unit Numbers and their corresponding FACS names

AU	Positive Examples	CNN		Ground-Truth Landmarks		Predicted Landmarks	
		Accuracy	AUC	Accuracy	AUC	Accuracy	AUC
0	1762	0.6677	0.7454	0.6123	0.7243	0.5744	0.5099
1	151	0.9806	0.8842	0.9754	0.7920	0.8956	0.5677
2	149	0.9627	0.7375	0.9544	0.7453	0.8777	0.544
4	145	0.9809	0.8454	0.9654	0.8233	0.9166	0.7055
5	169	0.9981	0.9479	0.9882	0.8188	0.9644	0.6767
6	2	0.9439	0.5841	0.933	0.5231	0.9222	0.5
7	8	0.9885	0.6289	0.9766	0.6455	0.9442	0.5656
8	19	0.9914	0.7367	0.9888	0.7111	0.9776	0.654
9	0	0.9940	0.4999	0.9912	0.4999	0.9434	0.4999
10	13	0.9787	0.7531	0.9677	0.7664	0.9540	0.7112
11	13	0.9792	0.5983	0.9554	0.5555	0.9222	0.6698
12	371	0.9959	0.9833	0.9911	0.9154	0.9321	0.7898
13	154	0.9842	0.8065	0.9777	0.7666	0.9122	0.6433
14	0	0.9924	0.4999	0.9888	0.4999	0.9646	0.4999
15	0	0.9806	0.4999	0.9666	0.4999	0.9891	0.4999
16	113	0.9900	0.6815	0.9664	0.6666	0.9232	0.5673
17	133	0.8703	0.8982	0.8666	0.8777	0.7843	0.7878
18	12	0.9768	0.5979	0.9666	0.5524	0.9234	0.5123
19	0	0.9900	0.4999	0.9545	0.4999	0.9757	0.4999
20	441	0.9881	0.9752	0.9777	0.9333	0.8954	0.6767
21	34	0.9365	0.7829	0.8867	0.6888	0.9187	0.6239
22	23	0.9842	0.7525	0.9775	0.6777	0.9543	0.6734
23	0	0.9859	0.4999	0.9988	0.4999	0.9322	0.4999
24	15	0.9689	0.5899	0.9333	0.5553	0.9543	0.5454
25	168	0.8712	0.6129	0.8777	0.5656	0.9623	0.5467
26	0	0.8868	0.4999	0.7777	0.4999	0.9883	0.4999
27	34	0.9842	0.6299	0.9646	0.5892	0.9432	0.5238
28	78	0.9885	0.5893	0.9776	0.6155	0.9567	0.5344
30	189	0.9761	0.5499	0.9666	0.5332	0.934	0.5099
38	0	0.9902	0.4999	0.9777	0.4999	0.9889	0.4999
45	248	0.9699	0.8868	0.921	0.7554	0.91	0.6935
Average		0.9754	0.7899	0.9423	0.6853	0.9332	0.5864

Table 4.2: Accuracy and AUC Score for CNN, Ground Truth Landmarks Model and Predicted Landmarks Model

Chapter 5

Emotion Detection

Detecting emotions in facial images has always been an interesting task. For this experiment, we try to classify facial images into seven basic emotions : anger, disgust, fear, happiness, sadness, surprise and neutral, by training/testing several different feature vectors to determine the feature set which is the most efficient at classifying emotions. We primarily use the JAFFE [18] dataset for classifying different emotions, while also using the network model a from previous experiment to detect AUs.

5.1 Classification Model and Evaluation

For all of our image representations, we classified images using a one-against-one multi class support vector machines (SVM) with a linear kernel. We chose this particular classifier since it returned the highest accuracy among the different classifiers we tried that were included in the scikit package for python [21]. In order to evaluate our results, we report the average score(accuracy) across 10 iterations of our classification method (using 10-fold cross validation) as well as the cumulative confusion matrix.

5.2 Pre-Processing the Dataset

Similar to the previous experiments to detect facial landmarks and AUs, we processed the images from the JAFFE dataset by using only the bounding box surrounding the face. We also used the CSRIO Face-Analysis SDK [2] to extract the facial landmarks from each image and processed them to reflect the bounding box crop as indicated in equation 3.1. We also re-sized all the images down to 227x227 for use in some of the methods described.

