Emotion Recognition for Affective Computing: Computer Vision and Machine Learning Approach

Arwa Mohmmed Basbrain

Thesis submitted for the degree of Doctor of Philosophy

School of Computer Science and Electronic Engineering

University of Essex

2021

الحمد لله حمداً كثيراً طيباً مباركاً فيه كما يحب ربنا ويرضى

عدد خلقة ورضاء نفسه وزنة عرشة ومداد كلماته

والصلاة والسلام على خير خلق الله سيدنا محمد وعلى آله وصحبة اجمعين

I dedicate my humble effort to my sweetness and love

Father & Mother

Muhammad Basbrain & Suad Basbrain

The memory of their affection, love, encouragement, and prayer day and night makes me able to achieve this success and honour. To their dreams that they always believed I could fulfil them.

To my loving husband, **Khalid Binothman,** *for letting me experience the kind of love that people dream about.*

To my beloved sons **Mohammed, Elyas, Albaraa, Danial** *for* their endless love, help and support.

ABSTRACT

The purpose of affective computing is to develop reliable and intelligent models that computers can use to interact more naturally with humans. The critical requirements for such models are that they enable computers to recognise, understand and interpret the emotional states expressed by humans. The emotion recognition has been a research topic of interest for decades, not only in relation to developments in the affective computing field but also due to its other potential applications.

A particularly challenging problem that has emerged from this body of work, however, is the task of recognising facial expressions and emotions from still images or videos in real-time. This thesis aimed to solve this challenging problem by developing new techniques involving computer vision, machine learning and different levels of information fusion.

Firstly, an efficient and effective algorithm was developed to improve the performance of the Viola-Jones algorithm. The proposed method achieved significantly higher detection accuracy (95%) than the standard Viola-Jones method (90%) in face detection from thermal images, while also doubling the detection speed. Secondly, an automatic subsystem for detecting eyeglasses, Shallow-GlassNet, was proposed to address the facial occlusion problem by designing a shallow convolutional neural network capable of detecting eyeglasses rapidly and accurately. Thirdly, a novel neural network model for decision fusion was proposed in order to make use of multiple classifier systems, which can increase the classification accuracy by up to 10%. Finally, a high-speed approach to emotion recognition from videos, called One-Shot Only (OSO), was developed based on a novel spatio-temporal data fusion method for representing video frames. The OSO method tackled video classification as a single image classification problem, which not only made it extremely fast but also reduced the overfitting problem.

ACKNOWLEDGMENTS

First of all, I would like to thank Almighty God (Allah) for giving me the strength, opportunity, and determination to complete this work. This journey has been miraculous for me due to the mercy of ALLAH Almighty; I don't have the words to express my gratitude.

I really appreciate the never-ending support of my parents. Whatever I am today is due to their determined efforts. I cannot pay them back for all their love, care and prayers. My parents are the most important people in my world, and I dedicate this thesis to their memories and dreams with loads of love.

I would like to express my deepest and sincere gratitude to my supervisor Prof. John Q Gan for his generous guidance, advice, help, enthusiasm and immense knowledge with which he supported me in this research and the encouragement, patience, understanding and support. I have learnt from him how to be focused, dedicated, and perseverance person. John is hugely responsible for transforming my outlook as a researcher and always motivated me to achieve high research standards. I could not have imagined having a better supervisor and mentor for my PhD study.

I would like to sincerely thank my second Co-supervisor, Dr Adrian F. Clark, for his guidance and advice and my supervisory panel members: Prof. Dongbing Gu and Dr Stephen J. Sangwine, for their helpful suggestions.

My deepest gratitude goes to my family for always treated me like a princess, my loving husband Khalid, for all the discussions and interests he showed to my research to help me think out loud and my sons Mohammed, Elyas, Albraa and little boy Danial the shine and joy of my life, with their love and support this thesis is accomplished.

I would like to thank my loving mother-in-law, Nafisa Ali, for her endless prays, and the boosting notes she constantly sent, which always lift me up and push me towards hard work.

I would also like to acknowledge the support of my brothers Sameer, Hisham, Omar and Abdul Majeed, my sisters Nawal, Prof. Sakeena, Abeer, Dalal, Areej and their families.

ACKNOWLEDGMENTS

I am also grateful to the King Abdul Aziz University in Jeddah, Saudi Arabia for offering me a scholarship, which helps me undertake my PhD study and the Computer Science Department at the university for supporting me through the course of my PhD research.

Finally, my deepest gratitude goes to all my friends at home and here in Colchester. Namely: Samah Felemban, Huda Altaisan, Nora Alkhamees, Manal Alghannam, Enas Jambi, Wafa Beheiri and Fatima Alshahrani. Thanks to many other family, friends, and colleagues who are too numerous to mention. I would like to thank all of my brothers, sisters, friends and colleagues in Saudi Arabia and in the UK.

I am truly grateful to everyone who so generously supported me and inspired me, to everyone who contributed to the work presented in this thesis.

CONTENTS

CONTENTS

CONTENTS

CONTENTS

TABLES

List of Tables

TABLES

FIGURES

List of Figures

FIGURES

FIGURES

Chapter 1

Introduction

1.1. Problem Statement and Motivation

Modern computer technology has certainly revolutionised and enriched our world, underpinning the creation of amazing tools that have been utilised across almost all areas of our lives. The increasing extent to which this technology has permeated our daily lives, however—whether in smart home and personal health devices or in multi-functional personal devices such as smartphones and smartwatches—has heightened the need for more natural interactions with their users. Affective computing aims to assist these natural interactions between computers and humans by creating reliable and intelligent models enabling computers to detect, recognise, understand and interpret the emotional states expressed by humans. Furthermore, since emotions have a substantial influence on a range of human cognitive processes, like learning, problem-solving, perception and memory, the incorporation of such models in computerised devices offers potential benefits in many areas, including in healthcare, education, social interaction and behavioural science, etc. This thesis, therefore, aims to develop novel methods to recognise emotion by using computer vision and machine learning techniques.

Previous work in this area has identified a number of significant challenges. Fundamentally, the development of an accurate and reliable emotion recognition

system is challenging because human beings may experience a combination of emotions or different emotions in different strengths at the same time. Currently-developed human emotion recognition systems neglect this fact and attempt to find only the strongest emotion at any one time [1]. In addition, the variable nature of the human emotional experience makes it challenging for computer systems to identify emotions reliably across a population. For example:

- Human emotions are not steady, and they occur to different degrees.
- The patterns of emotional change depend to some extent on people's cultural and linguistic background [1].
- The dominance of human emotions is based on a person's psychological type.

Another set of challenges arises from the fact that the performances of emotion recognition systems is highly dependent on the accuracy of their databases of emotions. Currently, however, these databases struggle to categorise emotions accurately, for the following reasons:

- ***Emotional interference***: This is considered to be one of the major obstacles facing both image and video-based emotion recognition systems and has a significant impact on recognition accuracy.
- ***Intra-class and inter-class variations***: The intra-class variation problem arises when samples (images/videos) of the same class (of objects, etc.) can have significantly different appearances. In the inter-class ambiguity problem, meanwhile, samples of different classes can show similar visual characteristics. In these circumstances, samples of the same class can be challenging to identify, and samples of different classes might also be easily

misclassified due to lack of clear distinction between features belonging to different classes.

- ***Data reliability***: One of the main challenges facing most of the research into emotion recognition systems is the reliability of the available real-world datasets, which have generally been collected from the Internet or from films. Since these datasets contain images or videos that have been captured in relation to real-life scenarios, they often present complex or even ambiguous emotions rather than prototypical or simple ones. This adds some difficulty and uncertainty to the annotation and labelling process. Also, due to the subjectivity and varied expertise of the labellers, there is sometimes disagreement among annotators, and this can lead to inconsistency in the dataset's labelling. This inconsistency demonstrates how difficult the task of distinguishing emotions is even for humans.
- ***Real-world conditions*:** Most early facial expression datasets do not represent real-world conditions because they were captured in a lab-controlled environment where the subjects were controlled and some of the other significant factors were simplified, eliminated or managed: i.e., illumination, lighting conditions, restrictions on clothing, eyeglasses, etc.
- ***Insufficient annotation of data***: Most recent studies that have focused on image and video-based recognition systems have embedded deep-learning models in their designs since these appear to exhibit the most efficient and promising results. To achieve this high accuracy, however, these models usually require an enormous number of annotated images/videos for training. The existing still-image datasets usually contain a very high number of

images compared to the number of videos in the video datasets constructed in relation to the same field (of recognition).

Finally, developing a system able to recognise general human emotions in real-time and with sufficiently high performance to meet the needs of the intended affective applications is considered an extremely challenging task. In short, in a complete human emotion recognition system, the difficulties in categorising emotions that have been summarised above have to sit alongside the regular challenges facing any automatic facial recognition system, such as illumination, face pose variation, face tracking and misalignment problems. When these two sets of challenges are put together, delivering reliable emotion recognition in real-time becomes very difficult. That is why the state-of-the-art algorithms which currently implement vision-based recognition system still fall far short of the requirements of real-time applications due to several kinds of challenges such as: dealing with the nature of human emotions.

To achieve the highest levels of accuracy and reliability, human emotion recognition systems, therefore, need to consider the above problems. Accordingly, this thesis focuses on developing methods for real-time emotion recognition, in which both the model performance and model complexity are taken into account.

