Northumbria University
NEWCASTLE

DEEP HUMAN FACE ANALYSIS AND MODELLING

GARY STOREY

PhD

2019

DEEP HUMAN FACE ANALYSIS AND MODELLING

GARY STOREY

A thesis submitted in partial fulfilment of the requirements of the University of Northumbria at Newcastle for the degree of Doctor of Philosophy

Faculty of Engineering and Environment

October 2019

Abstract

Human face appearance and motion play a significant role in creating the complex social environments of human civilisation. Humans possess the capacity to perform facial analysis and come to conclusion such as the identity of individuals, understanding emotional state and diagnosing diseases. The capacity though is not universal for the entire population, where there are medical conditions such prosopagnosia and autism which can directly affect facial analysis capabilities of individuals, while other facial analysis tasks require specific traits and training to perform well. This has lead to the research of facial analysis systems within the computer vision and machine learning fields over the previous decades, where the aim is to automate many facial analysis tasks to a level similar or surpassing humans. While breakthroughs have been made in certain tasks with the emergence of deep learning methods in the recent years, new state-of-the-art results have been achieved in many computer vision and machine learning tasks. Within this thesis an investigation into the use of deep learning based methods for facial analysis systems takes place, following a review of the literature specific facial analysis tasks, methods and challenges are found which form the basis for the research findings presented.

The research presented within this thesis focuses on the tasks of face detection and facial symmetry analysis specifically for the medical condition facial palsy. Firstly an initial approach to face detection and symmetry analysis is proposed using a unified multi-task Faster R-CNN framework, this method presents good accuracy on the test data sets for both tasks but also demonstrates limitations from which the remaining chapters take their inspiration. Next the Integrated Deep Model is proposed for the tasks of face detection and landmark localisation, with specific focus on false positive face detection reduction which is crucial for accurate facial feature extraction in the medical applications studied within this thesis. Evaluation of the method on the Face Detection Dataset and Benchmark and Annotated Faces in-the-Wild benchmark data sets shows a significant increase of over 50% in precision against other state-of-the-art face detection methods, while retaining a high level of recall. The task of facial symmetry and facial palsy grading are the focus of the finals chapters where both geometry-based symmetry features and 3D CNNs are applied. It is found through evaluation that both methods have validity in the grading of facial palsy. The 3D CNNs are the most accurate with an F1 score of 0.88. 3D CNNs are also capable of recognising mouth motion for both those with and without facial palsy with an F1 score of 0.82.

Contents

List of Figures

List of Tables

Acknowledgements

This thesis is dedicated to my parents, without their support I would not have had the opportunity to return to education and be in a position to complete this thesis. Also to my partner for her support and understanding throughout.

I would like to express my sincere appreciation to my supervisors Dr. Richard Jiang and Prof. Ahmed Bouridane for their continued encouragement, guidance and support during all stages of this research.

Declaration

I declare that the work contained in this thesis has not been submitted for any other award and that it is all my own work. I also confirm that this work fully acknowledges opinions, ideas and contributions from the work of others.

Any ethical clearance for the research presented in this thesis has been approved. Approval has been sought and granted by the *University Ethics Committee* on *01/03/2016*.

I declare that the Word Count of this thesis is 35,020 words.

Name: Gary Storey

Signature:

Date: 4 October 2019

Published Contributions

Storey, G., Jiang, R. and Bouridane, A. (2017), 'Role for 2D image generated 3D face models in the rehabilitation of facial palsy', Healthcare Technology Letters 4 (4), 145–148.

Storey, G., Jiang, R., Bouridane, A., Dinakaran, R. and Li, C. (2018), "Deep neural network based multi-resolution face detection for smart cities". International Conference on Information Society and Smart Cities.

Storey, G., Bouridane, A. and Jiang, R. (2018), 'Integrated Deep Model for Face Detection and Landmark Localization From "In The Wild" Images',IEEE Access 6, 74442-74452.

Storey, G. and Jiang, R. (2019), Face Symmetry Analysis Using a Unified Multi-task CNN for Medical Applications,in 'Intelligent Systems and Applications', pp. 451–463.

Storey, G., Jiang, R., Keogh S., Bouridane, A. and Li, C. (2019), '3DPalsyNet: A Facial Palsy Grading and Motion Recognition Framework Using Fully 3D Convolutional Neural Networks', IEEE Access 7, 121655-121664.

Chapter 1

Introduction

1.1 Background

The human face provides an extremely rich and complex source of visual data allowing humans to interact in complex social environments. While recognition is one element of this visual data allowing humans to identify individuals through facial characteristics, the complex neural innervation, musculature and flexibility of the human face has evolved to also provide a wide variety of motion signals (Tovée, 1995). These motion signals in isolation can convey conscious and subconscious information. Motion signals along with static recognition in many circumstances can be successfully interpreted by other humans through complex neural mechanisms within the brain (Simion and Giorgio, 2015).

Research into face perception concerns the study of the neural mechanisms and how these relate to the understanding and interpretation of the face has identified that humans are able to recognise other human faces are present from birth. During the infancy this neural system is continually developing and by 6 months a child displays the capacity to differentiate between familiar and non-familiar faces. This recognition capability is significant as it provides a survival mechanism warning a child to potential dangers (Simion and Giorgio, 2015). As people further develop their neural processes the capability to extract further recognition information from faces and add learnt labels to categorise this data is gained. The lowest level of recognition is the identification of a specific individual, while at higher levels there is also the capacity to estimate age, gender, ethnicity and elements of lifestyle. Humans are not only limited to these recognition skills, nonverbal

communication is an extremely important and well researched area of facial signalling, which has been of interest to the scientific community since the late 1800's when both William James and Charles Darwin published the books "Principles of Psychology" (James, 1913) and "The Expression of the Emotions in Man and Animals" (Darwin and Darwin, 2009) respectively. Nonverbal communication is used to signal the emotional and physical state of an individual (Riggio and Feldman, 2005), Paul Ekman a prominent researcher provided strong evidence that facial expressions are a universal phenomenon across all humans from his observations of nonverbal behaviour in secluded tribes in the South Pacific (Ekman, 2013). Facial expressions comprise one area of nonverbal communication which is believed to be a key to understanding the emotional state of an individual, with early studies finding 55% of messages relating to feeling are through facial expression (Sumathi et al., 2012). Facial expressions include both macro-expression and micro-expressions, where macro-expression have a duration of between 0.5 to 4 seconds and are visually obvious, examples include the defined six basic emotions these being anger, disgust, fear, happiness, sadness and surprise (Ekman, 2013). Micro-expressions are of a much shorter duration of less than 0.5 of a second and are not visually obvious and can be missed. Atypical motions are also important facial signals, these can be interpreted to indicate a potential medical condition (Lees, 2002). Within the medical domain there are many pathology's that affect the nerves and therefore the muscular motion of the face to varying degrees, for instance stroke and facial palsy have an immobilising affect on a single side of the face (Banks et al., 2015), while Parkinson's disease can affect the capability to produce facial expression through a condition called mask face (hypomimia) (Wu et al., 2014).

While not a exhaustive list of all facial signals that encompass the domain of face perception, the areas touched upon above highlight how integral these are to the complex social structures humans have developed and why this area of research has been prominent across many different disciplines for many decades.

1.2 Motivation

While face perception and the associated neural mechanisms in humans is intrinsic, dependant upon the specific task the successful interpretation can vary. Not all humans neural mechanisms develop identically with experience and neurological conditions impacting this development. Neu-

rological conditions such as face blindness (prosopagnosia) and autism affect the capability of the individuals to recognise identity and emotion respectively, making certain social interactions difficult (Corrow et al., 2016; Baron-Cohen et al., 2009). Experience based development is potentially wide ranging due to the way in which humans learn through experience from their environment, while there are many common and visible facial signal experiences the majority of humans will encounter, there are also less common or visible facial signals such as atypical motion or micro-expressions. To experience these facial signals and develop skills to interpret them requires specialised training which can be both expensive and time consuming. In the security domain for example micro-expressions have been show to be an indicator of a persons true intentions and have been applied in lie detection (Vrij, 2008), though even with training the recorded accuracy of an expert is only 47% for the detection of micro-expressions (Pfister et al., 2011). While in the medical domain and more specifically in the diagnosis and rehabilitation of those with facial palsy, regular tracking of the patients recovery and quality qualitative feedback produces the best clinical outcomes (Lindsay et al., 2010; Banks et al., 2015). This though relies on the availability and location of the trained medical specialists and patients which may not facilitate frequent checkups.

These areas where interpretation is difficult for humans provides a motivation to research into computer based facial analysis systems and their associated methods. At this point in time there has already been decades of research undertaken into various facial analysis tasks, much of this work has been dedicated to using computer vision and machine learning techniques for fundamental face analysis tasks which human are generally good at performing. Excellent progress has been made in tasks such as general face detection, facial recognition for bio-metric based security system and emotion recognition in the specific domain of macro-expressions (Ramanan et al., 2012; Wagner et al., 2012; Li et al., 2015). While other more difficult tasks such as micro-expression analysis and atypical face analysis have seen some initial success (Pfister et al., 2011).

In recent years the availability of large volumes of data and advancements in computer hardware specifically in Graphics Processing Units (GPUs) have led to the rise of deep learning methods. Within the computer vision field the CNN has ushered in significant advancements in a number of problem areas. The popularity of the CNNs was kick started with the introduction of the AlexNet (Krizhevsky et al., 2012) architecture in 2012 which comprehensibly beat all previous methods

on the ImageNet Large-Scale Visual Recognition Challenge (ILSVRC) by a margin of 10%. This object detection challenge consists of 1.2 million images and 1000 classes making it extremely difficult as a benchmark. Year-on-year deeper and more complex CNN architectures have been introduced including Visual Geometry Group (VGG) (Simonyan and Zisserman, 2015), Residual Networks (ResNets) (He et al., 2015) and Inception (Szegedy et al., 2015) which have been even more successful in the ILSVRC. While initially the complexity of training these models was restrictive due to the computational overhead required, over time the availability of pre-trained models and the use of transfer learning have allowed researchers access to the capabilities of these advanced CNN architectures. All of these recent advancements have opened up exciting possibilities to further advance the research in the area of facial analysis, specifically the opportunity to develop novel methods and applications.

1.3 Research Aims and Objectives

In the previous section the research motivations are explored, highlighting the deep diversity of facial analysis tasks. It is clear from the previous research that there are many computer vision and machine learning methods, this includes those that apply deep learning that have been proposed for a range facial analysis task. While some areas are well established the sheer diversity of the potential applications and the introduction of new state-of-the-art deep learning based methods opens up research opportunities. The aim of this research is to investigate the methods and applications of facial analysis systems, with the purpose of identifying a facial analysis application area where deep learning methods can be used to advance the current research. There is also scope to explore the current state-of-the-art methods commonly applied to facial analysis systems and identify challenges where further research could produce performance improvements. In order to achieve these aims, the following major objectives are defined:

1. Investigate the key techniques applied within face analysis systems. This includes the computer vision tasks of face detection and landmark localisation commonly used as the initial stages in facial analysis systems. The key methods from the previous research for the aforementioned tasks and the associated challenges, data sets and metrics used to evaluate task performance will be included within the investigation.

2. Investigate different face analysis system applications. This investigation will look at a variety of tasks that encompass the facial analysis domain from the previous literature, also identifying the methods applied and availability of data sets in these areas.

3. Identify from the investigation in objective 1 suitable area where a challenge exists, and propose novel methods to overcome the identified challenge. Evaluate the proposed using appropriate methodologies against the current state-of-the-art methods using suitable benchmark data sets.

4. Identify from the investigation in objective 2 a suitable facial analysis task and further investigate methods to perform this specific task. Produce suitable evaluations of the proposed methods on relevant data sets and produce a set of conclusions.

1.4 Research Contribution

To summarise the research presented within this thesis makes contributions in the following areas:

- A published contribution was made (Storey et al., 2017) which demonstrates a potential method for generating 3D face models and using these for facial palsy grading.

- A published contribution was made (Storey and Jiang, 2019) which demonstrates a initial approach to face detection and symmetry analysis using a unified CNN.

- A significant published contribution is made within the face detection task of facial analysis. Where a novel deep learning method is proposed and validated (Storey et al., 2018).

- A significant contribution is made investigating and proposing a novel method for using geometry-based features for the facial analysis task of facial palsy grading. The investigation also provides a insight into the issues with landmark localisation accuracy for a number of methods when applied to this specific task. This contribution has been adapted and is under review for publication.

- A significant contribution is made investigating and proposing novel methods for the facial analysis task of facial palsy grading and also mouth motion analysis. This contribution has been adapted and is under review for publication.

1.5 Thesis structure

The remainder of this thesis is structured as follows:

Chapter 2 - Literature Review: This chapter discusses many of the areas of facial analysis, starting with an overview of the typical structure of a facial analysis systems, then delving into the previous literature on the specific tasks and techniques researched in the fields of machine learning and computer vision for human face analysis and closely related areas. Specifically the tasks of face detection, landmark localisation, facial recognition, facial emotion recognition, medical application and general action recognition are detailed. For each of these tasks a descriptions of the challenging areas and the significant works to date is detailed.

Chapter 3 - Faster R-CNN: In recent years unified object detection networks have became popular within the computer vision research. In this chapter the focus is specifically on the Faster R-CNN method identified within the literature review and applied throughout the remaining chapter of this thesis. This chapter specifically provides a detailed overview of the model, including the composition of the architecture and the loss functions used in training. As Faster R-CNN is a general method for object detection a section is dedicated on how to re-train this model for the specific task of face detection.

Chapter 4 - Stacked Hourglass Network: This deep learning architecture was identified within the literature review as playing a significant role in the capability of providing accurate landmark localisation. Within this chapter the hourglass architecture is detailed, this is then expanded to the stacked hourglass. A specific section details the Face Alignment Network which applies the stacked hourglass architecture for landmark localisation and finally a overview of the network structure which can provide predicted 2D facial landmark depth mapping these to 3D landmarks is discussed.

Chapter 5 - Unified Multi-task Faster R-CNN: This chapter describes an initial investigation which proposes a unified multi-task Faster R-CNN based method for the task of face detection and facial symmetry analysis. The method is detailed and evaluated, where a number of conclusion and limitations are discussed which shaped the remaining research presented in this thesis.

Chapter 6 - Integrated Deep Model: A method named the Integrated Deep Model is proposed within this chapter for the task of face detection and landmark localisation. An overview of the method is given, detailing how the Faster R-CNN and stacked hourglass networks are cascaded. An evaluation is presented on a number of benchmark data sets, the evaluation shows that the proposed method provides state-of-the-art precision in the face detection task and a competitive level of recall. While it also maintains a high accuracy for the task of landmark localisation.

Chapter 7 - Geometry-based Symmetry Features for Facial Palsy Grading: In this chapter the specific task of facial palsy grading from video sequences is addressed. Investigations are detailed to understand the capacity of various landmark localisation techniques to generalise to those individual with facial palsy accurately. A method to extract geometry-based symmetry features from facial landmarks is also proposed and evaluated.

Chapter 8 - 3D CNNs for Facial Palsy Grading and Mouth Motion Recognition: The task of facial palsy grading from videos sequences is revisited in this chapter, while also addressing the identified challenge of mouth motion recognition. A proposed method using 3D CNNs is detailed and evaluated for both of the tasks.

Chapter 9 - Conclusion: This chapter critically assesses the methods proposed across the chapters of this thesis. The contributions of this thesis are outlined and discussed. Finally, some limitations and recommendations for future work are proposed.

Chapter 2

Literature Review

2.1 Introduction

This chapter focuses on the investigation of the literature surrounding computer based facial analysis systems, highlighting the significant previous research through a discussion of the stages, methods and tasks applied within both facial analysis and associated domains. The aim of this chapter is to identify relevant methods, challenges and tasks as defined in objectives 1 and 2 in section 1.3 the outcomes will then shape the research direction detailed within the following chapters of this thesis.

The chapter begins with an overview of facial analysis systems (section 2.2) detailing further motivation for their use of computer visions and machine learning methods, while providing an overview of the stages commonly applied in facial analysis systems. Following this the identified common computer vision tasks applied in each stage of a facial analysis system are expanded upon with a literature review highlighting the significant methods present within research. Face detection (section 2.3) is the first of these tasks covered, this section looks primarily at the challenges of face detection and the methods that have so far been applied within the research to overcome these, while also identifying challenging areas where there is still scope for improvement. The task of landmark localisation (section 2.4) is covered with a detailed review of both the traditional methods applied to this task and also the current state-of-the-art deep learning based methods. Facial analysis tasks are then explored, specifically the areas of facial recognition, facial expression recognition and medical tasks. For each of these specific areas the literature is reviewed

with respect to traditional and deep learning based methods, these include both image and video based systems. Finally a review of the general action recognition methods from videos are discussed with the aim to identify methods that have yet to be applied in facial analysis but have been successfully used in a related domain.

2.2 Facial Analysis Systems

Facial analysis is a term given to the visual analysis of the human face in both static and dynamic states. There are many goals of facial analysis including identification, behavioural predictions, emotional recognition and medical diagnosis. While the capacity to perform some face analysis tasks is intrinsic to humans there is strong motivation to apply computer vision and machine learning to automate these tasks to varying degrees. Two of the main motivational factors driving this are difficulty and cost. Certain face analysis task are difficult for even human observers, take for instance lie detection in the security domain where the presence of micro-expressions have shown to highlight if an individual is masking their true intentions. While it is possible to train a person to perform micro-expression facial analysis not all individuals can develop these skills adequately and even then the error rate of the human observer is still reportedly high (Pfister et al., 2011). In the medical domain certain diseases such as facial palsy require trained medical professionals for visual diagnosis, the availability in terms of time and location of both the patient and the clinicians, while there has also been some intra-observer discrepancies between medicals professionals. These motivations and many others have interested the research community and over the last three decades many computer vision methods have been applied within these domains in the form of facial analysis systems.

Dependant upon the specific goal a face analysis system will often employ a number of distinct stages (Figure 2.1 provides an interview of a typical face analysis system). In the scenario of unconstrained images in which the persons face may be located at any position, the first stage of a facial analysis system is the task of face detection. In some scenarios a face detection stage may be omitted, this would include situations where the images were taken in a constrained environment allowing the position of the face and the illumination to be controlled. Once the task of face detection has been completed an intermediate stage of landmark localisation is often performed. Performing landmark localisation provides a more detailed understanding of the individuals facial

structure. Where face detection simply provides a set of bounding box co-ordinates from the image space containing the face, landmark localisation provides specific location information of important facial features such as the eyes, nose, mouth and jaw line of the individual. In some facial analysis systems this level of facial information may not be required therefore this stage maybe omitted. The final stage of the facial analysis system is concerned with the prediction which is goal specific, for example facial identification would at this stage predict the name of the individual, while emotional recognition would predict a relevant emotion class label. At this stage facial information from the previous stages is used to perform feature extraction which are used for the final prediction. While there are numerous methods which can be applied to perform feature extraction, there are some factors which can affect which method can be used. One impacting factor is whether the input of the facial analysis system is a single image or a video in the form of a sequence of frames, where single images can be represented by spatial features while the added dimension of time with a sequence can benefit from spatio-temporal features (section 2.5 provides further discussion of feature types). Commonly facial analysis systems consists of the aforementioned stages which each apply different methods and are stacked in a linear fashion as shown in Figure 2.1.

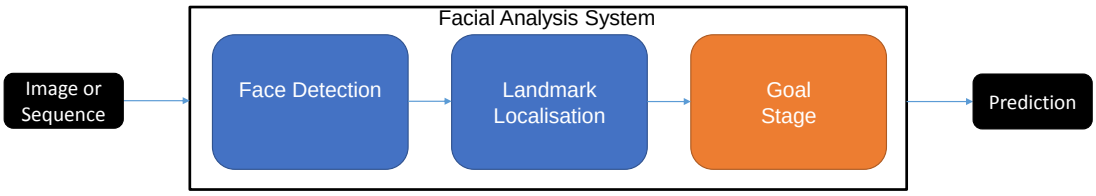

Figure 2.1: Facial Analysis System Framework Overview.

2.3 Face Detection

A distinctive subset of the object detection task, face detection is the process of identifying a human face within an image. This is a fundamental task which stands as the foundation for many facial analysis systems. While this task has a simple premise, there are many complications which make this challenging. The challenges associated with face detection are best highlighted in what are defined as 'in the wild' images, these are unconstrained meaning they contain many elements which effect facial appearance including extreme pose, scale variations, position within the frame,

number, occlusion, resolution variations, de-focus and illumination. Recently as face detection recall has increased there is also a challenge to increase the precision of methods through the reduction of false positive detection's. In this section each of these challenges is broken down to provide a brief overview of the issue.

A general overview of the process of face detection is as follows. A set of features are learnt from a training sample of data containing faces, these extracted features are representative of the human face. A large number of regions of interest from the search image are transformed to a feature representation which can be compared with the previously learnt face feature representation. The comparison output for a given region of interest is then output, in general this is either a probability or a score which is compared with a previously learnt threshold value.

As a problem domain that has been studied for decades there is a large body of research on this task and many different techniques have been proposed. This literature review mainly focuses on the historically important breakthroughs which provide context to the processes involved in face detection and more recent state-of-the-art approaches that have been documented since the dawn of deep learning and CNNs.

2.3.1 Data Sets and Evaluation Metrics

There are many publicly available labelled data sets that have been applied for both training and testing face detection methods across the literature, in the following text a brief summary is provided for the most popular within the current literature. AFW (Ramanan et al., 2012) was one of the initial face data sets that deals with faces in unconstrained 'in the wild' images, with images consisting of multiple faces at multiple scales and at many poses. The FDDB (Jain et al., 2010) is similar to the AFW in that it consists of unconstrained images, but has a far larger number of images, 2,845 in total containing 5,171 faces. The AFLW (Köstinger et al., 2011) data set differ in that it primarily has single faces though does contain sample with multiple faces in the. While sometimes used for evaluation purposes this method is also often used as a training set. The WiderFace data set is the most recently released (Yang et al., 2016*b*), this has 32,203 images with 393,703 labelled faces with large variability in scale, pose and occlusion. This is split between a testing, training and validation split. WiderFaces specifically consists of many images with small faces which makes this challenging.

The common evaluation metric applied to face detection to determine accuracy is that of Average Precision (AP). To calculate this first the Pascal Visual Object Classes (VOC) precision-recall protocol for object detection is used to determine between true positives detection and false positives (Everingham et al., 2010). Pascal VOC precision-recall protocol uses Intersection over Union (IoU) to calculate the overlap as:

$$overlap = \frac{area(B_p \cap B_{gt})}{area(B_p \cup B_{gt})} \tag{2.1}$$

where the overlap ratio $overlap$ between the prediction bounding box B_p and the ground truth bounding box B_{gt} is greater than 0.5 (50%) a true positive prediction is recorded otherwise a false positive is recorded. Given the true positive and false positive detection's both recall and precision can then be calculated. Recall is defined as the proportion of all positive examples ranked above a given threshold. Precision is the proportion of all examples above that threshold which are from the positive class. From this true positive and false positive predictions can be determined and these values are then used to calculate the AP. The AP summarises the shape of the precision/recall curve, and is defined as the mean precision derived from a set of eleven recall levels [0, 0.1,..., 1] which are equally spaced:

$$AP = \frac{1}{11} \sum_{r \in \{0, 0.1, \ldots, 1\}} Pinterp(r) \tag{2.2}$$

the precision at each recall interval r is then interpolated, this is calculated using the maximum precision measured for a method for which the corresponding recall exceeds r:

$$Pinterp(r) = \max_{\tilde{r}:\tilde{r} \geq r} p(\tilde{r}) \tag{2.3}$$

where $p(r)$ is the measured precision at recall r. Two methods for plotting results are the Precision Recall Curve and specifically the FDDB data set plots recall against the total false positives.

2.3.2 Position and Scale

When presented with an image a face could be located at any position or scale, Figure 2.2 provides examples of differing single face scenarios. A strategy is therefore required to search the image space to detect faces that is both robust but also efficient. The general principle for searching the image space is that of region proposal, in which n regions of interest from the image space are proposed by a given method and the features are extracted from this region and compared to detect a face or non-face. Further to these proposed regions an image pyramid has historically been used to deal with scale issues, where the original image is located at the bottom of the pyramid and a defined set of layers exist where the image is re-sized to be incrementally smaller in terms of height and width. For each of these scaled images in the pyramid regions of interest would then be proposed. The application of an image pyramid is method specific as sub-sampling an image at multiple scales can be computationally expensive when paired with exhaustive regions proposal methods (Dollár et al., 2014). There are several groups of methods that have been applied for region proposal and scaling which are described below.

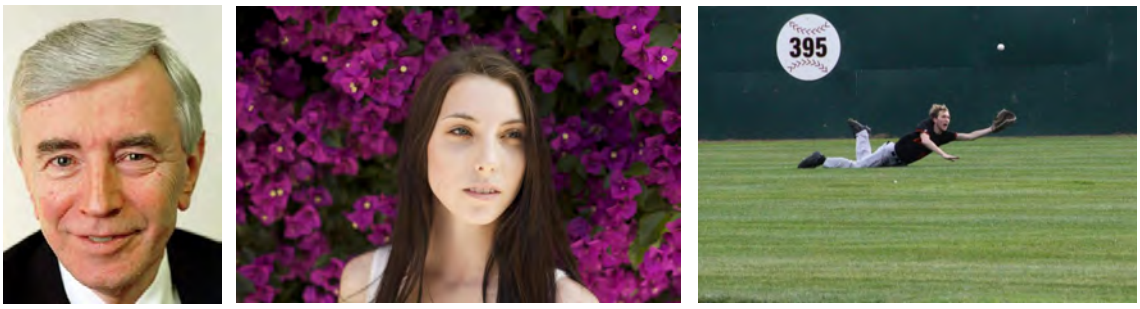

Figure 2.2: Face Position And Scale Examples.

Exhaustive search is the first group of methods and as the name implies they devise a strategy to search the entirety of the space. While these methods are exceptionally thorough they are also computationally expensive due to the large size of the search space and the use of an image pyramid to deal with scale, this problem has been further magnified as image resolution has continued to grow with new technology. Most of the exhaustive search methods are based upon an implementation of the sliding window, this involves incrementally sliding a n by m rectangular or square window across the entire image. The disposable part model based Tree Shape Model (TSM) (Ramanan et al., 2012) is an accurate face detector which applies both a exhaustive search

paired with an image pyramid which suffers from slow detection speed. One method to reduce the computational complexity of the exhaustive search is to use weak classifiers which are computationally inexpensive and offset the cost of the exhaustive search. One proposed method to speed up using a sliding window is the Voila-Jones face detector (Viola and Jones, 2004) which introduced a framework for real-time face detection for full frontal faces. This framework employs a sliding windows approach to generate the proposed regions of interest, which is sped up through the application of a cascade of weak classifiers. At the first stage of the cascade a compact set of Harr-like features are employed, at each subsequent stage of the cascade more detailed feature sets are applied to remaining regions of interest. This means that a large proportion of the sliding windows proposed regions are discarded early in the process with a face detection being registered only if a region has classified as positive at all levels of the cascade. The Harr-like features are integral to the computational speed, they use a subset of the proposed regions search space to calculate the summed pixel intensities of pre-defined rectangular regions and compare these which is both fast to compute and also scale-invariant which removes the need for a feature pyramid. The Harr-like feature type is only applicable in cascaded approaches where it has high classification accuracy as it is a weak classifier, in isolation each feature is only slightly better at classification than random guessing.

Segmentation based proposal methods use image segmentation algorithms as the basis to generate region proposals. Image segmentation algorithms are applied in computer vision to locate objects and boundaries in images by considering aspects of the image such as brightness, colour, texture similarity and/or edge strength, continuity, and closure (Hoiem et al., 2011). Notable methods proposed by Carreira and Sminchisescu (2010) and Endres and Hoiem (2014) generate numerous foreground and background segments. They then learn from these segments to predict the likelihood of a foreground segment being a whole object and use this information to provide a ranking value to each proposed segmentation region. Randomisation of the initial seeds (the pixels from which the segmentation algorithm initialises) provide a variety of locations from which to generate the segments. Both methods use a single strong segmentation algorithm where Carreira and Sminchisescu (2010) define a method based upon constrained parametric min-cut problems and Endres and Hoiem (2014) use the conditional random field model defined in Hoiem et al. (2011). Selective Search (Uijlings et al., 2013) is a segmentation based method which employs

similarity measures to construct region proposals. This method greatly reduces the amount of proposed regions when compared with exhaustive search and can be considered to generate these in a smarter manner, while also being computationally less expensive than previous segmentation based methods by not relying on a single strong segmentation algorithm. To generate the region proposals this method applies a bottom-up approach, initially the algorithm applies graph-based segmentation (Felzenszwalb and Huttenlocher, 2004) using pixel intensity to over-segmented the image. Unlike previous segmentation methods these are not explicitly used as proposals instead they are the seeds from which proposals are generated. Segmented areas of the image are processed in an iterative process taking smaller regions and joining them into larger regions based upon a similarity measures based on colour, texture, size and shape compatibility. The final set of region proposal taken from each stage are ordered in decreasing order of objectiveness, with an element of randomisation to offset the bias towards large regions proposal.

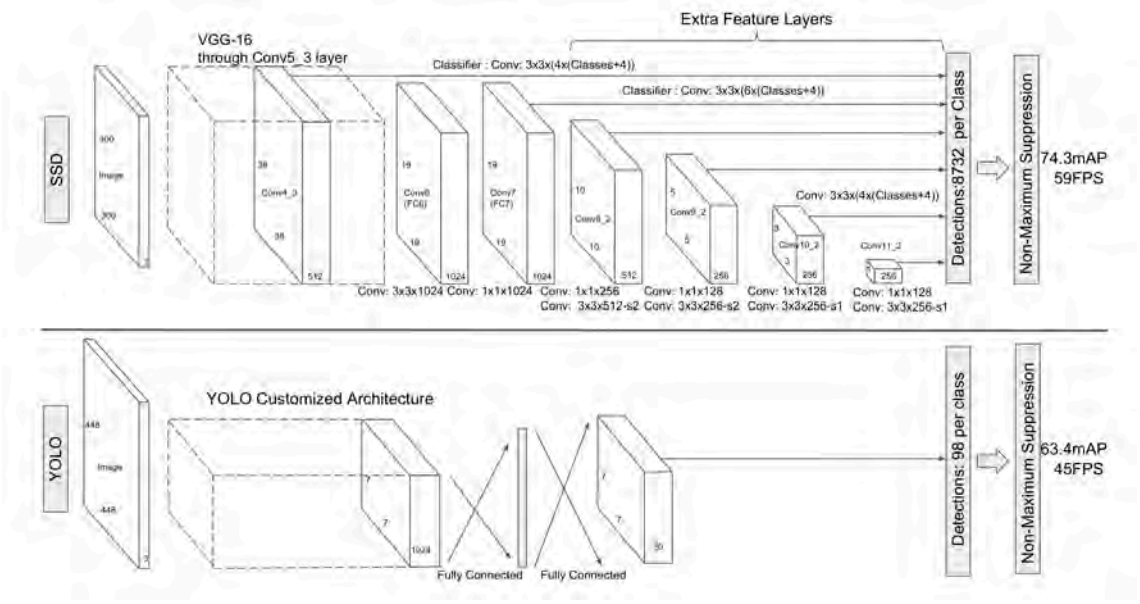

Figure 2.3: SSD and YOLO Architectures Overview (Liu et al., 2016).

As deep learning and more specifically CNN based models produced state-of-the-art results in object detection benchmarks such as ILSVRC (Krizhevsky et al., 2012) and Common Objects in Context (COCO) (Lin et al., 2014). Methods such as Region-CNN (R-CNN) (Girshick et al., 2014) and Fast R-CNN (Girshick, 2015) still relied upon external region proposal methods such as Selective Search (Uijlings et al., 2013), these method take as an input the part of the image defined by the region proposal method to learn feature mappings and perform object classification. While

these methods performed well there was still a significant computational bottleneck to this process of external region proposal generation as highlighted in Table 2.1. This all changed with the introduction of CNN architectures which were designed to provide internal region proposal. Three popular architectures have emerged to provide internal region proposal these being Faster R-CNN (Ren et al., 2015), YOLO (Redmon and Farhadi, 2016) and SSD (Liu et al., 2016). Both SSD and YOLO are defined as single shot detection methods and differ from Faster R-CNN which can be thought of as a two shot method, as it uses a Region Proposal Network (RPN) and object classifier. Figure 2.3 provides an graphical overview of the YOLO and SSD architectures while Figure 3.1 shows the Faster R-CNN architecture. Faster R-CNN incorporates a RPN into the Fast R-CNN architecture, the RPN consists of 3 convolutional filters and take the features generated from an initial shared set of convolutional layers. The RPN learns to determine if a proposed region is a background or a foreground object. The idea of anchors for generating proposal boxes is applied where each pixel of the scaled image has an associated proposal window at different scales and ratios with the anchor at the centre. Following foreground/background classification those foreground object and then classified as a specific object with an associated bounding box for the object. YOLO specifies a set of 24 convolutional layers for feature extraction for the standard implementation, where as both SSD and Faster R-CNN use pre-existing CNN network architectures for feature extraction, VGG-16 Simonyan and Zisserman (2015) in the original research papers. YOLO divides the input image into a 7×7 grid, each square in the grid is responsible for the detection of an object if the centre of the object is located there. The output of the network for each grid square are the offset co-ordinates of any detected objects bounding box from the grid square, a confidence value and the class probability of the given object. SSD differs in that it applies a preexisting CNN architecture for initial feature extraction (VGG-16 in the original paper), following this are a further 6 convolutional layers, making 11 layers in total. The extracted features at layers 4, 7, 8, 9, 10 and 11 are output from the end-to-end flow of the network and passed into a 3×3 conventional filter top produce a set of object bounding box and class predictions. The process of producing predictions at different layer can be thought of as a method to deal with object scale, where at layer 4 there the original image is a effectively a 38×38 grid in which 4 aspects ratios predictions are produced to make a total of 5776 predictions. At subsequent layers this is reduced until layer 11 where the image is represented as a 1×1 grid with 4 ratios and a total of 4 predictions. All three of these methods are significantly faster than previous methods even

CNNs which used Selective Search. The following comparative performances are summarised in Table 2.1. Selective Search with Fast R-CNN has a reported speed of 0.5fps, while Faster R-CNN with VGG-16 is 7fps, YOLOs reports 21fps for the normal implementation and 155fps for FastYOLO but this has a significant 15% drop in accuracy. Finally SSD variants SSD-300 and SSD-512 obtain 46 and 19 FPS respectively, where 300 and 512 refer to resolution of the scaled input image. The accuracy of each method in object detection benchmarks comparable, with SSD having a slight advantage. A number of factors influence these speed and accuracy results including training set used, data augmentation techniques, feature extraction architecture and initial image size. For example Faster R-CNN uses a larger initial image dimension of 1000×600 while SSD-512 uses image of 512×512 which requires less initial operation and therefore aids the speed. SSD also employed specific "zoom out" data augmentation to create more small training samples which helped increase the accuracy by 2%-3% which highlights how training strategy plays a role in the reported accuracy of a method. Faster R-CNN is described in further detail in the next chapter of this thesis.

Method	**Mean Average Precision**	**FPS**
Fast R-CNN* (Girshick, 2015)	70.0%	0.5
Faster R-CNN (Ren et al., 2015)	76.4%	7
YOLO (Redmon and Farhadi, 2016)	63.4%	21
SSD300 (Liu et al., 2016)	74.3%	46
SSD500 (Liu et al., 2016)	76.8%	19

Table 2.1: Object Detection Accuracy and Speed on the PASCAL VOC 2007 test set. * Method uses external region proposal.

2.3.3 Pose and Occlusion

The challenge of detecting faces becomes more difficult when dealing with different facial poses as seen in Figure 2.4 or when the face is partially covered as in the scenario of occlusion as shown in Figure 2.5. While not directly the same challenge, pose and occlusion share similarities in detecting faces which have large changes in their visual characteristics. Due to the overlap of both challenges the methods employed to overcome these are discussed in the section below.

While the original Voila-Jones face detector (Viola and Jones, 2004) was a landmark method for performing face detection on full frontal faces, it was not initially designed to handle multiple poses. An expansion to this method was proposed by the original authors using multi-view models

Figure 2.4: Facial pose examples from full frontal to profile.

(Jones and Viola, 2003), in which they trained multiple models for different poses. In order to reduce the added computational expense of having to evaluate each new model for a proposed region, a two-step approach was applied in which at the first level a model to detect the pose was applied, only then would the relevant pose model be applied using the same method as the original implementation. Following this the multi-view model became the primary method to solve the pose task, which uses multiple facial feature representations each specific to a pose angle of the face. Some notable examples include Chang Huang et al. (2005) who developed a tree based structure for determining the pose model and introduced a piece-wise function to approximate complicated distributions as features instead of Harr-like features, these increased the speed and accuracy for face detection. In (Jianguo Li et al., 2011) they replaced the Harr-like features with Speeded Up Robust Features (SURF) and produced state-of-the-art results on the benchmark FDDB data set.

One of the final key methods prior to deep learning was that of the TSM by Ramanan et al. (2012) which was a derivative of the Deformable Part Model (DPM) (Felzenszwalb et al., 2009) for face detection. The TSM not only applied a multi-view model approach but also a part-based model to describe the independent facial components. Each multi-view model consisted of 68 facial features which were extracted from the training data using Histograms of Oriented Gradients (HOG). To provide a shape correspondence between these parts a tree based method is applied to connect the parts allowing the application of efficient dynamic programming algorithms to find globally optimal solutions through branch to root backtracking. Uniquely the use of part based models also

Figure 2.5: Facial occlusion examples: (Left) - a more simple occlusion of a person wearing sunglasses. (Right) - a complex image with multiple faces being covered by other faces and objects such as flags to varying degrees of occlusion.

allowed the model to perform landmark localisation. While the accuracy in both face detection and landmarks localisation was very good at the time of publications this came at the price of speed. While the authors also introduced part sharing across models to reduce the computational overhead this came at a cost of reduced accuracy. While not specifically the intention of the TSM method using parts for detection faces also introduced some robustness to occlusions. The TSM was expanded further in the form of the Hierarchical DPM (Ghiasi Charless Fowlkes et al., 2014) which had the specific aim to better handle the occlusion problem. They extended the TSM with the addition of hierarchical model structures and explicit occlusion.

The arrival of CNNs has seen major advances over the previous methods in dealing with both pose and occlusion problems. Many methods using a variety of CNN based architectures have been proposed with increasing accuracy on benchmark data sets including (Taigman et al., 2014; Yang et al., 2016*a*; Farfade et al., 2015; Triantafyllidou et al., 2018; Zhang, Zhu, Lei, Shi, Wang and Li, 2017). While each of the proposed methods has specific contributions, it is more inherently the capability of these learnt features from CNNs in general which make them excel in these tasks that will be reviewed in this section. As can be observed from the methods proposed prior to deep learning the most successful used a large variety of features to model the face while explicitly training for a given challenge like occlusion. The primary issue with this was that more features equated to a large computational overhead and hand crafting features for all given face

detection scenarios was not feasible. CNN based features overcome both of these issues through self learning of the feature filters from the training data set and having the capacity to capture many different features representation without the additional computational overhead once the network training process is complete. The architecture of a CNN is structured in a multi-layered fashion, with typically four types of layers used, these being the convolutional, activation, pooling and fully connected layers. Convolutional layers compute take an input which is either the input image or output from other layers, it slides $n \times m$ sized filters across this input computing a dot product between the filters weight parameters for this region. This process results in output volume size $i \times j \times k$ for k filters of size $n \times m$. Activation layers are used to normalise the output of the convolutional layers to a certain value range which is dependant upon the specific activation function used. Pooling layers perform a down-sampling operation along the spatial dimensions. The fully connected layers are responsible for determining which features most correlate to a particular class. They are used at the end to produce the output volume prior to classification, for example the output size would be n for an n class problem.