5.3 Feature Sets

We explored several different methods to construct different feature sets to classify the given faces into the following seven facial expressions : anger, disgust,fear, happiness, sadness, surprise and neutral. Our first

approach involved using facial landmarks to classify different facial emotions. Since facial expressions are in fact just small facial muscle movements primarily centered around the important landmarks of the face, we hypothesized that knowing the bare landmark positions or a modification of them will characterize a facial expression accurately.

We also decided to use our results from the neural network which was used to detect AUs. Since each facial expression is basically a combination of certain AUs occurring concurrently, we hypothesized that feature sets extracted from the second last layer of weights from this network would represent the dynamics in different facial expressions more accurately than either plain or processed landmarks. We also studied to see if using a combination of this weight vector along with the plain or modified landmarks would make a better feature vector.

We also decided to explore eigenfaces [1] constructed from the JAFFE images to classify different emotions. Although eigenfaces are primarily used for facial recognition since they in fact eliminate variance arising some emotions, we wondered if they could detect this variance in a data-set of mostly similar looking people. Similarly we also used the weight vector extracted from the second last layer of weights to construct eigenvectors to determine if they performed better than the plain weight vectors.

5.3.1 Plain Facial LandMarks

Our first method involved using plain (and pre-processed) facial landmarks to classify different facial expressions. Thus the feature vector for each image consisted of a total of 66 x-y coordinates and the emotion label corresponding to the facial expression represented. We constructed the feature set by concatenating the feature vectors of each image. After running our classifier on the constructed feature set, we received an accuracy of 81.03% and the confusion matrix in Figure 5.1.

5.3.2 Modified LandMarks

For this approach, we decided to modify our existing landmarks to study the performance of classifying the different facial landmarks using a 12-dimensional feature set. We obtained our resulting feature vectors with single-valued approximations of the following for each image in the data-set: height/width ratio for mouth and eyes; slopes of the eyebrows; concavities of the eyelids, lips, and eyebrows. We then concatenated the feature vectors for each image to obtain the final feature set. After running our classifier on the feature set of processed landmarks, we received an accuracy of 64.48% and the confusion matrix in Figure 5.2.

5.3.3 Weight Vector From Neural Network Model

In order to determine how well our neural net model(which was trained to detect AUs), characterized the appearance of a face, we constructed a feature set by running each image in the JAFFE dataset through the neural network model and extracting the last layer of weights.

We first pre-processed each JAFFE image to convert into a 3-dimensional RGB image of only the cropped face. After running this image through the chosen network, we would extract the weights from the second last layer of the network to produce a feature weight vector of size 4096. This vector along with the corresponding expression label for the image made up the feature vector for each image. The final feature set was constructed by concatenating the feature vector for each image. After running our classifier on this feature set, we obtained an accuracy of 92.86% and the confusion matrix in Figure 5.3.

5.3.4 Weight Vector From Neural Network Model + Plain LandMarks

This feature set was made using the already constructed weight vector for each image, and concatenating this vector with the corresponding facial landmarks for that image. This experiment was mainly done to determine if using a combination of the appearance of the face as well the positions of landmarks would yield better results for classifying emotions.

Thus the final feature vector for each image consisted of its 4096 weight vector + 132 plain landmarks + the expression label for the image. The final feature set was constructed by concatenating the feature vector for each image. After running our classifier on this feature set, we obtained an accuracy of 91.43% and the confusion matrix in Figure 5.4.

5.3.5 Weight Vector From Neural Network Model + Processed LandMarks

This feature set was made using the already constructed weight vector from the chosen neural net, and concatenating each vector with the corresponding calculated modified 12 landmarks. This experiment was mainly done to determine if using a combination of the appearance of the face as well different dynamics in the positions of landmarks would be more effective in classifying different emotions.

Thus the final feature vector for each image consisted of its 4096 long weight vector + 132 plain landmarks + the expression label for the image. The final feature set was constructed by concatenating the feature vector for each image. After running our classifier on this feature set, we obtained an accuracy of 91.97% and the confusion matrix in Figure 5.5.

5.3.6 Eigenfaces

We then explored the method of eigenfaces, to test the performance of our classifier when each image was projected onto a lower dimensional space using Principal Component Analysis (PCA). We performed Principal Component Analysis(PCA) using only our designated training images, but use the resulting eigenvectors to project the training as well as testing data onto a lower dimensional space. We experimented by projecting the images onto different dimensional spaces, but found that we obtained the highest accuracy when the images were projected onto a 100 dimensional space.After running our classifier on the feature set of eigenfaces, we obtained an accuracy of 87.14% and the confusion matrix in Figure 5.6.