1.2. Research Objectives

The primary objective of the proposed research is to build a human emotion recognition system which can capture and recognise facial expressions and emotions from still images or videos in real-time by developing new techniques involving computer vision, machine learning and different levels of information fusion. Specifically, it subsumes the following objectives:

- To improve the robustness of the face detection method for thermal images that can be utilised for real-time emotion recognition applications.
- To address the facial occlusion problem in facial analysing systems by developing an effective and efficient methods for eyeglasses detecting.
- To increase the classification accuracy by developing new decision fusion methods in order to make use of multiple classifier systems.
- To develop novel methods for spatio-temporal data fusion that can be utilised for video-base recognition systems. For example, temporal features extracted from videos can be very useful for emotion recognition because they characterise the dynamic properties of emotional development.
- To meet the needs of real-time applications from the perspective of both the computational complexity and accuracy when designing each method of the framework of the emotion recognition system.

Psychologists describe the emotional state in terms of discrete categories, which include happiness, sadness, fear, anger, disgust and surprise. Most of the current automatic affect recognition studies focus on recognising these basic emotions, which may be expressed through a range of signals, including facial, gestural, postural, voice, and bio-potential signals. The existing research on the automatic recognition of

emotions attempts to determine the emotional state of a subject from the manifestation of emotions by using images and video clips.

1.3. Thesis Structure

Chapter 2: *Literature Review.*

This chapter presents a comprehensive and up-to-date review of computer vision-based approaches to recognising facial emotions and includes a detailed critical analysis of the frameworks designed to support image-based and video-based classification based on deep learning. The chapter also presents a survey of the available visual and thermal facial expression databases and compares the spontaneous and posed facial expressions databases. Some important challenges are highlighted in this chapter that inform the following chapters of the thesis.

Chapter 3: *Accuracy Enhancement of the Viola-Jones Algorithm for Thermal Face Detection.*

This chapter presents a method [2] for enhancing the Viola-Jones algorithm [3] for face detection by improving its performance in the thermal spectrum, allowing the detection of emotions in faces with or without eyeglasses. A performance comparison is undertaken of three different features, HOG, LBP and Haar-like, to find the most suitable one for face detection from thermal images. Additionally, to accelerate the detection speed, a pre-processing stage is added in both the training and detecting phases. Two pre-processing methods as tested and compared, together with the three features. The proposed enhancement process reduces the detection time of the Viola-Jones algorithm by roughly a factor of two while retaining high detection accuracy.

Chapter 4: *Shallow Convolutional Neural Network for Eyeglasses Detection.*

To improve the robustness of facial analysis systems and cope with real-world applications, this chapter [4] designs a rapid and highly accurate method for detecting eyeglasses, based on extracting deep features from a well-designed shallow convolutional neural network (CNN), called Shallow-GlassNet. To address the two essential challenges of CNN (the size of the training dataset required and the depth of the network architecture), we initialise the learning parameters of the shallow CNN by the parameters of a deep CNN which is fine-tuned on a small dataset. The depth of the neural network is then decreased by removing some convolutional layers after testing its performance on the validation dataset. Evaluation experiments are conducted on two large unconstrained facial image databases, LFW [5] and Celeb Faces[6]. The results demonstrate the superior performance of the proposed model for the detection of eyeglasses, both in terms of speed and accuracy.

Chapter 5: *A Neural Network Approach to Decision Score Fusion for Emotion Recognition*

This chapter presents an effective facial emotion recognition system [7] that classifies facial images to one of the six universal emotions (Anger, Disgust, Fear, Happiness, Sad & Surprise) and Neutral. The proposed system uses convolutional neural networks (GoogleNet-CNN) to detect eyeglasses and extract features, followed by a novel score fusion model. Nine different sets of emotional features are extracted from faces with and without eyeglasses by convolutional neural networks and classified by support vector machines (SVMs). Then, two neural network models are used and tested to accomplish decision fusion. The USTC-NVIE (NVIE) [8] database is used to evaluate the performance of the proposed system. Experimental results show that the

proposed facial emotion recognition system achieves a higher classification rate when using the eyeglass detector, while the multiple classifiers system increases the classification rates of the system.

Chapter 6: *One-Shot Only Real-Time Video Classification: A Case Study in Facial Emotion Recognition*

Previous work on video classification uses repeated evaluations of a CNN in order either to classify each frame separately or to combine several frames to be classified by means of a complex 3D-CNN. In this chapter [9], we present a new method called One-Shot Only (OSO), a novel approach to real-time video classification. The OSO method tackles video classification as a single image classification problem, spatially rearranging timeframes so as to form a simple storyboard and associate class probabilities to this. The method uses a single CNN which predicts the class probabilities directly in one evaluation from one full image which presents the complete sequence of the video frames. Since the whole classification pipeline is a single network, it can be optimised end-to-end directly as far as recognition performance is concerned. The proposed architectures are extremely fast, in terms of evaluation times, and this is appropriate to the real-time situation. Processing just this visual information, OSO still achieves superior or comparable classification accuracies (compared to repeated-evaluation based methods) on both image and video datasets, AffectNet [10], RAF-DB [11] and AFEW [12].

Chapter 7: *Conclusions and Future Work*

Finally, the contributions of my PhD work are summarised in this chapter, followed by a discussion of the limitations and some suggestions about potential directions for future research.

Chapter 2

Literature Review

2.1. Introduction

"Emotion represents the psychological state of the human mind and thought processes" [13]. Emotions are accompanied by internal and external bodily manifestations. External manifestations of emotions include facial expressions, body gestures and perturbations in verbal communications and even handwriting, while internal manifestations include changes in heart rate and body temperature. A number of researchers from several different domains have sought to study these manifestations of emotions, and a variety of innovative instruments have been developed for the purpose of measuring them accurately across differing modalities. For example, thermal infrared cameras have been used to measure the changes in thermal distribution occurring across blood vessels and the variations in facial skin temperature caused by a variety of emotions [14, 15]. Electroencephalograms (EEGs) can provide accurate measurements of the temporal changes which occur during emotional arousal [14, 16, 17].

As well as these sophisticated instruments, the kind of video-cameras which are now widely used in everyday life, due to the proliferation of function-rich mobile devices, can be utilised for such measurements. Their use has been further accelerated by the hitherto almost inconceivable increases in storage space and Internet bandwidth which have taken place; changes that have elevated images and videos to become an

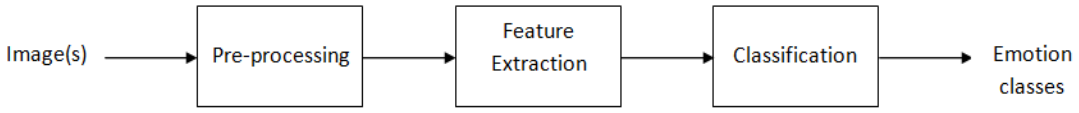


Figure 2-1: Basic stages of facial emotion recognition.

indispensable part of today's big data. These circumstances have provided an abundance of data, and this abundance has encouraged computer vision researchers to develop advanced techniques for a wide range of applications to interpret and understand video data, including in respect to the recognition of facial manifestations of emotions.

The typical approach of contemporary studies on recognising facial expressions and emotions based on visible-light and thermal images generally comprises three components: pre-processing, facial feature extraction and image classification (see Figure 2-1). Given an input image or image sequence, pre-processing is performed on it by detecting the area representing the face itself, then conducting normalisation, facial features localisation and face alignment. The second step, facial feature extraction, finds the relevant, strongly differentiated features from the various regions of the detected face. The final step is the machine classification of facial expressions. While most of the existing approaches have these components in common, they differ in terms of the exact methods used within each component.

In this chapter, we review the common models used for describing emotions and their expression and the state-of-the-art methods for identifying the different levels of facial expressions. We break down facial expression/emotion recognition systems into their basic components. We review the contemporary and state-of-the-art research regarding each component as this relates to dealing with the challenges of facial

analysis systems. Finally, we analyse the existing thermal and visible-light facial databases in detail, discussing their advantages and limitations.

Emotion explanations: The emotions we experience at any given moment have an influence on the actions we take, the choices we make and the perceptions we have. Therefore, it is important to have some ideas of what emotions are. Psychologists have attempted this, generally, by applying one of two different approaches: the discrete categorical model and the dimensional model.

*Emotions as discrete categorical models***:** Discrete emotion theory attempts to characterise human emotions by defining an innate set of basic emotions that are cross-culturally recognisable. A plethora of published studies in computer vision describes emotions as discrete categories. During the 1970s, psychologist Paul Ekman [1] suggested that people across all human cultures experience these basic emotions and that these are distinct and strong enough to be recognised or identified by an individual's facial expression and/or other biological processes. Ekman suggested that people across all human cultures experience a number of basic emotions (anger, disgust, fear, happy, sad and surprise) and that these are distinct. Each of these emotions associates with particular characteristics, allowing them to be expressed in varying degrees which are proportional to the strength of the emotion. Latterly, he made additions to his initial list of basic emotions, such as pride, excitement, shame and embarrassment. Another psychologist, Plutchik [18], further suggested that basic emotions are like colours and may be combined to create other shades/emotions. According to this theory, mixed emotions are something like a building which is created from different blocks of basic emotions.

Figure 2-2: Examples of six basic emotions and twelve compound emotions from RAF-DB database [11].