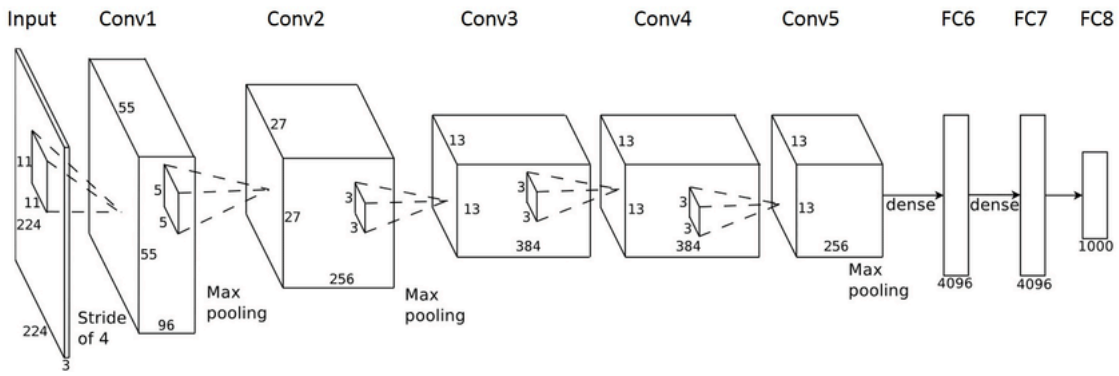

Figure 2.6: AlexNet CNN architecture overview (Krizhevsky et al., 2012)

An important early CNN architecture known as AlexNet (Krizhevsky et al., 2012) is shown in Figure 2.6, it is a 5 convolutional layer network, where at each of these layers during the training process a set of feature mapping filters are learnt which activate based upon input image. Visually once learnt each layer of the CNN can be viewed as a set of feature filters which at the first layer are very generalised to respond to common components of an image, such as edge shapes and straight lines. Deeper layers in the network have more object specific feature filters. A visual example of this is shown in Figure 2.7 which shows the filters at each of the 5 convolutional layers of the AlexNet CNN trained on the ImageNet data set (Krizhevsky et al., 2012). This highlights exactly

why CNN features have proven so adapt in the pose and occlusion problem, as there are many feature filters at each level which can learn features relating to different aspects of the human face. The data set used for training a model has a significant influence on the learnt features, specifically to learn features that deal with pose and occlusion the training data must include many examples of this type, a common approach to expand the data set samples is through augmentation. Examples of augmentation techniques include flipping, rotating and blanking sections of the face to introduce further occlusion.

2.3.4 Precision

The capability to detect faces has improved significantly with the application of deep learning methods as highlighted by the performance on commonly used benchmark data sets such as FDDB and AFW. The FDDB data set specifically is the most widely applied benchmark from which to compare both traditional methods and current deep learning, this data set consists of 5,171 faces in a set of 2,845 images. In Table 2.2 the recall percentage and total number of false positives are given for a number of the methods previously discussed in this section. In terms of the recall metric which refers to the n faces successfully detected form the total number of faces in the data set. In general there is a trend of increased recall with CNN based approaches showing the highest level including (Zhang, Zhu, Lei, Shi, Wang and Li, 2017) which sits at 99% recall. A similar trend though has yet to been seen in terms of precision which refers to the number of false positives faces detected. Many of the top performing methods in terms of recall also display a large number of false positives, while the methods such as (Triantafyllidou et al., 2018) which have higher precision through the use of hard negative mining completed during the training phase have a reduced recall rate. This leads to a question regarding if the method displaying very high recall are partially managing these by making a large number of total potential detection's.

2.4 Landmark Localisation

Landmark localisation is the process of detecting a set of facial landmarks which have significant meaning, the process is also often referred to as face alignment within the literature. There is no specific standard for the number of landmarks and the precise locations which vary in different data sets, though 68 landmarks is generally the most popular within the research. Figure 2.8 highlights

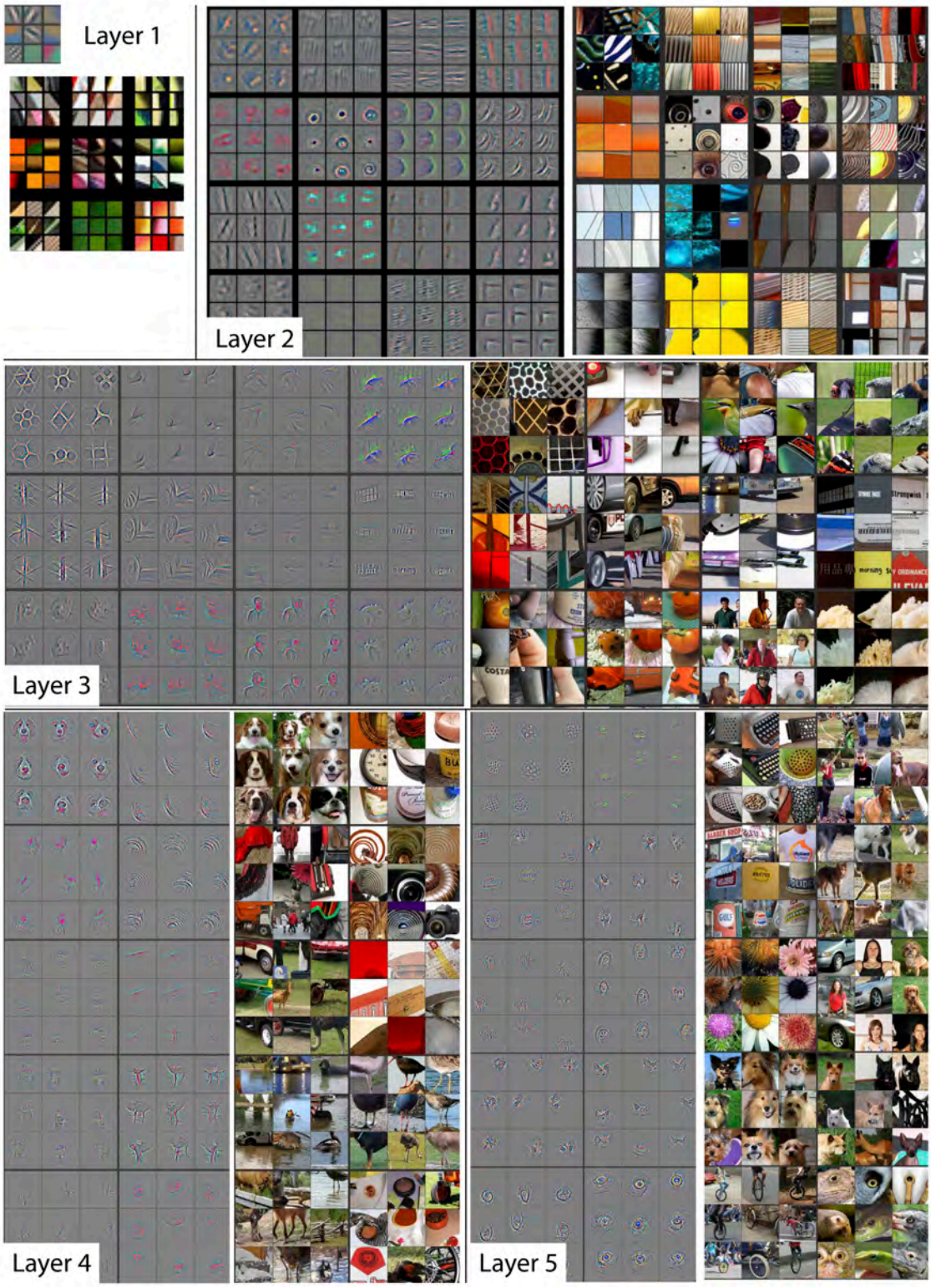

Figure 2.7: Convolutional Feature Filter Examples: (Layer 1) - Generic edges, (Layer 2) - Generic lines and curves, (Layer 3, 4, 5) - Domain specific shapes.

Method	Recall	False Positives
Viola-Jones (OpenCV implementation) (Viola and Jones, 2004)	82%	175932
TSM (Ramanan et al., 2012)	78%	3857
SURF (Jianguo Li et al., 2011)	78%	6978
Headhunter (Yang et al., 2016*a*)	91%	12408
HDPM (Ghiasi Charless Fowlkes et al., 2014)	87%	3747
FPS (Yang et al., 2016*a*)	91%	2000
SFD (Zhang, Zhu, Lei, Shi, Wang and Li, 2017)	99%	22773
Fast Mining (Triantafyllidou et al., 2018)	92%	759

Table 2.2: FDDB Benchmarks

the facial locations of the commonly applied 68 landmarks. The task of landmark localisation shares similarities to face detection with similar feature extraction techniques and challenges in areas like occlusion and pose. In the majority of the methods discussed there is consideration taken regarding both a texture feature describing the appearance of a specific landmarks and a global shape which takes into consideration the arrangement of these landmarks as a whole. The majority of landmark localisation methods are dependant upon prior face detection with a few notable expectations, therefore there is an expectation that the image provided is a face and has been scaled appropriately for the specific landmark localisation method. This landmark localisation literature review split between pre-deep learning methods which are termed traditional methods and deep learning methods.

2.4.1 Traditional Methods

Within the traditional methods the Active Shape Model (ASM) developed by Cootes et al. (2000) provided one of the first great breakthrough methods which could be applied to landmark localisation, they followed up this work an alternative method namely Active Appearance Model (AAM) (Cootes et al., 2001). Both methods while not specifically designed for face landmark localisation leverage the idea of defining statistically developed deform-able models. There are similarities and distinct differences between the methods, while both use a statically generated model consisting of both texture and shape components learnt from a training data set, the texture component and how it is applied in the landmark fitting process are distinct to each method. The shape model is composed through the alignment of the training images by using a variation of the Procustes method which scales, rotates and translates the training shapes so that they are aligned as closely

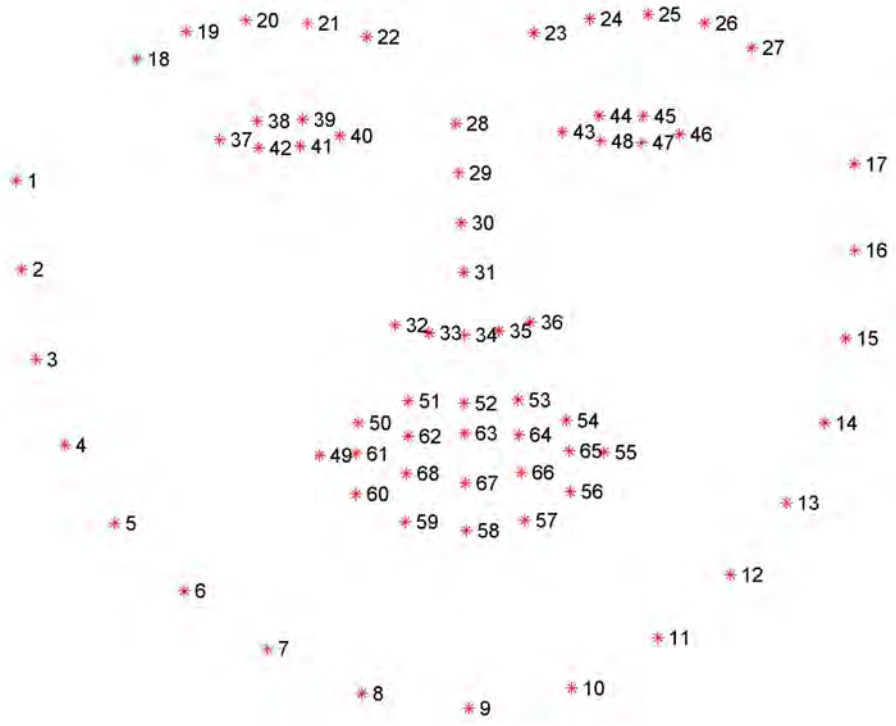

Figure 2.8: Reference Locations For IBUG 68 Facial Landmarks.

as possible. Principal Component Analysis (PCA) is then carried out on the training images reducing the dimensions of the features while retaining the variance in the shape data. A mean shape is also generated which is often used as a starting point for fitting to new images. The ASM is considered to be in the Constrained Local Model (CLM) group of methods, these model types use the texture model as local experts in which they are trained on texture information taken from a small area around each landmark. The local expert ASM uses a small set of gray-scale pixel values perpendicular to each landmark while other CLM techniques use a block of pixels around the landmarks or other feature descriptors such as Scale-Invariant Feature Transform (SIFT) (Cristinacce and Cootes, 2008; Milborrow and Nicolls, 2014). The fitting of the model is carried out via the optimisation of an objective function using the prior shape and the sum of the local experts to guide the alignment process. AAM differs from the CLM group of methods by using a texture model of the entire face rather than regions. To create this all face textures from the training images are warped to a mean-shape, transformed to grey scale and normalised to reduce global lighting effects. PCA is then applied to create the texture features. Alignment on an unseen image is carried out by minimising the difference between the textures of the model and the unseen image (Cootes et al., 2001).

Further advancements in accurate and computationally efficient landmark localisation arrived with the application of regression based fitting methods rather than sliding-windows based approaches. Regressors also provide detailed information regarding the local texture prediction criteria when compared with the classifier approach which is a binary prediction of match or not. Valstar et al. (2010) proposed a method named Boosted Regression coupled with Markov Networks, in which they apply Support Vector Regression and local appearance based features to predict 22 initial facial landmarks in an iterate manner, Markov Networks are then used to sample new facial locations to apply the regressor to in the next iteration. Cascaded regression was then applied by Dollár et al. (2010); Zhu, Li, Loy and Tang (2015*a*) in which a cascade of weak regressors is applied to reduce the alignment error progressively while providing computationally efficient regression methods. Different feature types have been applied these for example Cao, Wei, Wen and Sun (2014) have recently produced a face alignment method based up a multilevel regression using fern and boosting. This has been subsequently built upon in (Burgos-Artizzu et al., 2013) where a regression based technique named Robust Cascaded Posed Regression, which can also differentiate between landmarks that are visible and non-visible (occlusion) and estimate those facial landmarks that may be covered by another object such as hair or a hand proposed. Ren et al. (2014) have also applied a regression technique with local binary features and random forests to produce a technique that is both accurate and computationally inexpensive meaning that the algorithm can perform at 3000fps for a desktop PC and up to 300fps for mobile devices.

The previous methods predicted facial landmarks on faces in limited poses at most between ± 60 degrees, both the TSM of Ramanan et al. (2012) which was previously discussed within the face detection section of this chapter and Pose-Invariant Face Alignment (PIFA) (Jourabloo and Liu, 2017) are notable methods which could handle a greater range of face pose. The TSM (Ramanan et al., 2012) was unique amongst landmark localisation methods in that it did not use a regression or iterative methods for determining landmarks positions, instead this used the HOG parts to determine location based upon appearance and the configuration of all parts was scored to determine the best fit for a face. The final X and Y coordinates of the predicted landmarks are derived from the centre of a bounding box for that specific parts detection. Jourabloo and Liu (2017) proposed PIFA as a significant improvement in dealing with all face poses and determining the visibility of a landmark across poses for up to 21 facial landmarks. This method extended 2D cascaded

landmark localisation through the training of two regressors at each layer of the cascade. The first regressor predicts the update for the camera projection matrix which map to the pose angle of the face. The second is responsible for updating the 3D shape parameter which determines 3D landmarks positions. Using 3D surface normal's, visibility estimates are made based upon a z coordinate, finally the 3D landmarks are then projected to the 2D plane.

2.4.2 Deep Learning Methods

The initial deep learning CNN based landmark localisation methods while displaying high accuracy were limited to a very small set of sparse landmarks when compared with previous traditional methods. Sun et al. (2013) proposed a Deep Convolutional Network Cascade, this consisted of a 3 stage process for landmark localisation refinement, at each level of the cascade multiple CNNs were applied to predict the locations for individual and subsets of the landmarks. This method only considered 5 landmarks and the capability to expand this to further landmarks is computationally expensive due to the nature of using individual CNNs to predict each landmark. Zhang et al. (2014) applied multi-task learning to enhancement in which they trained a single CNN with not only facial landmark locations but also gender, smile, glasses and pose information. Linear and logistic regression were used to predict the values for each task from shared CNN features. When directly compared with the Deep Convolutional Network Cascade (Sun et al., 2013) they showed increased landmark accuracy with a significant computational advantage of using a single CNN. A Backbone-Branches Architecture was applied in Liang et al. (2015) which outperformed the previous methods in terms of both accuracy and speed for 5 facial landmarks. This model consisted of a multiple CNNs, a main backbone network which generates low-resolution response maps that identify approximate landmark locations, then branch networks produce fine response maps over local regions for more accurate landmark localisation.

The next generation of deep learning methods expanded on these initial methods increasing the number of landmarks detected to the commonly used 68. HyperFace (Ranjan et al., 2016) like (Zhang et al., 2014) applies a multi-task approach which also considered face detection. The idea of the multi-task approach is that inter-related tasks can strengthen feature learning and remove over-fitting to a single objective. HyperFace as shown in Figure 2.9 applies a single CNN originally AlexNet, but modified this by taking features from layers 1,3 and 5, concatenating these

into a single feature set, then passing these through a further convolutional layer prior to the fully connected layers for each task. At the same time the fully-convolutional network (FCN) (Liang et al., 2015) emerged as a technique, in which rather than applying regression methods to predict landmarks coordinates, they are based upon response maps with spatial equivalence to the raw images input. Convolutional and de-convolutional networks are used to generate a response map for each facial landmark, further localisation refinement applying regression was then used in (Lai et al., 2015; Xiao et al., 2016; Bulat and Tzimiropoulos, 2016). The stacked hourglass model proposed in Newell et al. (2016) for human pose estimation which applied repeated bottom-up then top-down processing with intermediate supervision has been applied to the landmark localisation in a method called the Face Alignment Network (FAN) (Bulat and Tzimiropoulos, 2017*b*) this has shown state-of-the-art performance on a number of evaluation data sets. Further more this method expanded the capability of detection from 2D to 3D landmarks through the addition of a depth predictions CNN which takes a set of predicted 2D landmarks and generates the depth. At the time of publication the FAN method outperformed previous methods for accurate landmark localisation. In chapter 4 the stacked hourglass architecture for landmark localisation is discussed in detail.

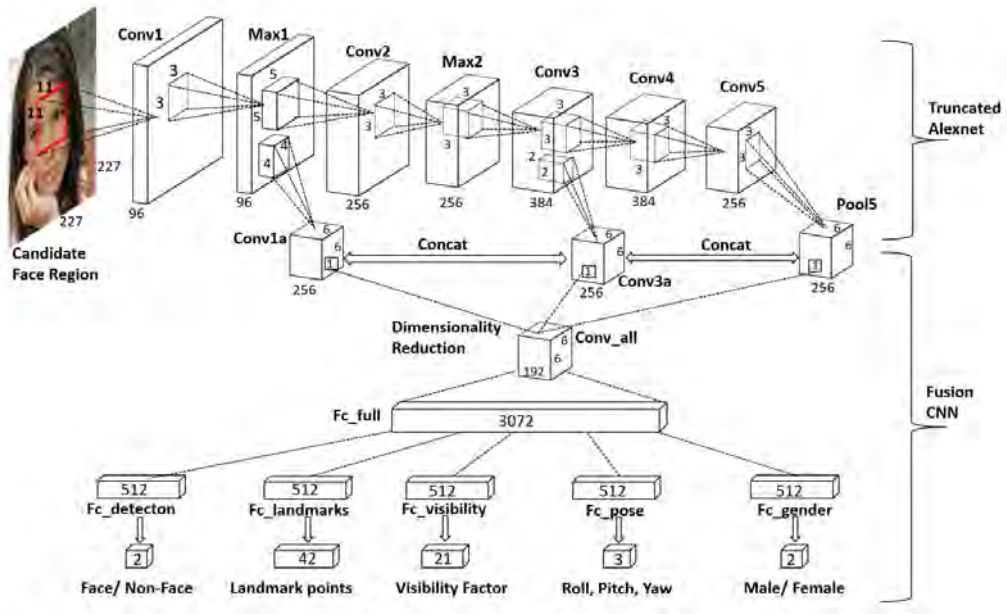

Figure 2.9: HyperFace architecture. Showing the approach taken by the authors to introduce multi-task learning into a single CNN framework, using multi-level feature concatenation and task specific fully connected layers (Ranjan et al., 2016).

2.4.3 Data Sets and Evaluation Metric

There are a number of data sets that have been made publicly available over the last 20 years for both training and evaluating landmarks localisation methods. One issue in the evaluation process is that the locations and number of landmarks annotated are not in correspondence across all data sets. In this section only data sets which are used in the most recent literature with annotated 68 landmarks are discussed. The 300-W test set consists of the 600 images originally applied to the 300-W challenge for evaluation purposes (Sagonas et al., 2013). This test set is split in two subsets of images these being indoor and outdoor respectively. A successor to the 300-W challenge the 300 Videos In-the-Wild (300-VW) challenge presented a large-scale face tracking data set (Shen et al., 2015), containing 114 videos and in total 218,595 frames. From the 114 videos, 64 are used for testing split into three categories of difficulty (A, B, and C) and 50 for training. Menpo is a newer data set introduced in Zafeiriou et al. (2017), containing landmark annotations for about 9,000 faces taken from the FDDB (Jain et al., 2010) and AFLW (Köstinger et al., 2011) face detection data sets. Within this data set only 39 facial landmarks were annotated for profile faces where as 68 where labelled to frontal faces. The AFW (Ramanan et al., 2012) data set mostly associated with face detection also has accompanying landmark annotations for all faces. For testing landmark localisation for occluded landmarks the Caltech Occluded Faces in the Wild (COFW) (Burgos-Artizzu et al., 2013) data set is often used, this consists of 500 training images with 1345 faces and 507 faces in the testing data set split. All 68 landmarks are annotated and a visibility classification is given to determine if the ground truth landmark is occluded or visible.

The most commonly used metric for determining the accuracy of landmark localisation method is that of Normalised Mean Error (NME) where initially normalisation was done via ocular distance (distance between the eyes) but has more recently moved to face size as ocular distance when dealing with faces at multiple poses (Bulat and Tzimiropoulos, 2017*b*). NME is defined as:

$$NME = \frac{1}{N} \sum_{k=1}^{N} \frac{\|x_k - y_k\|_2}{d} \tag{2.4}$$

where ground truth and predicted landmark for the k^{th} landmark are denoted as x_k and y_k respectively. The normalisation factor d is the square root of the ground truth bounding box calculated

as $d = \sqrt{width * height}$.

2.5 Facial Analysis Tasks

In this section we will review the literature regarding the facial analysis task areas of facial recognition, expression recognition and medical diagnosis. This is the final stage of a facial analysis system and produces the final task specific output. Generally this process takes the previously detected face and landmarks to extract discriminative features. When addressing the facial analysis tasks there are two distinct scenarios in the literature, those which perform analysis from a single image and those which use videos in the form of a sequence of images. The features used in these scenarios are significantly different, while image based analysis only consider spatial data, video sequences are also concerned with time and the affect this has on the spatial data. The feature extraction methods applied for video sequence analysis are called spatio-temporal features, while specific methods are discussed within the following section, it has been common in computer vision to extend successful spatial feature extraction methods to the spatio-temporal domain for example Local Binary Patterns (LBP) to Local Binary Patterns on Three Orthogonal Planes (LBP-TOP) and HOG to HOG-3D. Performing analysis on video sequences can provide a richer set of information than using an image, for some facial analysis tasks such as micro-expression recognition video is a necessity. Image analysis though remains popular as for certain tasks it is sufficient and also the collection of image data sets has been simpler and requires much less storage capacity than video. While the focus of this review is the deep learning methods, also included are sections that detail the traditional methods proposed prior to deep learning. The section regarding traditional methods are not a comprehensive analysis but are included to provide some background on how the tasks where originally approached. There is also still merit in many aspects of these methods which have the potential to be used in conjunction with deep learning.

2.5.1 Facial Recognition

Facial recognition is the task of identifying or verifying the identity of subjects in images or videos. This task is very well defined and has been a the focus of research in the computer vision field since the 1970's (Kanade, 1974). Specifically facial recognition comprises of two stages face representation and face matching. Face representation takes the detected face image data and

transforms it to a discriminative feature vector also known as a template. Ideally the feature vector for a subject should be similar. Face matching is the process of comparing two templates (feature vectors) from which a similarity score can be produced which indicates the likelihood that they belong to the same subject. The use of the human face as a biometric security has been one of the driving forces behind the research in this task, where due to them non-intrusive nature of image capture it is preferred over other biometric modalities such as iris and fingerprint (Trigueros et al., 2018). Like the base task of face detection face recognition is most challenging when deployed in unconstrained environments due to the high variability that face images present. The challenges include head poses, ageing, occlusions, illumination conditions, and facial expressions. The availability of large labelled data sets (Table 2.3 provides an overview of the common data sets for facial recognition) such as VGGFace2 (Cao et al., 2018) containing 3.31 million images of 9131 subjects and UMDFaces (Bansal et al., 2017) with 367,888 annotated faces of 8,277 subjects has accelerated this research field specifically with the application of CNN based methods.

Data set Name	**Images**	**Subjects**	**Images per subject**
CelebFaces+ (Sun et al., 2014)	202,599	10,177	19.9
UMDFaces (Bansal et al., 2017)	367,920	8,501	43.3
CASIA-WebFace (Yi et al., 2014)	494,414	10,575	46.8
VGGFace (Parkhi et al., 2015)	2.6M	2,622	1,000
VGGFace2 (Cao et al., 2018)	3.31M	9,131	362.6
MegaFace (Nech and Kemelmacher-Shlizerman, 2017)	4.7M	672,057	7
MS-Celeb-1M (Guo et al., 2016)	10M	100,000	100

Table 2.3: Facial Recognition Data Sets Overview

Traditional Methods

The original methods first applied to the face recognition task where that of geometry-based, at this time the data sets were very small and computational capacity was low, therefore the computational efficiency of these methods was beneficial. One successful geometry-based method used the Procrustes distance between two sets of facial landmarks and a method based on measuring ratios (Shi et al., 2006). As the field advanced one of the more seminal works in face recognition was that of eigenfaces (Turk and Pentland, 1991). The idea behind this technique is to project facial images onto a low-dimensional space, with the purpose of keeping the features important to face recognition while discarding those features of the data which are not important. The eigenfaces methods applies PCA to the training data discovering the eigenvectors that account for the most variance in the data distribution. As the discovered eigenvectors resemble faces these are

termed eigenfaces. The weights of a linear combination to reconstruct the face can be obtained through the projection of a new face to the subspace spanned by the eigenfaces. Face recognition was performed by weight comparison between the new face and a test set of faces. Later methods replaced PCA with Linear Discriminant Analysis (LDA) to overcome the potential that PCA derived eigenvectors may represent some features that are not important to the task such as facial expression. The most successful methods applied both PCA to reduce the dimensions of the data before applying LDA (Belhumeur et al., 1997). Following on from these PCA and LDA based methods there has been further significant research in this area, most of the proposed methods for face representation are similar to those used in other computer vision tasks described in this chapter. Common approaches include the use of appearance based texture feature from either the whole face or significant facial regions, where techniques such as LBP, SIFT and HOG have been applied. There have been a number of methods proposed for the process of face matching on these appearance based templates, this has included the use of Support Vector Machine (SVM) through the posing of matching as a classification problem where each subject is a class (Jonsson et al., 2002). Other matching metrics used have included logistic discriminant metric learning based on the principle that the distance between the same subject is smaller than between other subjects. While marginalised k-nearest neighbour was applied to find how many positive neighbour pairs can be formed from the neighbours of the two compared templates (Guillaumin et al., 2009).

Deep Learning Methods

The first end-to-end CNN based method for face recognition applied a siamese architecture and was trained using contrastive loss in Chopra et al. (2005). The results for this method were poor in contrast with traditional methods largely due to the shallow nature of the CNN and the limited data set used for training. The introduction of larger training sets and deeper model architectures saw state-of-the-art results in this task, one of the first examples was Facebook's DeepFace in Taigman et al. (2014). DeepFace applied locally connected layers within the CNN architecture containing which has the capability to learn different features from local facial regions, this is possible as landmark localisation is used to determine local facial regions. Training of DeepFace was conducted on a data set containing 4.4 million faces with 4,030 subjects and applied the softmax loss function which is used in many classification based CNN architecture. A hugely significant

increase of 27% accuracy was reported on the LFW benchmark data set over existing methods, with a 97.35% total reported accuracy. Similar results where also reported by Sun et al. (2014) where a system named DeepID applied a complex architecture using 60 CNNs each learning a different set of local features from 10 local face regions, at each of the local regions 3 scales and both RGB and gray scale image data was used. The learnt features from all CNNs was then concatenated to form the final feature vector from which facial recognition classification was performed. A data set containing 202,599 face images of 10,177 celebrities was applied for training DeepID (Sun et al., 2014). Following the distinct architectures of DeepFace and DeepID, many other approaches including (Liu et al., 2017; Wu et al., 2017; Hasnat et al., 2017) have adopted the CNN architectures that have produced state-of-the-art accuracy in the object detection field and specifically the ILSVRC namely VGG (Simonyan and Zisserman, 2015) and ResNets (He et al., 2015).

One key area most recently researched is that of the loss functions used to train these popular CNN architectures for the face recognition task. Figure 2.10 provides a visual overview of the significant loss functions described below. While fine-tuning pre-trained ImageNet models using the common softmax loss function produced accurate face recognition results, the capacity of this loss function to generalise well to subjects not present in the training data has been questioned. Softmax loss encourages the learning of features that increase inter-class differences but does not necessarily reduce intra-class variations (Trigueros et al., 2018). Therefore alternative loss functions have been proposed that have boosted the accuracy of these models. Triplet loss has proven to be a popular approach (Schroff et al., 2015), this loss function uses a margin principle to separate the distance positive and negative samples with the aim of learning more discriminantive features. One negative of triplet loss is that is slow to converge in comparison to softmax loss due to the requirements of creating triplet of samples from the training set. One common method used to overcome this is to first train the model with softmax and then fine tune with triplet loss. Center loss was proposed in Wen et al. (2016) with the the goal minimise the distances between features and their corresponding class centres. Center loss is often used in conjunction with softmax loss, as it has been shown to effectively increase inter-personal variations and reduce intra-personal variations, while also being less computationally complex as triplet loss. Another related learning method is range loss (Zhang, Fang, Wen, Li and Qiao, 2017) proposed for improving training

with unbalanced data sets. The range loss has intra-class and inter-class components, where intra-class loss minimises the largest distances between samples of the same class, and inter-class loss maximises the distance between the closest two class centres in each training batch. Through exploiting these extreme cases the range loss uses same information from each class, regardless of how many samples per class are available in the data set. Range loss requires combining with softmax loss to avoid the loss being degraded to zeros (Wen et al., 2016).

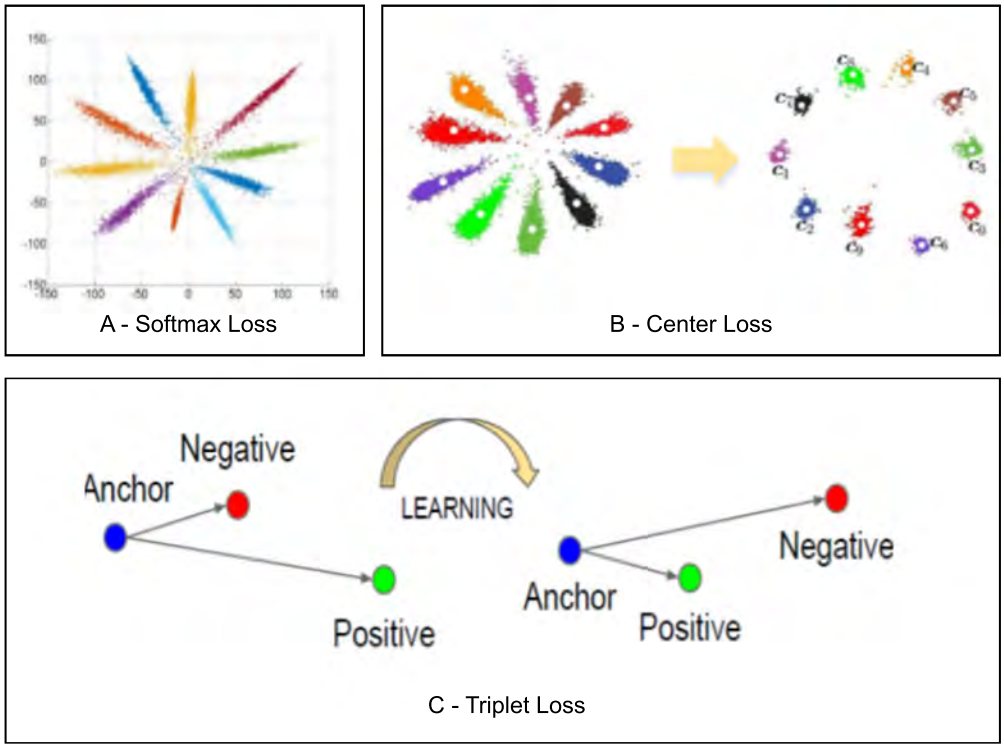

Figure 2.10: Loss Functions Visual Overview: (A) - Softmax loss learns inter-class differences to create discriminative features for each class (Trigueros et al., 2018), (B) - Center Loss defines a center for each class within feature space with the aim of clustering each class of features more closely to this center, thus increasing the discriminative nature of each feature (Wen et al., 2016), (C) - Triplet loss applies both a negative and positive sample to learn an anchor in feature space, where the distance to the negative sample should increase and distance to the positive sample should decrease during the learning process (Schroff et al., 2015).

2.5.2 Facial Expression Recognition

Facial Expression Recognition (FER) is a well established facial analysis task with a large body of research to date. The subcategory of macro-expression recognition makes up a majority of this research both for FER from images and videos. This is due to a large variety of labelled data sets (See Table 2.4) containing annotated images such as FER-2013 (Goodfellow et al., 2013)

and video such as the Extended Cohn-Kanade (CK+) Lucey et al. (2010). The most popular class labelling for FER is 6 basic expressions of anger, disgust, fear, happiness, sadness and surprise plus the neutral expression. Micro-expressions have also more recently been the subject of research though this is more limited primarily due to the lack of data set availability. This stems from the difficulty in labelling and collecting data in this area, as micro-expression must be captured using video and the context in which the expression is formed is crucial to labelling. Table 2.5 provides an overview of the currently available micro-expression data sets.

Data Set	**Samples**	**Subjects**	**Source**	**Labelling**
CK+ (Lucey et al., 2010)	593 videos	123	Lab	6 basic expressions plus contempt and neutral
MMI (Pantic et al., 2005)	740 images and 2,900 videos	25	Lab	6 basic expressions plus neutral
JAFFE (Lyons et al., 1998)	213 images	10	Lab	6 basic expressions plus neutral
FER-2013 (Goodfellow et al., 2013)	35,887 image	N/A	Web	6 basic expressions plus neutral
AFEW 7.0 (Dhall et al., 2017)	1,809 videos	N/A	Movie	6 basic expressions plus neutral
SFEW 2.0(Dhall et al., 2015)	1,766 images	N/A	Movie	6 basic expressions plus neutral
Multi-PIE (Gross et al., 2008)	755,370 images	337	Lab	Smile, surprised, squint, disgust, scream and neutral
BU-3DFE (Lijun Yin et al., 2006)	2,500 images	100	Lab	6 basic expressions plus neutral
Oulu-CASIA (Taini et al., 2008)	2,880 videos	80	Lab	6 basic expressions
RaFD (Langner et al., 2010)	1,608 images	67	Lab	6 basic expressions plus contempt and neutral
EmotioNet (Benitez-Quiroz et al., 2016)	1M images	N/A	Web	23 basic expressions or compound expressions
RAF-DB (Li et al., 2017)	29672 images	N/A	Web	6 basic expressions plus neutral and 12 compound expressions
AffectNet (Mollahosseini et al., 2017)	450,000 images	N/A	Web	6 basic expressions plus neutral
ExpW (Zhang, Luo, Loy and Tang, 2018)	91,793 images	N/A	Web	6 basic expressions plus neutral

Table 2.4: Facial Expression Recognition - Macro-Expression data sets overview.

Data Set	**Samples**	**Subjects**	**Labels**
SMIC (Li et al., 2013)	76	6	3
SMIC2 (Li et al., 2013)	164	20	3
CASME (Wen-Jing Yan et al., 2013)	195	35	7
CASME II (Yan, Li, Wang, Zhao, Liu, Chen and Fu, 2014)	247	35	7
York DDT (Warren et al., 2009)	18	9	4

Table 2.5: Facial Expression Recognition - Micro-Expression data sets overview.

Traditional Methods

There have been many traditional methods applied to FER in this section our primary focus is on those methods that have applied to the FER from videos. The rationale for this is that there are already many deep learning based methods applying FER on images but less literature on videos, while many of the spatial based methods are the same as those used in other areas such as face detection and landmark localisation discussed previously in this chapter. The spatio-temporal features extracted from face image data are the key to capturing both spatial and temporal data with the aim to produce a discriminative feature vector that can effectively describe a class of emotion.

In macro-expression recognition geometry-based features have proven successful, extending these feature types to the temporal domain is trivial. Both Ghimire and Lee (2013) and Ghimire et al. (2015) use an elastic bunch graph matching for face detection then track the face over the sequence. Normalization of the facial landmarks is then performed to bring them into scale alignment. In Ghimire and Lee (2013) geometric features based on the change in Euclidean distance and angle between pairs of facial landmarks, while Ghimire et al. (2015) applied a method using triangles formed between facial landmarks and the associated angles and side lengths of the triangles as features. Both methods use a feature selection method using AdaBoost using dynamic time warping and extreme learning machine respectively as the weak classifier to find the most discriminative features. SVM is used on the selected features for macro-expression recognition classification. Geometric deformation features were applied in Kotsia and Pitas (2007), these features were derived from the calculated displacement of 108 facial nodes between the initial frame of the expression sequence and the final full formed expression frame. SVM was applied as the classifier with results of 99.7% accuracy on the original CK dataset Kanade et al. (2000). Many methods have applied only appearance based features using many different texture descriptors, while some have used the whole face for feature extraction others have used local regions of interest based upon areas of key importance to the formation of the facial expression as described by facial action units (Ekman, 2013). When extracting features from the whole face an overlapping block pattern is often used, one method applying this technique was Pietikäinen et al. (2011) which utilised the popular LBP-TOP to extract features from 9×8 overlapping blocks of the whole facial region and a SVM classifier. In Zhang et al. (2012) gabor wavelets were used as appearance

features followed by PCA to reduce the feature vector dimensions with classification performed using a sparse representation classifier. Hybrid approaches where also successfully applied with Youssif and Asker (2011) applying geometric and appearance based features, with 19 geometric based features consisting of Euclidean distance and the angles of the mouth. They apply 'Canny' edge detection to generate an edge map and split the face into regions 4×4, edge orientation histograms are applied as the appearance features. An artificial neural network was then applied to classify the expression.

For micro-expression FER with the introduction of the CASME II data set in Yan, Wang, Liu, Wu and Fu (2014) saw a method using Local Weighted Mean (LWM) transformation for face warping and LBP-TOP for appearance based features extracted from a 5×5 grid was used. SVM was used as a classifier and set a benchmark of 63.41% accuracy on the data set. Subsequently Wang et al. (2015) proposed a variation of LBP-TOP namely LBP-Six Intersecting Points (LBP-SIP) with results showing both accuracy and computational advantages over LBP-TOP. A method of extracting features using Robust PCA and Local Spatio-temporal Directional Features from 16 specific region of interest of the face was proposed in Wang, Yan, Zhao, Fu and Zhou (2014). Liong et al. (2014) proposed an optical strain weighted feature extraction scheme. Optical strain magnitudes depict motion and are extracted and pooled spatio-temporally to obtain block-wise weights. Features are then scaled based upon the weighted importance of each block. In Liu et al. (2015) a new optical flow based feature type named Main Directional Mean Optical-Flow (MDMO) which is a variant of Histogram of Optical Flow (HOOF), this feature was applied to 36 separate regions of interest on the subject's face. This method produces a very compact feature vector with each region being described by only two values the direction and magnitude of the optical flow vector. A pre-processing technique was also proposed using an optical-flow-driven method to align all frames of a micro-expression video clip (Liu et al., 2015). State-of-the-art performance was realised in Li et al. (2015) providing an accuracy of 78.14% with a leave-one-video-out testing protocol on CASME II. A three stage pre-processing technique was used to overcome the subtle motion issue of micro-expressions. A LWM transformation is applied to align each face within the database to a reference face shape. Motion magnification is then applied to the sequence using the Eulerian video magnification method Wu et al. (2012) to boost visibility of the expression. A temporal interpolation model was used to overcome short video lengths through

the generation of extra frames and also to standardise the frame numbers of all videos. A spatio-temporal local texture descriptor was proposed which is a variant of HOG named Histogram of Image Gradient Orientation (HIGO) for feature extraction and a linear SVM for classification. Due to the subtle motions geometry-based currently do not posses the fidelity for application in micro-expression recognition.