5.3.7 Eigenfaces + Processed Landmarks

This feature set was made using the already constructed modified landmarks and eigenfaces. For each image, its feature vector was generated by concatenating its corresponding projection in a 100 dimensional space with its processed landmarks giving rise to a feature vector of size 112. After running our classifier on this feature set, we obtained an accuracy of 88.09% and the confusion matrix in Figure 5.7.

5.3.8 100-D Weight Vector From Neural Network Model

Since classifying different facial expressions using the weight vector generated from the second last layer neural network yielded a high accuracy, we tried to determine if projecting this vector onto a lower dimensional space using PCA would improve the model's performance. Similar to the method regarding Eigenface, we extracted the weight vector from the last layer of the network for each training image to obtain the top 100 eigenvectors, and projected both the training and testing images onto a 100-dimensional space.

Thus our final feature vector for each image consists of its weight vector projected onto a 100-dimensional space with its corresponding emotion label. The final feature set was constructed by concatenating the feature vector for each image.After running our classifier on this feature set, we obtained an accuracy of 91.09% and the confusion matrix in Figure 5.8.

5.3.9 100-D Weight Vector From Neural Network Model + Processed Landmarks

This feature set was made using the already constructed weight vectors projected down to a 100-dimensional space, along with the corresponding modified landmarks for each image, giving us a feature vector of size 112.We performed this experiment to test if using a combination of the appearance of a face projected to a

lower dimensional space, along the with a change in dynamics on the facial landmarks would perform better than the experiment with only the lower dimensional weight vector. We hypothesised that the modified landmarks do in fact convey some extra information to characterize the feature set, which should yield a better result. After running our classifier on this feature set, we obtained an accuracy of 92.28% and the confusion matrix in Figure 5.9.

5.4 Discussion

Table 5.1 summarizes our feature sets and the accuracy obtained for each one. Our initial results from the facial landmarks show that the geometric data of the position of facial features is very good for expression classification. Our processed landmarks had a significant reduction in accuracy, but it is surprising to note that we retain much of the information important for classification in the 132-dimensional data after reducing it to only 12-dimensions.

Our results also show that the weight vector derived from the last layer of weights from the network model which was used to detect AUs, yielded one of the best results in detecting facial expressions. We hypothesize that this is mainly because this network of weights represents the appearance of the face, learned from detecting AUs. Since facial expressions are basically a combination of AUs occurring concurrently, the weight vector would represent the appearance of the face with different AUs, which in turn would define the appropriate facial expression. Concatenating the plain and modified landmarks to this vector didn't yield a significant difference in results.

We found it quite unusual to discover that the eigenfaces methods produced such a high accuracy. Upon examination of the top eigenfaces of the entire dataset we noticed an interesting pattern. Many of the top eigenfaces had high components along the eyes, eyebrows, and mouth. These areas are of high importance to discerning a facial expression, and therefore representing an image by its position along these axes would likely be very useful for facial expression classification. However it must be noted that these are very atypical eigenfaces. The eigenface method became popular as a means of eliminating variance due to facial expression with the goal of classifying faces by the people to which they correspond. When using the eigenface method on a very diverse data-set, the top eigenfaces tend to correspond to directions of lighting along with features such as jaw width and nose size that represent variance in the appearance of different people. We hypothesize that it was due to two key qualities of our data-set that such atypical patterns were produced. One important factor is that each face is lit from roughly the same angle, eliminating variance due to lighting. The second is that the population from which the sample of people was drawn was a very narrow group. Variance in

facial appearance due to age, gender, culture, and race was minimized as each person in the data-set was an adult Japanese woman. Thus the only variance which was evaluated in this method is the difference in appearance due to different facial emotions.

We also noticed that while the modified landmarks feature set does not perform so well on its own, it performs better when combined with either eigenfaces or the 100-D vector constructed from the CNN feature vector. This could be because the combination of the features of the face and modified landmarks provide more information with regards to the dynamics of the face, compared to the modified landmarks by themselves. Modified landmarks don't include any information about the nose, cheeks and forehead, which contributes to its inferior performance. Ultimately combining these models will allow the classifier to learn the variances in facial expressions better.