A limitation of this categorical model of emotions is that mixed or complex emotions cannot always be appropriately described using a restricted set of basic emotions. To overcome this, some researchers have defined multiple compound categories of emotion, such as happily-surprised, fearfully-angry [11, 19]. Nonetheless, such compound emotion sets remain limited, and the intensity of an emotion cannot be described at all using the categorical model. Figure 2-2 shows samples of basic emotions and compound emotions taken from the Real-World Expression Database (RAF-DB) dataset [11].

*Emotions as dimensional models***:** Dimensional models attempt to explain human emotions in terms of where each specific emotion lies within a two- or three-dimensional conceptual space. In 1912, Wilhelm Wundt described emotions in terms of his proposed three-dimensional model, the conceptual dimensions being: (pleasurable-unpleasurable), (strain- relaxation) and (arousing-subduing) [20]. In 1954, meanwhile, Harold Schlosberg concluded that facial expressions and body changes complement each other in showing us dimensions along which emotions may vary. He named three dimensions: (pleasantness–unpleasantness), (attention–rejection) and (level of activation) [21].

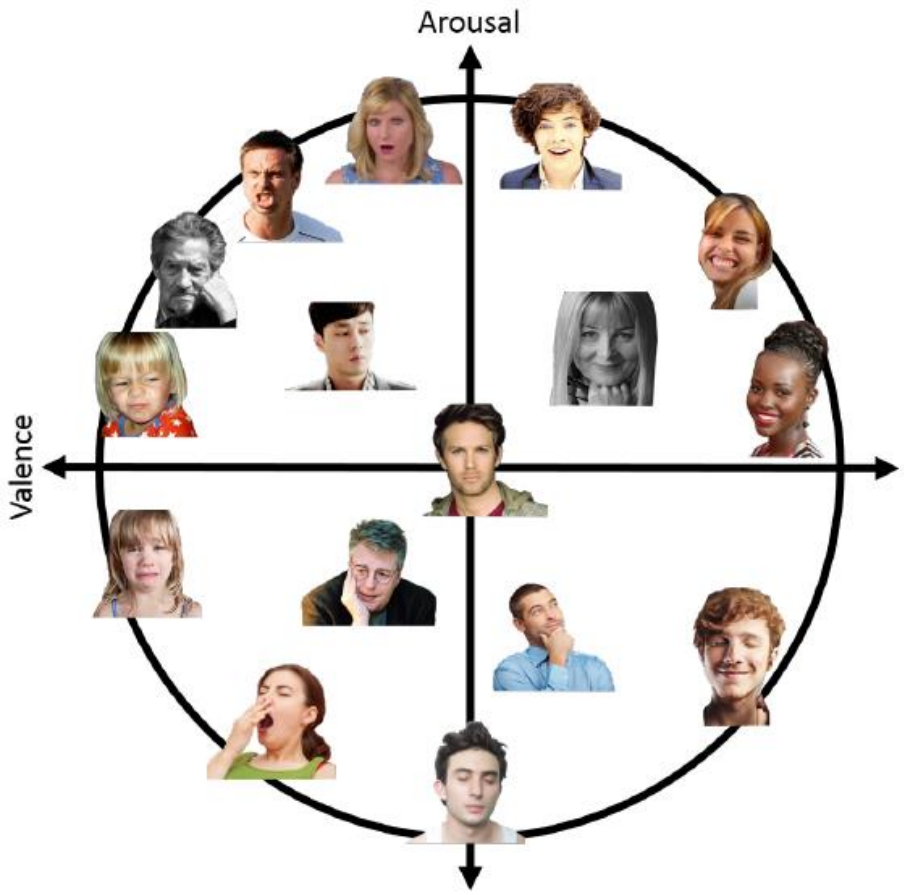


Figure 2-3: Examples of valence and arousal in the circumplex model [10]**.**

Several dimensional models of emotion have been proposed, such as the vector model, the circumplex model and the positive activation – negative activation model [22]. In contrast to categorical models, dimensional models can encode small changes in the intensity of each emotion and distinguish between slightly different displays of emotions on a continuous scale. Most dimensional models include valence and arousal or intensity dimensions. A valence scale reflects how positive or negative an event is, whereas arousal shows whether an event is exciting or calming, as shown in Figure 2-3.

***Descriptions of facial emotion*:** The most common modality of emotion recognition is that of facial expression analysis. State-of-the-art studies on the automatic analysis of facial expressions generally follow one of two main approaches to the description of different levels of facial expression: the message approach and the

sign judgement approach. The *message approach* attempts to derive the meaning conveyed by a facial display directly, whereas the *sign approach* attempts to study indirect expressions of emotion, such as a physical gesture (or sign) [1]. The Facial Action Coding System (FACS) [23] is a sign approach defined by Ekman and Friesen in 1978. FACS encodes the movements of specific facial muscles called Action Units (AUs), and these movements are taken to reflect distinct momentary changes in facial appearance [24]. The authors defined 32 AUs: with the upper face hosting nine; the lower face, eighteen; and a further five that could not be exclusively attributed to either the upper or lower face. Figure 2-4 shows examples of the AUs involved with the expression of some basic and some combined emotions. In addition, FACS encodes fourteen descriptors of various actions, such as those entailed in head pose and eye gaze direction, and for miscellaneous actions. This system, therefore, categorises every possible facial movement based on the changes manifested by that movement.

It is important to note that since FACS encodes facial actions without necessarily inferring the emotional state of the subject, it can be utilised to encode ambiguous and subtle facial expressions which result from emotions that cannot be easily categorised into one of the universal emotions. Furthermore, Del Giudice and Colle [25] demonstrated that the sensitivity of FACS to subtle differences in expression makes it capable of distinguishing between genuine and fake smiles. The system provides precise information concerning the actual facial movements made and has the advantage of high reliability [1, 19]. Despite the advantages of FACS in relation to the systematic analysis of facial expressions, however, it has a major limitation: since operators must be extensively trained, its application is time consuming and prone to bias due to subjectivity. All of this makes investigations of large samples difficult.

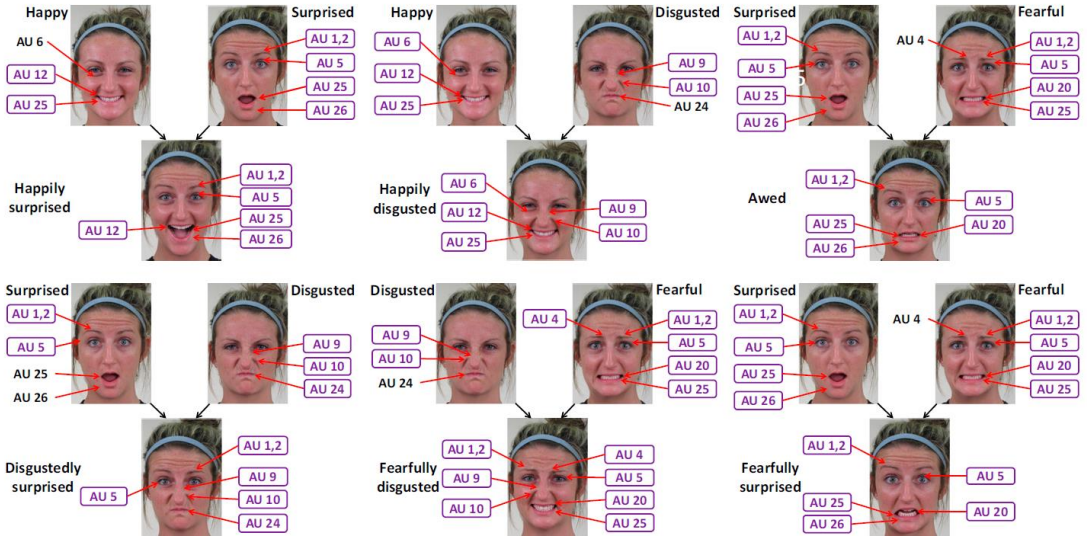


Figure 2-4: Examples of Action Units for some basic and combined emotions [19].

2.2. Methods for Image Pre-Processing

Pre-processing takes place using operations on images at the lowest level of abstraction: both input and output are intensity images. The man purpose of pre-processing is to provide improved image data (in terms of its subsequent processing) that has had unwanted distortions removed, and some image features enhanced. In this section, we briefly summarise the advances which have been made and noted in the literature as regards the pre-processing methods which are commonly utilised in face analysis applications such as facial recognition, facial orientations, head pose analysis and facial expression/emotion classification.

2.2.1. Face Detection

To build fully-automated systems that analyse the information contained in images of faces, robust and efficient face detection algorithms are required. Yang [26] gives a definition of face detection: “Given an arbitrary image, the goal of face detection is to determine whether or not there are any faces in the image and, if present, return the image location and extent of each face”. Most face analysis applications, whether

applied to videos or to still images, start with face detection as the first step of their pre-processing stage. According to Yang's definition, the goal of face detection is to determine all image regions which contain a face regardless of their position, orientation and the lighting conditions. Because face detection plays such an important role in automated face analysis systems, it has been studied intensively in computer vision research, and numerous techniques have been developed to detect the faces which exist in a single image. The most popular methods are those of Viola-Jones [3], Lienhart and Maydt [27] and Farfade et al [28].

The pioneering work of Viola-Jones [3] achieves high detection accuracy while minimizing the computation time. Their method is 15 times faster than previous algorithms at the time of publication with a 95% accuracy rate. The Viola-Jones approach is based on the quick evaluation of basic Haar-like features by using a new image representation. It produces a huge collection of features based on the integral image idea then utilises the cascade boosting technique AdaBoost to minimise the features set. The detector scans the grayscale images by using various sizes of the scanned windows to evaluate the Haar-like features quickly using a new image representation . The AdaBoost learning algorithm selects a set of critical features from a large set of evaluated features. This framework for face detection is capable of processing images incredibly quickly while maintaining a high detection rate.