Image Based Deep Learning Methods

In the subcategory of macro-expression there has been a significant amount of research with many deep learning methods proposed reaching state-of-the-art performance when performing FER on single images. Initially FER data sets were relatively small in sample size and it was found that training CNN architectures on these from randomly initiated weights was prone to overfitting. As demonstrated across many other task domains applying deep learning techniques researchers used pre-trained models of existing object detection CNN architectures like AlexNet, VGG and ResNets to overcome the overfitting issue. It has been shown that the pre-trained model used affected the FER accuracy, fine-tuning on a model pre-trained for the task of face recognition produced better FER accuracy than a model fine-tuned on an ImageNet pre-trained model (Li and Deng, 2018). More complex fine-tuning strategies have been proposed which have shown improved accuracy, one such method is the two-stage training algorithm of FaceNet2ExpNet (Ding et al., 2017). In this strategy the fine-tuned face net serves as a initialisation point from which the expression net is then trained for the convolutional layers only. The trained convolutional layers are then frozen and the fully connected layers of the network are trained from scratch.

A further technique employed to increase accuracy of emotion recognition is that of altering the existing CNN architectures with auxiliary blocks and layers. These extra blocks and layers are introduced with the purpose of enhancing the expression-related representation capability of learned features. HoloNet (Yao et al., 2016) was proposed specifically for emotion recognition, using a combination of ResNet and the concatenated rectified linear unit to create a network with increased depth to learn multi-scale features to capture expression variations. Supervised Scoring Ensemble (Hu et al., 2017), is another method which aimed to enhance the supervision degree, where three types of supervised blocks were embedded in the early hidden layers of the mainstream CNN for shallow, intermediate and deep supervision, respectively.

The use of multiple networks rather than a single network has shown to produce better performance in some circumstances. To gain performance benefit from multiple networks the first key aspect is to consider is the diversity of the networks. It has been shown to be important that the networks compliment each other and are not simply replications. To achieve diversity both data augmentation and architectural changes (e.g filter size, number of layers and multiple random seeds for weight initialisation) are strategies that have been successfully applied (Li and Deng, 2018). Different network types can also be used for example a CNN trained in a supervised way and a convolutional auto-encoder trained in an unsupervised way were combined for network ensemble, proposed in Hamester et al. (2015). The second aspect for consideration is where the multiple-networks are joined, generally either feature or decision level methods are used. Feature level joining is most commonly achieved through concatenation of the separate feature vectors derived from each network, these are then passed through a fully connected layer to produce a prediction (Bargal et al., 2016). Decision level joining usually applies a mechanism such as majority voting, simple average and weighted average to determine a final prediction. Weighted average produces a challenge of finding the optimum weighted values, a number of strategies have been proposed for this including an exponentially weighted average based on the validation accuracy which emphasise qualified individuals (Kim et al., 2015) while in Pons and Masip (2018) a further CNN is used to learn weights for each individual model.

Loss functions have been another area of research, like face recognition both triplet and center loss have been applied in this area. Center loss inspired two more variations which have been applied to emotion recognition these being island loss and locality-preserving loss. Island loss was designed to further increase the pairwise distances between different class centres, in practice the island loss is combined with softmax loss to produce the total network loss, with island loss being calculated at the feature extraction layer and the softmax loss calculated at the decision layer of the model (Cai et al., 2018). Locality-preserving loss (Li et al., 2017) was created to pull the locally neighbouring features of the same class together so that the intra-class local clusters of each class are compact.

Multi-task and cascaded networks have also been utilised for emotion recognition, where a multi-task method is trained on multiple associated tasks in addition to emotion recognition with the intention that the associated tasks assist in learning better feature representations. Cascaded net-

works alternatively consist of various modules for different tasks and are combined sequentially to construct a deeper network. Cascaded networks have applied other architectures such as multi-layer Restricted Boltzmann Machines and stacked autoencoder (Liu, Han, Meng and Tong, 2014; Rifai et al., 2012; Mengyi Liu et al., 2013), which at the time of publication the results where impressive the direction of the research since 2016 has moved to exclusively CNN and multi-task architectures. Multi-task approaches have included simultaneously training CNN models for both FER and facial recognition. The proposed identity-aware CNN (Meng et al., 2017) employs two CNNs with shared weight parameters, where the input to one network is expression based and the other identity based. Multiple loss functions are used, where auxiliary layers are introduced to use contrasitve loss for the individual task and softmax loss is applied to the concatenated features of the joint tasks. An all-in-one CNN model (Ranjan et al., 2017) was proposed to simultaneously solve a diverse set of face analysis tasks including face detection, landmarks localisation, pose estimation, gender recognition, smile detection, age estimation and face identification and verification. The network is first initialised using the weights from a pre-trained on model for face recognition, the features for each of the subsequent task is branched out a different levels of the CNN where there exists a fully connected layer and associated loss function for each task. Training for this network is conducted using a fully supervised method where 6 different training data sets where applied covering each task totalling almost 1 million images.

Video Based Deep Learning Methods

A range of techniques including frame aggregation, geometry-based features and multi-stream have been applied to perform FER on videos using deep learning methods. Frame aggregation techniques are a natural extension of image based FER, where the features or predictions from image based CNN methods are fused for different frames in a video. Methods for prediction level frame aggregation include simple concatenation of the probability vectors for each frame. When frame length differs from video to video which is common in the available data sets, frame averaging and frame expansion methods have been employed to create a standardised feature vector size. Statistical coding methods for frame aggregation do not require a fixed size, they use the average, max, average of square or average of maximum suppression vectors, to produce the final feature vector. When performing frame aggregation at feature level many methods have been used to fuse the features of each frame, these include concatenation of the mean, variance,

minimum, and maximum of the features over all frames (Bargal et al., 2016). Matrix-based models such as eigenvector, covariance matrix and multi-dimensional Gaussian distribution have also been proposed for this purpose (Ding et al., 2016; Liu, Wang, Li, Shan, Huang and Chen, 2014).

Geometry-based features have also been applied these methods relying upon accurate landmark localisation techniques to generate the features. In Yan et al. (2016) facial landmark points from frames over sequences were concatenated and normalised to provide the input to a CNN as image-like map. Instead of using all landmarks as a global feature Kim et al. (2017) split the landmarks into groups based upon the structure of the face, each group was then fed separately into the network hierarchically, this method proved to be efficient for feature encoding at both group and global level.

Recurrent Neural Network (RNN) and it's derivative the Long Short-Term Memory (LTSM) networks have been applied to FER (Li and Deng, 2018). The design of RNNs mean they naturally accept sequences as inputs and have the capability to model the temporal nature of data sequences like those found in video. LTSM is an improved RNN which offers flexibility to handle varying-length sequential data with lower computation cost. As a natural extension to the use of 2D CNNs in image FER tasks, using 3D CNNs for generation of spatio-temproal features has gained momentum in recent years. The C3D (Tran, Ray, Shou, Chang and Paluri, 2017) architecture was the first 3D CNN publicly released for video action recognition and has been applied in FER but this has primarily been as part of a multi-stream network where other input data is passed into an accompanying network that aids the predictions. Poor performance when using the C3D architecture as reported in Vielzeuf et al. (2017) is the primary factor for multi-stream methods. In Jung et al. (2015) they use both landmark geometry-based features and the video frames as the input to two separate networks, then propose an integrated final network layer which uses a proposed joint fine-tuning method. While in both Vielzeuf et al. (2017) and Fan et al. (2016) a C3D architecture was incorporated with a LTSM network and RNN respectively, booting the performance over using the C3D alone by 5% on the AFEW data set in Vielzeuf et al. (2017). In recent research by Kim et al. (2016) they introduced deep learning features for micro-expression recognition. They proposed a new method consisting of two parts. First, the spatial features of micro-expressions at different expression states (i.e., onset, onset to apex transition, apex, apex to offset transition and offset) are encoded using CNN. Next, the learned spatial features with expression-state constraints

are transferred to learn temporal features of micro-expression. The temporal feature learning encodes the temporal characteristics of the different states of the micro-expression using LTSM. The time scale dependent information that resides along the video sequences is consequently learned by using LTSM.

2.5.3 Medical Tasks

There are a number of medical conditions in which automated facial analysis systems have the potential to be applied for both diagnosis and rehabilitation tracking purposes, these include but are not limited to Parkinsons disease, facial palsy, stroke and dermatological disorders that present on the face. The motivation for such systems is to primarily aid the clinical decision making process to facilitate better patient outcomes. In general, systems that use facial analysis for medical tasks are less well researched than those covered in the previous sections. A primary reason for this is that medical data sets tend to be small which limits the use of deep learning and where data sets exist many are private and therefore not available across the research community. Further to this while medical conditions like Parkinson's disease affect facial motion there are other associated symptoms which have been prioritised as the primary diagnostic pathway. In this section there will be a brief overview of the current literature for Parkinsons disease, dermatological disorders, facial palsy and stroke detailing the areas where facial analysis systems can be beneficial and the computer vision and machine learning methods that have been applied to date in these areas.

Parkinson's Disease

Parkinson's Disease (PD) is a chronic, progressive, multi-lesion and neuro-degenerative condition which is caused by the loss of the neurotransmitter dopamine within the person (Pereira et al., 2016). The symptoms of PD present in number of ways primarily a individual will present involuntary muscle movements (dyskinesia) across the whole body which includes twitches, jerks, twisting or writhing movement (Li et al., 2018). Secondary symptoms include include masked face (hypomimia) which is the loss of facial expressions creating a mask-like appearance, shuffling gait when when walking, abnormally small or cramped handwriting (micrographia) and speech difficulties. Shuffling gait can dramatically increase the likelihood of falls resulting in further possible injury to the patients especially those that are elderly or frail. As there is no cure at present for PD, the major benefit that researchers seek is early identification of the disease which enables earlier

intervention so that the progress of PD can be managed effectively by a trained clinician. This desire has lead to various research paths that have involved machine learning though the majority of the research to date does not use the face analysis as a diagnostic tool. Speech analysis Zhang (2017), gate analysis (Li et al., 2018) and handwriting analysis in Pereira et al. (2016) have all applied deep learning to varying degrees of success. Research which has applied face analysis to PD includes Vinokurov et al. (2016) in which a method was proposed for quantifying hypomimia through the application of a depth sensing 3D camera. A commercial software product called Faceshift which tracks facial landmarks across a video sequence and outputs the intensity level of 51 facial action units. The data set used included 14 patients between the ages of 58 to 84 with varying levels of hypomimia and a further 15 Controls aged between 48 to 84. Each participant was recorded using a 3D camera while performing predefined actions such as facial expression and answering a set of standard questions. Each patient was graded using the Unified Parkinson's Disease Rating Scale which is the most common scale used in clinical studies to grade hypomimia level between 0 and 4, where 0 is normal and 4 represents complete loss of facial expression. Supervised linear regression was used to predict a hypomimia level from the extracted action unit features of the video sequences where the results present were over 90% correlated to the predictions provided by two experts. Further research was conducted in Rajnoha et al. (2018) where the focus of the study was in the possibility of detecting hypomimia using the simple static face analysis. The sample for this research included 50 PD patients and 50 age and gender matched healthy controls. Parameterization based on face recognition methods in combination with conventional classifiers including random forests were used to automatically identify PD hypomimia. Among the classifiers, the decision tree algorithm achieved the best accuracy (67.33%). The results suggest that automatic static face analysis can support PD hypomimia diagnosis, nevertheless it is not accurate enough to outperform the approaches based on video-recordings processing. At present there has been no published research applying deep learning methods for face analysis, this is possibly due to the limitations in obtaining data sets with a sufficient samples size.

Dermatological Disorders

Dermatological disorders are one of the most widespread diseases across the globe. While common the diagnosis process can prove difficult due to the complexities of skin tone, colour, presence of hair and the variety of specific conditions. While these conditions are not limited to the face

many dermatological disorders can present in the face some of which when not diagnosed correctly and treated can lead to terminal illnesses such as cancer Haofu Liao et al. (2016). There has been significant research conducted in this area using both traditional and CNN based computer vision techniques for skin analysis. For example in Sumithra et al. (2015) a method for skin lesion segmentation employing region-growing was proposed. Feature extraction was performed using facial skin texture and colour, this was then classified into 5 disease classes using both SVM and k-Nearest Neighbour (k-NN). One of the first CNN based approaches for diagnosis of dermatological disorders was presented in Haofu Liao et al. (2016), where an AlexNet (Krizhevsky et al., 2012) architecture was trained using transfer learning for predicting skin disease and lesion labels, respectively. The authors collated a data set of 75,665 skin disease images from six different publicly available dermatology atlantes from which 90/10% split was applied for training and evaluation. The experimental results showed top-1 and top-5 accuracy's for the disease-targeted classification are 27.6% and 57.9% with an AP of 0.42. Lesion-targeted skin disease classification showed more potential with a higher AP of 0.70. Both Zhang, Wang, Liu and Tao (2017) and Zhang, Wang, Liu and Tao (2018) investigate the diagnosis of four specific skin deases these being melanocytic nevus a common benign skin tumour, seborrheic keratosis a benign epidermal tumour caused by delayed cellular maturation, basal cell carcinoma which is the most common human skin cancer and psoriasis a common and easy recurrence of chronic inflammatory dermatological disorder. The treatment paths for each of the four classes disease are very different and therefore misdiagnosis or delayed diagnosis has the potential to lead to incorrect medical care, delayed treatment or no treatment. The data set applied in both papers was obtained from the Department of Dermatology, Peking Union Medical College Hospital. The 28,000 skin images within the data set were professionally collected using the MoleMax HD 1.0 dermoscopic device. The training and testing of the methods two data sets A and B applied a sub-sample of the total data set, with 1,067 images annotated in data set A, while data set B contained 132 images for each disease to avoid potential bias caused by the unbalanced distribution. In Zhang, Wang, Liu and Tao (2017) a pre-trained Inception v3 model was fine-tuned for the task skin disease classification task with the average accuracy using F-score as the metric was 86.54% for data set A and 85.86% for data set B. While in Zhang, Wang, Liu and Tao (2018) the initial paper was expanded to highlight areas of incorrect classification, the same model and data sets were used due to random test and training set generation using a 10 fold method, the results shown an improvement over the origi-

nal paper which were 87.25% for data set A and 86.63% for data set B. Recently Kumar Patnaik et al. (2018) have applied a number of deep learning models to the identification of dermatological disorders from images. The Inception V3, Inception ResNet V2 and Mobile Net (A lightweight architecture that can be run on a smart phone) CNN architectures were applied to the problem both individually and jointly. The joint method used a simple maximum voting scheme based upon the output of the individual networks. The fine-tuning of ImageNet based models was performed using a training set of containing 20 classes of dermatological disorders, the size of the data set is not specified in the paper. From the results they achieved a level of 88% accuracy for the 20 classes which they report to be higher than previous non-deep learning approaches (Kumar Patnaik et al., 2018).

Facial Palsy

Facial palsy is a condition which affect the facial nerve and impacts the motion of the muscles in the patients face to varying degrees, as this condition is predominantly diagnosed through facial motion analysis (secondary symptoms can also include affecting the taste and hearing of the individual) there is a potential to apply a facial analysis system to aid with the diagnosis of this medical condition (Wang, Dong, Sun, Zhang and Wang, 2014). To date there is a limited range of computer vision based research on facial palsy which is mostly based on the use of traditional methods. Previous work on facial paralysis and atypical faces includes Wang, Dong, Sun, Zhang and Wang (2014) which propose methods aimed at recognising the patterns of facial movement from facial paralysis patient images using specific facial landmarks and LBP features with a SVM classifier, in both an ASM (Cootes et al., 2001) was used for landmark localisation, fitting 68 facial landmarks which are used as guides for the feature extraction regions. To perform accurately the ASM is trained on the facial palsy data set used for evaluation proposes in the research. This work was furthered in Wang et al. (2016) where a novel method for automatically evaluating the degree of facial palsy was proposed. In this method they considered both static and dynamic facial asymmetries applying a technique which extracted differences between left and right half-faces with different facial expressions. Again the ASM method was used to predict 68 facial landmarks and then texture features were extracted using LBP for static analysis, further to this to analyse dynamic asymmetry facial landmarks were used to measure the change of speed in corresponding regions for each facial muscle movement. The static and dynamic features were combined to eval-

uate the final facial palsy grading. The class of grading related to the House-Brackmann grading system which consists of 6 scales (Fattah et al., 2015). 61 patients where used from a private data set where each patient was recorded performing 5 facial movements (raise eyebrows, close eyes, screw-up nose, plump cheeks and open mouth). The authors note that evaluation of the degree of facial palsy using only static facial asymmetry is not good enough and that dynamic motion is key to a high accuracy of prediction. The best results were found combining both static and dynamic where the highest accuracy was found using the close eyes videos which a high 97% accuracy. Sajid et al. (2018) have recently applied a CNN based method to a larger data set of 2000 images collected from various online sources. This method uses a VGG based network but also used two smaller initial networks consisting of 2 convolutional and 2 max pooling layers each. The facial image and a mirror image are then initially imputed to these 2 smaller networks and then the output is passed to the VGG part of the network. House-Brackmann grading was used as class labels and they reported a high level of accuracy of 92.60% in comparison with 88.90% obtained when applying a method using bilateral facial landmarks with SVM and LDA to the same data.

Stroke

Stroke is one of the leading causes of death and disability worldwide with approximately 15 million people suffering a stroke each year. It has been shown that reduction in the treatment time from the onset of stroke by as little as 15 minutes can have a significant decrease in death and disability but due to the suddenness of the onset identifying this is difficult. While machine learning and computer vision has been applied for early stroke detection systems Chin et al. (2017) these use brain scan images not facial analysis. In recent time facial muscle weakness has been highlighted as a possible pre-screening tool for stroke, in Bandini et al. (2018) they propose a method to aid stroke rehabilitation, this work is highly correlated to the research on facial palsy as they both look at face muscle and symmetry. In this work the authors recruited 23 volunteers where 12 had suffered a stroke and 11 were control subjects. Each participant was filmed performing 7 specific facial actions including the face at rest, jaw opening and the speaking of a common sentence. From these videos landmark localisation was performed using the supervised descent method Xiong and De la Torre (2013) and selected landmarks were used as features representing range of motion and asymmetry of face motion to train a SVM. The results show it was possible to discriminate patients post stroke from control subjects with high accuracy of 87%.

2.6 Video Action Recognition

While not specifically a facial analysis task general action detection is well established within the field of computer vision. Action recognition varies from identifying different sports based upon the motion of the participants (Karpathy et al., 2014) to identifying general human activities such as brushing teeth or blow drying hair (Soomro et al., 2012). The problem of action recognition is as a sequence of video frames. This section contains a review of the some further techniques that have been applied in general action recognition that could potentially be applied in the facial analysis domain.

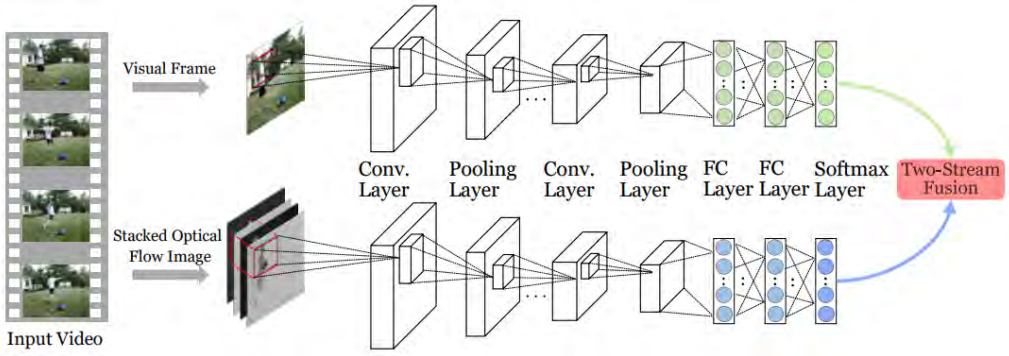

Figure 2.11: The two-stream architecture consists of two separate CNNs, the first takes an image from a video sequence as the input, the second a stack of pre-processed optical flow features of the video sequence. The output of both streams is then fused before a final prediction is made (Simonyan and Zisserman, 2014).

2.6.1 2D Convolutional Neural Networks

The two-stream 2D CNN-based approach to action recognition has proven a popular technique with this field. Originally proposed by Simonyan and Zisserman (2014) the two-streams refer to one stream which takes RGB images data for appearance features and the second stream takes stacked optical flow features to provide motion information (Figure 2.11 provides a visual overview of this approach). The combination of both appearance and motion information resulted in improved results in the benchmark action recognition performance at the time of publication on the UCF-101 (Soomro et al., 2012) and HMDB-51 (Kuehne et al., 2011) data sets. Following the proposal of the two-stream method this has been further studied to improve action recognition performance (Feichtenhofer, Pinz and Wildes, 2016; Feichtenhofer, Pinz and Zisserman, 2016; Feichtenhofer et al., 2017). As the stacked optical flow features are required to calculated as a

pre-processing step there is an computational cost to this architecture.

2.6.2 3D Convolutional Neural Networks

Recently 3D CNN-based approaches have begun to show promise in the task of action recognition as they have been able to leverage the introduction of large-scale training data sets. In contrast to the two-stream methods described previously these only have a single input to the network in the form of a video stacked as a set of individual frames. The extension to 3D convolutional kernels intuitively allows for the shift from the spatial domain to features in the spatio-temporal domain, where the 3^{rd} dimension captures motion across the temporal plane. One of the first fully 3D CNN based models was proposed in Tran, Ray, Shou, Chang and Paluri (2017) which they termed C3D, this archintecture is shown in Figure 2.12. The model used fully 3D convolutional kernels applying the Sports-1M data set (Karpathy et al., 2014) for training of the models parameters. Through model evaluations they found that $3 \times 3 \times 3$ convolutional filters produced the best level of performance. Expansion of the temporal length showed further improvements in recognition accuracy to the 3CD model were reported in Varol et al. (2018). In the same study it was reported that applying optical flows as inputs to the 3D CNN resulted in a higher level of performance than can be obtained from RGB inputs, but that the best performance could be achieved using a combination of RGB and optical flows. 3D CNN architectures using the Kinetics data set for training from scratch displayed results that were comparable with the results of an ImageNet trained 2D CNN architectures in Kay et al. (2017). Complex 3D CNN architectures have begun to be explored, while initial studies were limited to shallow ResNet architectures (Hara et al., 2017) more recently this has been expanded to much deeper ResNets with up to 152 layers and other architectures including ResNeXt-101 (Hara et al., 2018). ResNeXt-101 achieving the best performance on the Kinetics test set. The study also found a Kinetics data set pretrained simple 3D architectures outperforms complex 2D CNN architectures both on the UCF-101 and HMDB-51 data sets respectively.

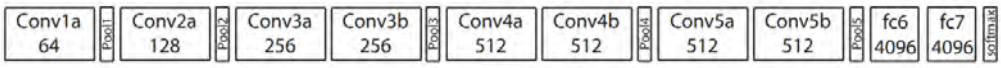

Figure 2.12: C3D architecture was the first method to introduce the use of 3D convolutions for action recognition from videos. The performance of the network was limited due to the shallow network depth limited to 8 total convolutional layers as shown and a lack of large scale training data (Tran, Ray, Shou, Chang and Paluri, 2017).

2.7 Summary

In this chapter an extensive literature review was presented with the aim to address objectives 1 and 2 as defined in section 1.3. In this section firstly, a summary of how the information gained from the literature review successfully completed objectives 1 and 2 is given. Secondly objectives 3 and 4 are further defined based upon the findings of the literature review. Finally a summary of key considerations discovered through the review which guide the direction of the research is presented.

2.7.1 Objective 1 Summary

Key face detection (section 2.3) and landmark localisation (section 2.4) methods applied across the previous literature were reviewed while also defining the common evaluation metrics and public data sets which can be applied to both training and evaluating models in both of these tasks. The introduction of deep learning and more specifically CNN based methods have seen an increased level of performance in both tasks, yet still challenges remain. Face detection specifically has a number of distinct challenges still to be addressed, these include detection of very small and out-of-focus faces, while increasing the face detection precision is also a key challenge.

2.7.2 Objective 2 Summary

Section 2.5 provides a review of 3 distinct application areas for facial analysis systems these being facial recognition, facial expression recognition and medical tasks. This review highlighted that numerous methods and techniques have been proposed, many of which are common across all tasks that have shown good performance, specifically it has been identified that deep learning methods have had a positive affect on the prediction accuracy. These advancements have been more significant when dealing with static image based systems, those that analyse video data have not yet advanced at the same rate and in some cases such as two-stream networks still also rely on traditional hand-crafted features to achieve a high level of accuracy for the given task. It was identified that there is a significant amount of research on the specific tasks of facial recognition and facial expression recognition while medical areas provide a set of tasks with much research potential if there is suitable data available for training and evaluating deep learning methods.

2.7.3 Defining Objective 3

The challenge of increased precision in the face detection task identified in this chapter forms the major research focus for objective 3. Currently the number of false positive face detection's predicted even in the state-of-the-art methods can be very high. Addressing this challenge is important as when developing specific facial analysis systems, there needs to be a high level of confidence that the detection is indeed a face and not some other object from within the image. This is specifically important in areas such as medical or security based systems where no detection of a present face is preferable to false detection which could cause analysis to be performed on non-face data with potential significant consequences. Even outside of these specific domains false positive detection's can be problematic as it introduces noise in the facial analysis pipeline which could affect accuracy of any system. While added computational overhead (for example landmark localisation could be performed on all false positives) and potentially require manual intervention to remove the incorrect detection's which is time consuming are also motivations for pursuing false positive reduction in face detection within this objective.

2.7.4 Defining Objective 4

Within the review carried out for objective 2, medical tasks were identified as an application area where there is a significant gap in the current research. More specifically asymmetrical facial motion is a key elements in both stroke and facial palsy analysis, while Sajid et al. (2018) highlighted that there is online availability of facial palsy data in sufficient volume to train a deep learning method. Therefore the focus of objective 4 is the investigation of asymmetrical facial motion and more specifically facial palsy due to the available training data and the current gaps in research within this specific application area.

2.7.5 Key Considerations

There are a number of key considerations in terms of strategies and areas that have been applied across the literature to increase performance in deep learning tasks which form the inspiration for the proposed methods described within the rest of this thesis. To summarise:

- CNN architecture and depth
- Cascaded and multiple-network models

- Multi-task learning
- Loss function selection
- Fine-tuning (Transfer learning)
- Data augmentation

When considering factors that affect the direction of research specifically when applying deep learning based methods, computational resource and software availability must also be considered when selecting applicable methods. In terms of computational resource as has been discussed within this chapter, as deep learning has progressed one of the driving factors of this is architectural complexity. This often equates to a deeper network with a larger number of parameters thus requiring more computational resources. It is therefore important that any methods applied within this research could be achieved on the available hardware. Regarding software availability there are many platforms/frameworks which allows for the training and running of deep learning alongside traditional methods including Matlab and other Python based deep learning frameworks such as the Microsoft Cognitive Toolkit (CNTK), PyTorch and TensorFlow. Across these options not all architectures, functionality or pre-trained models exist.

Given the findings of the literature research, the rest of this thesis applies this information to the further research of facial analysis tasks, specifically those of face detection and facial symmetry analysis. Chapters 5 and 6 of this thesis are concerned with the task of face detection and the challenge of increasing precision. While chapter 5, 7 and 8 detail proposed methods for facial symmetry analysis and more specifically facial palsy grading.

Chapter 3

Faster R-CNN

3.1 Introduction

In the previous literature review chapter an insight was given to the task and challenges of face detection and how this task forms a foundation stage within facial analysis systems. This review highlighted a number of state-of-the-art CNN architectures which provide real-time object detection one of these being Faster R-CNN. The focus of this chapter is to provide a comprehensive overview of Faster R-CNN method and the training protocol applied to re-purpose this to the specific task of face detection. The purpose for this overview is that Faster R-CNN is applied throughout the remaining chapter of this thesis.

This chapter is organised as follows: Initially in section 3.2 an overview of the Faster R-CNN method and the model architecture are presented. Section 3.3 details the multi-task loss functions used with in Faster R-CNN network training. Transfer learning specifically in relation to fine-tuning the general object detection model for the more specific task of face detection is then detailed within section 3.4. In section 3.5 Finally a conclusion of the method with a rationale of why this specific method has been chosen to be applied within the body of this research.

3.2 Network Architecture Overview

The Faster R-CNN network builds on the success of the Fast R-CNN and R-CNN object detection architectures (Girshick, 2015), both of these methods provided advancements to deep learning

based object detection at the time of publication. The primary motivation for the development of Faster R-CNN was the goal of achieving real-time object detection. To briefly recap from section 2.3.2 many of the successful object detection methods relied upon external region proposal, this was identified as a bottleneck which impacted the computational time, thus the idea for Faster R-CNN was to integrate region proposal into the deep learning architecture of Fast R-CNN, removing the bottleneck and allowing for real time detection. Architecturally Faster R-CNN can be described as containing two modules with a shared set features extracted via a CNN. The first module is the RPN a deep fully convolutional network responsible for region proposal. The second module is the object detector which determines the class of a proposed region of interest. The entire architecture is a single unified network for object detection as shown in figure 3.1.

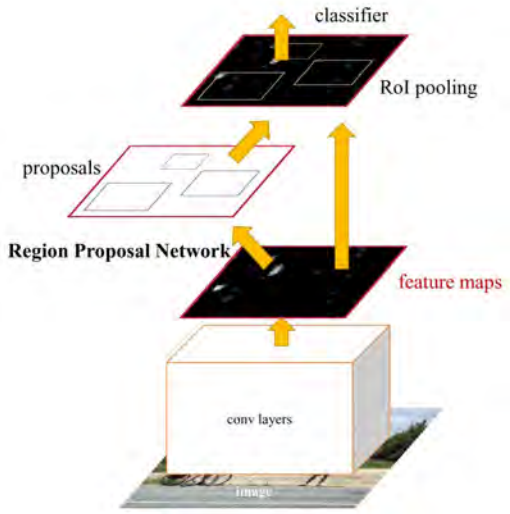

Figure 3.1: Faster R-CNN Architecture Overview (Ren et al., 2015).

3.2.1 Shared Convolutional Layer

The shared convolutional layer is the initial step in the Faster R-CNN architecture which takes as an input an image of $n \times m \times 3$ spatial size. The shared convolutional layers are not fixed to a specific architecture, in the original research both the VGG16 architecture with 5 convolutional layers (Simonyan and Zisserman, 2015) and the 13 layer Zeiler and Fergus model (Zeiler and Fergus, 2013) where applied and evaluated, while AlexNet (Krizhevsky et al., 2012) has also been used. Independent of the architecture applied these layers are trained to learn a feature map which is passed to both the RPN and object detector modules of the architecture.

3.2.2 Region Proposal Network

The task of the RPN module with the Faster R-CNN model is to output a set of Region Of Interests (ROIs) which contain a relevant objects. Given a set of features extracted from an image as an input the RPN output is a set of bounding box regions each given an objectness score. Objectness score is a measure of an object vs background of an image. The structure of the RPN consists of an intermediate convolutional layer which takes as an input an $n \times n$ spatial window from the shared convolutional layers feature map, where $n = 3$ in the original paper. The output from the intermediate layer is first fed into a Rectified Linear Unit (ReLU) layer then into two 1×1 convolutional layers, the first 1×1 layer being an objectness scoring layer and the second a directly related bounding box regression layer. The output of these layers for each proposed ROIs is the objectness score as a probability and a bounding box defined by a four-tuple (r, c, h, w) where (r, c) defines the top-left corner of a region while (h, w) represent a regions height and width. Anchors are a key concept behind the learning of ROIs. ROIs are learned through the application of a sliding spatial window technique, at the centre of each sliding window within the feature space there exists k anchor box, the paramter of the proposals are relative to the anchors position. Each of these k anchors has an associated scale and ratio to better detect different size and shape objects, where in the original paper $k = 9 = 3 \times 3$ where there are 3 scales and 3 aspect ratios. The total number of anchors is $W \times H \times k$, where W and H are the height and width of the convolutional feature map. Importantly the anchor and the proposal learning function are translation invariant therefore if an object is translated within an image, the proposal translates and prediction of the proposal should happen in either location.

3.2.3 Object Detector

The object detector module of Faster R-CNN is responsible for determining the final objects location and the associated classes. As an input this module takes both the feature map from the shared convolutional layers and also the ROIs from the RPN. The first step is to perform the task of ROI pooling Girshick (2015), this pooling process takes both inputs and for every valid object region proposed by the RPN the equivalent spatial region of the shared feature map is sampled. As the sampled spatial region can be a variety of sizes, to produce a uniform final pooled feature size the sampled spatial region is first divided into $H \times W$ areas where $H = W = 7$ is commonly

applied within the literature. Similar to the commonly applied max pooling the maximum feature map channel value from each area is taken as the final output value for that specific location in the final $H \times W$ output. ROI pooling is a key technique which allows the RPN to be integrated within the network as it allows for the sharing of the features. Following ROI pooling the $H \times W$ are then passed to a 3 fully connected layers, one which perform dimension reduction and then the other 2 which are responsible for predicting the final objects bounding box and predicting the object class.

3.3 Multi-task Loss

Within the Faster R-CNN architecture there are four tasks which require learning, two tasks relate to the RPN and region proposal learning for bounding box regression and a binary classification problem, while two similar objectives exist for the object detection layers, for the object bounding box regression and object classification for n classes respectively. The methodology applied to learn these tasks simultaneously is often described as multi-task loss in the literature and is defined as:

$$loss_{total} = \sum_{i=n} loss_i \lambda_i, \tag{3.1}$$

where the total network loss for all task $loss_{total}$ is the total loss of the individual loss corresponding to the ith task as $loss_i$. A weighting parameter λ_i is applied to balance the learning priorities between tasks.

The aim of RPN training is to learn the regions which contain an object for detection against those that represent the background of an image through the use of a random subset of all potential anchor points. Given a set of k anchors we assign a binary class label based upon the IoU of the region of interest and the ground truth object. An anchor that has an IoU overlap higher than 0.7 is assigned a positive/object detection label while those anchors registering an IoU of less than 0.3 are labelled as negative/background. Other anchors in which the IoU value lies between 0.7 and

0.3 are not used for training. The RPN classification loss is defined as:

$$rpn_loss_{cls} = \frac{1}{N_{cls}} \sum_{i=n} -(1 - p_i^*) \cdot log(1 - p_i) - p_i^* \cdot log(p_i) \tag{3.2}$$

where using the softmax loss function ROI classification is learnt for an object ($p_i = 1$) and a non-object ($p_i = 0$), where p_i^* is the ground truth class label and p_i the predicted class for the ith anchor respectively. This loss function is normalised by N_{cls} which is the mini-batch size. Bounding box regression for the RPN is defined as:

$$rpn_loss_{reg} = \frac{1}{N_{reg}} \sum_{i=n} p_i^* smooth_{L1}(t_i - t_i^*) \tag{3.3}$$

where for the ith anchor the $L1$ loss between the ground-truth box t_i^* and the predicted bounding box t_i is calculated. Both t_i^* and t_i are vectors representing the 4 parameterised coordinates of the predicted bounding box. Only positive anchors affect the loss as described by the term $p_i^* smooth_{L1}$. N_{reg} the total number of anchors normalises the loss function. The function $smooth_{L1}(t_i - t_i^*)$ is defined as:

$$smooth_{L1}(x) = \begin{cases} 0.5x^2 & \text{if } |x| < 1, \\ |x| - 0.5 & \text{if } otherwise. \end{cases} \tag{3.4}$$

where the input x is the distance between the prediction and the ground truth. Smooth L1 loss behaves like a combination of L1 loss and L2 loss. When the absolute distance x is high the loss behaves as L1 loss, while as x gets closer to 0 the behaviour is more correlated to L2 loss (Figure 3.2 presents a graph of this behaviour). Smooth L1 loss is critical for the network to effectively learn the regression mapping from anchor target to object detection bounding boxes (Girshick, 2015). The key attribute of Smooth L1 loss is the robustness to outliers.

The loss functions used to learn discriminative features to predict the final object classification are defined in the same manner as those for learning the RPN loss. The main difference is that while the RPN loss is posed as a binary class problem, object detection may have n classes where $n = 1000$ in the originally defined model. The detector loss is responsible for driving feature

learning to provide the capability to discriminate between the n classes.

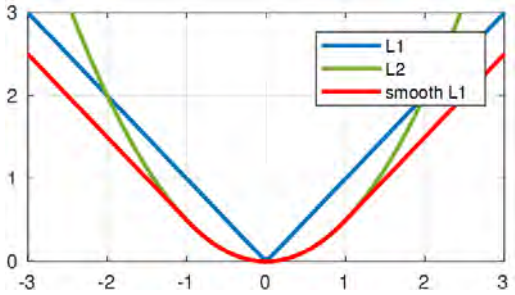

Figure 3.2: Smooth L1 loss function plotted against L1 and L2 loss (Girshick, 2015)

3.4 Faster R-CNN for Face Detection via Transfer Learning

To effectively apply the Faster R-CNN model for the task of face detection, there is a need to train a model for the specific task. The purpose of this is to provide a base model for the research conducted throughout this thesis. To achieve this the method of transfer learning was adopted, this technique enables the use of previous knowledge learned for a related task to be applied as the basis for learning a new task. Transfer learning is common across the research (Haofu Liao et al., 2016) and is sometimes referred to as model fine-tuning, although technically fine-tuning is a specific approach to finalising the training of the model the terms are sometimes used interchangeably. The popularity of transfer learning in deep learning stems from a lack of data, it is well understood that deep learning is most successful when the data sets used to train models contain a very large number of samples (e.g 1.2 million in ImageNet (Krizhevsky et al., 2012)). For many applications this size of corpus does not exist and therefore training a deep learning model from scratch is not a viable tactic. Even in situations where enough data does exists computational resource and time constraints can be an issue. Training from scratch is a term used to describe the training of a model from a set of randomly initialised weight parameters. It is this that has made transfer learning so popular, specifically in the domain of computer vision and image detection the availability of models trained on ImageNet provide an ideal base model. To apply transfer learning the pre-trained model is used as a base and then trained using the application specific data set. As early layer features within CNNs (as addressed previously in section 2.3.3) tend to generalise well to most image detection tasks, these early layers weight parameter's are often frozen and only the later application specific layers are trained further. Freezing layers require less

computational resource as back propagation does not need done for the whole model and thus can increase training speed.

To utilise Faster R-CNN for face detection, the object detector module of the network is redefined as a binary classification problem of face or not. In essence the training for both the RPN and object detector module loss functions are technically learning the same binary problem. There are a number large scale face data sets suitable for use in the training of a face detector, both the AFLW (Köstinger et al., 2011) and the WiderFace (Yang et al., 2016*b*) data are applied within the research. Within chapters 5 and 6 Faster R-CNN based methods are proposed for face detection, the specific training parameters for each model are discussed in the respective chapter.

3.5 Summary

In this chapter a more comprehensive overview of the Faster R-CNN method has been given. It was established in the previous chapter that face detection forms the base task for the majority of facial analysis systems, to over come the challenge of position and scale while also providing fast detection a number of methods exist, these being SSD, YOLO and Faster R-CNN. Faster R-CNN was selected to form the basis of the various face detection method applied in this thesis. There were a number of reasons for this choice, one main reason was the availability of an official software implementation of the method within CNTK deep learning framework which allowed for the possibly to train and customise the architecture. While SSD and YOLO have a speed advantage over in terms of how many frames per seconds they can process, Faster R-CNN still provides real-time detection which is suitable for the facial analysis tasks described in this thesis. Regarding the performance there is minimal difference in object detection accuracy (see section 2.3.2 for the review of these methods) the most significant boosts to accuracy are related to the type of CNN applied which is independent of the method in the case of Faster R-CNN, which provides flexibility to the design. In a number of the remaining chapters of this thesis Faster R-CNN is modified and used for face detection and also symmetry analysis, this chapter forms a reference point for the base Faster R-CNN method.

Chapter 4

Stacked Hourglass Network

4.1 Introduction

Within the literature review a state-of-the-art method for landmark localisation namely the FAN (Bulat and Tzimiropoulos, 2017*b*) was identified. The task of landmark localisation is an important aspect of facial analysis systems especially when applying geometry-based features or using specific local regions of interest for feature extraction. This is also of significance as the landmark features are re-purposed for the task of face detection within chapter 6. This chapter provides an in depth explanation of the FAN method and the underlying stacked hourglass architecture upon which the FAN is built. This chapter is organised as follows: Section 4.2 details the thinking behind the hourglass design, this is then followed by a discussion of the stacked hourglass architecture in section 4.3. The FAN method for landmark localisation is directly addressed in section 4.4 which is then expanded to include depth prediction for 3D landmark localisation in section 4.5. Section 4.6 provides a summary of the method.