We also received results similar to the feature consisting of the plain weight vector when we projected this weight vector from each image onto a lower dimensional space. We determined that the classifier is more adept at classifying emotions when presented feature sets which characterize the dynamics of different muscle movements in the face. The classifier would be able to more efficiently learn variances in appearances of faces by means of facial expressions. While this is largely because of the facial images in the data-set being quite similar to each other since they comprise of a very uniform group of subjects, we were able to determine that learning the presence of tiny muscle movements(AUs) on the face is an important precursor to determining facial expressions.

Feature Set	**Accuracy**
Plain LandMarks	81.03%
Modified LandMarks	64.48%
Weight Vector extracted from neural network	92.86%
Weight Vector extracted from neural network + Plain Landmarks	91.43%
Weight Vector extracted from neural network + Modified Landmarks	91.97%
Eigenfaces from plain images	87.14%
Eigenfaces + Modified landmarks	88.09%
CNN features projected to 100-D space	91.09%
CNN features projected to 100-D space + Modified Landmarks	92.28%

Table 5.1: Average Accuracy from classifying the following feature sets

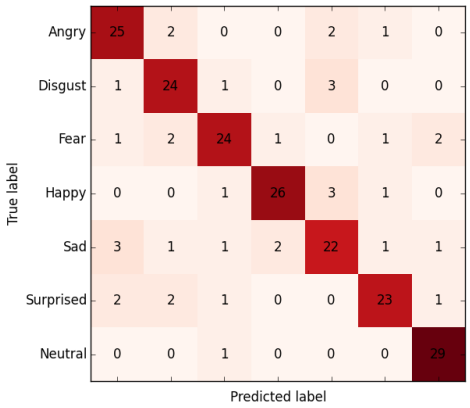


Figure 5.1: Plain Landmarks

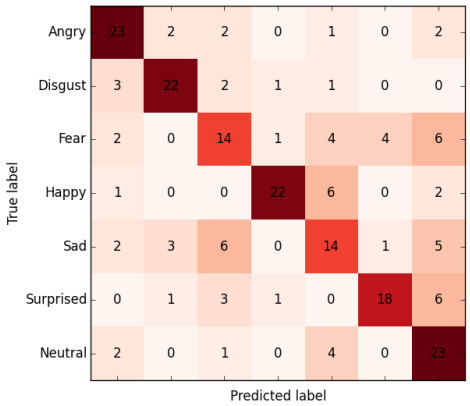


Figure 5.2: Modified Landmarks

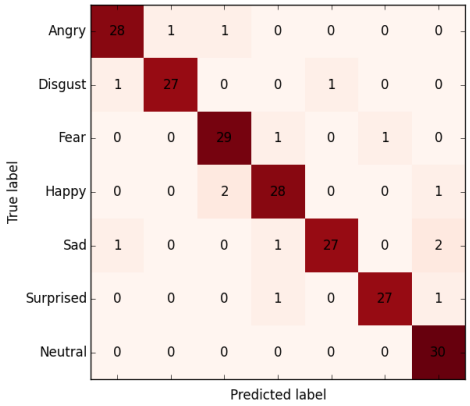


Figure 5.3: Weight Vector From Neural Net

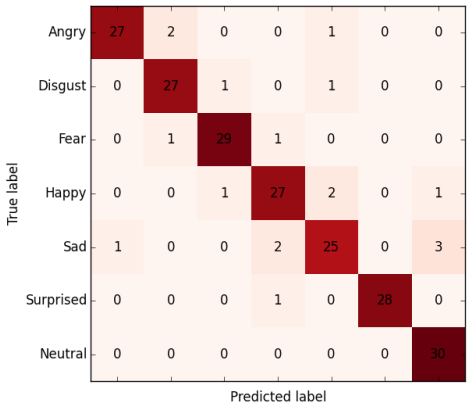


Figure 5.4: Weight Vector + Plain Landmarks

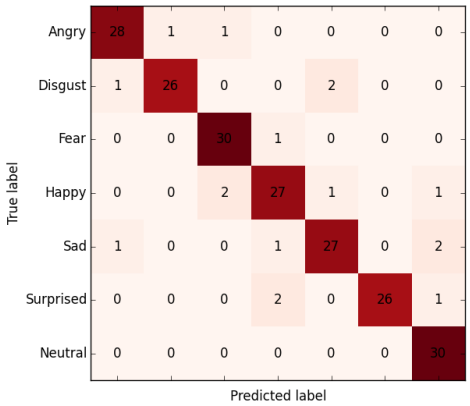


Figure 5.5: Weight Vector + Modified Landmarks

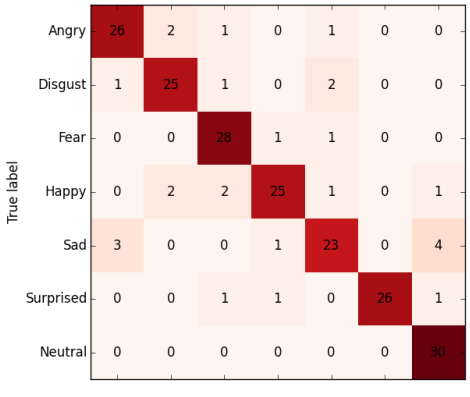


Figure 5.6: Plain Eigenfaces

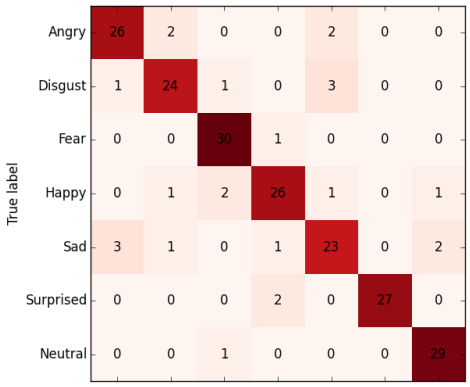


Figure 5.7: Plain Eigenfaces + Modified Landmarks

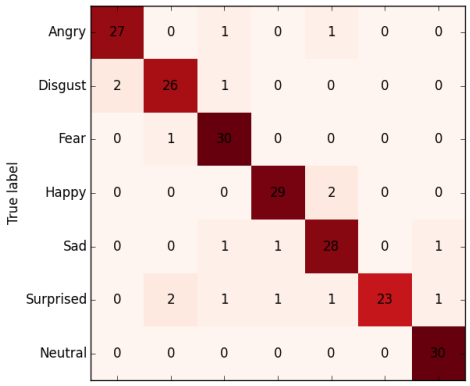


Figure 5.8: 100-Dimensional Weight Vector

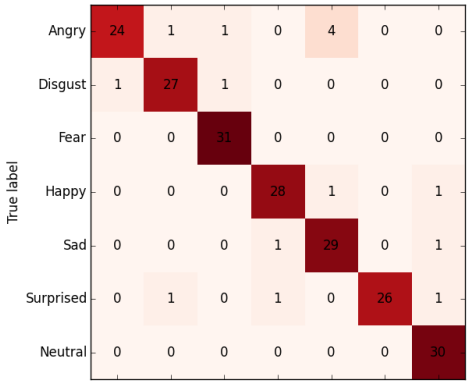