Lienhart and Maydt [27] extends the Viola-Jones face detector by introducing a novel set of rotated Haar-like features and a new post optimization technique for a given boosted cascade classifier. Their efficient set of 45" rotated features add further domain-knowledge to the leaning framework which shows off on average a 10% lower false alarm rate. Their post optimization procedure adds additional improvement to the

average false alarm rate further by 12.5%.

Farfade et al. [28] proposed a face detector method, called Deep Dense Face Detector (DDFD), that is based on a single deep learning model. Unlike Viola-Jones face detector, their method can detect faces in a wide range of orientations and does not require pose or landmark annotation. Similar to AlexNet [29], their model consists of 5 convolutional layers followed by 3 fully-connected layers. The complexity of the proposed method is low as it does not require additional processes such as segmentation, bounding-box regression, or SVM classifiers.

While face detection has attracted considerable attention in relation to dealing with visible-light images, this challenge remains unsolved in terms of thermal images [30]. The current thermal face detection algorithms operate on the basis of several critical conditions which must be fulfilled for the face detection process to take place. Section 3.2 describes these conditions.

2.2.2. Facial Features Localisation and Tracking

After determining the region of the image in which the face is represented, further localisation of facial components is required for other pre-processing steps such as face alignment. Although the facial features localisation step is optional, it yields rich geometric information which is important in relation to facilitating face registration and the selection of a region of interest (ROI) where the feature extraction step is performed. Facial landmarks, also known as facial feature points, are mainly located around facial components such as eyes, mouth, nose and chin.

Facial landmark detection usually begins from the start-point of a rectangular bounding box returned by a face detector. This bounding box is used to establish the

positions of facial feature points. The number of facial feature points labelled depends on the application, which is to use these points; 17-point models, 29-point models and 68-point models have all been used depending on the specific situation. What unites the models, however, is that the points labelled must cover the areas that carry the most important information present for both discriminative and generative purposes—the most commonly-labelled areas are the eyes, the nose and the mouth. The more points that are labelled, the richer is the information available; on the other hand, the more points there are, the more time-consuming is the process of detecting them [31, 32].

To localise and track 'landmarks', a deformable-face model is usually used, whereby a pre-trained face model is matched to the target face. The probability of aligning the target face's appearance with the underlying conceptual model is maximised by deforming the target using a pre-trained statistical model of face deformations. A number of proposals for deformable face models are well known: e.g., the Active Shape Model [33, 34], the Active Appearance Model [35] and the Constrained Local Model [36]. Recently, however, a number of more sophisticated models have been proposed: Zhou et al. [37] and Sun et al. [38] constructed a set of deep convolutional networks in a cascade manner for the purpose of detecting facial points. In addition, the facial landmarks can be used with classifiers such as SVMs [39] and/or Restricted Boltzmann Machines (RBMs) [40] in order to detect facial actions or recognize expression directly [41, 42].

Kotsia et al. [43], meanwhile, constructed a grid adaptation system by utilising deformable models [44] to extract geometrical information regarding the face from the first video frame; this information was then used for tracking through the rest of the video. For this, a grid-tracking algorithm was applied to produce a deformed facial grid

which could then be used to detect facial actions which were, in turn, used by a multi-class SVM to classify the corresponding facial expression—which appeared by the last frame of the video. In addition, Wang et al. [45] classified various different facial expressions presented by a near frontal face by tracking 26 facial feature points using a priori face shape models constructed based on RBM. Although the above approaches are successful at tracking facial expression, emotion recognition systems require more sophisticated methods of feature extraction and classification since, in real-world scenarios, facial actions often change both facial texture and geometry [23].

2.2.3. Region of Interest Selection

Turning to facial features localisation and region of interest selection from thermal images, most research focuses on either statistical temperature parameters (i.e. minimum, maximum, standard deviation, and mean) of regions of interest [46], or those same imaging features that are commonly used in the visible spectrum field for representing images [47]. The current research techniques [13, 48] have used two main methods for selecting regions of interest: holistic approaches in which the ROI is defined to be the entire face, and modular or facial feature-based approaches, where information is extracted from specific ROIs. Some researchers have located the facial features manually [49, 50] while others have automated the process using both special operators.

Standard feature extraction techniques include Principal Component Analysis (PCA) [51]. L. Trujillo et al., in [52], proposed Eigen-image representation, based on PCA, for each of the recognised facial regions. The construction of Eigen-images is a local and global automatic feature localisation procedure. In the Eigen-image representation, PCA is used to reduce the dimension and interest point clustering in

order to facilitate the estimating of feature facial localisation. Jarlier et al. [53], meanwhile, used a spatial pattern detection procedure, also based on PCA, to extract the features from the representative temperature maps of nine AUs. Yoshitomi [54] transformed the greyscale values of each block of the facial area of an image into frequency components; these were then used to create feature vectors, via a two-dimensional Discrete Cosine Transformation (2D-DCT), and this, in turn, was used to recognise the expressions. Wang et al. [47] employed statistical parameters (minimum, maximum, standard deviation and mean) held in three special matrices — the horizontal, vertical and sequential difference grid-feature matrices — to compute the statistical differences between the onset expression and the apex expression of the same subject. According to Wang et al. [47], however, there are currently few features that have been specifically designed for use with thermal images.

Researchers have focused particularly on four regions of interest where temperatures increase or decrease significantly when the emotion being expressed/felt changes (the forehead, the two eyeholes, and the cheekbone) [46, 55]. Wang et al. and Nakanishi et al. [56, 57] constructed a system which automatically locates four points — the centres of the eyes (two points), the tip of the jaw and the tip of the nose — in thermal images. Based on these points, the facial region is divided into a number of grids, all with the same size. Asada et al. and Yoshitomi et al. [58-60] defined the horizontal and vertical centrelines of the face region and used them to estimate the deviation of the facial image from the standard frontal view. Sugimoto et al. [61], meanwhile, define regions of interest corresponding to the areas surrounding the nose, mouth, cheek and eye regions of the face by using template matching for appropriate localisation.

In order to specify the sections of the face that represent regions of interest, Trujillo et al. [52] apply automatic procedures to localise a local-global feature found in a thermal image by using interest point clustering to estimate facial feature localisation. Hernández et al. [48] proposed a visual learning technique based on evolutionary computation (EC) in order simultaneously to select the region of interest and extract features. The Grey-Level Co-occurrence Matrix was then used to compute region descriptors as well as to select the best subsets of descriptors.

Since each facial expression generates specific facial muscle contractions which produce fluctuations in facial temperature patterns, some researchers have studied and analysed the facial heat patterns concomitant with a particular expression. Khan et al. [62] sought to identify sets of particular Facial Thermal Feature Points (FTFPs) on human faces. The FTFPs were mapped onto the underlying facial muscles, which fluctuate in temperature during a change in expression. These FTFPs were then used as reference points for comparisons between the normal face and the intentional expression. Jarlier et al. [53] used the Facial Action Coding System (FACS), which utilises all visible facial movements to describe facial activity in terms of muscle action units (AUs). They discriminated according to the contraction of particular muscles related to the production of muscle AUs, or combinations of (AUs) that determine a specific expression.

2.2.4. Face Alignment / Registration

Face alignment is a specific topic of image registration which is considered to be an important component in a typical automatic face recognition system. Image registration is the process of adjusting differing images so that they can fit into the same coordinate system. In image processing, this is an essential task which is used to reduce subject

variation, such as the differences in facial configuration. The goal of the face alignment step is to account for variations in head pose, and for inter-subject differences [63].

Face analysis applications use differing registration methods on the detected face to remove rigid motions such as translation, head rotations, and differences in scale. Several facial alignment methods bypass exact localisation of facial landmarks and use only the midpoints of the eyes and mouth to roughly align the faces, while other methods depend on accurate landmark locations. In general, facial alignment techniques can be classified into one of two main approaches: 2D and 3D. One of the 2D methods used is coarse registration. In this method, the distances between the inner facial components such as that between the eyes, are set to be equal in all faces, so as to remove the differences due to translation and scales. The drawback of this simple approach is that it is still sensitive to head rotation and subject variation [64]. To address this problem, another approach uses dense facial points around the eyes and other facial landmarks to register each face with a reference face. The facial points which are not affected by facial expressions are used to learn the transformation. Then the transformation is applied to all the facial points [65].

2.2.5. Face Normalisation

The goal of face normalisation is to remove the gross differences between images or to reduce computational complexity. Normalisation takes place via processes such as rotation [58], resizing [46], and cropping into predefined sizes [57]. The most typical face normalization methods are illumination Since the illumination can vary in different images even in consecutive video frames, especially in real-world environments, this uncontrolled illumination conditions can cause large intraclass variances. Several normalisation algorithms have been used for illumination

normalization such as isotropic diffusion , difference of Gaussian, discrete cosine transform [66] and homomorphic filtering-based normalization [67].

Several normalisation methods have been used specifically in relation to thermal images. Wang et al. [47] removed the baseline temperature, i.e. the one with the highest frequency in the histogram, in order to minimise the influence of temperature differences across the environment and the temperature shift exhibited by the thermal infrared cameras. Wang et al. [47] used four methods to normalise the grey-level values of the images: histogram equalisation, regional histogram equalisation, gamma transformation and regional gamma transformation.