4.2 Hourglass Design

The importance of capturing information at every scale across an image was the primary motivation for Newell et al. (2016) design of the hourglass network. Originally designed for the task of human pose estimation where the key components of the human body such as head, hands and elbow are best identified at different scales. The design of the hourglass provides the capability to

capture these features across different scales and bring these together in the output as pixel-wise predictions. The name hourglass is taken from the appearance of the networks down sampling and up sampling layers which are shown in figure 4.1. Given an input image to the hourglass, the network initially consists of down sampling convolutional and max pooling layers which are used to predict features down to a very low resolution. During this down sampling of the input the network branches off prior to each max pooling step and further convolutions are applied on the pre-pooled branches, this is then fed back into the network during up sampling. The purpose of network branching is to capture intermediate features across scales, without the application of these branches rather than learn features at each scale the network would behave in a manner previously shown in figure 2.7 where initial layer learn general features and deeper layers learn more task specific information. Following the lowest level of convolution the network then begins to up sample back to the original image resolution through the application of nearest neighbour up sampling and element wise addition of the previously branched features. Each of the cuboids in figure. 4.1 is a residual module also known as bottleneck blocks as shown in figure 4.2. These blocks are the same as those used within the ResNet architectures. In section 8.3.3 a more detailed description of ResNet and the block design is given.

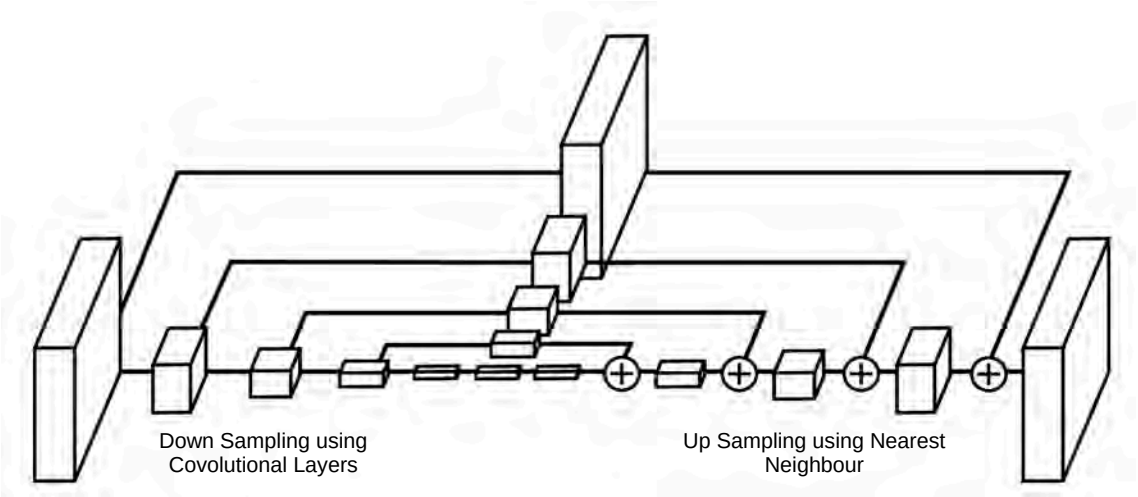

Figure 4.1: Hourglass Design (Newell et al., 2016).

4.3 Stacked Hourglass with Intermediate Supervision

The final architecture proposed by Newell et al. (2016) took the hourglass design and stacked n hourglasses in an end-to-end fashion, where in the best performing configuration for human pose estimation was $n = 8$. Each of these hourglass's is independent in terms of the weight parameters. The purpose of this stacked approach it to provide a mechanism in which the predictions derived from a single hourglass can be evaluated at multiple stages within the total network. A key technique in the use of this stacked design is that of intermediate supervision, in which at the end of each individual hourglass a heat-map output is generated to which a Mean Square Error (MSE) loss function can be applied. This process is similar to the iterative processes found in other landmark localisation methods, where each hourglass further refines the features and therefor the predictions as they move through the network. Following the intermediate supervision the heat-map, intermediate features from the hourglass and also the feature from the previous hourglass are added. To do this a 1×1 convolutional layer is applied to remap the heat-map back into feature space.

4.4 Face Alignment Network

FAN takes the stacked-hourglass design and trains this for the task of facial landmark localisation. Landmark localisation has similar challenges to that of human pose estimation, where the face landmarks are represented at different local scales within the context of the global context human face. In comparison with other landmark localisation methods FAN has archived state-of-the-art results on all major test data sets as shown in Table 4.1.

Data Set Name	**FAN**	**MDM**	**iCCR**	**TCDCN**	**CFSS**
300VW-A (Shen et al., 2015)	72.1%	70.2%	65.9%	NA	NA
300VW-B (Shen et al., 2015)	71.2%	67.9%	65.5%	NA	NA
300VW-C (Shen et al., 2015)	64.1%	54.6%	58.1%	NA	NA
Menpo (Zafeiriou et al., 2017)	67.5%	67.1%	NA	47.9%	60.5%
300W (Sagonas et al., 2013)	66.9%	58.1%	NA	41.7%	55.9%

Table 4.1: Comparative AUC benchmark results for FAN against MDM (Trigueros et al., 2018), iCCR (Sánchez-Lozano et al., 2016), CFSS (Zhang et al., n.d.) and TCDCN (Zhu, Li, Loy and Tang, 2015*b*).

Architectural changes are made to the network design where FAN reduces the total number of stacked hourglass's from 8 to 4. Also the structure of the convolutional blocks are changed from

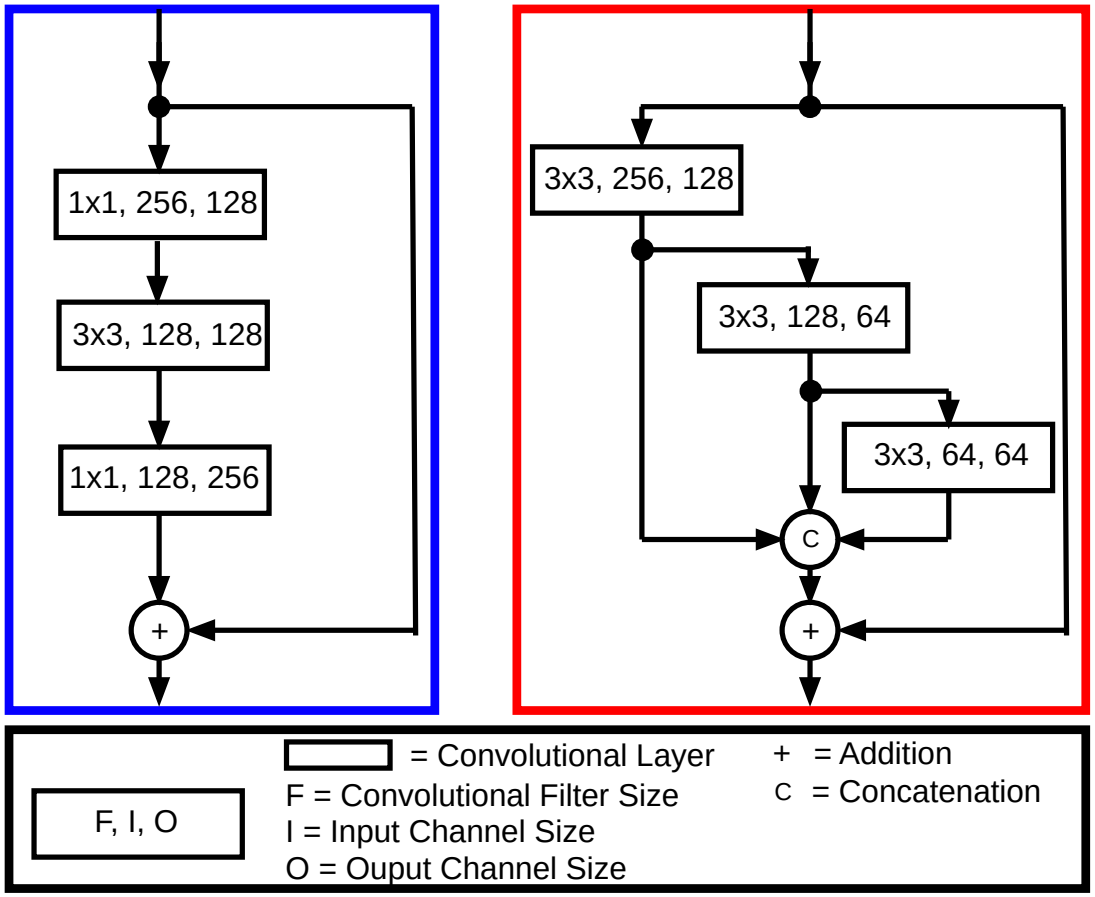

Figure 4.2: Block Design: (Blue) The basic bottleneck block Newell et al. (2016). (Red) The hierarchical, parallel and multi-scale block of FAN (Bulat and Tzimiropoulos, 2017*b*).

bottle necks to a hierarchical, parallel and multi-scale block, which performs three levels of parallel convolution alongside batch normalisation before outputting the concatenated feature map (Figure 4.2). It was shown in Bulat and Tzimiropoulos (2017*a*) that when the total parameter number is equal this block type outperforms the bottle neck design. The parameters of the 1×1 convolutional layers are changed to output heat-maps of dimension $H \times W \times m$, where H and W are the height and width of the input volume and m is the total number of facial landmarks predicted where $m = 68$.

Training of FAN was completed using a synthetically expanded version of the 300-W (Sagonas et al., 2013) named the 300-W-LP (Zhu, Lei, Liu, Shi and Li, 2015), while the original 300-W was also used to fine-tune the network. Data augmentation was applied during training, this employed random flipping, rotation, colour jittering, scale noise and random occlusion. The training applied

a learning rate of 10^{-4} with a mini-batch size of 10. At 15 epoch intervals the learning rate was reduced to 10^{-5} then again to 10^{-6}. A total of 40 epochs were used to fully train the network. The MSE loss function is used to train the network:

$$MSE = \frac{1}{n}\sum(Y_i - \hat{Y}_i)^2 \tag{4.1}$$

where Y_i is predicted heat-map for the i^{th} landmark and $\hat{Y}_i$ is a ground truth heat-map consisting of a 2D Gaussian centred on the landmark location of the i^{th} landmark.

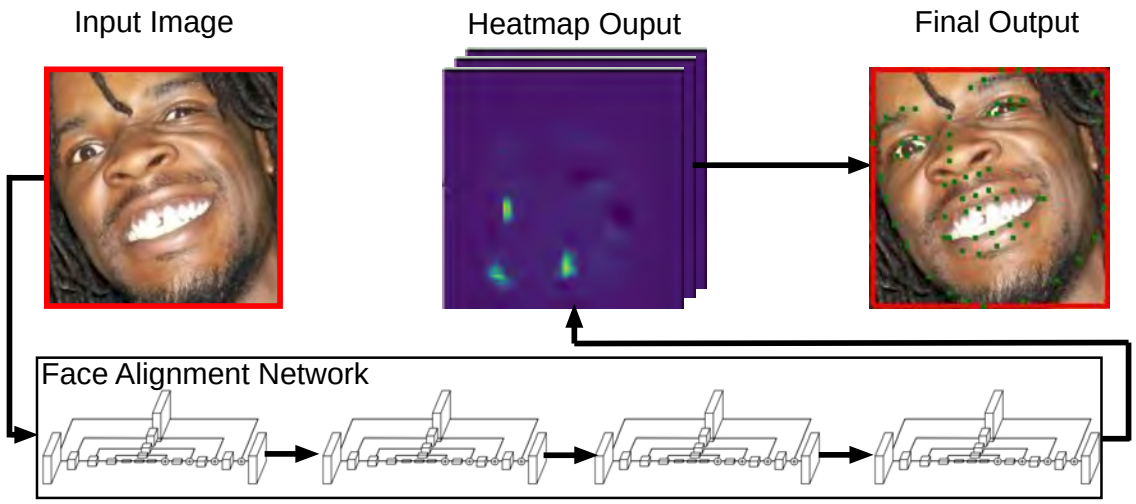

Figure 4.3: Face Alignment Network Architecture Overview.

4.5 Depth Network for 3D landmarks

A further extension to the FAN method is the capability to extend the 2D facial landmarks to 3D, this is achieved through the application of a second network. This second network takes as the input the predicted heat maps from the original 2D landmark localisation and the face image. The heatmaps guide the networks focus on areas of the image at which depth should be predicted from. This network is not hourglass based but instead a adapted ResNet-152, where the input takes $3+N$ where 3 is the RGB channels of the image and N is the heatmap data where $N = 68$. The output of the network is $N \times 1$. Training applied 50 epochs using similar data augmentation as the 2D model training, with a learning rate of 10^{-3} and an L2 loss function.

4.6 Summary

A detailed overview of the stacked hourglass network architecture is the focus of this chapter. The stacked hourglass is an important architecture in the task of landmark localisation, more precisely the FAN method based upon this architecture has shown state-of-the-art results on a number of facial landmark localisation data sets. For this reason the FAN is applied within this thesis for facial analysis tasks. While in some areas of the research the FAN method is used only for landmark localisation, within chapter 6 the local facial features are also applied to aid the task of face detection.

Chapter 5

Unified Multi-task Faster R-CNN method for Face Detection and Facial Symmetry Analysis

5.1 Introduction

This chapter describes a preliminary investigation into the use of a unified Faster R-CNN architecture to perform multiple tasks within a single CNN. The tasks were that of face detection and also binary facial symmetry analysis which acts as a precursor to the more specific task of facial palsy grading investigated in later chapters of this thesis. The primary objectives of the research conducted within this chapter was to establish the potential for multi-task learning while also evaluating the feasibility to apply a CNN for classifying facial symmetry from static images. The method and evaluations presented in this chapter formed a published contribution to the Science and Information IntelliSys 2018 conference on artificial intelligence where the research won the best student paper award (Storey and Jiang, 2019).

This chapter is organised as follows: Section 5.2 provides a insight into the motivation behind the method proposed within this chapter. The method and training protocol are described in Section 5.3, while the results and discussion are presented in section 5.4. Finally a conclusion is given in section 5.5.

5.2 Motivation

Motivated by the findings of the literature review, both multi-task learning and facial symmetry analysis were identified as potential avenues for the direction of the research. The task of facial symmetry analysis was selected as a gateway to specific medical application such as stroke or facial palsy as discussed in section 2.5.3. Multi-task learning provides a method to potentially incorporate further tasks into the Faster R-CNN face detector.

5.2.1 Multi-Task Learning

The concept of multi-task learning is that within a single method n correlated tasks can be effectively learnt using shared features and the prediction accuracy can potentially be boosted. One of the first major breakthroughs in facial analysis which employed multi-task learning was that of the TSM (Ramanan et al., 2012) were the tasks of face detection, landmark localisation and pose estimation were carried out in a single method. This method gained attention at the time of publication and the performance was state-of-the-art on the benchmark data sets including AFW for face detection and the 300-W for landmark localisation. This method applied tree based models which represented the global face feature which was comprised from multiple HOG parts features representing the local facial landmarks. As the research community moved towards CNNs in recent years the idea of multi-task learning has been adopted in a number of successful ways. These methods leverage the capability of CNNs to learn features that can be shared across tasks, which is why there needs to be a level of correlation between tasks. Faster R-CNN as described in chapter 3 applies this strategy to be able to learn a shared feature set used for distinguishing both region of interest in an image and also specific object classification. Multi-task learning using CNNs was recently studied in Zhang and Zhang (2014) where the method shows an improvement in landmark localisation by also learning gender attributes. Finally in Ranjan et al. (2016) the authors present using a single modified Fast R-CNN architecture called HyperFace for the tasks of face detection, landmark localisation, gender recognition and pose estimation. HyperFace leverages the way in which features at each layer of the network are distributed, with the lower level features being exploited for landmarks localisation and pose estimation, while the features from deeper layers are used for the more global tasks of face detection and gender recognition. To successfully use these different features, the features at different layers of the network are extracted and then

concatenated with those from later levels to firm the final feature representations which is are used for the predictions of each of the tasks. An important conclusion from the research conducted on Hyperface was that all of the face analysis tasks saw benefit from applying multitask learning. The reason for this improvement was concluded to be from the ability of the network to learn more discriminative features than when using a single task network for the same set of facial analysis tasks.

At the time this method was proposed there was no research into applying further multi-task goals into the Faster R-CNN architectures. In HyperFace a modified Fast R-CNN architecture was applied this took external region proposal and therefore suffered from slower performance due to the recognised bottleneck. Therefore part of the motivation of this investigation was to establish how further tasks could be added to the Faster R-CNN architecture, which has performance speed benefits over Fast R-CNN.

5.2.2 Facial Symmetry Analysis

Levels of facial symmetry has been shown to indicate important information in areas like perceived attractiveness (Grammer and Thornhill, 1994) and also in medical diagnosis and rehabilitation for conditions such as facial palsy or stroke (Wang, Dong, Sun, Zhang and Wang, 2014; Bandini et al., 2018). Specifically within the medical domain which is the primary interest of this research, the importance of diagnosis and rehabilitation tracking for the best possible outcomes in patients with facial paralysis such as facial palsy and stroke has been highlighted within recent research (Ishii, 2016; Guerreschi et al., 2016; Monini et al., 2016). The capability of applying computer vision methods within this domain has previously been identified and discussed in section 2.5.3, it has been highlighted that there is potential to aid clinical diagnosis with such systems but at the time of the investigation the body of research is relatively small in comparison to other domains such as facial recognition. Traditional methods have been applied in Wang, Dong, Sun, Zhang and Wang (2014), Wang et al. (2016) and Bandini et al. (2018) to analysis symmetry in faces, but CNNs had not been applied at the time this research was undertaken. This provided the motivation to investigate the potential to apply state-of-the-art deep learning methods to the task of symmetry analysis, due to the success they have achieved in other facial analysis application.

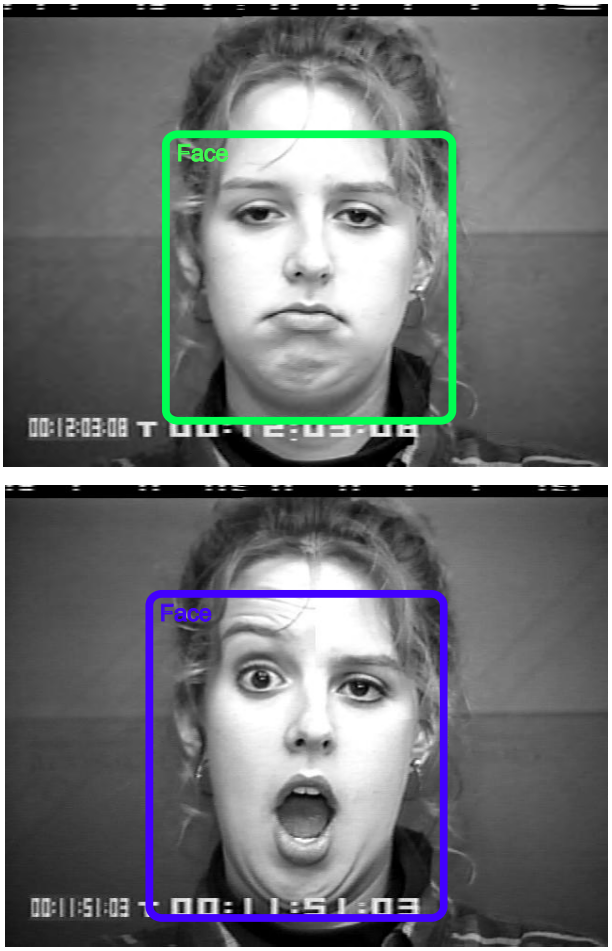

Figure 5.1: Unified Multi-task Faster R-CNN Output Examples: (Top Image) - This image shows an example of the output from the proposed method depicting a positive face detection and a symmetrical face, (Bottom Image) - This image shows an example of an asymmetrical face as detected by the proposed method. (Images © Jeffrey Cohn)

5.3 Proposed Method

In this section a novel unified multi-task CNN framework is presented for simultaneous object proposal, face detection and face symmetry analysis from an input image (see Figure 5.1). The basis of this framework is the Faster R-CNN network detailed in chapter 3, in this section the proposed method of modifying Faster R-CNN to incorporate the task of face symmetry analysis is detailed. Figure 5.2 provides an overview of the modified Faster R-CNN model, highlighting the fully connected layer for the task of face symmetry analysis which is added for this proposed method. To successfully train the proposed model, the topic of synthesised data is addressed as a method to overcome small training sets.

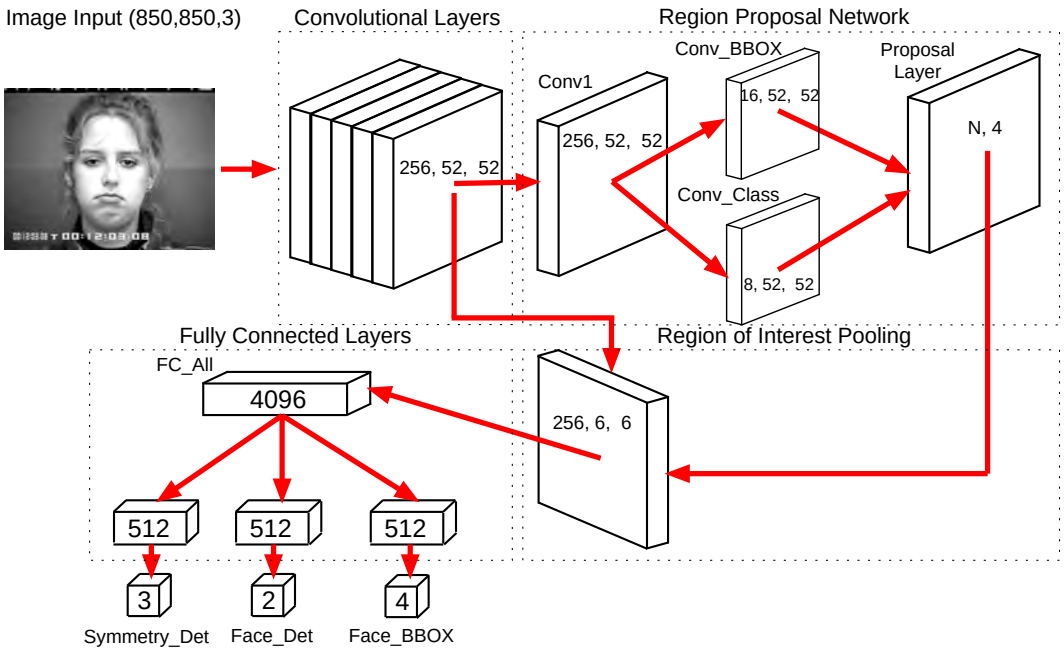

Figure 5.2: Overview of the proposed unified Multi-task Faster R-CNN architecture for face detection and face symmetry analysis (Face Image © Jeffrey Cohn). Showing the original Faster R-CNN network modified with a further fully connected layer for the face symmetry analysis task.

5.3.1 Faster R-CNN with Face Symmetry

The Faster R-CNN network Ren et al. (2015) introduces a multi-task loss function to train the RPN and object detection tasks in a single CNN network (see chapter 3), the proposed framework further expands this architecture with the further task of face symmetry analysis. The proposed method benefits from sharing convolutional layers across all tasks with additional fully connected layers (Figure 5.2) defined at the output layer specifically for symmetry analysis task. A loss function is defined to provide the capability to learn the symmetry analysis task simultaneously with the face detection tasks. The total loss of the network is defined as:

$$loss_{total} = \sum_{i=n} loss_i \lambda_i \tag{5.1}$$

where the individual loss corresponding to the i^{th} task is defined as $loss_i$. A weighting parameter λ_i is applied to balance the learning priorities. Symmetry analysis loss is defined as a single task within the proposed framework. As discussed in the previous section asymmetry can be a indi-

cation of a potentially non-diagnosed medical condition. The task is specified as a classification problem within the network. Symmetrical or asymmetrical is nominally posed as a binary class problem. When incorporating this task into a single unified framework to leverage the computational benefits of the RPN module, the binary classification problem requires modification. The RPN module outputs a set n regions, where only m regions are labelled as positive face detection's, the other regions are labelled as background. As the face symmetry tasks are only associated with the m regions, a further background class label is added to the task which is associated with the background regions. This is required to keep the n regions dimensions equal for the region of interest pooling layer. Symmetry analysis loss is defined as:

$$loss_{sym} = \frac{1}{N_{box}} \sum_{i=n} -(1 - s_i^*) \cdot log(1 - s_i) - s_i^* \cdot log(s_i) \tag{5.2}$$

where the softmax loss function for symmetry analysis for the ith region of interest where $s_i = 1$ defines an asymmetrical face while $s_i = 0$ is a symmetrical face. The ground truth label for a given face is given by s_i^* and the loss is normalised by N_{box} which is the total number of regions of interest.

5.3.2 Training Protocol

A network architecture that incorporates a RPN layer has been shown to be optimally trained using an alternating training scheme applying a number of sequential training steps (Ren et al., 2015). The proposed method applies a dual stage training protocol, where stage 1 trains the models RPN layer for region proposal and the single task of face detection using the sequential training detailed by Ren et al. (2015). Stage 2 applies a novel training method using synthesised data to learn the multi-task problem of both face detection and symmetry analysis. This proposed training method in stage 2 overcomes issues with limited training data while also providing a method for balancing the training of the multiple tasks simultaneously.

Stage 1: Initially the AFLW (Köstinger et al., 2011) database is used to to perform transfer learning specifically for the task of face detection. It contains 21,997 real-world 'in the wild' images with 25,993 fully annotated faces. Annotations for the face bounding-box are used in the training process and to further supplement the database size, random horizontal flip-

ping is applied as a data augmentation technique. The convolutional layers of this specific Faster R-CNN based model uses the AlexNet architecture (Krizhevsky et al., 2012). The 1^{st} step is to initialise the weights of the network with pre-trained values learnt on Imagenet. The AFLW database is then used to train the model initially end-to-end for the region proposal task. In the 2^{nd} step, a face detection network is trained using the proposals generated by the RPN learnt in the 1^{st} step. At this point the two trained networks do not share convolutional layers. A 3^{rd} step is then used in which the face detector network of the 2^{nd} step is used to initialise final RPN training, the shared convolutional layers are frozen and only fine-tuning commences on the layers unique to RPN. Now the two networks share convolutional layers. The 4^{th} and final step of this stage involves keeping the shared convolutional layers frozen, the unique layers specific for face detection are then fine-tuned. At this point a full Faster R-CNN network for face detection is trained.

Stage 2: A second data set namely the CK+ (Lucey et al., 2010) is then used for learning the face symmetry task. As this data set does not contain asymmetrical faces a method is applied to synthesise these. The capability to synthesise samples is due to the format of the CK+ data set, where each image sequence incorporates the neutral face through to the end range of a facial expression. The process to create the large data set involved retaining both quarter and half portions of frames from the same expression sequence, these facial areas were then merged with frames later in the sequence to synthesise the appearance of asymmetry. The facial areas used correspond with typical areas of facial palsy due to the facial nerve placement and image examples are provided in highlighted in Figure 5.3. The training process consisted of loading the Faster R-CNN network for face detection from stage 1, the fully connected layers for the face symmetry task are added and randomly initialised. The RPN layer weights are frozen and the network is trained in an end-to-end process for both the face symmetry and face detection tasks. Within the multi-task loss the values for λ were $\lambda = 1$ for the face detection task and $\lambda = 1$ for facial symmetry loss also. Treating both tasks equally in the total loss function allowed the loss to convergence.

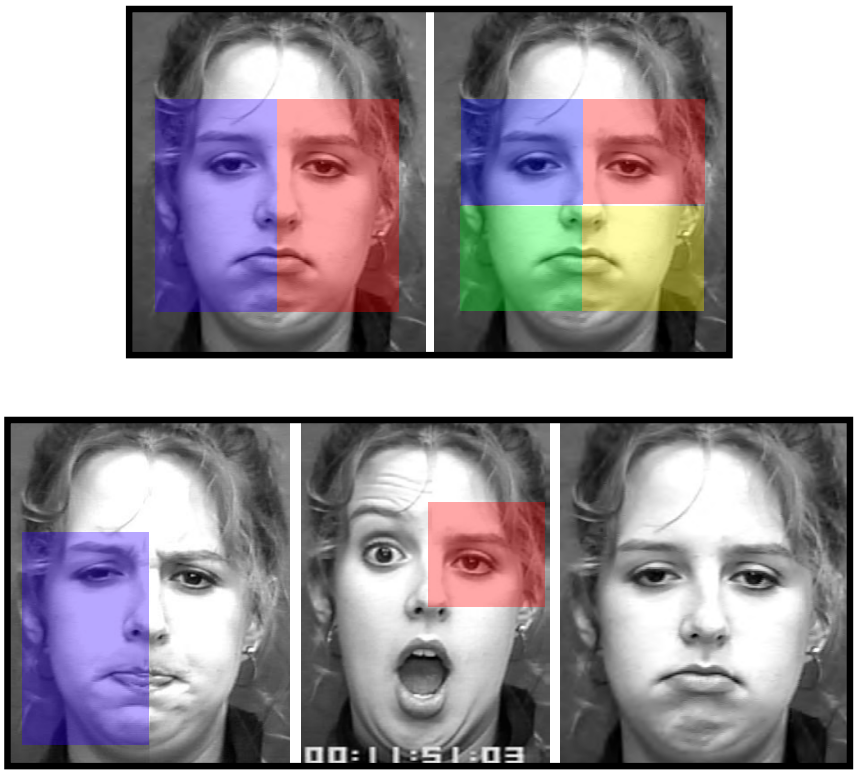

Figure 5.3: Synthesised Asymmetry Images Examples: (Top Row Images) - These show the six regions of asymmetry synthesised for the training set, (Bottom Row Images) - The left and middle image are examples of synthesised images with half face and quarter face asymmetry present respectively. The right image is a symmetrical face example. (Images © Jeffrey Cohn)

5.4 Evaluation

Evaluation of the proposed method used the CK+ database Lucey et al. (2010) testing of the proposed method. The data set consists of 593 sequences containing 10,201 individual frames from 123 subjects. Following the data synthesis process a total 18,786 unique training images where used to train the model, 50% of which were synthesised and labelled asymmetrical while the original images were labelled as symmetrical. The evaluation data set applied 1,808 images not used within training, with a split of 839 symmetrical and 950 asymmetrical faces. A further test set of 27 individuals diagnosed with facial palsy was also applied during the evaluation's. All experiments where conducted using the Microsoft CNTK framework in Windows 10 with a Nvidia GTX-1080 GPU.

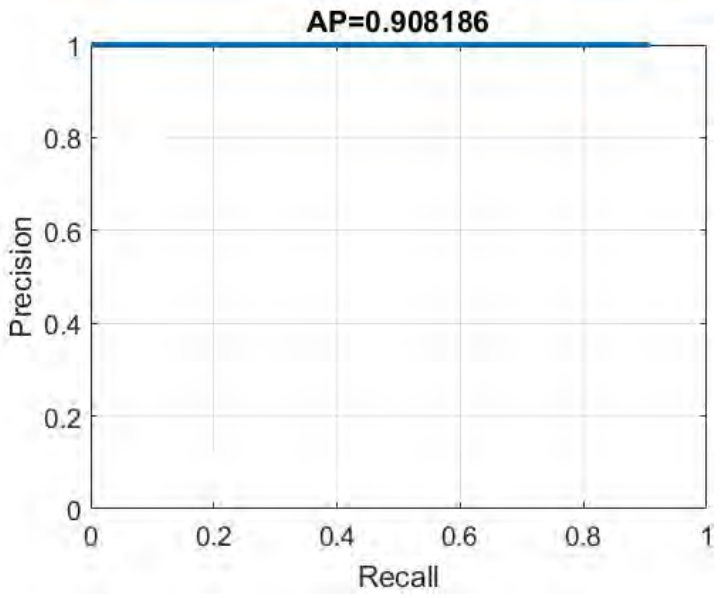

Figure 5.4: Face Detection Precision/Recall Curve - Synthesised CK+ test set.

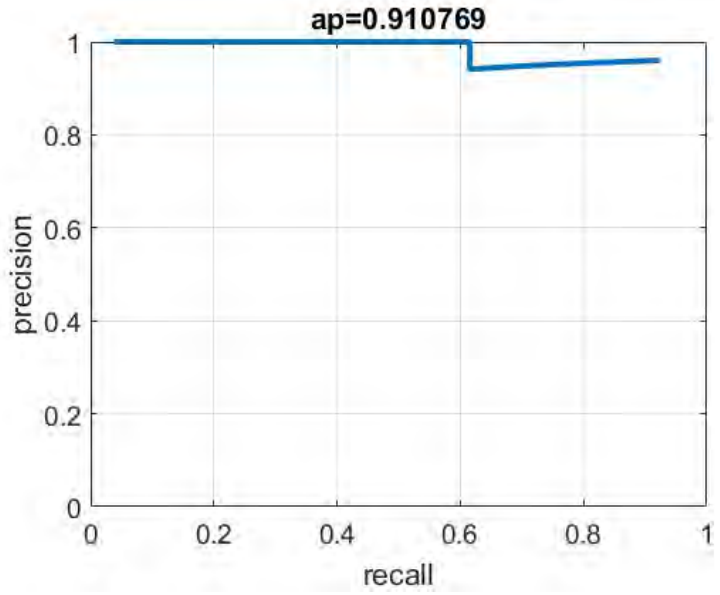

Figure 5.5: Face Detection Precision/Recall Curve - Facial Palsy test set.

5.4.1 Face Detection

The capability of the model to perform accurate face detection was analysed in this evaluation. The PASCAL VOC precision-recall protocol was adopted for the evaluation, where for a positive face detection a threshold of a 50% overlap was applied as is common across object recognition literature (see section 2.3.1 for further details). The precision-recall curve for the synthesised CK+ test results are shown in Figure 5.4. The proposed model show a very high degree of precision on

these images with no false positives, while the recall is also high at 90%. Surprisingly the reduced rate of recall is on samples labelled as symmetrical, with 950 asymmetrical faces being detected from a total of 961 and 712 symmetrical faces from 839. A similar level of performance is seen on the facial palsy test set as shown in Figure 5.5 which suggests the model generalises well for the task of face detection to unseen data. Though the lack of false positives suggest there may be a level of over-fitting as it is unexpected when compared with other face detection research.

5.4.2 Symmetry Classification

This evaluation investigates the classification accuracy of whether a face shows symmetrical or asymmetrical features. The CK+ synthesised test set and facial palsy test set were used to generate the results. Note that only samples which were correctly found by face detection where applied in this evaluation reducing the total samples to 1668 and 25 for the CK+ synthesised test set and facial palsy test set respectively. Results for the synthesised CK+ are displayed as a confusion matrix within Table 5.1, these show that the proposed method produces highly accurate detection accuracy on the test set, in total there are only 23 miss classified samples from the total of 1,662. For the facial palsy set the results are shown in Table 5.2, it was found that model performs well for those with prominent facial palsy classifying these as asymmetrical, it has more issues with subtle levels of palsy. This could be due to the examples used in the training of the model and the nature of the problem where individuals naturally have varying levels of symmetry and there maybe crossover in the populations.

Table 5.1: Symmetry Classification Confusion Matrix - Synthesised CK+ test set.

	Symmetrical	Asymmetrical	Background
Symmetrical	699	10	3
Asymmetrical	7	940	3
Background	3	3	0

Table 5.2: Symmetry Classification Confusion Matrix - Facial Palsy test set.

	Symmetrical	Asymmetrical	Background
Asymmetrical	17	7	1

5.5 Conclusion

Within this chapter a novel unified multi-task CNN for face detection and facial symmetry analysis was proposed. From the evaluation and training process a significant amount of was learnt highlighting both the potential and limitations of the method and also the evaluation.

The results are promising on the two test data sets applied within the evaluations, where a high level face detection accuracy and symmetry classification were obtained. The model proved to be highly computationally efficient with the average time taken to process an image being 0.045 seconds on a Nvidia GTX 1080 GPU. This highlights the potential of unified models to give real-time feedback and when compared with architectures which use external region proposal the processing time is reduced by up to 2 seconds per image as reported in Ranjan et al. (2016).

The facial symmetry analysis provided by the proposed method has limitations, while it has the capacity to detect asymmetry within a facial image with good accuracy, for use in medical applications there is a need to further refine the symmetry analysis task. A limitation is the data set used, both in terms of labelling and synthesised images. Regarding labelling for more accurate analysis the data set requires labelling with a level of asymmetry and potentially the area. Even better would be labelling which directly relates to a specific medical condition such as facial palsy. Synthesised data allowed a large increase in sample size which enabled the training of a CNN, the limitation of this method is that the faces created are not truly representative of individuals with specific medical conditions. A preferable method would be to have real samples for those displaying asymmetrical faces. Finally applying static images of an individual may be limiting, while these can provide information regarding symmetry the capacity to use video sequences would present the opportunity to further understand facial range of motions. These findings shaped the approach taken in the rest of this thesis, where real data and videos sequences became the primary methods for researching facial analysis for asymmetrical faces.

Training a multi-task network also proved to be difficult and volatile, specifically with balancing the competing loss function λ. While it was found that $\lambda = 1$ performed best, when the tasks were given unequal values the loss would not decrease incrementally as is normal with training and would fluctuate greatly. These difficulties with training a multi-task Faster R-CNN resulted in other methods being investigated in the rest of the research.

Chapter 6

Integrated Deep Model for Precise Face Detection and Landmark Localisation

6.1 Introduction

The task of face detection is the focus of this chapter in which a novel method named the Integrated Deep Model is proposed and validated on a series of benchmark data sets. Within Section 2.3 of the literature review one specific challenge identified in the face detection task is that of increasing precision through reducing false positive detection's. In any end-to-end deep learning facial analysis systems reducing false positives is extremely beneficial in ensuring that irrelevant data is not extracted and forwarded to later stages of the pipeline. The proposed method integrates the Faster R-CNN (Chapter 3) and the stacked hourglass (Chapter 4) specifically leveraging the local features of landmark localisation to aid the face detection process, due to the methods applied the Integrated Deep Model also performs landmark localisation. The method contributed within this chapter was published with in the IEEE Access journal under the title "Integrated Deep Model for Face Detection and Landmark Localisation From 'In The Wild' Images" (Storey et al., 2018).

This chapter is organised as follows: The rationale behind the research and the method proposed in this chapter is discussed in section 6.2. Section 6.3 details the method describing the how the Faster R-CNN and the stacked hourglass where integrated. Experimental evaluation including benchmarks against other face detection methods is detailed in section 6.4, while a conclusion is present in section 6.5.

6.2 Motivation

The research conducted within this chapter was shaped by a number of factors. These factors were the identification of an existing challenge within the face detection task regarding increasing precision which was discussed within section 2.3.4, the cascaded approaches which have resulted in improved FER accuracy and the findings of previous research on multi-task learning and the importance of both local and global feature representations.

Increase in face detection precision through the reduction in false positive detection's is vital in specific facial analysis applications. This is specifically true in medical applications used for diagnosis where there is a need for certainty that a true face has been detected. In these systems though ideally all faces would be correctly detected it is preferable to miss a face detection rather than to present a false positive from which incorrect diagnosis would occur. As highlighted in section 2.3.4 while recall in the face detection task has increased significantly the same is not true for precision, considering the medical applications investigated in the remaining chapters of this thesis and the more general needs of facial, this motivated the research to focus presented in this chapter to focus on false positive reduction while retaining a high level of recall.

Across the literature there are many examples where different techniques have been combined in a cascaded approach including those in landmark localisation discussed in section 2.4 and FER section 2.5.2, the potential of such system is that it can combine the strengths to provide an improvement over the individual components. In the area of deep learning while many the popular architectures such as AlexNet, VGG and ResNet share similar principles which have proven excellent at object recognition, other significantly different architectures like the stacked hourglass have shown greater performance in other related image based recognition tasks. Therefore it is feasible that an approach which integrates different architecture types could provide better performance

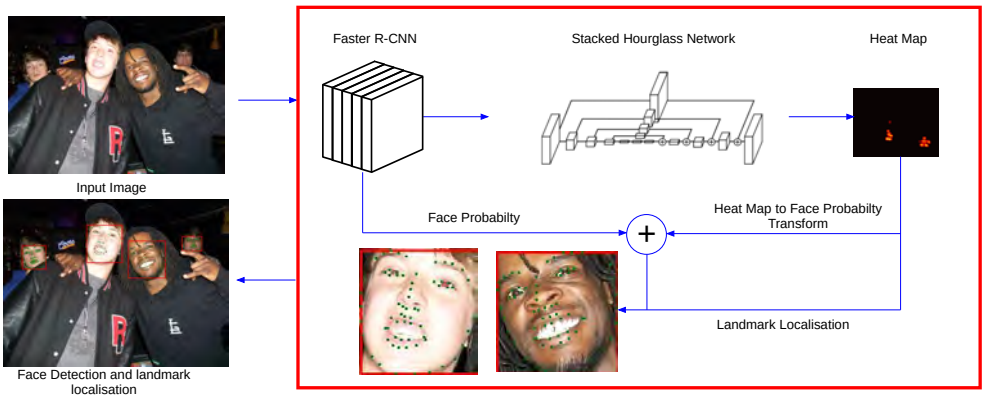

Figure 6.1: Integrated Deep Model (IDM) a step-by-step overview. Note the FAN component is depicted as a single hourglass network when in reality in has 4 stacked together.

than a single architecture.