Figure 5.9: 100-Dimensional Weight Vector + Modified Landmarks

Chapter 6

Conclusion

In this paper, we explored different feature sets and models to predict facial landmarks, AUs and emotions, and used the results from intermediate experiments to try and improve baseline performances. We learn that CNNs are highly effective in learning the appearance as well as the different dynamics which characterize a face, and are quite successful in predicting facial landmarks and classifying AUs. Landmark models also worked fairly well in detecting AUs and emotions, but we learnt their feature sets are not robust enough to detect tiny facial muscle movements in areas not concentrated around the principle regions of the face. We also learnt from our experiments with detecting emotions in faces that feature sets which can effectively outline the different properties of a face are more useful in classifying facial emotions than either images or facial landmarks.

References

[1] P. N. Belhumeur, J. a. Hespanha, and D. J. Kriegman. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. pages 45–58, 1996. URL http://dl.acm.org/citation.cfm?id=645309.648965.

[2] M. Cox, J. Nuevo, J.Saragih, and S. Lucey. Csiro face analysis sdk. 2013.

[3] P. Ekman, W. Friesen, and J. Hager. Facial Action Coding System (FACS): Manual. 2002.

[4] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874, 2008.

[5] B. Fasel. Head-pose invariant facial expression recognition using convolutional neural networks, 2002.

[6] S. Ghosh, E. Laksana, S. Scherer, and L.-P. Morency. A multi-label convolutional neural network approach to cross-domain action unit detection. 2015.

[7] R. B. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. *CoRR*, abs/1311.2524, 2013. URL http://arxiv.org/abs/1311.2524.

[8] A. Gudi, H. E. Tasli, T. M. den Uyl, and A. Maroulis. Deep learning based facs action unit occurrence and intensity estimation. 06:1–5, May 2015. doi: 10.1109/FG.2015.7284873.