2.3. Methods for Feature Extraction and Classification

Most existing methods to recognise facial emotions can be generally categorised into two approaches: the handcrafted feature approach and the deep learning approach. In the handcrafted feature approach, expert knowledge is used to develop a manually-predefined algorithm to extract features from images. Traditional facial emotion recognition systems rely on many such handcrafted features to recognise both individual features and parts of the face. In contrast, in the deep learning approach, the features are derived from a training image dataset using deep learning methods which effectively use the feedback information to investigate the suitability of the extracted features. Most of the facial emotion recognition methods that are being used currently rely on the deep learning approach, with convolutional neural networks being one example of deep neural networks that can be used to learn deep features. The sub-sections below consider both handcrafted and deep learning-based feature extraction methods in more detail.

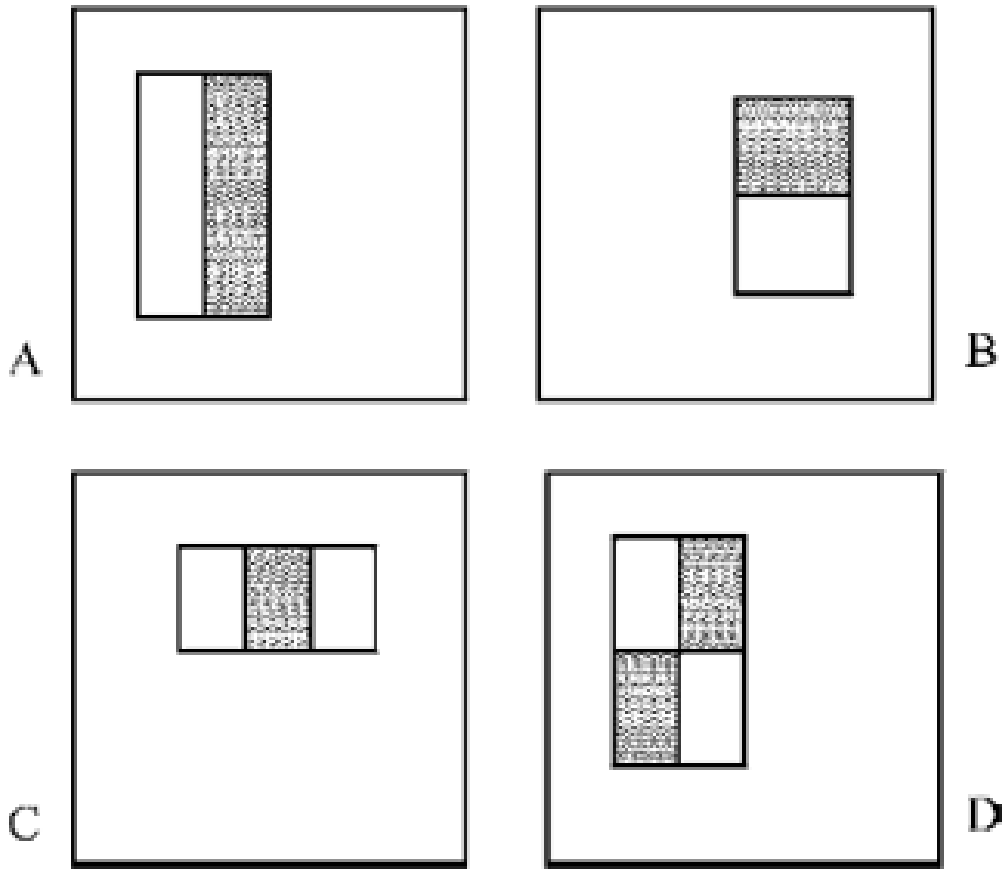


Figure 2-5 : Haar-like rectangle features defined by Viola-Jones: A and B are two-rectangle features, C is a three-rectangle feature, and D is a four-rectangle feature [3].

2.3.1. Handcrafted Feature Extraction

In order to understand what have been done in the literature to recognize facial emotion, a set of well-known handcrafted feature descriptors are briefly described in this section. These descriptors are mainly designed to extract the characteristics of the images such as texture, gradient magnitudes and orientations.

2.3.1.1. Haar-like Features

The basic idea of Haar-like features is to make use of the differences between the summed pixel intensities of rectangular image regions. A rectangle with white and grey

areas is moved over the original image, and the difference between the sum of the pixel values within the grey area and the sum of the pixel values in the white area is calculated. Features from rectangles that have one white area and one grey area are called two-rectangle features. Viola-Jones [3] defined three-rectangle features and four-rectangle features, as shown in Figure 2-5. The Haar-like features indicate certain characteristics of a particular area of the image, such as the existence or absence of edges or changes in texture.

2.3.1.2. Local Binary Patterns (LBP) Features

Local Binary Patterns (LBP) was first proposed as a grey level invariant texture primitive [68]. LBP features describe each pixel by its level of greyness relative to its adjacent pixels. Each centre pixel is represented as a binary string, and its grey value is compared with the grey values of its eight neighbourhood pixels. If the value of the centre pixel is greater than all its neighbours' values, then the value of the centre pixel is set to zero, otherwise to one. The combination of the ones and zeros of the eight neighbouring values are represented as an 8-bit binary number, resulting in there being 28 distinct values for the binary pattern.

Local Binary Patterns from Three Orthogonal Planes (LBP-TOP) [69] is an extension of the LBP which is created by concatenating local binary patterns on three orthogonal planes: XY, XT and YT. The XY plane represents the spatial texture information, while the XT and YT planes represent information about the space-time transitions. Widely used in ordinary texture analysis, LBP-TOP offers efficient representations of dynamic image texture, which is an extension of texture to the sequential domain. Both LBP and LBP-TOP have been successfully applied to facial expression recognition [70-72].

Sun et al. [71] divided the sequences of face images into 4×4 blocks and extracted the LBP features from each block, then concatenated them into an enhanced feature vector which represents the appearance and motion of the facial expression sequence.

2.3.1.3. Local Phase Quantisation (LPQ) Descriptor

The Local Phase Quantisation LPQ [73] descriptor is a texture analysis method based on the Fourier transformation and using phase information computed locally for a window in every image position. The phases of the four low-frequency coefficients are represented in an eight-dimensional space. A histogram of the resulting code is generated and utilised as a feature in texture classification. Since only phase information is used, the method is robust to image blurring and invariant to uniform illumination changes.

Local Phase Quantisation from Three Orthogonal Planes (LPQ-TOP) [74] is an extension to the LPQ operator used for spatial texture analysis. LPQ-TOP is based on the binary encoding of the phase information of the local Fourier transform at low-frequency points. As with the LBP-TOP feature of Sun et al. [71], the sequences of face images are divided into 16 blocks of volumes, then the LPQ-TOP features on each block are extracted and concatenated together.

2.3.1.4. Histograms of Oriented Gradient (HOG) Descriptors

In the context of the detection of human figures/faces, Dalal and Triggs [75] proposed image descriptors that describe a local object's appearance and shape by computing a dense grid of histograms of oriented gradients. Their method divides an image into blocks of various sizes, where each block consists of a number of cells. A local 1-D cantered orientation histogram of gradients is calculated from the gradient orientations

of sample pixels from within each cell. Depending on the values found in the gradient, each pixel within the cell casts a weighted vote into the orientation histogram. Each histogram splits the gradient angle range into a pre-defined number of bins.

Histogram of Oriented Gradients from Three Orthogonal Planes (HOG_TOP) [76] is an extension to the (HOG) to represent the dynamic spatial-temporal features of image sequences. Like LBP-TOP, each location in a sequence has a 3-D (XY, XT and YT). HOG_TOP description is obtained by calculating the gradients along with the 3-D. Then the histograms obtained from the planes (XY, XT and YT) are concatenated to form a global description.

Pyramid of Histograms of Orientation Gradients (PHOG) [77] is a spatial shape descriptor which consists of a HOG over each image subregion at deferent resolution levels. The histograms (vectors) for all levels are concatenated to represent the final PHOG vector.

Bag-of-Words (BoW) Model

The BoW model treats an image as a document by representing image features as words and counting the sparse histogram over the words (local image features). As with document classification, a bag of words is a sparse vector of occurrence counts of words. The BoW model is the most commonly-used of the handcrafted feature extraction methods, both for object recognition [78, 79] and facial expression recognition [80, 81]. It usually comprises three modules: feature extraction, feature encoding and feature pooling. In feature extraction, local features, such as SIFT, HOG and SURF, are used to characterise the local regions. Next, feature encoding is used to make image representation more robust, such as Locality-constrained Linear Coding

(LLC) [82]. To summarise the results of feature encoding, feature pooling reduces the image representation to the common variances. The commonly-used functions for feature pooling are average and maximum pooling.

2.3.2. Deep Learning and Convolutional Neural Networks

Most early methods of object and scene recognition started by applying some well-engineered features to describe the image and then combined these features to produce a feature vector which was subsequently fed into a general-purpose classifier. These methods relied significantly on the researchers being able to design good feature descriptors and ways to combine them. In contrast, given a large amount of image data, deep learning methods learn better feature descriptors and better ways of combining them. Deep learning methods attempt to learn features automatically at multiple levels of abstraction, allowing a system to map the input to the output directly from the data, without depending completely on features designed by the researchers [83, 84]. The most well-known algorithm among various deep learning models is the convolutional neural network (CNN) due to its tremendous success in computer vision applications. CNNs have a powerful learning ability due to the use of multiple feature extraction phases that automatically and adaptively learn the spatial hierarchies of features from the raw data through a backpropagation algorithm.