When considering the tasks of face detection and landmark localisation, it is most common for these to be performed as entirely separated tasks. Even with the introduction of multi-task methods where the task are combined in a single architecture (Ramanan et al., 2012; Ranjan et al., 2016), the state-of-the-art performance still resides with methods that deal with a single task. At the time of writing these were S^3FD (Zhang, Zhu, Lei, Shi, Wang and Li, 2017) for face detection and FAN (Bulat and Tzimiropoulos, 2017*b*) for landmark localisation. The primary difference between the tasks is that face detection concerned with learning a set of features that can describe the faces at a global level, while the focus of landmark localisation is on the learning of smaller local features which surround specific landmarks of the face such as the corners of the mouth and eyes. While the global facial features have shown good performance for recall in the face detection task also using the local features from a different architecture has the potential to increase precision. This hypothesis forms the basis for the method proposed and evaluated in this chapter.

6.3 Method

The proposed Integrated Deep Model for both face detection and landmark localisation from 'in the wild' images integrates two state-of-the-art architectures namely the Faster R-CNN network architecture (Ren et al., 2015) (Chapter 3) and the stacked hourglass based Face Alignment Network (Chapter 4) as defined in (Bulat and Tzimiropoulos, 2017*b*) respectively. The Faster R-CNN is traditionally applied to object detection tasks, this is modified and the model trained solely for

Figure 6.2: Example of tiny false positive detection's (A) - Original Image, (B) - Tiny detection's from the Faster Face, (C) - The remaining detection's.

the task of face detection. FAN (Bulat and Tzimiropoulos, 2017*b*) has previously been used for 68 point landmark localisation with an independent face detector in a linear based framework. While face detection recall is high in modern deep learning based techniques there is still room for further increases in precision, this is specifically true when attempting to reduce the detection of false positives (Mathias et al., 2014; Zhang, Zhu, Lei, Shi, Wang and Li, 2017; Triantafyllidou et al., 2018). The primary aim is to exploit the learnt features in both networks to provide a more precise face detection method while still providing accurate landmark localisation and maintaining computationally efficiency. Within this section novel Integrated Deep Model method is described with Figure 6.1 showing a visual overview of the method.

6.3.1 Integrated Deep Model

The hypothesis is that the features learnt for landmark localisation have inter-connectivity with the task of face detection, given the two independent architectures the proposed novel method is presented to combine the networks in a cascade creating the Integrated Deep Model. To achieve this integration techniques are proposed for heat map transformation, integrated loss function using a joint probability for face detection and size scaling techniques, while adding minimal computational overhead when compared to using the two architectures in a linear framework.

Given the heatmap output of the FAN as $H = h_1, h_2, ..., h_n$ where each h_i corresponds to a specific facial landmark as an n x m matrix, each value is a probability of that facial landmark relating to that specific pixel within a given face image. The proposed method to transform the heat map H to a probability score that can be applied to the task of face detection and integrate this with the loss function of the Faster R-CNN face detector is defined as:

$$p_{fan} = \frac{1}{N} \sum_{i=n} \max(H_i)\gamma_i \tag{6.1}$$

given the maximum probability $\max(H_i)$ for the i^{th} facial landmark a specific scaling factor γ_i is applied for that that landmark. The sum of the scaled probability is then normalised and can be considered as the probability of a face detection derived from the FAN network defined as p_{fan}. The scaling value γ is primarily introduced to deal with wide ranging face poses in which certain landmarks retain visibility across all poses where others become occluded, the values of γ_i used are reported within the intermediate results in section 6.4. The next step is to define the joint probability of a region being a face termed as p_{face} and defined as:

$$p_{face} = \frac{(p_{fan} + (p_{faster}\delta))}{2} \tag{6.2}$$

where p_{faster} is the probability based upon the output of the trained Faster Face features. The penalisation factor δ is specifically introduced for situations where extremely small detection's are classed in the very high 90% probability range as being faces when they are not (Figure 6.2 provides examples of this). The value of δ is determined by:

$$\delta = \begin{cases} 0.7 & \text{if det * (100 / img)} \leq 2 \\ 1 & \text{otherwise} \end{cases} \tag{6.3}$$

where det is the width of the face detection box and img is the width of the image, probability penalisation is only used when a face width is less that 2% of the total image width. Finally the

p_{face} is used within the loss function for the face detection classification as described as:

$$loss_{face} = \frac{1}{N_{cls}} \sum_{i=n} -(1 - p_i^*) \cdot log(1 - p_{face,i}) \\ -p_i^* \cdot log(p_{face,i}) \tag{6.4}$$

6.3.2 Model Training

The IDM method specifically trains the layers of a Faster R-CNN architecture (Ren et al., 2015) with images containing multiple faces from the Wider Face training set (Yang et al., 2016*b*). The popular augmentation method of flipping is employed to further the available training data. The faster R-CNN implementation applies a VGG-16 Simonyan2015 architecture for the convolutional layers and the weight parameters are initiated using a pre-trained imageNet model prior to transfer learning for face detection. The publicly available PyTorch pre-trained model of the FAN of Bulat and Tzimiropoulos (2017*b*) is used as the stacked hourglass component of IDM. The IDM model was trained for 15 epochs with a learning rate of 10^{-3} for the initial 10 epochs then 10^{-4} for the final 5.

6.4 Evaluation

This section presents a through experimental evaluation of the proposed IDM method in the areas of face detection and landmark localisation. All experiments are conducted using PyTorch 0.4 on Windows 10 with a Nvidia GTX 1080 GPU.

6.4.1 Face Detection Evaluation

To evaluate the effectiveness the proposed IDM method for the task of face detection a set of intermediate and benchmark evaluations were conducted. For robust evaluation three face detection test sets were applied, these being the AFW database Ramanan et al. (2012), the AFLW Köstinger et al. (2011) and the FDDB Jain et al. (2010). The AFW database consists of 205 images where each image contains at least a single face, in total there are 468 faces located within the database. The AFLW database test set contains 10,001 images of annotated faces in real-world images cap-

turing multiple viewpoints, different expressions and illumination conditions. Finally the FDDB consists of 5,171 faces in 2,845 images from unconstrained environments. The PASCAL VOC precision-recall protocol for object detection is adopted as the evaluation metric requiring 50% IoU for positive detection of a face (see section 2.3.1 for further details of this metric).

Intermediate Results

For the intermediate results four different methods are evaluated, these being three variations of the IDM method using different parameters and a standalone Faster R-CNN only face detector given the name Faster Face. The three IDM variations are as follows, IDM Mean where the $\gamma = 1$ for all 68 landmarks, IDM Scaled which applies varying values for γ, the landmarks that are visible across all facial poses such as the nose are given a value of 1 while other less visible landmarks are given γ values of 0.75. Finally IDM Scale and Sized adds the box size weighting factor penalty as described within the previous section.

Table 6.1: AFW Results Benchmark

Method	True Positives	False Positives
Faster Face	468	361
IDM Mean	468	176
IDM Scaled	468	125
IDM Scaled + Size	468	73

Table 6.2: FDDB Results Benchmark

Method	True Positives	False Positives
Faster Face	4926	882
IDM Mean	4876	484
IDM Scaled	4876	336
IDM Scaled + Size	4876	336

Table 6.3: AFLW Results Benchmark

Method	True Positives	False Positives
Faster Face	1183	638
IDM Mean	1181	471
IDM Scaled	1181	368
IDM Scaled + Size	1181	352

The results for AFW test set are given in Table 6.1, 100% recall of the faces for all methods is shown. While all IDM methods dramatically reduce the number of false positives, the greatest reduction being from 361 to 73 which is significant. The IDM Scaled and Size also has success

in correctly identifying the small detection's (examples of this are shown within Figure 6.2). For the FDDB test set the results are highlighted in Table 6.2, again a large reduction of over 50% is shown in the false positives for IDM methods. A small decrease in recall is noticed in comparison with Faster Face, when analysed this drop is almost entirely for very blurred faces (see Figure 6.5). Finally AFLW test set results are shown in Table 6.3, this follows a similar pattern to the previous results with a minimal drop in recall of 2 faces, but a significant drop of 286 false positives between the Faster Face method and the full IDM method. The overall observations from all three tests set is that the proposed IDM method and most specifically the IDM Scale and Sized variation has a large impact on reducing false positive. The negative aspects is that there is a small reduction in recall from analysis of the images, finding that the primary source of decreased recall is blurred faces. The results also highlights the effectiveness of the Faster Face architecture alone for detecting faces with 'in the wild' images in terms of recall, where this has issue is that it also provides a high amount of false positives.

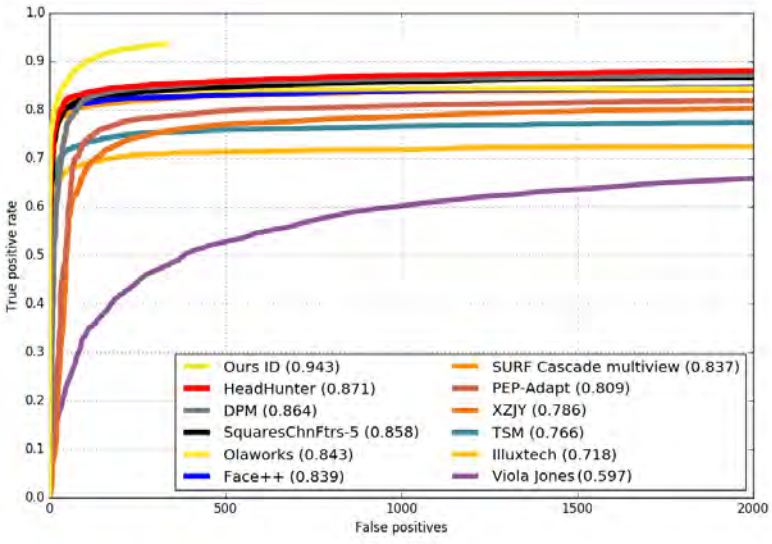

Figure 6.3: FDDB Benchmark Results.

Benchmark Results

To benchmark against other face detection methods the AFW and FDDB test sets are applied, as the AFLW test set is not a standardised image set. The evaluation tool provided by Mathias et al. (2014) is used for results generation and provides benchmarks against a number of methods

including Headhunter (Mathias et al., 2014), Structured Models (Yan, Zhang, Lei and Li, 2014) and TSM (Ramanan et al., 2012). Note this software uses an alternative set of ground truth boxes for the AFW explaining the difference in average precision compared with the previous intermediate evaluation. For the FDDB and AFW test set the method outperforms the methods included within the evaluation tool, both higher recall and also less false positives by a significant margin are reported in both cases as shown in Figure 6.3 and Figure 6.4. Furthermore when comparing the method to other recent state-of-the-art CNN based methods the Fast Deep Convolutional (FD-CNN) (Triantafyllidou et al., 2018) and the impressive S^3FD (Zhang, Zhu, Lei, Shi, Wang and Li, 2017) very competitive results in terms of recall are reported, while also showing a much lower rates of false positive detection's. For the FDDB test set FD-CCN has a recall rate of 92.6% while S^3FD has a 98.2%, the method is in between at 94.3%. Where the method excels is in the false positives which are significantly lower than both methods at 336 compared to 700 and over 1000 for FD-CNN and S^3FD respectively. Only the S^3FD provides benchmarks for the AFW in which they report at 99.8% average precision compared to the 99.6%, again the IDM method has greater precision. As identified in the intermediate results the main recall issue for the model is severely blurred faces (see Figure 6.5 for examples), primarily due to the training set not containing blurred faces to this degree.

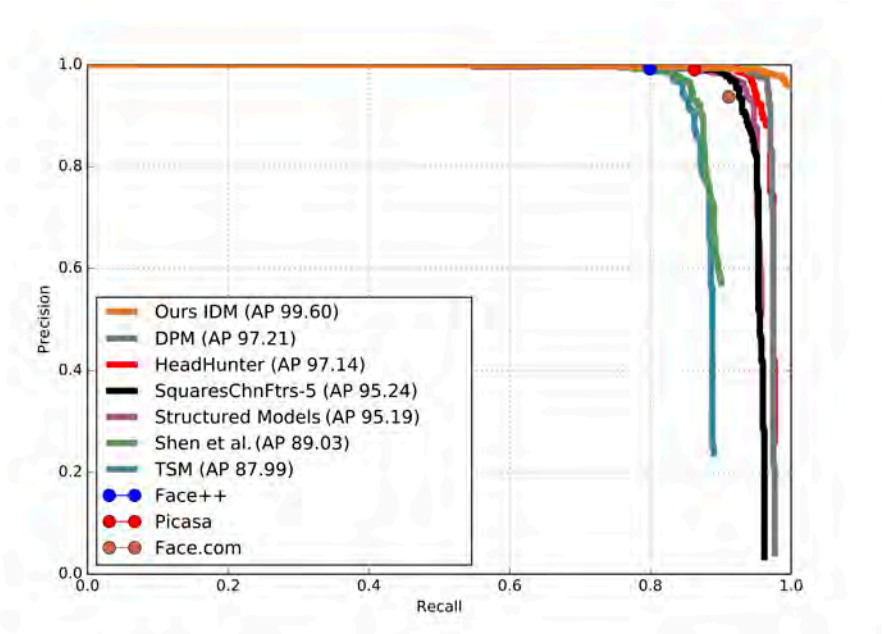

Figure 6.4: AFW Benchmark Results.

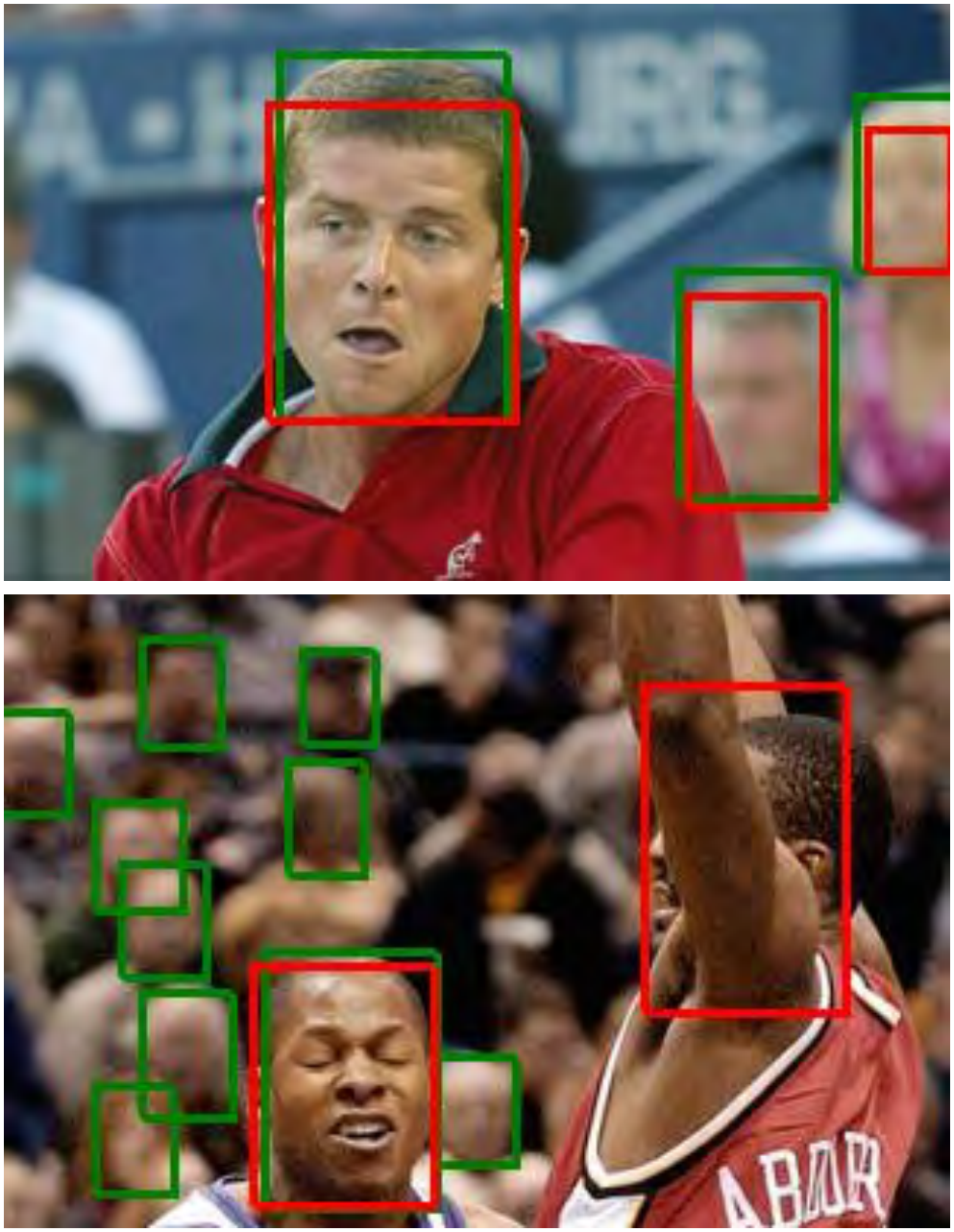

Figure 6.5: Blurred Face Detection Examples: (Red boxes represent a detection by IDM, green boxes are the ground truth face data). (Top) - image displays example of moderately blurred faces which the model successfully detects, (Bottom) - image highlights extreme blur where the IDM method misses 9 faces, while also detecting a face not accounted for in the ground truth data.

6.4.2 Comparative Face Detection Study

In the previous chapter a method for face detection was also proposed as a Unified Multi-task Faster R-CNN, which highlighted good initial results on a limited set of data. In this comparative study, the two facial palsy data sets evaluated on the previous model are also evaluated on the IDM. These are the Synthesised CK+ with 1,809 images and the Facial Palsy set with 27 images. The results shown in Table 6.4, highlight that the IDM provides a 100% recall while returning minimal false positives. The 10% difference on the Synthesised CK+ data is significant as this represents 180 missed face detection's. The results backup the rationale for moving from the unified multi-task method which was also difficult to train in this thesis for the IDM method which provides the most accurate face detection.

Table 6.4: Comparative Face Detection - Results from the evaluation between the IDM (Chapter 6) and Unified Multi-task Faster R-CNN (Chapter 5).

Method	Data Set	Average Precision	False Positives
Unified Multi-task Faster R-CNN	Synthesised CK+	0.91	0
IDM	Synthesised CK+	1	2
Unified Multi-task Faster R-CNN	Facial Palsy	0.91	1
IDM	Facial Palsy	1	0

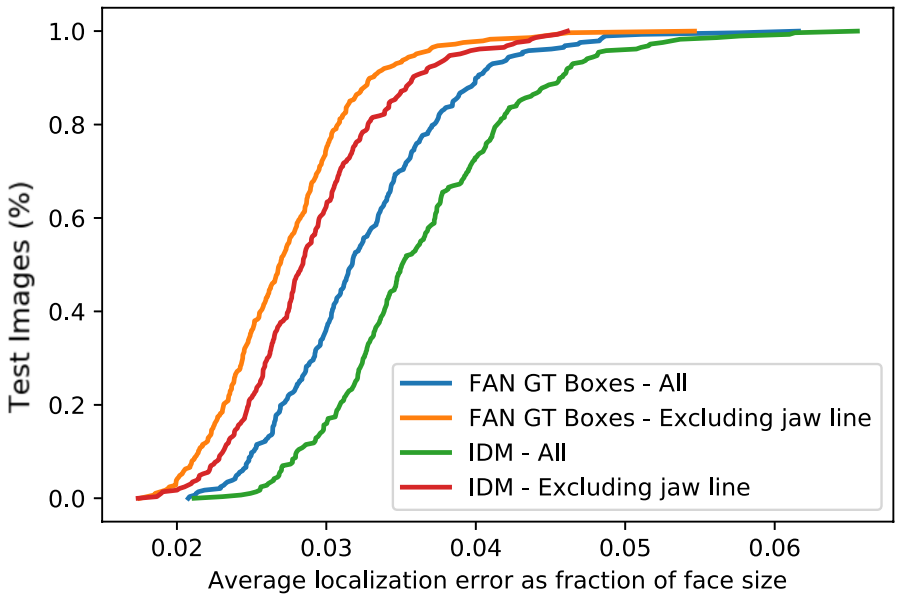

Figure 6.6: Cumulative Localisation Error Distribution - 300-W test set.

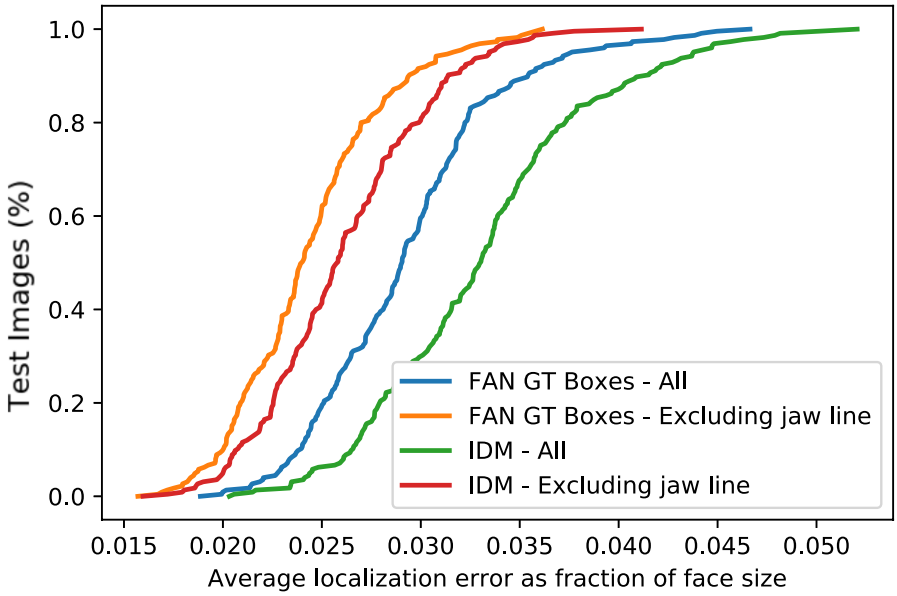

Figure 6.7: Cumulative Localisation Error Distribution - AFW test set.

6.4.3 Landmark Localisation Evaluation

The primary objective for the proposed method is to improve the accuracy of face detection as analysed in the previous experiments, as the landmark localisation features of the method are not re-trained, increased accuracy over the baseline presented in Bulat and Tzimiropoulos (2017*b*) is not expected to be observed. Instead evaluation of the performance of face alignment accuracy in IDM against the base line experiments was conducted, for the purpose of understanding how face bounding box affects landmark localisation accuracy and to what degree. This mimics a more real world application of the landmark localisation where ground truth face bounding boxes are not provided. Evaluation of the face landmark predicted by the proposed IDM on the 300-W test set (Sagonas et al., 2013), which consists of 600 fully annotated faces and the AFW test set previously used in the face detection evaluation using the IBUG annotations. NME using face size normalisation as described in section 2.4.3 is used as the evaluation metric. Cumulative localisation error is shown in Figure 6.6 and Figure 6.7. For both evaluations the results are similar, there is a slightly larger error margin when using the bounding boxes from the IDM method compared to the using ground truth boxes. This outcome is not surprising but highlight's the importance bounding box accuracy even with state-of-the-art landmark localisation techniques. Accuracy with the IDM method is in general high, at the largest there is a 0.005 reduction in error

as a fraction of the face size. This suggests the IDM method has good bounding box accuracy and the landmark localisation is somewhat robust to initialisation. The largest error is for those landmarks that make up the jaw line which seem to be the most influenced by bounding box placement. One reason for this is in cases where the predicted face detection bounding box does not cover the entire face. An example of the effect of bounding box on landmark localisation can be found in Figure 6.8.

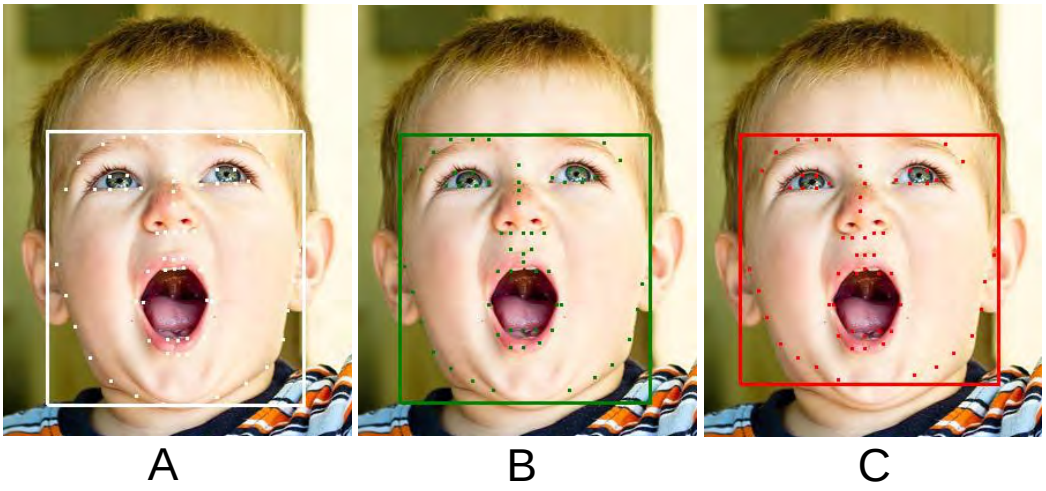

Figure 6.8: An example of Face Bounding Box affecting landmark localisation accuracy: (A) - Ground truth data, (B) - FAN using ground truth bounding box, (C) - IDM (Predicted bounding box does not cover the point of the chin which affects the landmark accuracy around the jaw line).

6.5 Conclusion

In this chapter the goal was to propose and evaluate a robust and precise face detection method that can be applied in facial analysis systems. Specifically the intention was to reduce the number of false positive detection's returned, to achieve this the Integrated Deep Model is introduced. Unlike the multi-task approach in chapter 5 which proved difficult to train a cascaded approach integrates both Faster R-CNN and FAN architectures to aid the face detection while also providing landmark localisation. To achieve this integration a novel method for heat map transformation and an integrated loss function using a joint probability for face detection and size scaling techniques were defined.

The evaluation of the IDM highlights a significant reduction of over 50% on false positive detection's against current state-of-the-art methods on the AFW and FDDB test sets. There is also a

Figure 6.9: Example of False Positive Detection's (Red boxes represent a detection by the IDM, green boxes are the ground truth face data): (A) - Faces unlabelled in the ground truth data, (B) - Overlapping boxes on a single face, (C) - Detected face box to small in comparison to the ground truth to be within the metrics applied, (D) - Non face detection's.

minimal reduction in the recall where only blurred faces provided a significant issue, primarily due to the lack of training on this type of face. In a comparative study with the approach proposed in chapter 5 the IDM is significantly more accurate for face detection, achieving a AP of 100 % against the 90% of the previous method. Figure 6.9 highlights some of the examples of situation were the IDM returns incorrect detection's, which while not addressed in this thesis will form the basis of future work. The analysis of landmark localisation performed by the IDM on the 300-W and AFW test sets highlights high accuracy though when compared with landmark localisation using ground truth face bounding boxes there is a small increase in error specifically for the landmarks which border the face such as the jaw line. The main cause of error is the precision of the face bounding box.

This is the last chapter in which the face detection task is approached, the remaining chapters of this thesis investigate the task of facial palsy grading. Where stated the IDM method is applied as the face detection stage of the facial analysis system in the remaining chapters.

Chapter 7

Geometry-based Symmetry Features for Facial Palsy Grading

7.1 Introduction

In this chapter an investigation is undertaken to look at the potential for using geometry-based features for the specific medical facial analysis task of facial palsy grading. Facial palsy grading is an important method used within the medical community in the diagnosis and rehabilitation tracking of facial palsy patients. This investigation is split in two parts, the initial investigation conducted chronologically near the initiation of this research and then a further investigation which introduces a larger data sample and applies more recent deep learning methods to the problem. A method for transforming these facial landmarks to geometry-based symmetry features is then proposed and evaluated. An investigation into generating 3D face models is also undertaken. The initial part of the research conducted in this chapter was published in IET Healthcare Technology Letters under the title "A Role for 2D Image Generated 3D Face Models in The Rehabilitation of Facial Palsy".

This chapter is organised as follows: Section 7.2 provides a overview of the rationale for both the application domain of facial palsy and the methods applied in this research. Section 7.3 details the methodology applied, while the method evaluation including the results and discussion are presented in section 7.4. Finally a conclusion is given in section 7.5.

7.2 Motivation

Early in the research time line for this contribution it was discovered that a potential facial analysis system to pursue was that of facial symmetry and more specifically facial palsy, as the previous literature established there was a medical opportunity to improve diagnosis and rehabilitation (Banks et al., 2015) and some initial computer vision and machine learning methods had been published (Wang, Dong, Sun, Zhang and Wang, 2014). In chapter 5 an initial method for conducting facial symmetry analysis highlighted limitations of that approach. In this chapter these limitation were taken into account with the use of real video sequences rather than synthesised static images and expanding the task to facial palsy grading rather than basic binary classification. In this section further discussion focuses on a more in-depth look at the condition of facial palsy, while also providing a rational for the initial study of geometry-based features as a method for predicting facial palsy grading from videos of facial palsy patients. 3D face models and their generation from 2D images is also discussed as a potential avenue for research in facial palsy grading.

7.2.1 Facial Palsy

Facial palsy is a medical condition affecting the 7^{th} cranial nerve resulting in the loss of facial muscle motion on one side of the face at varying degrees of severity. This nerve damage produces an extreme asymmetrical appearance which can be especially significant in the eyes, brow and mouth regions of the face both when at rest and during the forming of facial expressions. The most common aetiology is Bell's palsy which is associated with viral damage of the nerve and accounts for 50% of all facial palsy cases, affecting 25 in every 100,000 people annually (Chang et al., 2016). While various treatment and rehabilitation paths exist dependant on the specific aetiology of the facial palsy, the aim is to restore a degree of facial muscle movement to the patient. Previous medical research (Ishii, 2016; Guerreschi et al., 2016; Monini et al., 2016) has highlighted the correlation between patient outcomes and the diagnosis and rehabilitation prescribed by trained medical professionals. Lindsay et al. (2010) completed a comprehensive study over a 5 year period of the rehabilitation process and outcomes for 303 facial paralysis patients, the key finding was the need for specialised therapy plans tailored via regular feedback which resulted in the best patient outcomes. The most common diagnostic procedure is through the assessment of facial muscle motion observed through a range of facial expressions. There are a number of grading

Grade	Impairment
1	Normal.
2	Mild dysfunction (slight weakness, normal symmetry at rest).
3	Moderate dysfunction (obvious but not disfiguring weakness, normal symmetry at rest) Complete eye closure w/ maximal effort, good forehead movement.
4	Moderately severe dysfunction (obvious and disfiguring asymmetry) Incomplete eye closure, moderate forehead movement.
5	Severe dysfunction (barely perceptible motion).
6	Total paralysis (no movement).

Table 7.1: House-Brackmann Facial Paralysis Grading Scale.

systems that are used within the medical community globally to then determine the level of paralysis from the motion assessment, these include House-Brackmann, Chavier and the Yanagihara grading systems. House-Brackmann is considered the most widely applied method and consists of 6 scales of grading (detailed in Table 7.1) where grade 1 is normal face function and progresses in stops to grade 6 which represents total paralysis of the affected side of the face (Fattah et al., 2015).

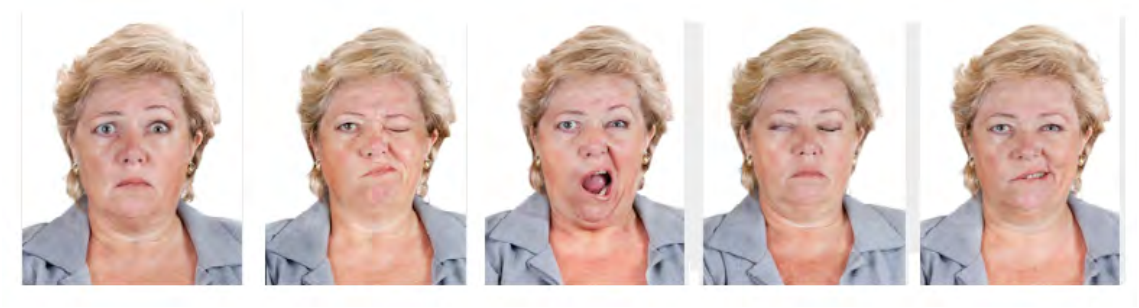

Figure 7.1: Examples of the range of facial asymmetry displayed by a facial palsy patient.

The motivation for researching the application of computer vision and machine learning for the task of aiding the facial palsy diagnosis and rehabilitation process is driven by the number of potential benefits this could bring to both clinicians and patients. The need for quality qualitative feedback to shape the development of a suitable rehabilitation plan for the best patient outcomes as established by Banks et al. (2015) is a key motivation. The United Kingdom based charity Facial Palsy UK undertook a Delphi study in 2017 which included the identification of the top 10 health professional and researcher priorities for facial palsy research, this includes investigating the most reliable set of measures (both functional and psychological) for assessing facial palsy and treatment outcomes while also identifying a method to standardise the methods and grading

systems applied in clinical settings to assess patients with facial palsy (*Identifying the research priorities for facial palsy - Facial Palsy UK*, n.d.). With the current established grading methods there is the potential for intra-observer reliability issues when a patient is assessed by different clinicians. Objective qualitative measures from computer based systems could remove these issues while providing continuity of care if for instance the clinician changes during the rehabilitation period. The potential to use such as system on a smart device such as a phone could provide the clinician with more regular objective feedback on the condition and tailor therapy without always needing to physically see the patient given. This is especially beneficial in scenarios where the distance between clinician and patient is large or the availability of either party to physically meet is limited. As the face plays a major role during interpersonal communication and facial expression, the onset of facial palsy can have a significant psychological impact upon the patients. The capability to track rehabilitation privately within a comfortable setting like their own home may also provide a benefit to certain patients.

7.2.2 Geometry-based Features

Geometry-based features have been shown over many decades to be both a accurate and computationally fast especially in task such as FER (Kotsia and Pitas, 2007; Ghimire and Lee, 2013; Ghimire et al., 2015). This is specifically true in scenarios like facial palsy where facial motion is to be considered and the capture of temporal data is important. The argument to pursue the research on video sequences rather than static images in the analysis of facial palsy is primarily based upon facial morphology. Facial morphology is defined by the bone structure and musculature of the human face, where there are significant difference between individuals, when considering a sample of facial palsy patients range of motion is defined by both the physical structure and the level of palsy. This is why in the clinical setting the patient is guided by the clinician through a range of facial motions from which the clinician can establish the normal range of motion from the unaffected side and factor this into the assessment (Fattah et al., 2015).

The key method for predicting the facial landmarks from which geometry-based features are derived from is through landmark localisation. As discussed in chapter 2, landmark localisation is a classic computer vision problem which is applied to determine key facial features locations such as the mouth and eyes when given an image containing a human face. It has been utilised fre-

quently in many application domains that perform facial analysis tasks such as medical diagnosis in Wang, Dong, Sun, Zhang and Wang (2014) and FER in Zhao et al. (2007); Li et al. (2015). Computer vision methods have previously been identified as a method for predicting facial palsy grading, challenges remain though due to the atypical nature of the features present in those with facial paralysis (Wang, Dong, Sun, Zhang and Wang, 2014). The previous work on facial paralysis and atypical faces is at present limited, research in Wang, Dong, Sun, Zhang and Wang (2014) and Wang et al. (2016) is aimed at recognising the patterns of facial movement from facial paralysis patients using specific facial landmarks and LBP features with a SVM classifier, in both of these works ASM (Cootes et al., 2001) are used for landmark localisation fitting 68 facial landmarks which are used as guides for the feature extraction regions. It was reported that the ASM model applied for landmark localisation performed poorly at fitting landmarks accurately when not trained directly on the asymmetrical faces present within the data set and also used for palsy classification. This highlights an issue with the generalisation of the ASM method for unseen faces which could be problematic for medical based systems where unseen faces would be common. Therefore prior to applying geometry-based features there is a requirement to understand the accuracy of landmark localisation techniques when evaluated on the unseen asymmetrical faces of individuals with facial palsy. The impact on the accuracy of the landmark localisation will impact on the accuracy of any geometry-based features and therefore part of this chapter investigates the accuracy of landmark localisation on unseen faces.

7.2.3 3D Face Models

There is a potential capacity to apply 3D face modelling to an automated framework for tracking of facial palsy rehabilitation. The motivation for this idea is that 2D landmarks are limited to only 68 locations on the X and Y plane, while a dense 3D mesh consists of thousands of landmarks providing a far richer model of the face and has the added Z plane for depth. There are a number of methods to generate these dense 3D meshes, the 3D Morphable Model (3DMM) (Blanz and Vetter, 1999) which aims to fit a 3D face shape to a 2D face image across a large range of poses. The 3D face model is modelled using a linear subspace such as Tensor (Cao, Hou and Zhou, 2014) or PCA (Blanz and Vetter, 2003) and achieves fitting through the minimisation of the difference between the model appearance and the image. Often these techniques leverage a set of previously localised 2D landmarks and project these to relevant key-points on the dense mesh to determine

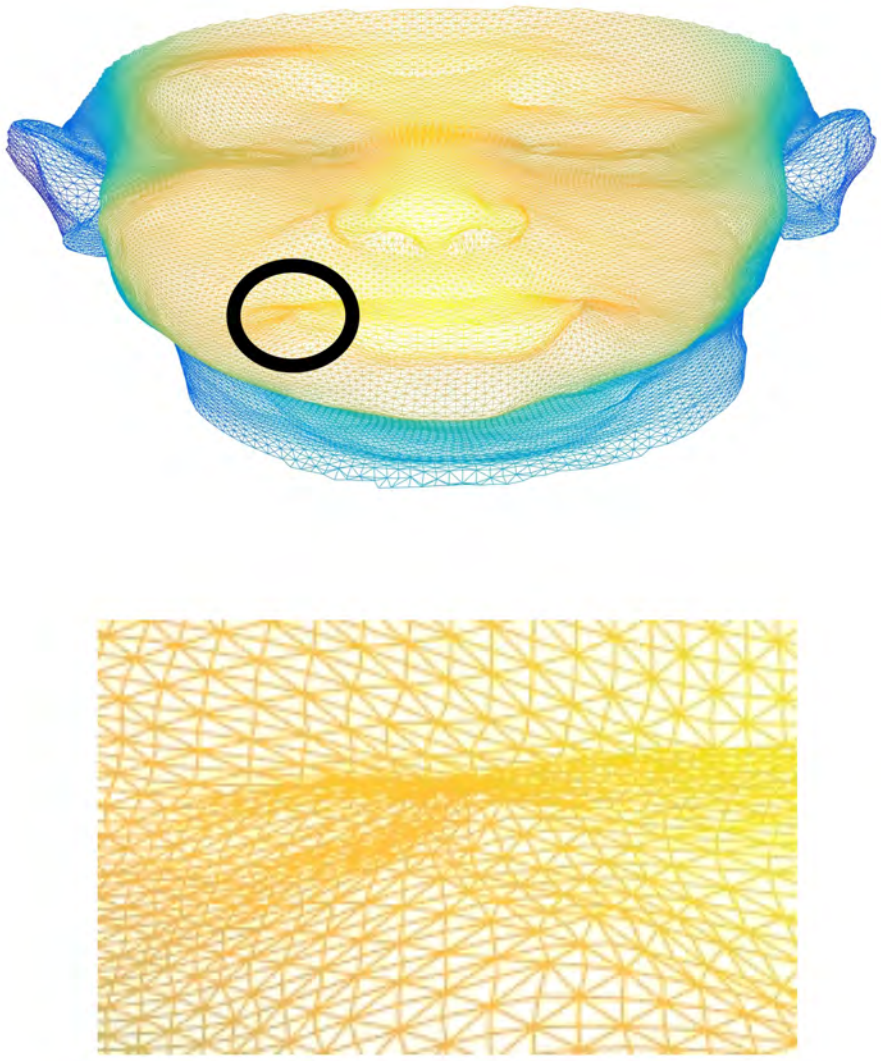

Figure 7.2: Example of a 3D face models dense mesh of vertices that describe the face geometry

shape and expression parameters. These techniques can also suffer from a high computational cost though recently regression based 3DMM fitting (Cao, Hou and Zhou, 2014; Yu et al., 2016; Jeni et al., 2015) have improved on the efficiency. Methods have more recently introduced deep networks for more accurate regression of these parameters (Tran, Hassner, Masi, Paz, Nirkin and Medioni, 2017). A method for 3D mesh generation is discussed in the next section and some initial qualitative results are presented.