[9] S. Jaiswal and M. F. Valstar. Deep learning the dynamic appearance and shape of facial action units. pages 1–8, 2016. doi: 10.1109/WACV.2016.7477625. URL http://dx.doi.org/10.1109/WACV.2016.7477625.

[10] H. Jung, S. Lee, S. Park, I. Lee, C. Ahn, and J. Kim. Deep temporal appearance-geometry network for facial expression recognition. *CoRR*, abs/1503.01532, 2015. URL http://arxiv.org/abs/1503.01532.

[11] S. E. Kahou, C. Pal, X. Bouthillier, P. Froumenty, c. Gülçehre, R. Memisevic, P. Vincent, A. Courville, Y. Bengio, R. C. Ferrari, M. Mirza, S. Jean, P.-L. Carrier, Y. Dauphin, N. Boulanger-Lewandowski, A. Aggarwal, J. Zumer, P. Lamblin, J.-P. Raymond, G. Desjardins, R. Pascanu, D. Warde-Farley, A. Torabi, A. Sharma, E. Bengio, M. Côté, K. R. Konda, and Z. Wu. Combining modality specific deep neural networks for emotion recognition in video. pages 543–550, 2013. doi: 10.1145/2522848.2531745. URL http://doi.acm.org/10.1145/2522848.2531745.

[12] P. Khorrami, T. L. Paine, and T. S. Huang. Do deep neural networks learn facial action units when doing expression recognition? pages 19–27, 2015. URL http://dblp.uni-trier.de/db/conf/iccvw/iccvw2015.htmlKhorramiPH15.

[13] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. pages 1097–1105, 2012. URL http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf.

[14] Y. li Tian, T. Kanade, and J. E. Cohn. Recognizing action units for facial expression analysis.

[15] P. Liu, S. Han, Z. Meng, and Y. Tong. Facial expression recognition via a boosted deep belief network. June 2014.

[16] 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.

[17] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba. Coding facial expressions with gabor wavelets. pages 200–, 1998. URL http://dl.acm.org/citation.cfm?id=520809.796143.

[18] M. J. Lyons, S. Akemastu, M. Kamachi, and J. Gyoba. Coding facial expressions with 214 gabor wavelets. *Coding Facial Expressions with 214 Gabor Wavelets, 3rd IEEE International Conference on Automatic Face and Gesture Recognition*, 1998.

[19] L. Maat, R. Sondak, M. Valstar, M. Pantic, and P.Gaia. Mmi face database. URL http://www.ai.rug.nl/conf/bnaic2004/cp/c61b.pdf.

[20] M. H. Mahoor, M. Z. andKevin L. Veon, S. M. Mavadati, and J. F. Cohn. Facial action unit recognition with sparse representation. *Automatic Face Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference.* URL http://ibug.doc.ic.ac.uk/media/uploads/documents/cvpr2010$_w$.pdf.

[21] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in python. *J. Mach. Learn. Res.*, 12:2825–2830, Nov. 2011. ISSN 1532-4435. URL http://dl.acm.org/citation.cfm?id=1953048.2078195.

[22] D. Ramanan. Face detection, pose estimation, and landmark localization in the wild. pages 2879–2886, 2012. URL http://dl.acm.org/citation.cfm?id=2354409.2355119.

[23] G. Sandbach, S. Zafeiriou, and M. Pantic. Local normal binary patterns for 3d facial action unit detection. pages 1813–1816, Sept 2012. ISSN 1522-4880. doi: 10.1109/ICIP.2012.6467234.

[24] Y. Sun, X. Wang, and X. Tang. Deep convolutional network cascade for facial point detection. *CoRR*, abs/1406.4773, 2014. URL http://arxiv.org/abs/1406.4773.

[25] Y.-l. Tian. Evaluation of face resolution for expression analysis. pages 82–, 2004. URL http://dl.acm.org/citation.cfm?id=1032636.1032973.

[26] A. Vedaldi and K. Lenc. Matconvnet – convolutional neural networks for matlab. 2015.

[27] Y. Wu, T. Hassner, K. Kim, G. Medioni, and P. Natarajan. Facial landmark detection with tweaked convolutional neural networks. *arXiv preprint arXiv:1511.04031*, 2016. URL http://www.openu.ac.il/home/hassner/projects/tcnn$_l$*andmarks*.