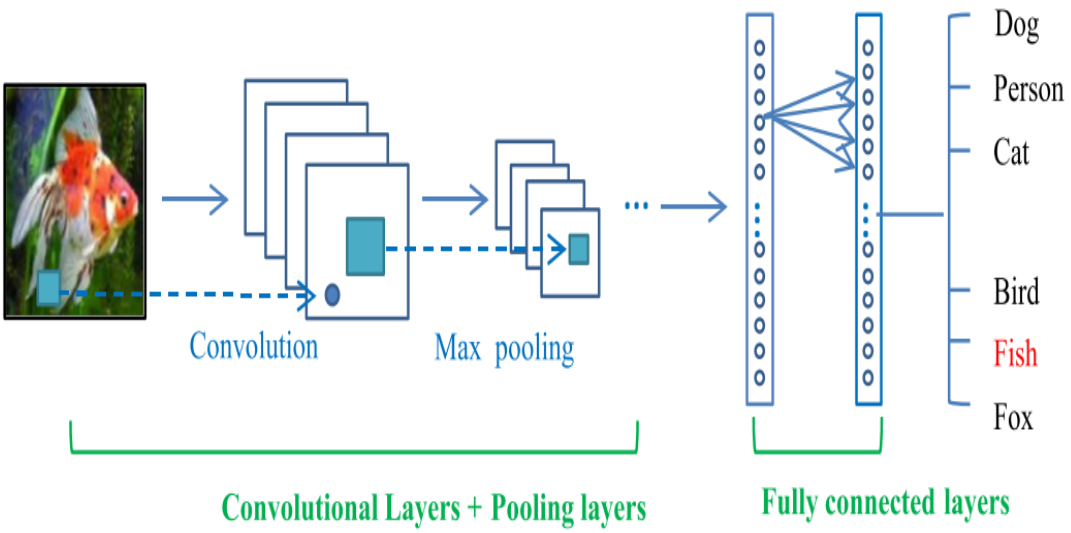


Figure 2-6: The general convolutional neural network architecture pipeline [85].

A CNN is a special model of neural network that uses a mathematical operation called convolution, which is a specialised kind of linear operation. In a traditional neural network, the matrix multiplication operation is applied between the network layers, a matrix of parameters, and separate parameters. This means that there is an interaction between every output unit and every input unit. In contrast, convolutional networks typically have sparse interactions by making the sparse connectivity, called a kernel, smaller than the input. This leveraged the CNN with three important ideas that improved the machine learning system: sparse interactions, parameter sharing and equivariant representations. Moreover, convolution can work with inputs of variable size [86-88]. A CNN architecture comprises three main types of neural layers: convolutional layers, pooling layers and fully connected layers. Each type of layer plays a different role. Figure 2-6 illustrates the three main layers and general CNN architecture pipeline.

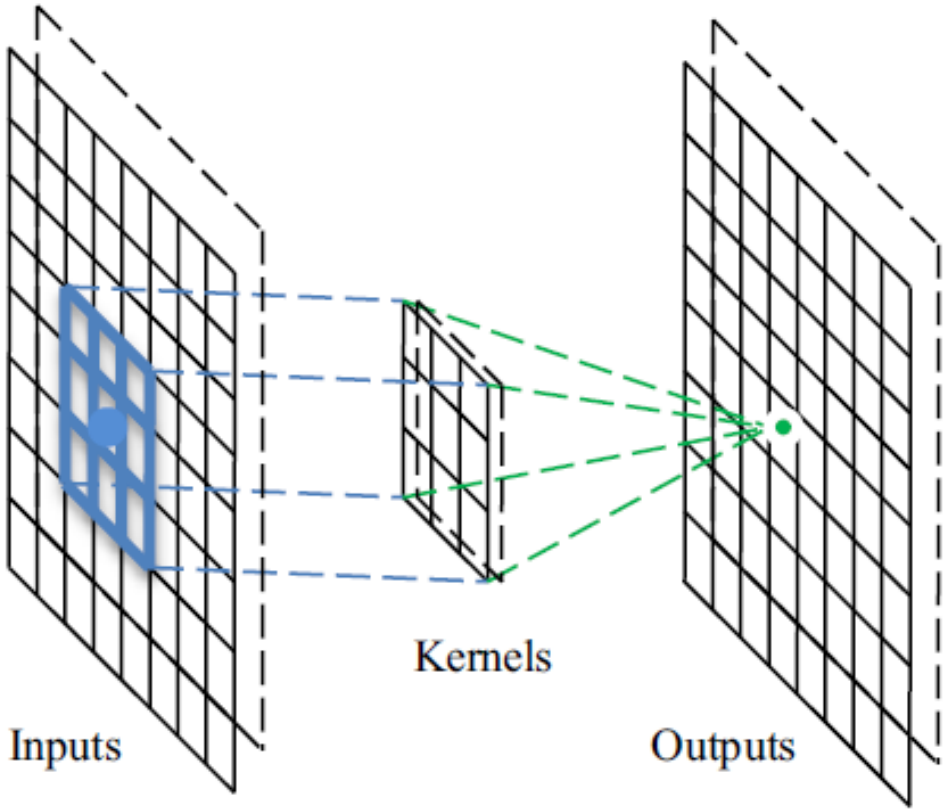


Figure 2-7: How the convolutional layer operates [85].

In *convolutional layers*, a CNN convolves the whole image or the intermediate feature maps by utilising a set of convolutional kernels, or filters, to generate various feature maps, as shown in Figure 2-7. Convolutional kernel operation divides the image into small parts, known as receptive fields. The kernel has a specific set of weights which are multiplied with the corresponding elements of the receptive field to extract its feature patterns [89]. Every convolution layer is followed by an activation function, which is a mathematical equation used to determine whether the output should be activated or not. This process helps normalise the output to a range between (1 and 0) or (-1 and 1). There are several activation functions in the literature, such as sigmoid, tanh, maxout, SWISH and ReLU. The most widely-used function, however, is ReLU, together with its variants (leaky ReLU, ELU and PReLU), because these are better at overcoming the vanishing gradient problem [90, 91].

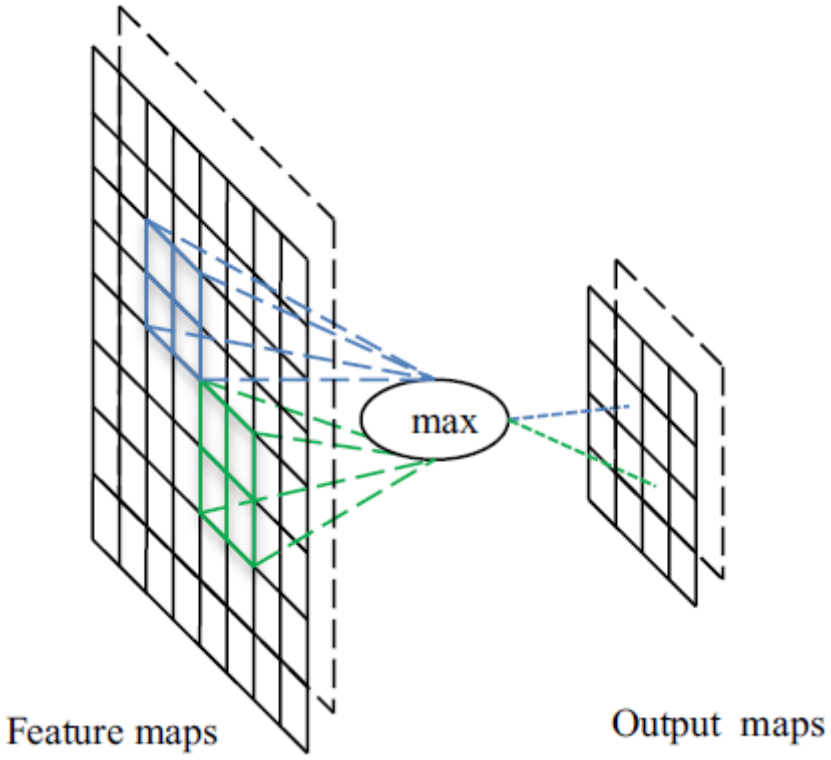


Figure 2-8: How the max-pooling layer operates [85].

In the *pooling layer* subsampling or down sampling operations are performed in order to reduce the spatial dimensions (width × height) of the input for the next convolutional layer without affecting the depth dimension of the volume. Figure 2-8 shows an example of a max-pooling layer which reduces the size of the output map. While this leads to a loss of information, it remains beneficial for the network overall since the decrease in size minimises the computational overhead for the next layers of the network, while also reducing overfitting. The most frequently-used pooling strategies in CNNs are max-pooling and average-pooling [85, 89].