7.3 Method

While the task of landmark localisation provides a fundamental underpinning in the method proposed in this chapter, the geometry-based symmetry features proposed are agnostic to the method of landmark localisation. In this section the focus is on the geometry-based symmetry feature extraction method applied to landmarks and used for facial palsy grading, with the accuracy of specific landmark localisation methods on unseen faces evaluated in the next section. Also presented within this section is an initial proposal for the potential use of 3D face models to further the available landmarks.

7.3.1 Geometry-based Symmetry Features Extraction Method

The transformation of the n detected facial landmarks from a series of images into a set of geometry-based symmetry features, described within this section exploits the idea of the facial symmetry present within people with facial palsy. These extracted features can be used to provide a quantitative facial palsy grading to aid or inform the clinician of the rehabilitation process. Given a video sequence of a patient of with n frames defined as $V = \{F_1, F_2, ..., F_n\}$. Each of the n frames of the video sequence V are inputted for feature extraction, where the value of n differs on a video to video basis due to varying lengths. For every frame F_i in the sequence V firstly they are processed through a face detection task which predicts a corresponding face detection bounding box FD_i to F_i where $FD_i = \{x, y, w, h\}$ with x and y the coordinates referring to the upper left position of the box with the video sequence frame and w and h being the width and height of the box respectively. Prior to the landmark localisation task F_i is cropped to the bounding box FD_i. Landmark localisation predicts L_i as a n x m matrix, where $n = 68$ represents the different facial landmarks and $m = 2$ is the x and y coordinates planes. Once the total sequence V has been processed there exists $LL = \{L_1, L_2, ..., L_n\}$ which represents all predicted facial landmarks coordinates for each of the n frames of a video sequence.

To transform the predicted LL into a set of features that can be applied to the task of facial palsy grading, a number of separate steps are proposed to generate the quantitative grading measure. The first step is head motion removal, this is an important step to remove any affects of noise being introduced in the data. While many facial landmarks move significantly both on the x-axis and y-axis in the 2D coordinate space during the formation of facial expressions, there is a small set of

landmarks which remain static and can then be applied to differentiate between facial expression formation and head motion. The head motion removal function is defined as:

$$M(L_1, L_i, S) = L_i - (\frac{1}{n}\sum_{j=1}^{n}(L_{1j} - L_{ij})) \tag{7.1}$$

where $M(L_0, L_i, S)$ takes as inputs the predicted landmarks first frame L_1, the landmarks for the ith frame of the sequence and a set of static facial landmarks $S = \{x_1, \ldots, x_n\}$. The mean motion shift is calculated as the sum of the difference between L_{1j} and L_{ij} where j refers to a static landmarks from set S. $S = \{2, 16, 28, 29, 30, 31, 34\}$ as defined by the FAN method (Bulat and Tzimiropoulos, 2017*b*) were used within this research. This mean motion shift is then removed from all landmarks in L_i.

The features are extracted on the basis that facial symmetry/asymmetry is a crucial aspect of the clinical assessment procedure. A subset of the facial landmarks can be paired such as the left and right mouth corners, the method uses these paired landmarks as a crucial measure for determining the palsy grading. Step two involves shifting each facial landmark coordinate so L_1 is now positioned at the origin $(0, 0)$, all subsequent n frames L_n are shifted in accordance to L_1. Landmarks located on the left side of the face have the x-axis values negated this aids direct comparison to the right hand side paired landmark. From the transformed LL a subset of data for a single landmark j is taken as $D_j = \{L_{1j}, L_{2j}, ..., L_{nj}\}$ for all n landmarks and the maximum absolute value is determined as LD_j. LD_j holds two floating point numbers which describes the maximum movement from the origin a landmark makes during a facial motion on the x and y axis. The final features for all landmarks $LM = \{LD_1, LD_2, ..., LD_j\}$ this is then reduced through paired comparison, note that those non-parable landmarks are not used. Paired comparison simply requires finding the difference in motion between landmarks pairs e.g $LD_{49} - LD_{55}$ for the corners of the mouth. The value determined by paired comparison is then scaled. This scaling value is applied to remove the differences in face size. These values are the final quantitative values that can be used for facial palsy grading and tracked across time to evaluate rehabilitation.

For the purpose of evaluating the proposed method the derived quantitative values are transformed to House-Brackmann facial palsy grading scale. This process applies a min-max normalisation

range defined as:

$$x_{norm} = (b - a)\frac{x - min(x)}{max(x) - min(x)} + a \tag{7.2}$$

where the value x_{norm} is a House-Brackmann grades, b and a represent the House-Brackmann grades possible range therefor $a = 1$ and $b = 6$ respectively. $min(x)$ is defined as 0 which would represent a perfectly symmetrically face motion, while $max(x)$ is defined as the largest observed value from the data set.

7.3.2 Generating 3D Face Model

Generation of a 3D face model from any 2D image and 68 facial landmarks fitted during the landmark localisation process can be achieved through the application of a 3DMM fitting method. The 3DMM is used to represent a dense 3D shape of an individuals face which in this method applies the fitting technique as described by Zhu, Lei, Yan, Yi and Li (2015). The 3DMM is defined as:

$$S = \bar{S} + A_{id}\alpha_{id} + A_{exp}\alpha_{exp} \tag{7.3}$$

where S describes the 3D face, $\bar{S}$ is the mean shape, A_{id} and A_{exp} are the principle axes trained on the 3D face scans with neutral expression and expression scans respectively. While α_{id} are the shape parameters and α_{exp} the expression parameters. A_{id} and A_{exp} are provided by the Basel Face Model (Paysan et al., 2009) and Face-Warehouse (Chen Cao et al., 2014) respectively. A Weak Perspective Projection is used to project the face model to the image plane for the fitting of the 3DMM to a face image as:

$$s_{2d} = fPR(S + t_{3d}) \tag{7.4}$$

where s_{2d} are the 2D positions of 3D points on the image plane, f denotes the scaling factor, P is the orthographic projection matrix $\left(\begin{smallmatrix}1 & 0 & 0\\ 0 & 1 & 0\end{smallmatrix}\right)$, $R = (\alpha, \beta, \gamma)$ is the 3×3 rotation matrix constructed with pitch(α), yaw(β) and roll(γ) and t_{3d} is the translation vector. The fitting of this model is

defined by:

$$\underset{f,R,t_{3d},\alpha_{id},\alpha_{exp}}{\arg\min} \quad \|s_{2dt} - s_{2d}\| \tag{7.5}$$

where the 2D landmarks predicted through landmark localisation are defined here as s_{2dt}, the associated 3D points and the model parameters are estimated by minimising the distance between s_{2d} and s_{2dt}. A fitted 3D face model S is a dense mesh consisting of m vertices where $m = 53215$ in this method. The geometry-based symmetry features as described in the previous section could then be extended from using 2D landmarks to 3D landmarks of the dense 3D face model mesh with the potential to leverage the depth information for more accurate facial palsy grading. In this chapter an initial study investigates the feasibility of generating 3D face model of facial palsy patients.

7.4 Evaluation

Within this section there are three distinct areas of evaluation, the first is to establish the accuracy of landmark localisation methods when applied to faces which display various grades of facial palsy. The evaluation of landmark localisation accuracy is split into an initial and a further study, the initial study was conducted at the inception of this research when landmark localisation methods were still dominated in terms of performance by traditional methods and the deep learning methods were in their infancy, the further study revisits this following the introduction of further deep learning methods which showed significant improvements in accuracy on benchmark data sets when applied to non-palsy faces. The second distinct evaluation is concerned with the task of facial palsy grading using extracted geometry-based features derived from the landmarks localisation methods as described in the previous section. Finally an investigation on the ability to generate accurate 3D face models is discussed.

7.4.1 Landmark Localisation - Initial Investigation

In this initial experiment the aim was to determine the accuracy of landmark localisation methods that were state-of-the-art at the time of this initial investigation. Based upon the literature at the time a number of top performing landmark localisation methods where identified, these were

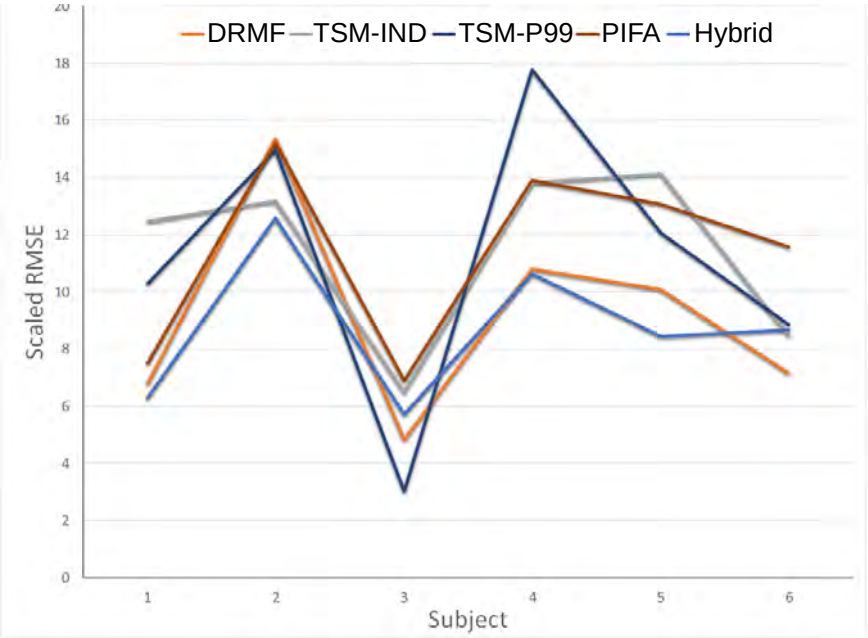

Figure 7.3: Root Mean Square Error Per Subject

Discriminative Response Map Fitting (DRMF) (Asthana et al., 2013), PIFA (Jourabloo and Liu, 2016) and TSM (Ramanan et al., 2012) where both a fully independent part model and also a shared part model using 99 parts released by the authors were used in the evaluation. Also an initial deep learning based method PIFA was evaluated along with a hybrid method which applied a PIFA to the mouth landmarks only and DRMF for the other landmarks. All methods were evaluated using codes released by the respective authors and were used off-the-shelf to ascertain how well the methods generalised to asymmetrical faces presented within the data set. A data set of 6 individuals who have a confirmed diagnosis of facial palsy were used to conduct the evaluation. In this evaluation each image was a cropped full frontal facial image which had been manually marked with 68 facial landmarks to be used as the ground truth landmark positions. As different landmark localisation methods use different landmark schemes for evaluation purposes a subset of landmarks are applied that are uniform across the methods. The root mean square error (RSME) with ocular scaling to deal with the images having different size faces as used in Asthana et al. (2013) to measure the accuracy of the techniques.

Figure 7.3 highlights that across the test data set the accuracy of methods on a subject to subject basis can vary significantly. The TSM shared model is especially volatile giving the best fit for subject 3 but by far the worst for subject 4. While not performing the best for any subject the

Method	Total RMSE
DRMF	9.17
TSM Independent	11.42
TSM 99 Part Shared	11.16
PIFA	11.36
Hybrid	8.72
FAN	6.81

Table 7.2: Total root mean square error per method.

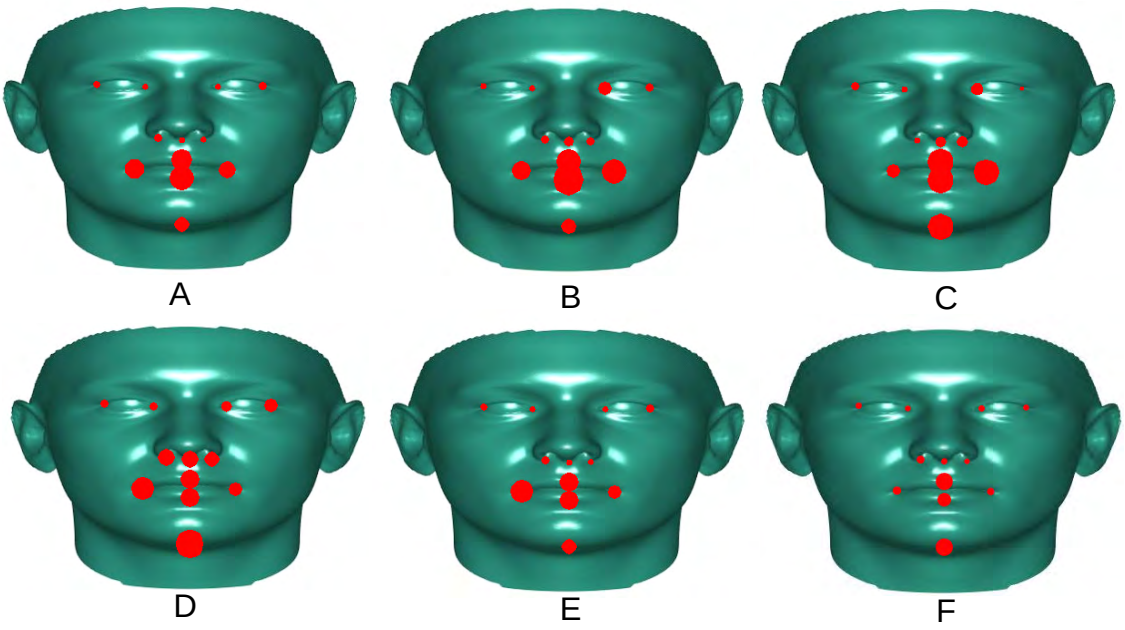

Figure 7.4: Evaluation showing root mean square error per landmark with larger landmarks representing a larger error: (A) - DRMF, (B) - TSM Independant, (C) - TSM 99 Part Shared, (D) - PIFA (E) - Hybrid and (F) FAN.

Hybrid method is the most consistent in terms of accuracy across the data set. When examining the mean RMSE as shown in Table 7.2 it can seen that the Hybrid performs above all other methods with DRMF also showing a distinct advantage over the other methods including the CNN based PIFA method. For comparison the FAN method was added retrospectively highlighting the improvement over the previous methods on this specific test set.

When further examining the results shown in 7.4, the accuracy for certain key facial landmarks using DRMF has very high accuracy for all landmarks except for the mouth area. Where as the PIFA method fits mouth landmarks better but struggles on this data set with accuracy in other areas. The hybrid method provides the best accuracy though there is still some issues when fitting the corners of the mouth. This is likely due to the asymmetrical nature of the mouth location on the

test set and that none of the models tested have been trained on any data specifically relating to this condition. This initial investigation highlights a large level of error across all methods highlighting a problem with then generalising to the asymmetrical faces presented within the data set, even when combining the best performing methods for different landmarks the error is still significant. The conclusion from this initial investigation is that these landmark localisation methods are bot accurate enough to extracted discriminative geometry-based feature. For further comparison the FAN method was added retrospectively highlighting the improvement over the previous methods on this specific test set.

7.4.2 Landmark Localisation - Further Investigation

This further investigation extends the previous study in two specific ways. Firstly the PIFA Jourabloo and Liu (2016) is exchanged for the state-of-the-art FAN method (Bulat and Tzimiropoulos, 2017*b*) for landmark localisation. Secondly an expanded data set is introduced which provides a more robust set of results than in the initial investigation. The methods evaluated in order of publication are the TSM (Ramanan et al., 2012), DRMF (Asthana et al., 2013) and the deep learning based FAN (Bulat and Tzimiropoulos, 2017*b*).

The evaluation of the landmark localisation accuracy uses two separate data sets both containing images of individuals with varying grades of facial palsy. Data set A consists of 47 facial images which have 12 ground truth landmarks. Data set B consists of a further 40 images which are annotated with 18 ground truth landmarks per image. NME using face size normalisation as described in 2.4.3 was used as the evaluation metric. Different methods of landmark localisation have variance in both the number and specific locations of the landmarks predicted, a subset of facial landmarks are used which are common across all methods which allows for a comparative analysis.

The cumulative localisation NME error for data sets A and B are shown in Figure 7.5 and Figure 7.6 respectively. The results show that the deep learning based FAN method displays a consistently higher level of accuracy across both data sets. DRMF performs accurate landmark prediction for certain test samples but specifically in test set B where there is high degree of facial asymmetry there is a percentage of the sample for which the error increases by a substantial amount. Finally TSM performs poorly in general comparatively and this error grows substantially as the level of fa-

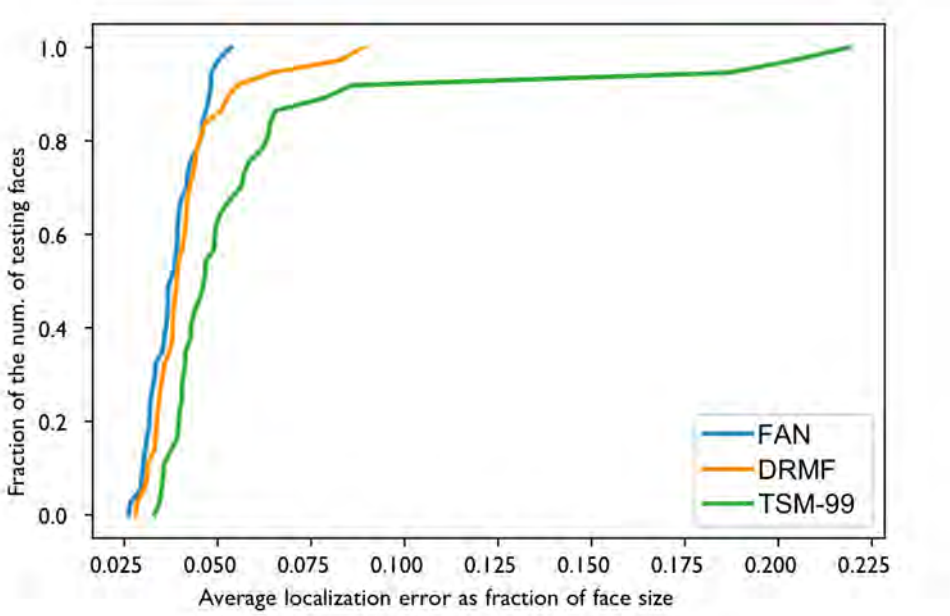

Figure 7.5: Cumulative Localisation Error Distribution - Facial Palsy Test Set A.

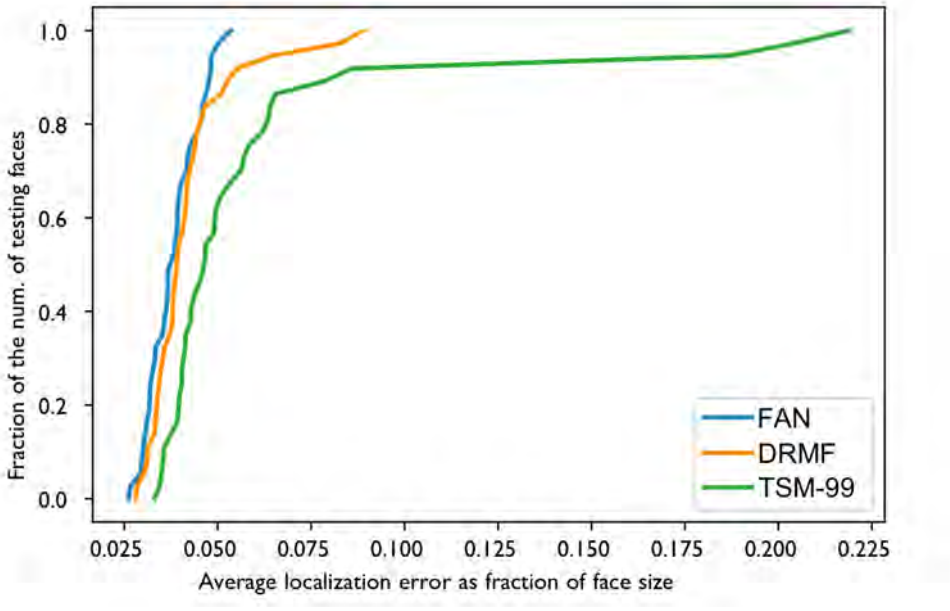

Figure 7.6: Cumulative Localisation Error Distribution - Facial Palsy Test Set B.

cial asymmetry increases. Analysing the NME error for a specific selection of landmarks as shown in Figure 7.7, the results show that while FAN and DRMF have similar levels of accuracy for the eye and nose landmarks, the mouth which has the largest range of asymmetrical deformation is where the deep learning based FAN excels. Figure 7.8 provides a visual example of the landmark

localisation output, this highlights the capability of the deep learning FAN method to provide a high level of accuracy when fitting landmarks to the face and specifically the mouth region when compared with previous techniques.

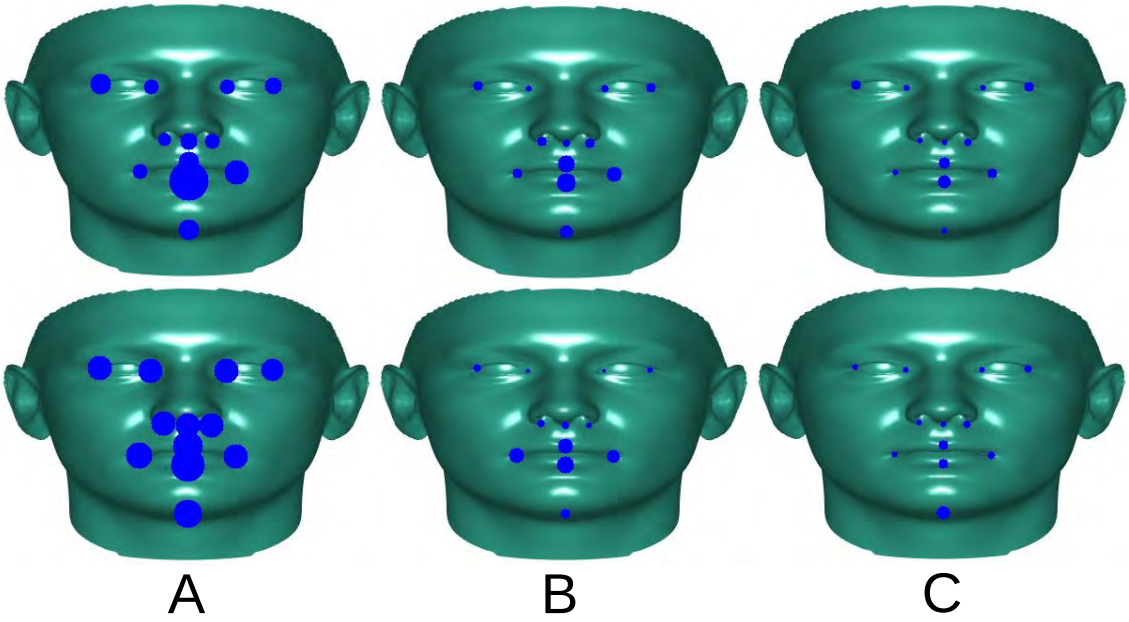

Figure 7.7: Normalised Mean Error Per Landmark: (Top) - Facial Palsy Test Set A (Bottom) - Facial Palsy Test Set B, (A) - TSM 99 Part Shared, (B) - DRMF, (C) - FAN

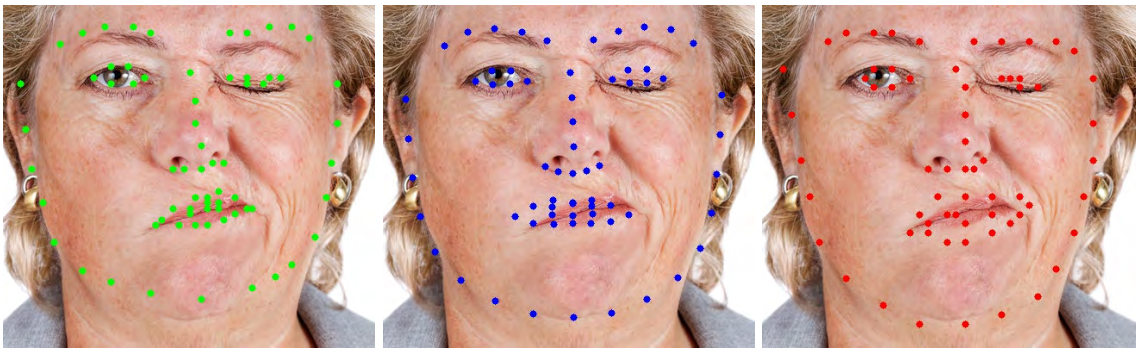

Figure 7.8: Landmark Localisation fitting example for each evaluated method. (Left) - FAN, (Centre) - DRMF, (Right) - TSM.

7.4.3 Geometry-based Symmetry Features Evaluation for Facial Palsy Grading

To validate the proposed facial palsy grading method, an evaluation was undertaken applying a data set consisting of facial palsy video sequences labelled with ground truth facial palsy grading using the House-Brackmann (Fattah et al., 2015) method as described within Table 7.1. The method applied to extract and transform the predicted facial landmarks into quantitative metrics then for evaluation purposes converted these to the House-Brackmann grading system is described in detail within section 7.3. The evaluation data set consists of 25 samples where each sample consists of a lower and upper facial motion sequence, these being a smile and eye brow raise respectively, this equates to a total of 50 video sequences. In total 7 different individuals comprise the 25 samples, 2 male and 5 female respectively, each of the individuals have multiple samples taken a different times where the grading of their facial palsy changes due to the effects of rehabilitation. To provide further analysis on the importance of facial landmark accuracy, facial landmarks predicted using the FAN, DRMF and TSM methods are applied to generate the geometry-based symmetry features. In the experiments two subsets of landmark pairs are used. For a lower sequence a subset of mouth landmarks were applied that best described that motion, these being landmarks pair (49,55) using both x-axis values. The upper motion applied landmark pairs (19,26) on the y-axis. The final quantitative facial palsy grading was the mean of the pairs. These values transformed to House-Brackmann scale using min-max normalisation as a range of 1-6 and rounded to the nearest integer. This normalisation was possible once all features had been extracted from the facial landmarks across the whole data set the data consisted of all House-Brackmann grades.

The results are displayed in Figure 7.9 for the FAN, TSM and DRMF methods within confusion matrix's indexed by a House-Brackmann grading level. The results are broken down to highlight the findings of both the upper and lower facial movement sequence for each individual and also the total grading. The total is derived from the mean of the sum of the upper and lower grading respectively. It is clear from these results that the less accurate landmarks localisation methods of TSM and DRMF are not capable of generating features that can be applied to this tasks. The results for the FAN extracted symmetry features though show a capability to effectively grade the palsy especially within a ± 1 grade range, although with caveats. An F1 Score of 0.4 is reported achieved though this is significantly increased to 0.87 if those within 1 grade are classified as correct. There

are two individuals where the grading for one specific sequence type is over 4 grades out. On inspection both highlight a small landmark fitting error for that specific sequence which induced an error. Table 7.3 provides a further overview of the results showing the breakdown between each sample.

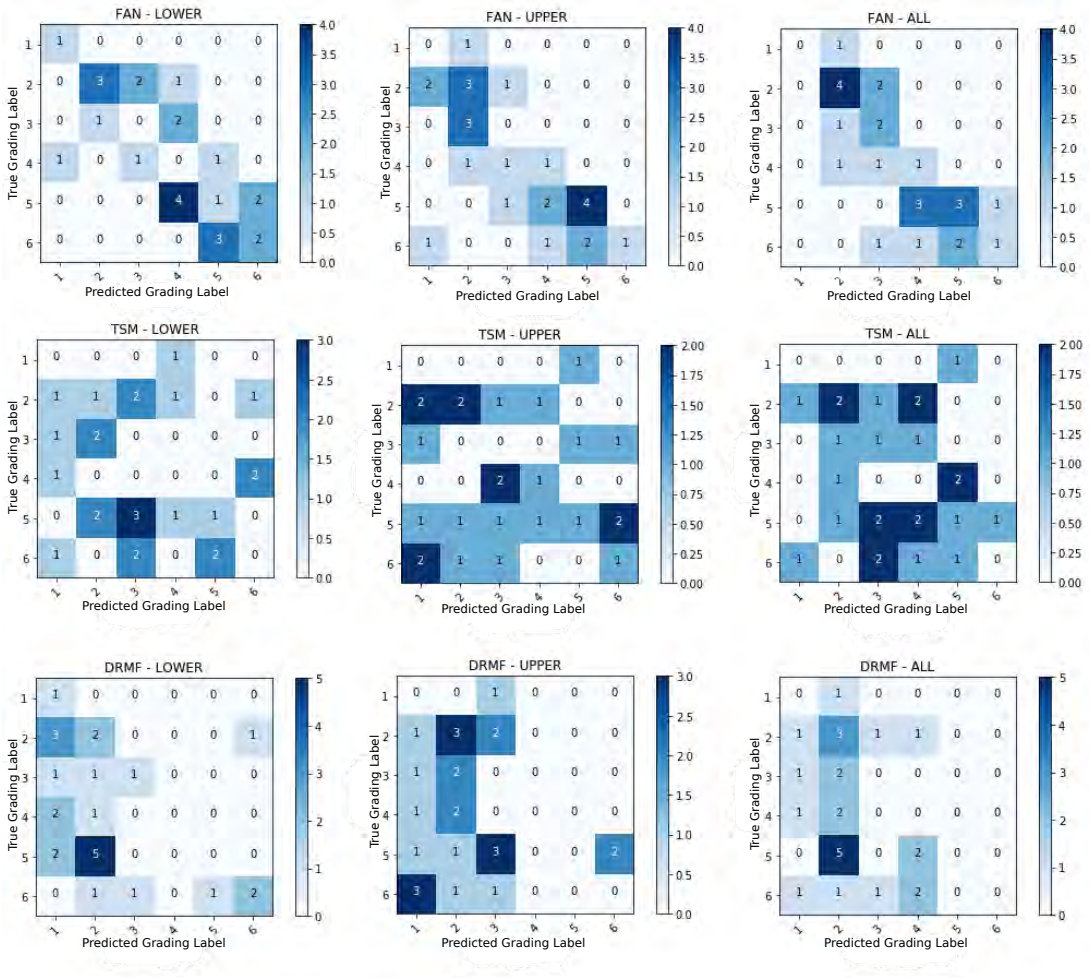

Figure 7.9: Facial Palsy Grading Validation confusion matrices (TOP) - FAN, (Middle) - TSM, (Bottom) - DRMF. (Left) - Lower, (Centre) - Upper and (Right) - All

Table 7.3: Expanded results with sequence information for evaluating the validating the proposed Facial Palsy Quantitative Grading method using FAN

Sample	**Subject**	**Ground Truth Grade**	**Upper Grade**	**Lower Grade**	**Total Grade**
1	1	5	4	5	4
2	1	4	1	4	2
3	1	3	2	2	2
4	1	2	2	2	2
5	2	6	5	4	4
6	3	2	3	3	3
7	4	6	6	5	5
8	4	4	3	3	3
9	4	2	3	2	2
10	4	1	1	2	2
11	5	5	4	3	4
12	5	5	4	5	5
13	5	4	5	2	4
14	5	5	6	5	5
15	6	2	2	1	2
16	7	6	5	1	3
17	7	6	5	5	5
18	7	5	5	4	5
19	7	3	4	2	3
20	8	6	6	6	6
21	8	5	6	5	6
22	8	3	4	2	3
23	8	2	4	1	3
24	9	5	4	4	4
25	9	2	2	2	2

7.4.4 3D Model Generation Investigation

The approach to this initial investigation into generating 3D models of individuals with facial palsy is qualitative. This is because at present there is no 3D ground truth data to produce a quantitative measure of model accuracy. In this approach the landmarks detected in section 7.4.1 were used to generate 3D face models using the method previously described. Figure 7.10 shows an example of the 3D models generated, models generated from the more accurate FAN method, provides the most accurate model when inspected visually. Significant differences exist between the shape of the face from the original image and that of the 3D models, for example the width and length of the face are not accurate. Due to this level of 3D model generation error, it was decided not to extend the geometry-based symmetry features to use landmarks from 3D models until a time when more accurate models can be generated.

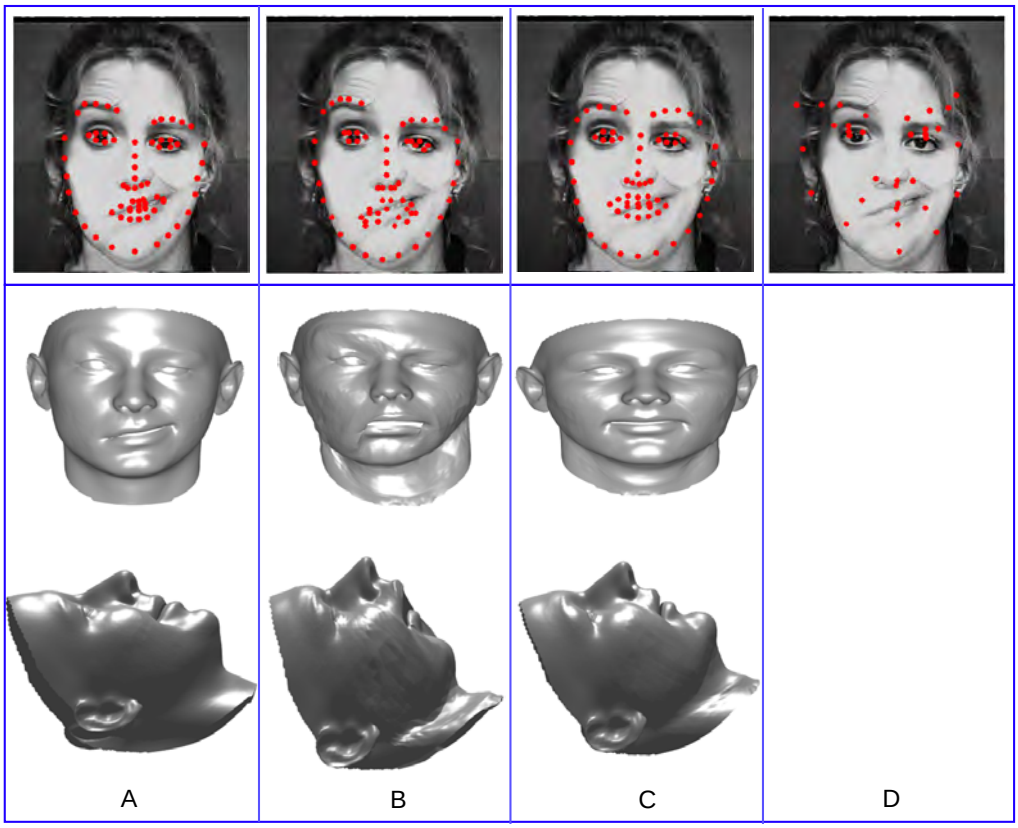

Figure 7.10: Examples of landmark localisation and the subsequent generated foundation 3D face mesh. (A) - FAN, (B) - TSM, (C) - DRMF, (D) - PIFA (Note a model for this method was not generated due to the sparse set of landmarks.

7.5 Conclusion

The focus of this chapter was to investigate the possibility of facial palsy grading through the application of geometry-based features. The key foundation generating such feature is the accuracy of landmark localisation which had in previous research been shown to be inaccurate when using the ASM method on unseen faces, which are those not in the training data of the model. The initial investigation in this chapter was to evaluate significant landmark localisation methods on unseen individuals with facial palsy. It was found that only the state-of-the-art FAN method could accurately predict landmarks especially on the difficult mouth landmarks. A method of transforming these predicted facial landmarks to a set of geometry-based symmetry features was proposed and evaluated for the task of facial palsy grading, where it was found that while these features show initial promise they are not as accurate as required for a medical system. Specifically they tend to be ± 1 grade out, while the samples which show greater variance highlight that small landmark localisation error affects the predictions significantly. The application of 3D face models is also initially proposed, this used the 2D image and landmarks to generate a 3D face model using the 3DMM method, while this has potential to extend the number and dimensions of geometry-based features, from the evaluation it is clear that the current capacity to generate 3D models from 2D images is not accurate. Though this may be a viable option in the future as methods evolve.

The overall conclusion is that while features derived from facial landmarks show initial success for facial palsy grading, there exists a level of error in landmark fitting even in the current state-of-the-art which can introduce error in the prediction process. Within FER systems it has been highlighted that methods which apply other feature types, such as texture based features can perform well which provides an avenue for further research. The next chapter further examines the potential for the use of computer vision and machine learning in the task of facial palsy grading by introducing the use of a fully 3D CNN for feature generation instead of the geometry-based features presented in this chapter.

Chapter 8

3D CNNs for Facial Palsy Grading and Mouth Motion Recognition

8.1 Introduction

In the previous chapter the task of facial palsy grading was approached through the application of geometry-based symmetry features, in this chapter the task of facial palsy grading prediction is approached again. This time applying a proposed framework which uses a fully 3D convolutional neural network to learn features and produce predictions. In addition to facial palsy grading a secondary task of mouth motion recognition is also evaluated using the same proposed framework. Further evaluation of the proposed framework is conducted concerning the loss function choice and the frame duration parameters and there affect on classification accuracy of the aforementioned tasks of facial palsy grading and mouth motion recognition. The research conducted in this chapter was published in IEEE Access under the title "3DPalsyNet: A Facial Palsy Grading and Motion Recognition Framework using Fully 3D Convolutional Neural Networks".

This chapter is organised as follows: Section 8.2 provides a overview of the rationale for the research undertaken within this chapter, while an overview of the methods applied within the proposed framework including model architecture, training and pre-processing are described in section 8.3, the method evaluation including the results and discussion are presented in section 8.4. Finally a conclusion is given in section 8.5.

8.2 Motivation

The direction of the research within this chapter is a direct result of the findings of the previous chapter and the information regarding methods gathered during the literature review. In the previous chapter research was conducted into using specific facial landmarks extracted from a sample of individuals with facial palsy as geometry-based symmetry features, these features were used to produce a facial palsy grading prediction. While the results were promising, the use of facial landmarks presented a number of issues. The main issue is that of fitting error, while the FAN method provided a significant increase in accuracy over previous non-deep learning methods error still exists which can affect results.

Further to this to realise an automated system that can assist medical professionals in the tracking and planning of a facial palsy rehabilitation plan, there are a number of additional challenges. Two key functions of a potential system are the capability to recognise facial motions and grade the facial palsy level accurately. In the clinical setting the medical professional would guide the patient through a range of specific facial motions for facial palsy grading, in this scenario the medical professionals supervision would ensure correct motions were performed (Wang et al., 2016). The challenge of recognising specific facial motions for example a smile have been heavily research in the domain of facial expression recognition but when introducing facial palsy the asymmetrical nature of the facial motion add a further challenge (Li and Deng, 2018). Figure 8.1 provides an example this specific challenge, where it can be seen that smile of a individual with facial palsy is different from someone without this condition, while also sharing a resemblance to other asymmetrical facial expression that would not be classified as a smile. It can be noted also that as the grading level of facial palsy reduces the level of asymmetry is also reduced. The challenge is to be able to recognise a smile from a patients when either no or low grade of asymmetry up to the most severe of grades.

As discussed in the previous chapters it is known that the temporal information such as that from video data can provide further information to ascertain a facial expression (Cohen et al., 2003), this temporal information also has the potential to boost facial palsy grading providing the capability to examine the range of motion across an entire action rather than a single frame. Within the task of FER discussed within section 2.5.2 of the literature review it was that highlighted that geometry-based features (as proposed in the previous chapter) while computationally fast were outperformed

Figure 8.1: Mouth Motion Smile Examples: (Left) - Smile with facial palsy, (Centre) - Smile without facial palsy, (Right) - Not a smile without facial palsy.

by other feature types. Action recognition tasks from video sequences has remained difficult and has not yet seen the dramatic increase in performance accuracy that has occurred in detection tasks from static images with deep learning approaches. While approaches applying deep learning based methods have been proposed such as RNN (Yu et al., 2018), Two-stream (Simonyan and Zisserman, 2014) and C3D (Tran, Ray, Shou, Chang and Paluri, 2017) each method has shown some limitations. RNN based networks have been shown to be incapable of capturing the powerful convolutional features for recognition tasks (Li and Deng, 2018). Two-stream methods use both image data and optical flow features to represents the spatial and temporal data respectively, while producing some of the most promising results they require pre-processed optical flow features which adds additional computational overhead. While the C3D method uses 3D convolutional layers to learn spatio-temporal features and proved good accuracy on the sport action data set, it does not generalise well to other recognition task (Hara et al., 2018). Partially the reason for this is down to the relatively small video data sets available for optimising the large number of parameters in 3D CNNs. In addition the C3D network is shallow in comparison to the state-of-the-art architectures used in image detection, where deeper networks have generally performed better. Recently the introduction of the Kinetics data set (Kay et al., 2017) which contains 300,000 videos has provided a large scale data set that has the potential to train deep 3D CNNs that have the capability to generalise well to other general action recognition tasks (Hara et al., 2018). This motivated the method defined in this chapter in which a 3D CNN pre-trained on the Kinetics data set applies transfer learning to ascertain the potential to apply this method for facial analysis systems.