Finally, *fully connected layers* are an essential component of CNNs, performing the high-level reasoning in the neural network. While convolution and pooling layers impart the image into features and analyse them independently, after several iterations of the convolutional and pooling layers, fully connected layers are responsible for the

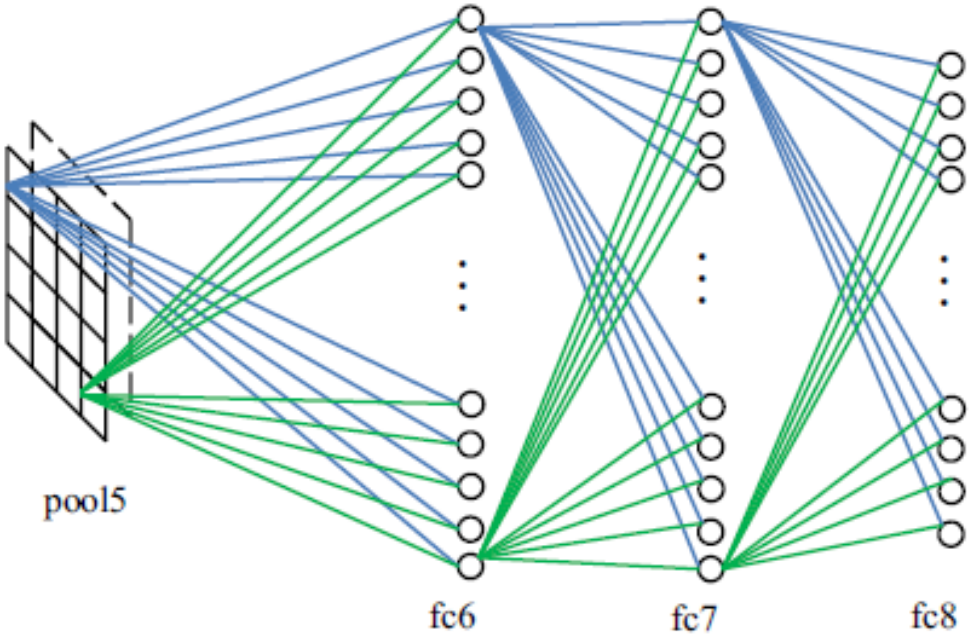


Figure 2-9: How the fully connected layer operates [85].

final classification decision. Fully connected layers eventually convert the two-dimensional feature maps of the previous layer into a one-dimensional feature vector, as seen in Figure 2-9. As their name implies, every neuron in a fully connected layer has full connections to all activations in the previous layer [86, 87]. CNN models are commonly used in computer vision algorithms as both feature extractor and classification mechanisms. CNNs can be utilised as a very effective feature extractor by converting the features map resulting from convolution or pooling layers to a flattened features vector. There are many well-known CNN models which have been key in building computer vision algorithms. The following sub-sections give brief descriptions of the core architecture of those models that we use in our experiments: GoogLeNet, VGGNet and ResNet.

Inception Architecture and GoogLeNet: In the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) 2014 [92], GoogLeNet [93], also known as Inception-V, was the winner, achieving a 6.67% error rate, which was close to human

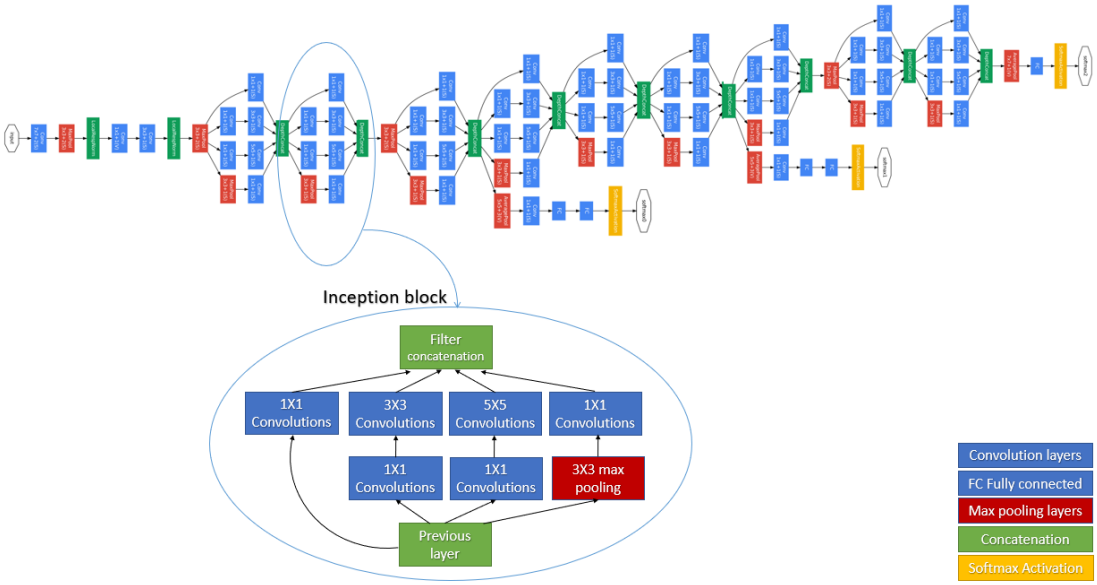


Figure 2-10: The GoogLeNet and its inception block architectures [93].

performance. The GoogleNet model achieved not only high accuracy but also reduced computational cost. Its architecture combines a novel element which is dubbed an inception block. The conception of the inception block is to capture spatial information at different scales by combining filters of different sizes (1x1, 3x3 and 5x5). These multi-scale convolutional layers apply split, transform and merge concepts to overcome problems related to the variations in the resolutions of images present in the same category. In the GoogleNet architecture, not all output feature-maps have a connection to all input feature-maps, hence omitting redundant information and reducing the computational cost. This leads to a drastic reduction in the feature space of the next layer, however, and thus may cause loss of useful information [87, 89]. Figure 2-10 illustrates the architecture of the GoogLeNet and the inception block.

VGGNet: The second place in the 2014-ILSVRC competition was taken by the VGG [94] models proposed by Simonyan and Zisserman. The central concepts of VGG architecture are an increase in the network depth to 16–19 weight layers and the use of very small (3×3) convolution filters. These small size filters significantly improve the

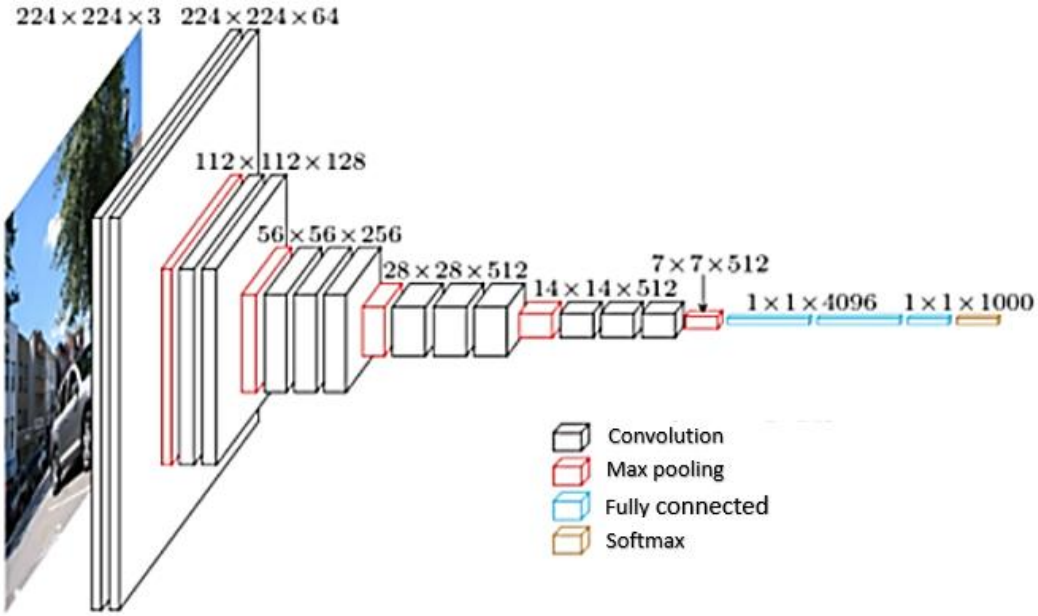


Figure 2-11: The VGG16 architecture [94].

performance of the CNN and reduce the computational complexity. Simonyan and Zisserman experimentally demonstrated that replacing large size filters (11x11 and 5x5) with a stack of small size filters (3x3) could induce the same effect. These findings initiated a new trend in CNN architecture towards smaller size filters. The main drawback of VGG models, however, is the large number of parameters (138 million), which make it difficult to deploy in systems with limited resources [85, 89]. Figure 2-11 shows the VGG architecture.

ResNet: Kaiming He et al. [95] proposed the Residual Neural Network (ResNet), which in fact beat human-level performance by achieving a top-5 error rate of 3.57% at the ILSVRC 2015. They proposed a novel architecture, called shortcut connections, and devised an efficient methodology for the training of deep networks. With 152 layers, ResNet is eight times deeper than VGG while still having lower complexity. Figure 2-12 shows the gated units or gated recurrent units, which are also known as shortcut connections.

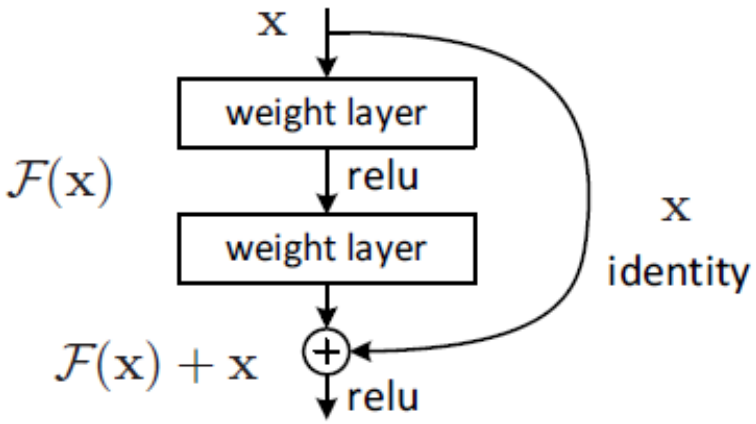


Figure 2-12: Residual learning: a building block [95].