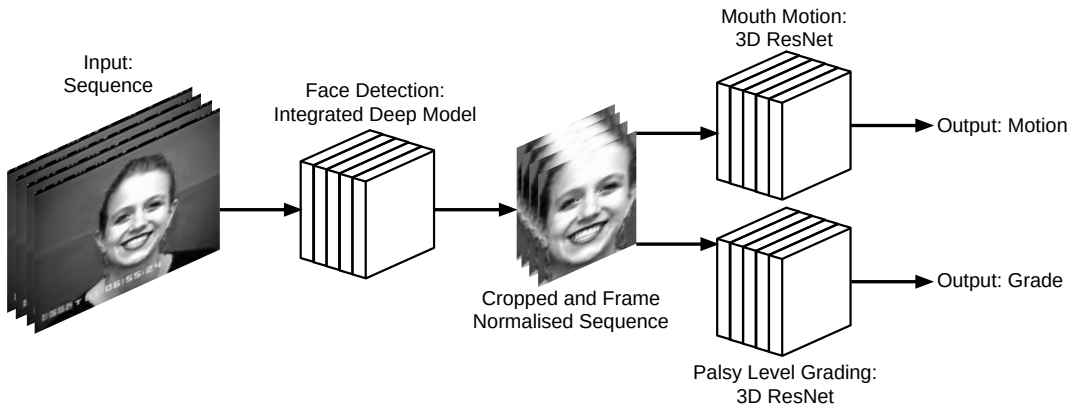

Figure 8.2: 3D CNN Framework Overview.

8.3 Method

In this chapter a framework is presented for the novel tasks of mouth motion recognition and facial palsy grading. The framework leverages the Integrated Deep Model (Storey et al., 2018) to initially perform face detection on video sequence frames, then a fully 3D end-to-end CNN network with a ResNet backbone for each of the recognition tasks. Furthermore a center loss strategy is incorporated into the training due to the advantages it has shown for learning discriminative features in other recognition image based tasks over Softmax loss alone (Li and Deng, 2018). The facial analysis system proposed for the task of mouth motion recognition and facial palsy grading comprises of two distinct stages the pre-processing stage and motion analysis stage, the proposed framework is shown in Figure 8.2.

8.3.1 Face Detection and Video Sequence Pre-processing

The IDM (Storey et al., 2018) provides accurate face detection which has shown to provide both high recall and precision as detailed in chapter 6. The requirement for accurate face detection is essential to ensure that faces from each frame of the video sequences are extracted for the second stage of the framework. To accurately address head motion across the frames, instead of extracting the face directly from the corresponding bounding box for that specific frame, a method is used which calculates the global bounding box for a sequence. Using the minimum and maximum values for the X and Y bounding box coordinates the global bounding box is determined. This global bounding box is then used to extract the face from across each frame of the sequence.

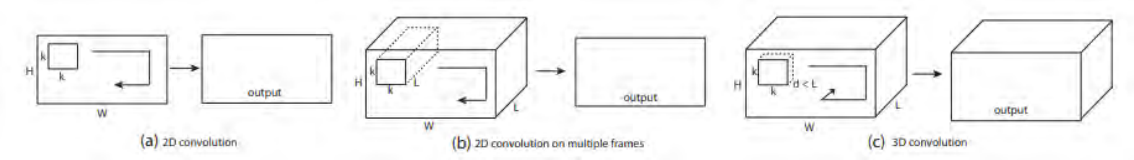

Figure 8.3: An example of how a 3D convolution differs from a 2D convolution. Where the 3D convolution outputs a 3D spatio-temporal volume compared with the 2D spatial output of the 2D convolution (Tran, Ray, Shou, Chang and Paluri, 2017).

8.3.2 3D Convolutional Neural Networks

3D convolutional architectures are a natural extension of the 2D architectures that have been widely applied to great affect in image classification tasks. Where 2D CNN apply a 2D convolutional filters and pooling which learning discriminative features in the spatial domain the application of 3D convolutional filters and 3D pooling operations provide the capacity to learn features in the spatio-temporal domain. Figure 8.3 illustrates the difference between 2D and 3D convolutions, significantly when the input is either an image or a sequence of images the output from a 2D convolution is still an image, where only the spatial information is preserved and the temporal information lost. To also capture the temporal data which is important in action recognition task from video data, only 3D convolution naturally achieves the preservation of spatio-temporal data. This procedure is also true of 2D and 3D pooling techniques applied within CNN architectures. While not being the first research to use 3D CNN's the so called C3D architecture (Tran, Ray, Shou, Chang and Paluri, 2017) was the first to use fully end-to-ends 3D convolutions. The network consisted of 8 3D convolutions, 5 3D max-pooling and 2 fully connected layers with a softmax output layer for classification. While the C3D method provided an early indication on the potential for fully 3D CNN based architectures for action recognition from video sequences, when applied to tasks involving facial analysis the results were not state-of-the-art (Li and Deng, 2018). With the proposal of deeper 3D CNNs in Hara et al. (2018) and the introduction of the Kinetics data set (Kay et al., 2017), 3D CNNs are now potentially viable for facial analysis tasks. The 3D CNN method used within the proposed framework adopts the ResNet He et al. (2016) architecture as the backbone of the network, these architectures have been highly successful for image classification tasks. ResNet is described further in the next section.

8.3.3 Residual Networks

The Residual Network commonly abbreviated as ResNet is a ground breaking architecture allowing for very deep CNNs (He et al., 2015). AlexNet an initial key CNN architecture consisted of only 5 layers, as subsequent architectures were defined by researchers, further layers were added in a stacked approach which produced performance improvements. There became a point at which simply stacking further layers negatively affected the performance. This issue is known in the research community as the vanishing gradient problem, as architectures get deeper when the gradient is back-propagated to earlier layers the repeated multiplication may make the gradient infinitely small, this lead to inadequate changes to parameter weights affecting the training of the network. ResNet was designed to overcome this vanishing gradient and allow for very deep networks. Given a traditional CNN with stacked layers each layer learn a function $y = H(x)$ where x is the input from the previous layer, the ResNet redefines this as $y = F(x) + x$ where the original input mapping is recast into the output of $F(x)$. The hypothesis is that it is easier to optimise the residual mapping function $F(x)$ than $H(x)$, in practice shortcut connections are the key design idea allowing the data signal to bypass one layer and move to the next layer in the sequence, this allows gradients to flow from later layers to the early layers. A basic ResNet block consists of two convolutional layers and each convolutional layer is followed by batch normalisation and a ReLU. A shortcut pass connects the top of the block to the layer just before the last ReLU in the block (Figure 8.4 displays a ResNet-34 architecture). Where the dimensions of the input and output do not match a 1×1 convolution is applied to reshape the output to match the input volume to allows for $F(x) + x$. In the proposed method a ResNet-18 is applied primarily due to the computational demands of the deep 3D network which means deeper network are to large to train on the available hardware.

8.3.4 Loss Function

Previous research using ResNet based 3D CNNs presented in Hara et al. (2018) applied only softmax loss, it has been identified that this may not provide the optimal operation for learning the most discriminative features. Instead joint supervised learning of both softmax loss and center loss is adopted. Previous literature indicates this will produce better features representation, as it has in the spatial domain CNNs and therefor it could do the same for learning spatio-temporal

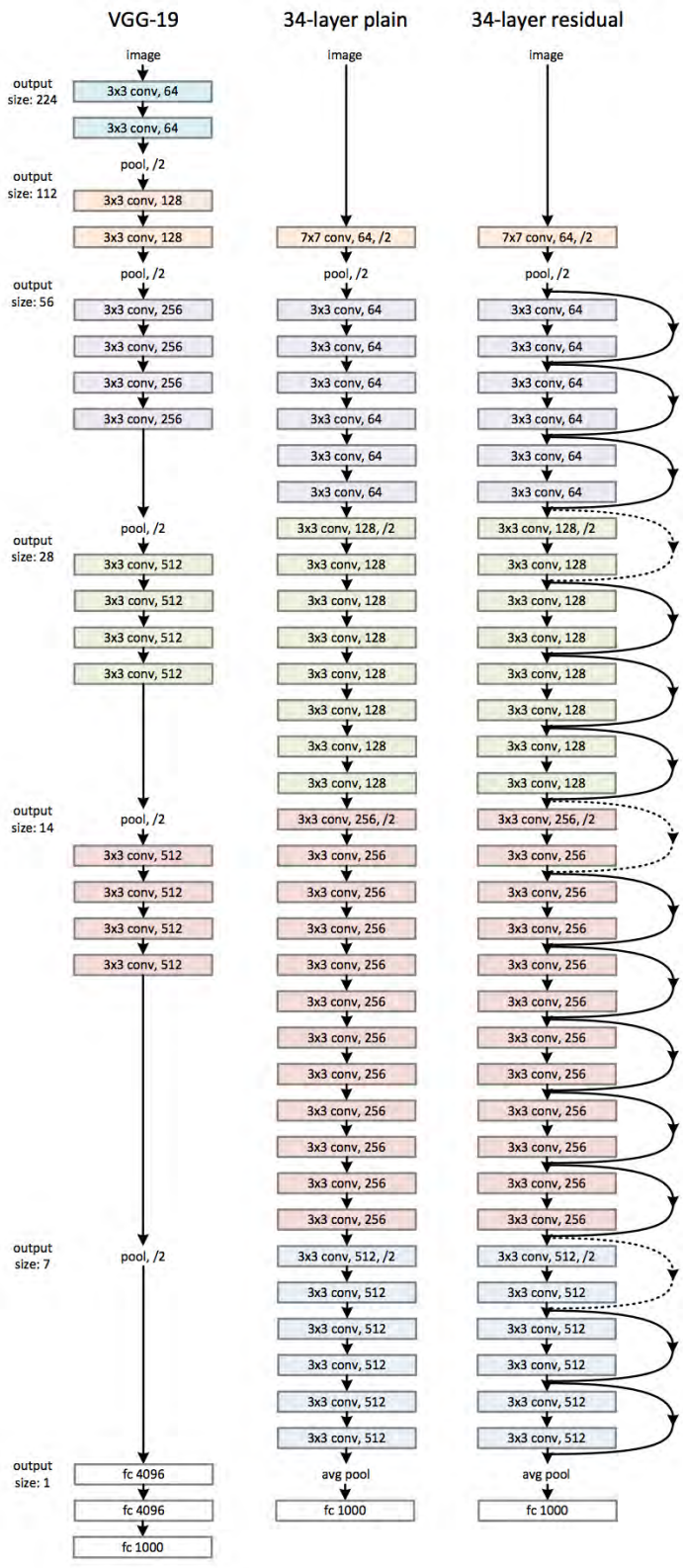

Figure 8.4: ResNet overview showing the shortcut connections in a ResNet-34 architecture in comparison to other network architectures (He et al., 2015).

features. This decision is validated in the evaluation section of this chapter.

It has been shown in other facial analysis tasks that using softmax loss only can result in large intra-class variations in the learned features (Wen et al., 2016). The adoption of center loss in addition aids inter-class dispensation and intra-class compactness as much as possible. Softmax loss is calculated as:

$$loss_{softmax} = \sum_{i=1}^{m} \log \frac{e^{W_{y_i}^T x_i + b_{y_i}}}{\sum_{j=1}^{n} e^{W_{y_j}^T x_i + b_j}} \tag{8.1}$$

where $x_i \in R^d$ is the i^{th} feature of the y_i^{th} class. The feature dimension is defined by d. $W_j \in R^d$ denotes the j^{th} column of the weights $W \in R^{d \times n}$ in the last fully connected layer and $b \in R^n$ is the bias term. Mini-batch size and the total number of class are defined as m and n, respectively. Center loss is defined as:

$$loss_{center} = \frac{1}{2} \sum_{i=1}^{m} \parallel x_i - c_{y_i} \parallel_2^2 \tag{8.2}$$

where $c_{y_i} \in R^d$ is the y_ith class centre of the learnt feature. The feature centres are updated after each mini-batch of training data. The total loss of the network is calculate as:

$$loss_{total} = loss_{softmax} + \lambda loss_{center} \tag{8.3}$$

where λ is used for balancing the two loss functions. Center loss is significantly larger value and therefore requires scaling down, based upon the experimentation in Wen et al. (2016) a value of $\lambda = 0.001$ is used within the proposed 3D CNN.

8.3.5 3D CNN Model Training Protocol

Both of the proposed 3D CNN models are trained with the following protocols for their specific task. Initially ResNet18 model is pre-trained on the Kinetics data set for the task of action recognition Hara et al. (2018). Transfer learning is then used to train the models for their respective facial analysis task, specifically the initial layers weight parameters are frozen, with only the last convolutional layers parameters and the fully connected layers trained. These layers are trained using a hybrid data set by combining samples from the CK+ emotion and a facial palsy data set with relevant labels for the associated task (Section 8.4 details the breakdown of the data set and the associated class labelling). To address the class imbalance within the data set weighted sampling is employed. Weighted sampling is a technique applied in training which balances the class distribution of class labels for each mini-batch, this ensures the model does not overfit on data of the majority class in the data set. Prior to the training process the video sequences in the training set are first passed through the face detection stage of the framework and the faces are extracted. The extracted face sequences are then re-sized spatially to 112×112 pixels and temporally to n total frames. In this work different values for n is considered. When a sequence is less than n frames, duplicate frames are interpolated into the sequence while those greater than n have frames removed at equally spaced intervals. Data augmentation techniques are applied to increase the total samples. To help avoid over-fitting random flipping, rotation, colour jitter and with 50% probability are employed. Two stochastic gradient descent optimisers are then applied to train the network. The training parameters include a learning rate of 0.1, with a weight decay of 0.001 and 0.9 for momentum. Each model was trained for 50 epochs which was sufficient to minimise model loss.

8.4 Evaluation

This section covers the results and discussion regarding the experimental evaluations conducted on the proposed 3D CNN methods. The evaluations consider classification accuracy on the primary tasks of facial palsy grading and mouth motion recognition. Further to this the key model parameters of frame duration is evaluated and the a final study is presented looking at the affect the choice of loss function has on the accuracy. All experiments are conducted using PyTorch 0.4 on Windows 10 with a Nvidia GTX 1080 GPU.

For the evaluation of the proposed method data the CK+ Lucey et al. (2010) and a Facial Palsy data set were used. The CK+ database consists of 593 sequences generated from 113 subjects, while the facial palsy data set consists of 696 different sequences with 17 subjects collected from online sources. Since the CK+ sequences range from a neutral face and end at the full expression, they are aligned with the facial palsy dataset by adding reversed frames so that the last frame is also a neutral expression. While all samples from the CK+ data set are posed, the facial palsy set contains both posed motions and also general motions such as talking. In the case of mouth motion recognition each sequence is labelled into as follows: no motion, smile, mouth open and other mouth motions. For facial palsy grading the labelling follows the House-Brackmann (Table 7.1) commonly applied by medical professionals.

To test the models accuracy for the two classification tasks a Leave-One-Subject-Out (LOSO) protocol is adopted. This allows for testing on unseen faces reducing any potential overfitting to previously seen faces. In practice models to test all subjects in the data set are not trained. 10 subjects have been used for the evaluation process; they are split equally into 5 having facial palsy (Subjects 1 to 5) and 5 who do not have facial palsy (Subjects 6 to 10). The 10 selected subjects cover the total range of labels for both tasks. Therefore, in total there are 397 samples used for the evaluation.

8.4.1 Mouth Motion Recognition

Figure 8.5 provides the overall results for the mouth motion recognition, where the findings show that the proposed model has a good predictive capability in this task producing an F1 Score of 82%. In the Figure 8.8 the results for each of the LOSO test sets are given, finding that all subjects perform reasonably well with F1 Scores close to 80% with the exception of subject 5. On inspection (Figure 8.6 show the confusion matrix for this sample) this subject and the sample which prove difficult to classify correctly are reflective of issue which reduce the accuracy for all subjects. This surrounds the overlap in motions that occur between those labelled as others and the rest. There are motions which are similar to smiles, due to the frame normalisation there is a possibility of losing the frames which can differentiate these motions. As this method also uses the global features to learn features it is possible that other motions such as those from the movement of the eyes and brows show overlap across classes therefore reducing accuracy.

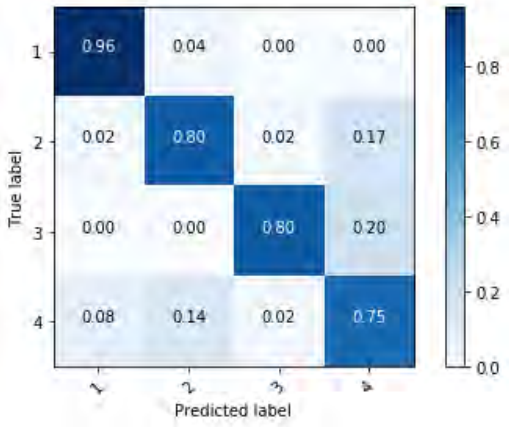

Figure 8.5: Overall Mouth Motion Confusion Matrix.

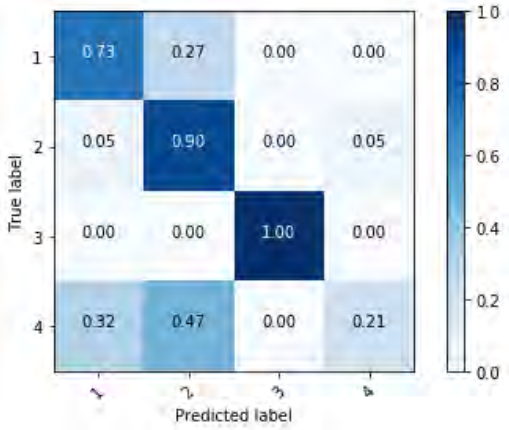

Figure 8.6: Subject 5 Mouth Motion Confusion Matrix.

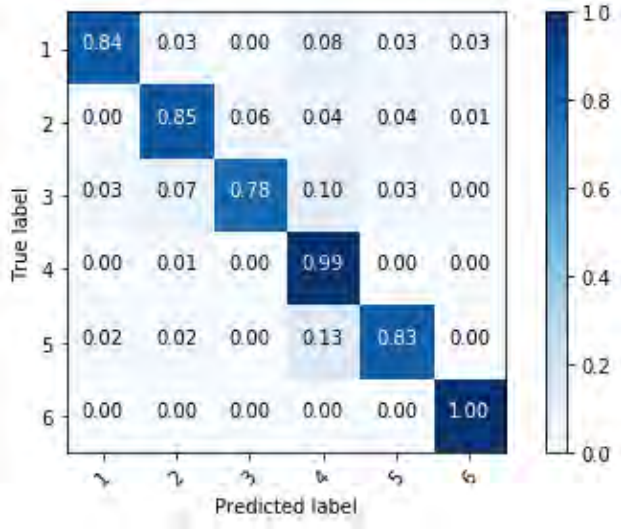

Figure 8.7: Overall Palsy Level Grading Results.

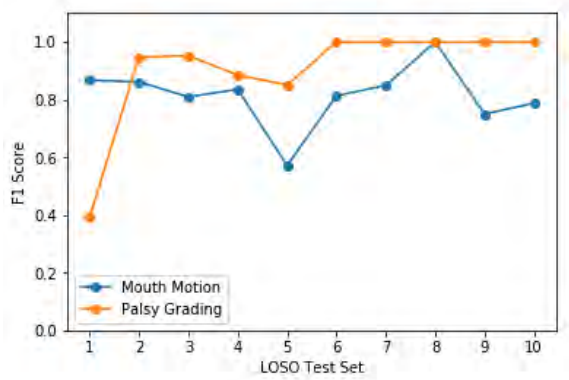

Figure 8.8: F1 Score by Subject Test Set.

Figure 8.9: Example of a sample sequence with incorrect palsy grading with the associated frame number below. This figure highlights the large range of facial motions present across the entire sequence when a subject is engaged in general conversation rather than a specific posed motion.

8.4.2 Facial Palsy Grading

The overall results for the facial palsy grading evaluation are shown in Figure 8.7. The proposed model provides a high level of accuracy with a F1 Score of 88%. In the Figure 8.8 the results for each of the LOSO test sets are shown, finding that all subjects from the CK+ data sets which all have a palsy grading label of 1 are all correctly classified. Subject 1 shows a very poor accuracy in comparison to all other subjects. Subject 1 has 29 samples, out of the 20 incorrect grading 16 are within 1$\pm$ grades. Subject 1 is a specifically difficult set of sequences as most of the facial expression are not posed but of the individual during normal conversation. Also the combined average frame duration is 23 frames per sequence meaning many frames are not used due to limitations on the frame duration. Highlighted in Figure 8.9 is an example of an incorrect palsy grading for subject 1 in which the ground truth palsy label was 3 and the prediction was 2, it can be seen there are a large and varied amount of motions across the 8 frames used from a total of 27 for this sequence, while general conversation does not always display then end range of motion which provides a more accurate understand of facial muscle motion for palsy grading.

Frame Duration	F1 Score	Training Time
8	86%	2h 10m
12	71%	3h 20m
16	79%	4h 25m

Table 8.1: Frame Duration Results

8.4.3 Ablation Study - Frame Duration

Frame duration is a potentially significant parameter when processing video sequences. Reducing the sequences to a short frame duration can remove important features while long frame duration's may add redundant information resulting in more computational overhead of the method. In action recognition work of Hara et al. (2018) a frame duration of 16 was found to work well. As the task of face motion are typically shorter in duration it is proposed to evaluate shorter frame duration's. In this experiment the performance effect on the frame duration is evaluated Table 8.1 illustrates the F1 scores achieved for each duration over the test sets for frame duration of 8, 12 and 16. From the results it can be seen that a frame duration of 8 seems to give the best performances. It is to be noted that there are samples which are correctly classified in the larger frame duration but incorrectly graded in the 8 frame duration. This is due to the lack of uniformity across motion duration in these tasks. Not only does this parameter have an affect on the accuracy presented by the model, it also has a large effect on the computational overhead of the framework. This can be seen in Table 8.1 where the use of an additional 4 frames adds about 1 hour to the time the model took to train for 50 epochs of the data set.

8.4.4 Ablation Study - Loss Function

The authors in Wen et al. (2016) have demonstrated the advantage of using both center Loss and softmax loss in joint supervision method to aid model the training process by finding a more discriminative feature representation rather than using using softmax loss alone. Primarily these findings were obtained for image recognition task using 2D CNNs. In this chapter this experiment is revisited for 3D CNN and specifically modified it and adapted it for the proposed framework. This study only used Subjects 1 to 5 and the results obtained are shown in Figure 8.10. For the 366 samples in the facial palsy test it was found that F1 scores of 86% and 82% for center and softmax loss and softmax loss alone, respectively. This has resulted in a small improvement of the performances as might be expected in image recognition problems. On the other hand, the results

obtained for mouth motion recognition have shown to more difficult to improve as demonstrated by a significant decrease of F1 score going from 82% to 49% when also applying center loss.

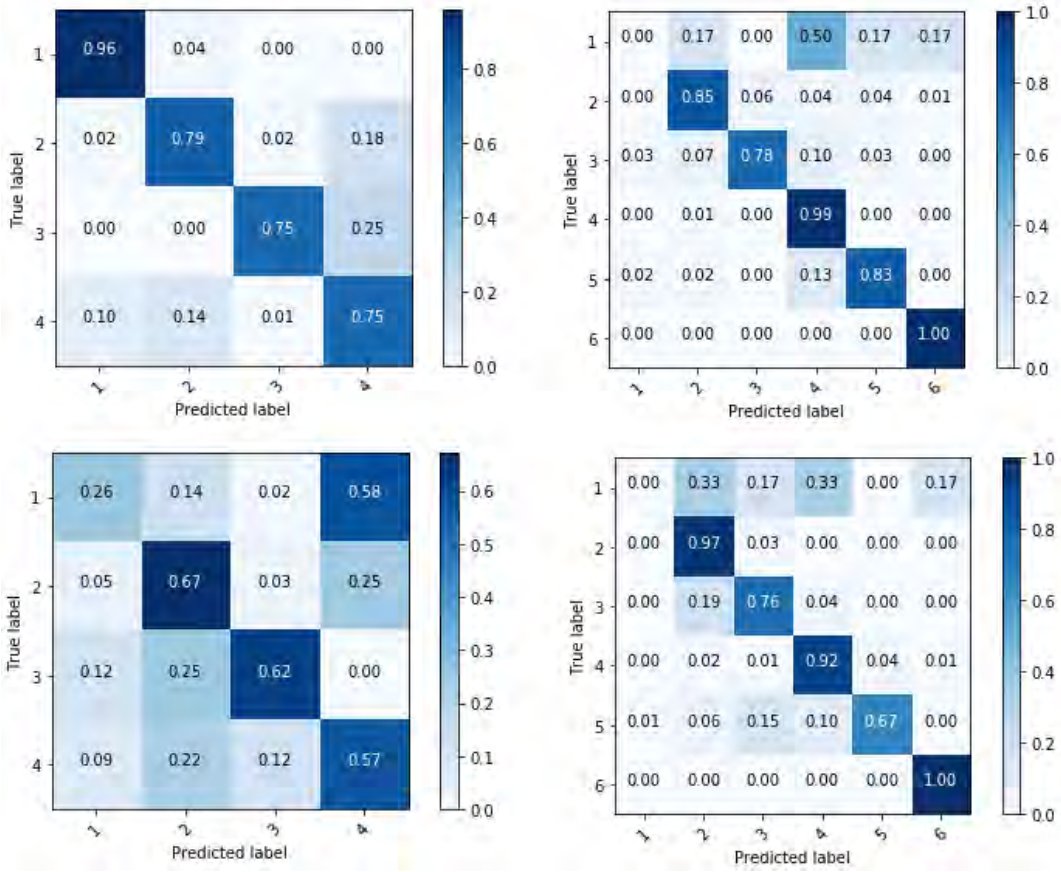

Figure 8.10: Loss Function Evaluation: (Top) - Center and Softmax Loss, (Bottom) - Softmax Loss

8.5 Conclusion

In this chapter further research was conducted on the facial analysis task of facial palsy grading. Unlike the previous approach which used geometric-based symmetry features as fully 3D CNN was applied to learn the spatio-temporal features from video sequences. Added to this was the evaluation of a second task, this was mouth motion prediction with the aim of classifying for example a smile in both facial palsy patients and those without facial palsy. The purpose of investigating this second task was based upon the evaluation process of facial palsy patients as used in clinical settings, where the clinician directs the patient to perform specific facial actions. Any automated systems would also require the capability to determine if the patient has performed the correct motion. The mouth motion task presented here is an initial approach to this. Encouraging

results were found for both tasks with an F1 score of 82% and 88% for the mouth motion and facial palsy grading respectively.

The fully 3D CNN using a deep backbone such as ResNet has little previous application outside of general action recognition Hara et al. (2018). The training and evaluation of the methods proposed in this chapter have indicated that transfer learning for facial analysis tasks that require features to be extracted from video sequences is a feasible with this approach. This opens the potential to adopted the 3D CNN for other dynamic facial analysis tasks. This chapter also establishes that center loss is a worthwhile addition to the learning of spatio-temporal features, as it provides an increase in the accuracy of both tasks. While the study of frame duration provides a less obvious outcome with regard to an optimal parameter. When considering the nature of the video sequences applied within the data sets used the length of the specific motion differs significantly, when normalisation is used to adopt a uniform duration, loss of data in likely as temporal differences are discarded. This leads to there not being a optimal parameter for all sequences in the approach taken in this chapter. Larger frame duration do add a significant computational load where using 16 instead of 8 frames doubles the training time.

While the results are promising there are many areas in which this research can be taken forward. Firstly there is the potential to investigate more complex backbone and deeper networks, which in this work is limited due to the computational overheads of 3D CNNs. As shown in the frame duration experiment this parameter plays a part in the model accuracy but there is not universal best parameters when sequences length can vary within the data set, there is potential to look at other pre-processing rather than simple frame duplication or reduction. Specifically for the task presented in this work a larger set of labelling for mouth motion is required to better separate similar facial motions.

Chapter 9

Conclusion

9.1 Introduction

The aim of this research was to investigate the use of deep learning techniques within facial analysis system. Based upon the information gathered during the literature review process two areas of specific research area were identified. The first area was that of face detection, this task is fundamental to a majority of facial analysis systems and a challenge to increase the precision of the current methods was identified and researched. The second area revolved around the specific facial analysis task in which asymmetrical faces and more specifically the medical condition of facial palsy grading was investigated. While some initial researched had been conducted in this area there was a significant gap in the research from which the application of deep learning to this task could contribute to the base of knowledge. The literature research also highlighted a number of possible techniques that had been applied across the computer vision domain that provided inspiration for the methods proposed across the research presented in this thesis.

9.2 Contributions

Based upon the research conducted throughout this thesis in both face detection and asymmetrical face analysis and more specifically the medical condition of facial palsy grading a number of significant contributions were made as follows:

1. In chapter 5 a unified Faster R-CNN architecture was proposed for the tasks of both face

detection and also binary facial symmetry analysis. This involved modifying the original Faster R-CNN architecture to integrate multi-tasking learning.

2. Furthermore in chapter 5 a training protocol was proposed which applied synthesised data to generate training samples and allow training of the multi-task network simultaneously for both tasks. Evaluation of this method in the face detection task were positive at 90% and 91% AP.

3. In chapter 6 a novel method for face detection and landmark localisation called the IDM was proposed. This method leveraged the Faster R-CNN network specifically trained for the task of face detection and re-purposed the local facial landmark localisation features learnt in the FAN (Bulat and Tzimiropoulos, 2017*b*) to support the face detection task in a novel cascaded network architecture.

4. A benchmark was conducted in chapter 6 for the proposed IDM against other state-of-the-art face detection methods. It was found that the IDM method reduced the number of false positives by over 50%. This is a significant increase and highlights how the leveraging of features from different CNN architectures using a cascaded approach can be a highly beneficial to increasing the precision.

5. Chapter 7 presented a comparative study of facial landmark accuracy methods when applied to asymmetrical faces, showing that only the FAN (Bulat and Tzimiropoulos, 2017*b*) method generalised well to this population. Specifically it vastly outperformed the other evaluated methods in fitting the difficult mouth landmarks.

6. A geometry-based feature extraction method for facial palsy analysis was also proposed within chapter 7. This method extracted quantitative values based upon the motion difference of paired facial landmarks and mapped these to a House-Brackmann facial palsy grading for evaluation purposes.

7. In chapter 8 a fully 3D CNN with a ResNet backbone was proposed for the tasks of facial palsy grading. The proposed 3D CNN performs better than geometry-based symmetry features with an F1 score of 88% to 40% for the facial palsy grading task. This level of accuracy was achieved with a significantly larger sample size of 397 to 25.

8. Chapter 8 also applied the fully 3D CNN for the novel task of mouth motion recognition. This novel task was designed around detecting a smile, for instance, in both those with facial palsy and those without. The proposed method achieved an F1 score of 82% with potential to be higher with more nuanced data set labelling.

9.3 Research Limitations

There are some limitations regarding the research presented in this thesis. Time and computational resource were a source of some of these limitations, as this restricted the CNN architecture that could be investigated. The deeper the CNN model the more trainable parameters exist and therefore memory and computation increase. A specific example is that of the 3D CNN applied in chapter 8 where only ResNet-18 was applied due to hardware limitations, while frame duration also affected the computational load and therefore 16 was the largest before the memory on the GPU full. It is possible that deep networks could have produced better levels of accuracy in this task as shown in Hara et al. (2018), where deeper architectures performed best on general action recognition.

The data sets applied to evaluate the facial palsy tasks also mean the results can guide the community rather than be verified as state-of-the-art. Unlike the contribution to face detection where large scale and well established benchmark data sets exist which allows for evaluation against previous methods, as a new field of study facial palsy does not have such data sets. While there exists public images and video of individuals with the condition most of the previous work is conducted with private unreleased data sets. Therefore the results can only be analysed in isolation.

9.4 Future Research

The research presented within this thesis includes both a novel method for face detection which improves the precision with minimal impact on recall and also investigates potential methods for the tasks of facial palsy and facial motion analysis. However there are still many challenges which exist and questions that form the basis for future research these include:

1. While face detection false positives were significantly reduced through the novel IDM method, there is still scope to for further reduction. A more complex integration of the

stacked hourglass and Faster R-CNN architectures rather than the cascaded approach applied in the IDM method. One potential integration would be the fusion of the feature extraction layers of both the Faster R-CNN and stacked hourglass allowing direct feature map sharing between architectures which would enable end-to-end training of all feature layers.

2. As discussed in chapter 6 the IDM also has a slight reduction in recall, particularly in the area of blurry faces and badly positioned detection's for occluded faces. Further data augmentation techniques could be applied for the blurry face issue or further research in deblurring techniques such as through the application of Generative Adversarial Network.

3. This thesis established a basis for using both geometry-based features and also those learnt using a 3D CNN for facial palsy grading. A joint framework which leverages both geometry based features and 3D CNN could be leveraged to boost prediction accuracy. This could be achieved through data fusion of the frame data and geometry features.

4. To realise an automated facial palsy system the capability to extend mouth motion recognition to other facial motions is a necessity for example to recognise eye brow raising, eye closing. This leads to a potential issue in that requiring a model for each of these tasks is computationally expensive. Future work would concentrate on how multi-task learning framework can be developed so that a single 3D CNN can be leveraged for all tasks. Not only does this have the the potential to have a significant reduce on the computational time, but also boost classification accuracy by learning from inter-related training samples.

5. It was shown in chapter 8 that applying specific frames duration is not the best strategy to maximise accuracy when performing video analysis tasks. A more flexible approach which has the capacity to smartly reduce/increase the frames to a uniform size could be a possible solution to this. Previous research has shown 1×1 convolutional layers have been regularly applied for learnt dimension reduction within the ResNet architectures. As frame duration can be posed as a dimension reduction/expansion issue, application of 1×1 convolutional layers as pre-processing method is one possible approach to the frame duration issue.

Acronyms

300-VW	300 Videos In-the-Wild
300-W	300 Faces In-the-Wild
3DMM	3D Morphable Model
AAM	Active Appearance Model
AFLW	Annotated Facial Landmarks in the Wild
AFW	Annotated Faces in-the-Wild
AP	Average Precision
ASM	Active Shape Model
CK+	Extended Cohn-Kanade
CLM	Constrained Local Model
CNN	Convolutional Neural Network
CNTK	Microsoft Cognitive Toolkit
COCO	Common Objects in Context
COFW	Caltech Occluded Faces in the Wild
DPM	Deformable Part Model
DRMF	Discriminative Response Map Fitting
FAN	Face Alignment Network
FDDB	Face Detection Dataset and Benchmark
FER	Facial Expression Recognition

GPU	Graphics Processing Unit
HIGO	Histogram of Image Gradient Orientation
HOG	Histograms of Oriented Gradients
HOOF	Histogram of Optical Flow
IDM	Integrated Deep Model
ILSVRC	ImageNet Large-Scale Visual Recognition Challenge
IoU	Intersection over Union
k-NN	K-Nearest Neighbour
LBP	Local Binary Patterns
LBP-TOP	Local Binary Patterns on Three Orthogonal Planes
LDA	Linear Discriminant Analysis
LOSO	Leave-One-Subject-Out
LTSM	Long Short-Term Memory
MDMO	Main Directional Mean Optical-Flow
MSE	Mean Square Error
NME	Normalised Mean Error
PCA	Principal Component Analysis
PD	Parkinson's Disease
PIFA	Pose-Invariant Face Alignment
R-CNN	Region-CNN
ReLU	Rectified Linear Unit

ResNet	Residual Network
RNN	Recurrent Neural Network
ROI	Region Of Interest
RPN	Region Proposal Network
SIFT	Scale-Invariant Feature Transform
SSD	Single Shot Detection
SURF	Speeded Up Robust Features
SVM	Support Vector Machine
TSM	Tree Shape Model
VGG	Visual Geometry Group
VOC	Visual Object Classes
YOLO	You Only Look Once

References

Asthana, A., Zafeiriou, S., Cheng, S. and Pantic, M. (2013), Robust discriminative response map fitting with constrained local models, *in* 'Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition', pp. 3444–3451.

Bandini, A., Green, J., Richburg, B. and Yunusova, Y. (2018), Automatic Detection of Orofacial Impairment in Stroke, *in* 'Interspeech 2018', ISCA, ISCA, pp. 1711–1715.

Banks, C. A., Bhama, P. K., Park, J., Hadlock, C. R. and Hadlock, T. A. (2015), 'Clinician-Graded Electronic Facial Paralysis Assessment', *Plastic and Reconstructive Surgery* **136**(2), 223e–230e.

Bansal, A., Nanduri, A., Castillo, C. D., Ranjan, R. and Chellappa, R. (2017), UMDFaces: An annotated face dataset for training deep networks, *in* '2017 IEEE International Joint Conference on Biometrics (IJCB)', IEEE, pp. 464–473.

Bargal, S. A., Barsoum, E., Ferrer, C. C. and Zhang, C. (2016), 'Emotion Recognition in the Wild from Videos using Images', *Proceedings of the 18th ACM International Conference on Multimodal Interaction - ICMI 2016* pp. 433–436.

Baron-Cohen, S., Golan, O. and Ashwin, E. (2009), 'Can emotion recognition be taught to children with autism spectrum conditions?', *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* **364**(1535), 3567–74.

Belhumeur, P., Hespanha, J. and Kriegman, D. (1997), 'Eigenfaces vs. Fisherfaces: recognition using class specific linear projection', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **19**(7), 711–720.

Benitez-Quiroz, C. F., Srinivasan, R. and Martinez, A. M. (2016), EmotioNet: An Accurate, Real-Time Algorithm for the Automatic Annotation of a Million Facial Expressions in the Wild, *in* '2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)', IEEE, pp. 5562–5570.

Blanz, V. and Vetter, T. (1999), A morphable model for the synthesis of 3D faces, *in* 'Proceedings of the 26th annual conference on Computer graphics and interactive techniques - SIGGRAPH '99', ACM Press, New York, New York, USA, pp. 187–194.

Blanz, V. and Vetter, T. (2003), 'Face recognition based on fitting a 3D morphable model', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **25**(9), 1063–1074.

Bulat, A. and Tzimiropoulos, G. (2017*a*), 'Binarized Convolutional Landmark Localizers for Human Pose Estimation and Face Alignment with Limited Resources', *2017 IEEE International Conference on Computer Vision (ICCV)* pp. 3726–3734.

Bulat, A. and Tzimiropoulos, G. (2017*b*), 'How far are we from solving the 2D & 3D Face Alignment problem? (and a dataset of 230,000 3D facial landmarks)', *2017 IEEE International Conference on Computer Vision (ICCV)* pp. 1021–1030.

Bulat, A. and Tzimiropoulos, Y. (2016), Convolutional aggregation of local evidence for large pose face alignment, *in* 'Procedings of the British Machine Vision Conference 2016', British Machine Vision Association, pp. 1–86.

Burgos-Artizzu, X. P., Perona, P. and Dollar, P. (2013), Robust Face Landmark Estimation under Occlusion, *in* '2013 IEEE International Conference on Computer Vision', IEEE, pp. 1513–1520.

Cai, J., Meng, Z., Khan, A. S., Li, Z., OReilly, J. and Tong, Y. (2018), Island Loss for Learning Discriminative Features in Facial Expression Recognition, *in* '2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)', IEEE, pp. 302–309.

Cao, C., Hou, Q. and Zhou, K. (2014), 'Displaced dynamic expression regression for real-time facial tracking and animation', *ACM Transactions on Graphics* **33**(4), 1–10.

Cao, Q., Shen, L., Xie, W., Parkhi, O. M. and Zisserman, A. (2018), VGGFace2: A Dataset

for Recognising Faces across Pose and Age, *in* '2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)', IEEE, pp. 67–74.

Cao, X., Wei, Y., Wen, F. and Sun, J. (2014), 'Face alignment by explicit shape regression', *International Journal of Computer Vision* **107**(2), 177–190.

Carreira, J. and Sminchisescu, C. (2010), Constrained parametric min-cuts for automatic object segmentation, *in* '2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition', IEEE, pp. 3241–3248.

Chang Huang, Haizhou Ai, Yuan Li and Shihong Lao (2005), Vector boosting for rotation invariant multi-view face detection, *in* 'Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1', IEEE, pp. 446–453.

Chang, Y.-S., Choi, J. E., Kim, S. W., Baek, S.-Y. and Cho, Y.-S. (2016), 'Prevalence and associated factors of facial palsy and lifestyle characteristics: data from the Korean National Health and Nutrition Examination Survey 2010-2012.', *BMJ open* **6**(11), e012628.

Chen Cao, Yanlin Weng, Shun Zhou, Yiying Tong and Kun Zhou (2014), 'FaceWarehouse: A 3D Facial Expression Database for Visual Computing', *IEEE Transactions on Visualization and Computer Graphics* **20**(3), 413–425.

Chin, C.-L., Lin, B.-J., Wu, G.-R., Weng, T.-C., Yang, C.-S., Su, R.-C. and Pan, Y.-J. (2017), An automated early ischemic stroke detection system using CNN deep learning algorithm, *in* '2017 IEEE 8th International Conference on Awareness Science and Technology (iCAST)', IEEE, pp. 368–372.

Chopra, S., Hadsell, R. and LeCun, Y. (2005), Learning a Similarity Metric Discriminatively, with Application to Face Verification, *in* '2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)', Vol. 1, IEEE, pp. 539–546.

Cohen, I., Sebe, N., Garg, A., Chen, L. S. and Huang, T. S. (2003), 'Facial expression recognition from video sequences: temporal and static modeling', *Computer Vision and Image Understanding* **91**(1-2), 160–187.

Cootes, T., Baldock, E. and Graham, J. (2000), 'An introduction to active shape models', *Image Processing and Analysis* pp. 223–248.

Cootes, T., Edwards, G. and Taylor, C. (2001), 'Active appearance models', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **23**(6), 681–685.

Corrow, S. L., Dalrymple, K. A. and Barton, J. J. (2016), 'Prosopagnosia: current perspectives.', *Eye and brain* **8**, 165–175.

Cristinacce, D. and Cootes, T. (2008), 'Automatic feature localisation with constrained local models', *Pattern Recognition* **41**(10), 3054–3067.

Darwin, C. and Darwin, F. (2009), *The expression of the emotions in man and animals*.

Dhall, A., Goecke, R., Ghosh, S., Joshi, J., Hoey, J. and Gedeon, T. (2017), From individual to group-level emotion recognition: EmotiW 5.0, *in* 'Proceedings of the 19th ACM International Conference on Multimodal Interaction - ICMI 2017', ACM Press, New York, New York, USA, pp. 524–528.

Dhall, A., Ramana Murthy, O., Goecke, R., Joshi, J. and Gedeon, T. (2015), Video and Image based Emotion Recognition Challenges in the Wild, *in* 'Proceedings of the 2015 ACM on International Conference on Multimodal Interaction - ICMI '15', ACM Press, New York, New York, USA, pp. 423–426.

Ding, H., Zhou, S. K. and Chellappa, R. (2017), FaceNet2ExpNet: Regularizing a Deep Face Recognition Net for Expression Recognition, *in* '2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)', IEEE, pp. 118–126.

Ding, W., Xu, M., Huang, D., Lin, W., Dong, M., Yu, X. and Li, H. (2016), Audio and face video emotion recognition in the wild using deep neural networks and small datasets, *in* 'Proceedings of the 18th ACM International Conference on Multimodal Interaction - ICMI 2016', ACM Press, New York, New York, USA, pp. 506–513.

Dollár, P., Appel, R., Belongie, S. and Perona, P. (2014), 'Fast Feature Pyramids for Object Detection.', *IEEE transactions on pattern analysis and machine intelligence* **36**(8), 1532–45.

Dollár, P., Welinder, P. and Perona, P. (2010), Cascaded pose regression, *in* 'Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition', pp. 1078–1085.

Ekman, P. (2013), *Emotion in the Human Face*, 2nd edn, Malor, USA.

Endres, I. and Hoiem, D. (2014), 'Category-Independent Object Proposals with Diverse Ranking', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **36**(2), 222–234.

Everingham, M., Van Gool, L., Williams, C. K. I., Winn, J. and Zisserman, A. (2010), 'The Pascal Visual Object Classes (VOC) Challenge', *International Journal of Computer Vision* **88**(2), 303–338.

Fan, Y., Lu, X., Li, D. and Liu, Y. (2016), Video-based emotion recognition using CNN-RNN and C3D hybrid networks, *in* 'Proceedings of the 18th ACM International Conference on Multimodal Interaction - ICMI 2016', ACM Press, New York, New York, USA, pp. 445–450.

Farfade, S. S., Saberian, M. and Li, L. J. (2015), 'Multi-view Face Detection Using Deep Convolutional Neural Networks', *International Conference on Multimedia Retrieval 2015 (ICMR)* p. 19.

Fattah, A. Y., Gurusinghe, A. D. R., Gavilan, J., Hadlock, T. A., Marcus, J. R., Marres, H., Nduka, C. C., Slattery, W. H., Snyder-Warwick, A. K. and Sir Charles Bell Society (2015), 'Facial Nerve Grading Instruments', *Plastic and Reconstructive Surgery* **135**(2), 569–579.

Feichtenhofer, C., Pinz, A. and Wildes, R. P. (2016), Spatiotemporal Residual Networks for Video Action Recognition, *in* 'Proceedings of the 30th International Conference on Neural Information Processing Systems.', pp. 3476–3484.

Feichtenhofer, C., Pinz, A. and Wildes, R. P. (2017), Spatiotemporal Multiplier Networks for Video Action Recognition, *in* '2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)', IEEE, pp. 7445–7454.

Feichtenhofer, C., Pinz, A. and Zisserman, A. (2016), Convolutional Two-Stream Network Fusion for Video Action Recognition, *in* '2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)', IEEE, pp. 1933–1941.

Felzenszwalb, P. F., Girshick, R. B., McAllester, D. and Ramanan, D. (2009), 'Object Detection with Discriminatively Trained Part Based Models', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **32**(9), 1627–1645.

Felzenszwalb, P. F. and Huttenlocher, D. P. (2004), 'Efficient Graph-Based Image Segmentation', *International Journal of Computer Vision* **59**(2), 167–181.

Ghiasi Charless Fowlkes, G. C., Ghiasi, G. and Fowlkes, C. C. (2014), 'Occlusion Coherence: Localizing Occluded Faces with a Hierarchical Deformable Part Model'.
URL: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6909641http://openaccess.thecvf.com/content_cvpr_2014/papers/Ghiasi_Occlusion_Coherence_Localizing_2014_CVPR_paper.pdf

Ghimire, D. and Lee, J. (2013), 'Geometric feature-based facial expression recognition in image sequences using multi-class AdaBoost and support vector machines.', *Sensors* **13**(6), 7714–7734.

Ghimire, D., Lee, J., Li, Z.-N., Jeong, S., Park, S. H. and Choi, H. S. (2015), 'Recognition of Facial Expressions Based on Tracking and Selection of Discriminative Geometric Features', *International Journal of Multimedia and Ubiquitous Engineering* **10**(3), 35–44.

Girshick, R. (2015), Fast R-CNN, *in* '2015 IEEE International Conference on Computer Vision (ICCV)', IEEE, pp. 1440–1448.

Girshick, R., Donahue, J., Darrell, T. and Malik, J. (2014), Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation, *in* '2014 IEEE Conference on Computer Vision and Pattern Recognition', IEEE, pp. 580–587.

Goodfellow, I. J., Erhan, D., Carrier, P. L., Courville, A., Mirza, M., Hamner, B., Cukierski, W., Tang, Y., Thaler, D., Lee, D.-H., Zhou, Y., Ramaiah, C., Feng, F., Li, R., Wang, X., Athanasakis, D., Shawe-Taylor, J., Milakov, M., Park, J., Ionescu, R., Popescu, M., Grozea, C., Bergstra, J., Xie, J., Romaszko, L., Xu, B., Chuang, Z. and Bengio, Y. (2013), Challenges in Representation Learning: A Report on Three Machine Learning Contests, Springer, Berlin, Heidelberg, pp. 117–124.

Grammer, K. and Thornhill, R. (1994), 'Human (Homo sapiens) facial attractiveness and sexual selection: the role of symmetry and averageness.', *Journal of comparative psychology (Washington, D.C. : 1983)* **108**(3), 233–242.

Gross, R., Matthews, I., Cohn, J., Kanade, T. and Baker, S. (2008), Multi-PIE, *in* '2008 8th IEEE International Conference on Automatic Face & Gesture Recognition', IEEE, pp. 1–8.

Guerreschi, P., Gabert, P.-E., Labbé, D. and Martinot-Duquennoy, V. (2016), 'Paralysie faciale chez l'enfant', *Annales de Chirurgie Plastique Esthétique* **61**(5), 513–518.

Guillaumin, M., Verbeek, J. and Schmid, C. (2009), Is that you? Metric learning approaches for face identification, *in* '2009 IEEE 12th International Conference on Computer Vision', IEEE, pp. 498–505.

Guo, Y., Zhang, L., Hu, Y., He, X. and Gao, J. (2016), 'MS-Celeb-1M: A Dataset and Benchmark for Large-Scale Face Recognition'.

Hamester, D., Barros, P. and Wermter, S. (2015), Face expression recognition with a 2-channel Convolutional Neural Network, *in* '2015 International Joint Conference on Neural Networks (IJCNN)', IEEE, pp. 1–8.

Haofu Liao, Yuncheng Li and Jiebo Luo (2016), Skin disease classification versus skin lesion characterization: Achieving robust diagnosis using multi-label deep neural networks, *in* '2016 23rd International Conference on Pattern Recognition (ICPR)', IEEE, pp. 355–360.

Hara, K., Kataoka, H. and Satoh, Y. (2017), Learning Spatio-Temporal Features with 3D Residual Networks for Action Recognition, *in* '2017 IEEE International Conference on Computer Vision Workshops (ICCVW)', IEEE, pp. 3154–3160.

Hara, K., Kataoka, H. and Satoh, Y. (2018), Can Spatiotemporal 3D CNNs Retrace the History of 2D CNNs and ImageNet?, *in* '2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition', IEEE, pp. 6546–6555.

Hasnat, A., Bohne, J., Milgram, J., Gentric, S. and Chen, L. (2017), DeepVisage: Making Face Recognition Simple Yet With Powerful Generalization Skills, *in* '2017 IEEE International Conference on Computer Vision Workshops (ICCVW)', IEEE, pp. 1682–1691.

He, K., Zhang, X., Ren, S. and Sun, J. (2015), 'Deep Residual Learning for Image Recognition'.

He, K., Zhang, X., Ren, S. and Sun, J. (2016), Deep Residual Learning for Image Recognition, *in*

‘2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)’, IEEE, pp. 770–778.

Hoiem, D., Efros, A. A. and Hebert, M. (2011), ‘Recovering Occlusion Boundaries from an Image’, *International Journal of Computer Vision* **91**(3), 328–346.

Hu, P., Cai, D., Wang, S., Yao, A. and Chen, Y. (2017), Learning supervised scoring ensemble for emotion recognition in the wild, *in* ‘Proceedings of the 19th ACM International Conference on Multimodal Interaction - ICMI 2017’, ACM Press, New York, New York, USA, pp. 553–560.

Identifying the research priorities for facial palsy - Facial Palsy UK (n.d.).
URL: https://www.facialpalsy.org.uk/research/identifying-the-research-priorities-for-facial-palsy/

Ishii, L. E. (2016), ‘Facial Nerve Rehabilitation’, *Facial Plastic Surgery Clinics of North America* **24**(4), 573–575.

Jain, V., Jain, V. and Learned-miller, E. (2010), ‘FDDB: A benchmark for face detection in unconstrained settings’, p. 13.

James, W. (1913), *The principles of psychology.*, Henry Holt and Company.

Jeni, L. A., Cohn, J. F. and Kanade, T. (2015), Dense 3D face alignment from 2D videos in real-time, *in* ‘2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)’, Vol. 1, IEEE, pp. 1–8.

Jianguo Li, Tao Wang and Yimin Zhang (2011), Face detection using SURF cascade, *in* ‘2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)’, IEEE, pp. 2183–2190.

Jones, M. and Viola, P. (2003), ‘Fast Multi-view Face Detection’, *Mitsubishi Electric Research Lab TR2000396* .

Jonsson, K., Kittler, J., Li, Y. and Matas, J. (2002), ‘Support vector machines for face authentication’, *Image and Vision Computing* **20**(5-6), 369–375.

Jourabloo, A. and Liu, X. (2016), Large-Pose Face Alignment via CNN-Based Dense 3D Model Fitting, *in* '2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)', IEEE, pp. 4188–4196.

Jourabloo, A. and Liu, X. (2017), 'Pose-Invariant Face Alignment via CNN-Based Dense 3D Model Fitting', *International Journal of Computer Vision* **124**(2).

Jung, H., Lee, S., Yim, J., Park, S. and Kim, J. (2015), Joint Fine-Tuning in Deep Neural Networks for Facial Expression Recognition, *in* '2015 IEEE International Conference on Computer Vision (ICCV)', IEEE, pp. 2983–2991.

Kanade, T. (1974), 'Picture processing system by computer complex and recognition of human faces'.

Kanade, T., Tian, Y. and Cohn, J. F. (2000), 'Comprehensive Database for Facial Expression Analysis', p. 46.

Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R. and Fei-Fei, L. (2014), Large-Scale Video Classification with Convolutional Neural Networks, *in* '2014 IEEE Conference on Computer Vision and Pattern Recognition', IEEE, pp. 1725–1732.

Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., Suleyman, M. and Zisserman, A. (2017), 'The Kinetics Human Action Video Dataset'.

Kim, B.-K., Lee, H., Roh, J. and Lee, S.-Y. (2015), Hierarchical Committee of Deep CNNs with Exponentially-Weighted Decision Fusion for Static Facial Expression Recognition, *in* 'Proceedings of the 2015 ACM on International Conference on Multimodal Interaction - ICMI '15', ACM Press, New York, New York, USA, pp. 427–434.

Kim, D. H., Baddar, W. J. and Ro, Y. M. (2016), Micro-Expression Recognition with Expression-State Constrained Spatio-Temporal Feature Representations, *in* 'Proceedings of the 2016 ACM on Multimedia Conference - MM '16', ACM Press, New York, New York, USA, pp. 382–386.

Kim, D. H., Lee, M. K., Choi, D. Y. and Song, B. C. (2017), Multi-modal emotion recognition using semi-supervised learning and multiple neural networks in the wild, *in* 'Proceedings of the

19th ACM International Conference on Multimodal Interaction - ICMI 2017', ACM Press, New York, New York, USA, pp. 529–535.

Köstinger, M., Wohlhart, P., Roth, P. M. and Bischof, H. (2011), Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization, *in* 'Proceedings of the IEEE International Conference on Computer Vision', pp. 2144–2151.

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

Krizhevsky, A., Sutskever, I. and Hinton, G. E. (2012), 'ImageNet classification with deep convolutional neural networks'.
URL: https://dl.acm.org/citation.cfm?id=2999257

Kuehne, H., Jhuang, H., Garrote, E., Poggio, T. and Serre, T. (2011), HMDB: A large video database for human motion recognition, *in* '2011 International Conference on Computer Vision', IEEE, pp. 2556–2563.

Kumar Patnaik, S., Singh Sidhu, M., Gehlot, Y., Sharma, B. and Muthu, P. (2018), 'Automated Skin Disease Identification using Deep Learning Algorithm', *Biomedical and Pharmacology Journal* **11**(3), 1429–1436.

Lai, H., Xiao, S., Pan, Y., Cui, Z., Feng, J., Xu, C., Yin, J. and Yan, S. (2015), 'Deep Recurrent Regression for Facial Landmark Detection'.

Langner, O., Dotsch, R., Bijlstra, G., Wigboldus, D. H. J., Hawk, S. T. and van Knippenberg, A. (2010), 'Presentation and validation of the Radboud Faces Database', *Cognition & Emotion* **24**(8), 1377–1388.

Lees, A. J. (2002), 'Odd and unusual movement disorders.', *Journal of neurology, neurosurgery, and psychiatry* **72 Suppl 1**(suppl 1), I17–I21.

Li, M. H., Mestre, T. A., Fox, S. H. and Taati, B. (2018), 'Vision-based assessment of parkinsonism and levodopa-induced dyskinesia with pose estimation', *Journal of NeuroEngineering and Rehabilitation* **15**(1), 97.

Li, S. and Deng, W. (2018), 'Deep Facial Expression Recognition: A Survey'.

Li, S., Deng, W. and Du, J. (2017), Reliable Crowdsourcing and Deep Locality-Preserving Learning for Expression Recognition in the Wild, *in* '2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)', IEEE, pp. 2584–2593.

Li, X., HONG, X., Moilanen, A., Huang, X., Pfister, T., Zhao, G. and Pietikainen, M. (2015), 'Towards Reading Hidden Emotions: A Comparative Study of Spontaneous Micro-expression Spotting and Recognition Methods', *IEEE Transactions on Affective Computing* pp. 1–1.

Li, X., Pfister, T., Huang, X., Zhao, G. and Pietikainen, M. (2013), A Spontaneous Micro-expression Database: Inducement, collection and baseline, *in* '2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)', IEEE, pp. 1–6.

Liang, Z., Ding, S. and Lin, L. (2015), 'Unconstrained Facial Landmark Localization with Backbone-Branches Fully-Convolutional Networks', *arXiv:1507.03409 [cs]* **1**.

Lijun Yin, Xiaozhou Wei, Yi Sun, Jun Wang and Rosato, M. (2006), A 3D Facial Expression Database For Facial Behavior Research, *in* '7th International Conference on Automatic Face and Gesture Recognition (FGR06)', IEEE, pp. 211–216.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P. and Zitnick, C. L. (2014), Microsoft COCO: Common Objects in Context, Springer, Cham, pp. 740–755.

Lindsay, R. W., Robinson, M. and Hadlock, T. A. (2010), 'Comprehensive Facial Rehabilitation Improves Function in People With Facial Paralysis: A 5-Year Experience at the Massachusetts Eye and Ear Infirmary', *Physical Therapy* **90**(3), 391–397.

Liong, S.-T., See, J., Phan, R. C.-W., Le Ngo, A. C., Oh, Y.-H. and Wong, K. (2014), Subtle expression recognition using optical strain weighted features, *in* 'Computer Vision-ACCV 2014 Workshops', Springer, pp. 644–657.

Liu, M., Wang, R., Li, S., Shan, S., Huang, Z. and Chen, X. (2014), Combining Multiple Kernel Methods on Riemannian Manifold for Emotion Recognition in the Wild, *in* 'Proceedings of the 16th International Conference on Multimodal Interaction - ICMI '14', ACM Press, New York, New York, USA, pp. 494–501.

Liu, P., Han, S., Meng, Z. and Tong, Y. (2014), Facial Expression Recognition via a Boosted Deep Belief Network, *in* '2014 IEEE Conference on Computer Vision and Pattern Recognition', IEEE, pp. 1805–1812.

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y. and Berg, A. C. (2016), SSD: Single Shot MultiBox Detector, *in* 'ECCV 2016. Lecture Notes in Computer Science', Springer, Cham, pp. 21–37.

Liu, W., Wen, Y., Yu, Z., Li, M., Raj, B. and Song, L. (2017), SphereFace: Deep Hypersphere Embedding for Face Recognition, *in* '2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)', IEEE, pp. 6738–6746.

Liu, Y.-J., Zhang, J.-K., Yan, W.-J., Wang, S.-J., Zhao, G. and Fu, X. (2015), 'A Main Directional Mean Optical Flow Feature for Spontaneous Micro-Expression Recognition', *IEEE Transactions on Affective Computing* **3045**(c), 1–1.

Lucey, P., Cohn, J. F., Kanade, T., Saragih, J., Ambadar, Z. and Matthews, I. (2010), The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression, *in* '2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops', IEEE, pp. 94–101.

Lyons, M., Akamatsu, S., Kamachi, M. and Gyoba, J. (1998), Coding facial expressions with Gabor wavelets, *in* 'Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition', IEEE Comput. Soc, pp. 200–205.

Mathias, M., Benenson, R., Pedersoli, M. and Van Gool, L. (2014), Face Detection without Bells and Whistles, Springer, Cham, pp. 720–735.

Meng, Z., Liu, P., Cai, J., Han, S. and Tong, Y. (2017), Identity-Aware Convolutional Neural Network for Facial Expression Recognition, *in* '2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)', IEEE, pp. 558–565.

Mengyi Liu, Shaoxin Li, Shiguang Shan and Xilin Chen (2013), AU-aware Deep Networks for facial expression recognition, *in* '2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)', IEEE, pp. 1–6.

Milborrow, S. and Nicolls, F. (2014), 'Active Shape Models with SIFT Descriptors and MARS', *Proceedings of the 9th International Conference on Computer Vision Theory and Applications* (i), 380–387.

Mollahosseini, A., Hasani, B. and Mahoor, M. H. (2017), 'AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild'.

Monini, S., Buffoni, A., Romeo, M., Di Traglia, M., Filippi, C., Atturo, F. and Barbara, M. (2016), 'Kabat rehabilitation for Bell's palsy in the elderly', *Acta Oto-Laryngologica* pp. 1–5.

Nech, A. and Kemelmacher-Shlizerman, I. (2017), Level Playing Field for Million Scale Face Recognition, *in* '2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)', IEEE, pp. 3406–3415.

Newell, A., Yang, K. and Deng, J. (2016), Stacked Hourglass Networks for Human Pose Estimation, *in* 'Computer Vision – ECCV 2016', Springer, Cham, pp. 483–499.

Pantic, M., Valstar, M., Rademaker, R. and Maat, L. (2005), Web-Based Database for Facial Expression Analysis, *in* '2005 IEEE International Conference on Multimedia and Expo', IEEE, pp. 317–321.

Parkhi, O. M., Vedaldi, A. and Zisserman, A. (2015), Deep Face Recognition, *in* 'Procedings of the British Machine Vision Conference 2015', British Machine Vision Association, pp. 1–41.

Paysan, P., Knothe, R., Amberg, B., Romdhani, S. and Vetter, T. (2009), A 3D Face Model for Pose and Illumination Invariant Face Recognition, *in* '2009 Sixth IEEE International Conference on Advanced Video and Signal Based Surveillance', IEEE, pp. 296–301.

Pereira, C. R., Weber, S. A. T., Hook, C., Rosa, G. H. and Papa, J. P. (2016), Deep Learning-Aided Parkinson's Disease Diagnosis from Handwritten Dynamics, *in* '2016 29th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)', IEEE, pp. 340–346.

Pfister, T., Li, X., Zhao, G. and Pietikäinen, M. (2011), 'Recognising spontaneous facial micro-expressions', *Proceedings of the IEEE International Conference on Computer Vision* pp. 1449–1456.

Pietikäinen, M., Hadid, A., Zhao, G. and Ahonen, T. (2011), *Computer Vision Using Local Binary Patterns*, Vol. 40.

Pons, G. and Masip, D. (2018), 'Supervised Committee of Convolutional Neural Networks in Automated Facial Expression Analysis', *IEEE Transactions on Affective Computing* **9**(3), 343–350.

Rajnoha, M., Mekyska, J., Burget, R., Eliasova, I., Kostalova, M. and Rektorova, I. (2018), Towards Identification of Hypomimia in Parkinson's Disease Based on Face Recognition Methods, *in* '2018 10th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT)', IEEE, pp. 1–4.

Ramanan, D., Xiangxin Zhu and Ramanan, D. (2012), 'Face detection, pose estimation, and landmark localization in the wild', *2012 IEEE Conference on Computer Vision and Pattern Recognition* pp. 2879–2886.

Ranjan, R., Patel, V. M. and Chellappa, R. (2016), 'HyperFace: A Deep Multi-task Learning Framework for Face Detection, Landmark Localization, Pose Estimation, and Gender Recognition'.

Ranjan, R., Sankaranarayanan, S., Castillo, C. D. and Chellappa, R. (2017), An All-In-One Convolutional Neural Network for Face Analysis, *in* '2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)', IEEE, pp. 17–24.

Redmon, J. and Farhadi, A. (2016), 'YOLO9000: Better, Faster, Stronger'.

Ren, S., Cao, X., Wei, Y. and Sun, J. (2014), Face Alignment at 3000 FPS via Regressing Local Binary Features, *in* '2014 IEEE Conference on Computer Vision and Pattern Recognition', IEEE, pp. 1685–1692.

Ren, S., He, K., Girshick, R. and Sun, J. (2015), 'Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **39**(6), 1137–1149.

Rifai, S., Bengio, Y., Courville, A., Vincent, P. and Mirza, M. (2012), Disentangling Factors of Variation for Facial Expression Recognition, *in* 'Proceedings of the 12th European conference on Computer Vision - Volume Part VI', Springer-Verlag, pp. 808–822.

Riggio, R. E. and Feldman, R. S. R. S. (2005), *Applications of nonverbal communication*, L. Erlbaum Associates, Publishers.

Sagonas, C., Tzimiropoulos, G., Zafeiriou, S. and Pantic, M. (2013), 300 Faces in-the-Wild Challenge: The First Facial Landmark Localization Challenge, *in* '2013 IEEE International Conference on Computer Vision Workshops', IEEE, pp. 397–403.

Sajid, M., Shafique, T., Baig, M., Riaz, I., Amin, S., Manzoor, S., Sajid, M., Shafique, T., Baig, M. J. A., Riaz, I., Amin, S. and Manzoor, S. (2018), 'Automatic Grading of Palsy Using Asymmetrical Facial Features: A Study Complemented by New Solutions', *Symmetry* **10**(7), 242.

Sánchez-Lozano, E., Martinez, B., Tzimiropoulos, G. and Valstar, M. (2016), Cascaded continuous regression for real-time incremental face tracking, *in* 'Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)'.

Schroff, F., Kalenichenko, D. and Philbin, J. (2015), FaceNet: A unified embedding for face recognition and clustering, *in* '2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)', IEEE, pp. 815–823.

Shen, J., Zafeiriou, S., Chrysos, G. G., Kossaifi, J., Tzimiropoulos, G. and Pantic, M. (2015), The First Facial Landmark Tracking in-the-Wild Challenge: Benchmark and Results, *in* '2015 IEEE International Conference on Computer Vision Workshop (ICCVW)', IEEE, pp. 1003–1011.

Shi, J., Samal, A. and Marx, D. (2006), 'How effective are landmarks and their geometry for face recognition?', *Computer Vision and Image Understanding* **102**(2), 117–133.

Simion, F. and Giorgio, E. D. (2015), 'Face perception and processing in early infancy: inborn predispositions and developmental changes.', *Frontiers in psychology* **6**, 969.

Simonyan, K. and Zisserman, A. (2014), Two-Stream Convolutional Networks for Action Recognition in Videos, *in* 'Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 1', pp. 568–576.

Simonyan, K. and Zisserman, A. (2015), 'Very Deep Convolutional Networks for Large-Scale Image Recognition', *International Conference on Learning Representations (ICRL)* pp. 1–14.

Soomro, K., Zamir, A. R. and Shah, M. (2012), 'UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild'.

Storey, G., Bouridane, A. and Jiang, R. (2018), 'Integrated Deep Model for Face Detection and Landmark Localization From "In The Wild" Images', *IEEE Access* **6**.

Storey, G. and Jiang, R. (2019), Face Symmetry Analysis Using a Unified Multi-task CNN for Medical Applications, *in* 'Intelligent Systems and Applications', pp. 451–463.

Storey, G., Jiang, R. and Bouridane, A. (2017), 'Role for 2D image generated 3D face models in the rehabilitation of facial palsy', *Healthcare Technology Letters* **4**(4), 145–148.

Sumathi, C. P., Santhanam, T. and Mahadevi, M. (2012), 'AUTOMATIC FACIAL EXPRESSION ANALYSIS A SURVEY', *International Journal of Computer Science & Engineering Survey (IJCSES)* **3**(6).

Sumithra, R., Suhil, M. and Guru, D. S. (2015), Segmentation and classification of skin lesions for disease diagnosis, *in* 'Procedia Computer Science'.

Sun, Y., Wang, X. and Tang, X. (2013), Deep Convolutional Network Cascade for Facial Point Detection, *in* '2013 IEEE Conference on Computer Vision and Pattern Recognition', IEEE, pp. 3476–3483.

Sun, Y., Wang, X. and Tang, X. (2014), Deep Learning Face Representation from Predicting 10,000 Classes, *in* '2014 IEEE Conference on Computer Vision and Pattern Recognition', IEEE, pp. 1891–1898.

Szegedy, C., Wei Liu, Yangqing Jia, Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V. and Rabinovich, A. (2015), Going deeper with convolutions, *in* '2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)', IEEE, pp. 1–9.

Taigman, Y., Yang, M., Ranzato, M. and Wolf, L. (2014), DeepFace: Closing the Gap to Human-Level Performance in Face Verification, *in* '2014 IEEE Conference on Computer Vision and Pattern Recognition', IEEE, pp. 1701–1708.

Taini, M., Zhao, G., Li, S. Z. and Pietikainen, M. (2008), Facial expression recognition from

near-infrared video sequences, *in* '2008 19th International Conference on Pattern Recognition', IEEE, pp. 1–4.

Tovée, M. J. (1995), 'Face Recognition: What are faces for?', *Current Biology* **5**(5), 480–482.

Tran, A. T., Hassner, T., Masi, I., Paz, E., Nirkin, Y. and Medioni, G. (2017), 'Extreme 3D Face Reconstruction: Seeing Through Occlusions'.

Tran, D., Ray, J., Shou, Z., Chang, S.-F. and Paluri, M. (2017), 'ConvNet Architecture Search for Spatiotemporal Feature Learning'.

Triantafyllidou, D., Nousi, P. and Tefas, A. (2018), 'Fast Deep Convolutional Face Detection in the Wild Exploiting Hard Sample Mining', *Big Data Research* **11**, 65–76.

Trigueros, D. S., Meng, L. and Hartnett, M. (2018), 'Face Recognition: From Traditional to Deep Learning Methods'.

Turk, M. and Pentland, A. (1991), 'Eigenfaces for Recognition', *Journal of Cognitive Neuroscience* **3**(1), 71–86.

Uijlings, J. R. R., Van De Sande, K. E. A., Gevers, T. and Smeulders, A. W. M. (2013), 'Selective Search for Object Recognition', *International Journal of Computer Vision* **104**(2), 154–171.

Valstar, M., Martinez, B., Binefa, X. and Pantic, M. (2010), Facial point detection using boosted regression and graph models, *in* 'Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition', pp. 2729–2736.

Varol, G., Laptev, I. and Schmid, C. (2018), 'Long-Term Temporal Convolutions for Action Recognition', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **40**(6), 1510–1517.

Vielzeuf, V., Pateux, S. and Jurie, F. (2017), Temporal multimodal fusion for video emotion classification in the wild, *in* 'Proceedings of the 19th ACM International Conference on Multimodal Interaction - ICMI 2017', ACM Press, New York, New York, USA, pp. 569–576.

Vinokurov, N., Arkadir, D., Linetsky, E., Bergman, H. and Weinshall, D. (2016), Quantifying Hypomimia in Parkinson Patients Using a Depth Camera, Springer, Cham, pp. 63–71.

Viola, P. and Jones, M. J. (2004), 'Robust Real-Time Face Detection', *International Journal of Computer Vision* **57**(2), 137–154.

Vrij, A. (2008), *Detecting lies and deceit: pitfalls and opportunities. Wiley Series in the Psychology of Crime, Policing and Law*.

Wagner, A., Wright, J., Ganesh, A., Zhou, Z., Mobahi, H. and Ma, Y. (2012), 'Toward a practical face recognition system: Robust alignment and illumination by sparse representation', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **34**(2), 372–386.

Wang, S.-J., Yan, W.-J., Zhao, G., Fu, X. and Zhou, C.-G. (2014), Micro-expression recognition using robust principal component analysis and local spatiotemporal directional features, *in* 'Computer Vision-ECCV 2014 Workshops', Springer, pp. 325–338.

Wang, T., Dong, J., Sun, X., Zhang, S. and Wang, S. (2014), 'Automatic recognition of facial movement for paralyzed face', *Bio-Medical Materials and Engineering* **24**, 2751–2760.

Wang, T., Zhang, S., Dong, J., Liu, L. and Yu, H. (2016), 'Automatic evaluation of the degree of facial nerve paralysis', *Multimedia Tools and Applications* **75**(19), 11893–11908.

Wang, Y., See, J., Phan, R. C.-W. W. and Oh, Y.-H. H. (2015), 'Efficient spatio-temporal local binary patterns for spontaneous facial micro-expression recognition.', *PloS one* **10**(5), e0124674.

Warren, G., Schertler, E. and Bull, P. (2009), 'Detecting Deception from Emotional and Unemotional Cues', *Journal of Nonverbal Behavior* **33**(1), 59–69.

Wen-Jing Yan, Wu, Q., Yong-Jin Liu, Su-Jing Wang and Fu, X. (2013), CASME database: A dataset of spontaneous micro-expressions collected from neutralized faces, *in* '2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)', IEEE, pp. 1–7.

Wen, Y., Zhang, K., Li, Z. and Qiao, Y. (2016), A Discriminative Feature Learning Approach for Deep Face Recognition, Springer, Cham, pp. 499–515.

Wu, H.-Y., Rubinstein, M., Shih, E., Guttag, J., Durand, F. and Freeman, W. (2012), 'Eulerian video magnification for revealing subtle changes in the world', *ACM Transactions on Graphics* **31**(4), 1–8.

Wu, P., Gonzalez, I., Patsis, G., Jiang, D., Sahli, H., Kerckhofs, E. and Vandekerckhove, M. (2014), 'Objectifying facial expressivity assessment of Parkinson's patients: preliminary study.', *Computational and mathematical methods in medicine* **2014**, 427826.

Wu, Y., Liu, H., Li, J. and Fu, Y. (2017), Deep Face Recognition with Center Invariant Loss, *in* 'Proceedings of the on Thematic Workshops of ACM Multimedia 2017 - Thematic Workshops '17', ACM Press, New York, New York, USA, pp. 408–414.

Xiao, S., Feng, J., Xing, J., Lai, H., Yan, S. and Kassim, A. (2016), Robust Facial Landmark Detection via Recurrent Attentive-Refinement Networks, Springer, Cham, pp. 57–72.

Xiong, X. and De la Torre, F. (2013), Supervised Descent Method and Its Applications to Face Alignment, *in* '2013 IEEE Conference on Computer Vision and Pattern Recognition', IEEE, pp. 532–539.

Yan, J., Zhang, X., Lei, Z. and Li, S. Z. (2014), 'Face detection by structural models', *Image and Vision Computing* **32**(10), 790–799.

Yan, J., Zheng, W., Cui, Z., Tang, C., Zhang, T., Zong, Y. and Sun, N. (2016), Multi-clue fusion for emotion recognition in the wild, *in* 'Proceedings of the 18th ACM International Conference on Multimodal Interaction - ICMI 2016', ACM Press, New York, New York, USA, pp. 458–463.

Yan, W.-J. J., Li, X., Wang, S.-J. J., Zhao, G., Liu, Y.-J. J., Chen, Y.-H. H. and Fu, X. (2014), 'CASME II: An improved spontaneous micro-expression database and the baseline evaluation', *PLoS ONE* **9**(1), 1–8.

Yan, W. J., Wang, S. J., Liu, Y. J., Wu, Q. and Fu, X. (2014), 'For micro-expression recognition: Database and suggestions', *Neurocomputing* **136**, 82–87.

Yang, S., Luo, P., Loy, C. C. and Tang, X. (2016*a*), 'From facial parts responses to face detection: A deep learning approach', *Proceedings of the IEEE International Conference on Computer Vision* **11-18-Dece**(3), 3676–3684.

Yang, S., Luo, P., Loy, C. C. and Tang, X. (2016*b*), WIDER FACE: A face detection benchmark, *in* 'Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition', Vol. 2016-Decem, pp. 5525–5533.

Yao, A., Cai, D., Hu, P., Wang, S., Sha, L. and Chen, Y. (2016), HoloNet: towards robust emotion recognition in the wild, *in* 'Proceedings of the 18th ACM International Conference on Multimodal Interaction - ICMI 2016', ACM Press, New York, New York, USA, pp. 472–478.

Yi, D., Lei, Z., Liao, S. and Li, S. Z. (2014), 'Learning Face Representation from Scratch'.

Youssif, A. A. A. and Asker, W. A. A. (2011), 'Automatic Facial Expression Recognition System Based on Geometric and Appearance Features', *Computer and Information Science* **4**(2), 115–124.

Yu, X., Huang, J., Zhang, S. and Metaxas, D. N. (2016), 'Face landmark fitting via optimized part mixtures and cascaded deformable model', *IEEE Transactions on Pattern Analysis and Machine Intelligence* .

Yu, Z., Liu, G., Liu, Q. and Deng, J. (2018), 'Spatio-temporal convolutional features with nested LSTM for facial expression recognition', *Neurocomputing* **317**, 50–57.

Zafeiriou, S., Trigeorgis, G., Chrysos, G., Deng, J. and Shen, J. (2017), The Menpo Facial Landmark Localisation Challenge: A Step Towards the Solution, *in* 'IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops'.

Zeiler, M. D. and Fergus, R. (2013), 'Visualizing and Understanding Convolutional Networks'.

Zhang, C. and Zhang, Z. (2014), Improving multiview face detection with multi-task deep convolutional neural networks, *in* 'IEEE Winter Conference on Applications of Computer Vision', IEEE, pp. 1036–1041.

Zhang, S., Zhao, X. and Lei, B. (2012), 'Robust facial expression recognition via compressive sensing', *Sensors* **12**(3), 3747–3761.

Zhang, S., Zhu, X., Lei, Z., Shi, H., Wang, X. and Li, S. Z. (2017), 'S$ˆ3$FD: Single Shot Scale-invariant Face Detector'.

Zhang, X., Fang, Z., Wen, Y., Li, Z. and Qiao, Y. (2017), Range Loss for Deep Face Recognition with Long-Tailed Training Data, *in* '2017 IEEE International Conference on Computer Vision (ICCV)', IEEE, pp. 5419–5428.

Zhang, X., Wang, S., Liu, J. and Tao, C. (2017), Computer-aided diagnosis of four common cutaneous diseases using deep learning algorithm, *in* '2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)', IEEE, pp. 1304–1306.

Zhang, X., Wang, S., Liu, J. and Tao, C. (2018), 'Towards improving diagnosis of skin diseases by combining deep neural network and human knowledge.', *BMC medical informatics and decision making* **18**(Suppl 2), 59.

Zhang, Y. N. (2017), 'Can a Smartphone Diagnose Parkinson Disease? A Deep Neural Network Method and Telediagnosis System Implementation.', *Parkinson's disease* **2017**, 6209703.

Zhang, Z., Luo, P., Loy, C. C. and Tang, X. (2014), Facial landmark detection by deep multi-task learning, *in* 'Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)', Vol. 8694 LNCS, pp. 94–108.

Zhang, Z., Luo, P., Loy, C. C. and Tang, X. (2018), 'From Facial Expression Recognition to Interpersonal Relation Prediction', *International Journal of Computer Vision* **126**(5), 550–569.

Zhang, Z., Luo, P., Loy, C. C. and Tang, X. (n.d.), 'Facial Landmark Detection by Deep Multi-task Learning'.

Zhao, G., Pietikäinen, M. and Pietikainen, M. (2007), 'Dynamic texture recognition using local binary patterns with an application to facial expressions.', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **29**(6), 915–928.

Zhu, S., Li, C., Loy, C. C. and Tang, X. (2015*a*), Face alignment by coarse-to-fine shape searching, *in* 'Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition', Vol. 07-12-June, pp. 4998–5006.

Zhu, S., Li, C., Loy, C. C. and Tang, X. (2015*b*), Face alignment by coarse-to-fine shape searching, *in* 'Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition'.

Zhu, X., Lei, Z., Liu, X., Shi, H. and Li, S. Z. (2015), 'Face Alignment Across Large Poses: A 3D Solution', p. 11.

Zhu, X., Lei, Z., Yan, J., Yi, D. and Li, S. Z. (2015), High-fidelity Pose and Expression Normalization for face recognition in the wild, *in* 'Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition', Vol. 07-12-June, pp. 787–796.