2.4. Methods for Pattern Classification

Once features have been extracted, they can be used as the input to the classification process of a facial expression recognition system. A number of different methods have been proposed for the classification of both visible and thermal images in relation to facial expression recognition. Vyas et al. [96] divided facial expression classification methods into frame-based (image-based) and sequence-based (video-based) methods**,** depending on how the classification was performed.

Image-based methods generally employ static multi-class classifiers to classify emotions into six basic categories; these methods include Artificial Neural Network (ANN), SVM [97], K-nearest Neighbours (KNN) [50, 53] and Linear Discriminate Analysis (LDA) [62, 98]. Image-based methods are mostly utilised to classify spatial features, which refer to the data features extracted from one frame at a time and neglect the temporal features, where the data correlates with a specific time. Video-based methods**,** on the other hand, usually classify both temporal and spatial features from several consecutive frames at a time, such as with Recurrent Neural Network (RNN) [99], Long short-term memory (LSTM) [100], Bidirectional LSTM [101] and 3D-CNN [102, 103].

Zhang et al. [104] propose a deep learning framework called spatial-temporal recurrent neural network (STRNN) to recognise facial emotion. They employed CNN to extract spatial information from frames, and multidirectional RNN to classify the discriminative features characterising the temporal dependencies of the sequences. To extract facial features from temporal sequences, Zhang et al. [105] utilised the Part-based Hierarchical Bidirectional Recurrent Neural Network (PHRNN), which they fed by extracting facial landmarks from four parts based on the facial physical structure. The frameworks of the image-based and video-based recognition systems are described in Section 2.6.

2.5. Facial Expression Databases

A facial expression database is the most important component of facial emotion recognition systems, and a close relationship exists between the advances in emotion recognition algorithms and the availability of facial expression databases which comprehensively represent, in a controlled manner, the varying factors which affect the expression of emotions. Most researchers currently construct their datasets by asking subjects to demonstrate a series of emotional expressions in front of the camera and within tightly controlled environments. Within these limited environments, their recognition systems have attained near-perfect performance but this very lack of diverse subjects and conditions, in fact, hinders the progress of emotion recognition capable of operating ‘in the wild’. Recently, a number of publicly available databases have elicited images and videos encapsulating emotions from the web in order to provide a more comprehensive dataset. Due to the nature of facially-expressed emotion, however, many of these publicly available databases are severely limited [10, 11]. Specifically:

- They provide a limited number of facial images/sequences/videos labelled with accurate expression information such as SFEW 2.0 and AFEW [12].
- Unlike the simple and prototypical emotions labelled in posed emotion databases, the emotions captured in real-life images/videos often present compound, complex [11] or even ambiguous emotions [10]. This is why most of the current emotion databases include only seven very general categories: the six basic emotions (surprise, fear, disgust, happiness, sadness and anger) and neutral.
- The number of labellers available to work on these databases is too small, which reduces the reliability and validity of the emotion labels [12].

The following section focuses on discussing two important dimensions of the existing databases: posed vs spontaneous expressions, and the lab-controlled environment. Then it presents a review of the existing thermal and visible-light facial databases — listed in Table 2-1 and Table 2-2, respectively.

2.5.1. Spontaneous vs Posed Facial Expressions

Although the automatic recognition of emotions in posed, controlled audio-visual displays can achieve reasonably high levels of accuracy, detecting emotions via expressions in less controlled settings is still a very challenging problem because the intentional performance of an emotion (as in the posed settings) differs from spontaneous behaviour in terms of visual appearance, audio profile and timing. Accordingly, the main criticism of the existing facial expression recognition systems is that the methods depend on posed data. Many facial expression databases have used hired “actors” or “portrayers”, instructed to express single-label emotions, sometimes using scripts or restricted scenarios [106]. This inevitably means that the resulting posed facial expressions are exaggerated compared to those that would typically occur

in real life, as there is a lot of variation around the neutral in terms of emotions as they are actually expressed. In short, posed expressions typically operate via exaggerated changes where only slight changes in facial expression would be more natural.

For the above reasons, most of the facial expression recognition systems that utilise databases fail in real-life applications. As a result, researchers have come to realise that, when designing a system for the automated recognition of facial expressions in the real world, the differentiation between spontaneous and posed facial expressions is an issue that must be taken account of. Accordingly, the research in the field started to focus on the automatic analysis of spontaneously occurring behaviour [107, 108]. Furthermore, state-of-the-art research has shown that spontaneous facial expressions provide valuable information from such things as general appearance, timing, head movements and other bodily gestures. Specifically, research is now demonstrating that the temporal information related to how an expression makes its appearance, and the geometric features of some facial regions, those produced by facial action units, can all be applied to discriminate expressions [64, 109, 110].

2.5.2. Lab-Controlled Environment Databases

Most early facial expression databases do not represent real-world conditions as they were captured in a lab-controlled environment where the subjects were controlled, and some important other factors were simplified, eliminated or managed: i.e., illumination, lighting conditions, restrictions on clothing, eyeglasses, etc.

- The number of samples and subjects in such databases is limited. The diversity is also limited, so the samples have low levels of variation in face shape, texture, colour. Also, it must be remembered that facial and scalp hair varies with sex, ethnic background and age. Consequently, most of these databases are unsuitable

for use with deep learning models which need a huge amount of training data.

- Moreover, due to the limited number of subjects, the database might have several samples generated by one particular subject who sequentially acted/posed the same expression. When this sample-similarity is found in training data, the deep learning models are more likely to face overfitting problems. Also, when both the training and testing datasets have this sample-similarity, there is more potential for bias to influence test accuracy.
- Relatively little attention has been paid to the problem of pose invariance in relation to the lab-controlled environment databases. In contrast, in real-world situations, large variations in head position and facial orientation are common and often accompany changes in expression. These natural movements make facial expression recognition systems based on lab-controlled environment databases more difficult to use for real-world applications.
- Another restriction/condition that most lab-controlled environment databases apply is scene simplicity. Static backgrounds are usually de-rigour and/or the use of a consistent pattern and the requirement that only a single person is present. This can influence the accuracy of face detection, tracking, and expression recognition. In natural environments, many people may be present together and interacting with each other. This variation should be represented in training data so to develop and test algorithms that are robust to such variation.

From all of the above, it is evident that lab-controlled environment databases do not represent the wide variety of real-world conditions. To address these issues, researchers have recently started paying attention to databases constructed from images from 'the wild' [11] [10, 111].

2.5.3. Thermal Facial Expression Databases

Table 2-1 lists current thermal facial expression databases along with information regarding the name, the number of subjects, and the expression description and elicitation method related to this.

The first two, the NIST Equinox [112] and IRIS [113] Thermal/Visible Face Database include posed thermal expression images which have been captured by asking subjects to perform a sequence of emotional expressions in front of a camera. Thirdly, the USTC-NVIE [8] database has been used in many types of research focused on thermal images. It is a visible-light, and infrared facial expression database which includes good posed and also spontaneous thermal images.

Table 2-1: Thermal facial expression databases.

Database	Size	Education	Expr. Descript.	Elicitation Method
NIST Equinox[112]	600 subjects 1919 IR images	Posed	Smiling, Frowning, Surprise	asking the subjects to perform expressions
IRIS [113]	30 subjects, 4228 pairs of thermal and visible images	Posed	Surprise, Laughter, Anger	asking the subjects to perform expressions
USTC-NVIE [8]	215 subjects	Posed and spontaneous (30 visible-25 infrared frames per second)	Six basic emotions plus neutral	spontaneous expressions induced by film clips posed images obtained by asking the subjects to perform expressions
Naturalistic Database of Thermal Emotional Facial Expressions[114]	49 subjects 120,000 images from each camera (both the visual and thermal one)	Spontaneous snapshot (one frame per second)	Sadness, Disgust, Happiness, Surprise and Neutral	expressions induced by watching movies with strong emotional content and by playing a memory game
Multimodal Databases for Emotion Analysis [115]	36	Posed and spontaneous (180 visible-60 infrared frames per second)	Neutral, Anger, Amusement, Disgust Fear, Sadness	expressions induced by watching movies with strong emotional content, pictures interviewed
KTFE (A Kotani Thermal Facial Emotion) [116].	26 subjects	Posed and spontaneous (5 frames per second)	Six basic emotions plus neutral	watching movies with strong emotional content

On the other hand, this database embeds inaccuracies due to the procedure used to induce emotions during the data acquisition. Only two-minute gaps were allowed between each emotion clip, and this is too short a time for subjects to re-establish a neutral emotional status before producing the next non-neutral expression. Moreover, changes in skin temperature occur subsequently to changes in emotion, which increases the potential for interaction with the different emotions shown by the other emotional video clips [8].

Fourth, the Naturalistic Database of Thermal Emotional Facial Expressions [56] is very useful for expression recognition, particularly because it captures one visible and one thermal image per second for each subject. Fifth, the Multimodal Databases [115] provide a corpus of emotional responses whereby researchers can discover the internal emotional states of subjects. This database can be used by different types of professionals because it is composed of eight different, but synchronised, kinds of recording: videos (four cameras are used to record the motions of the face, the full body and the temperatures of shoulders and face) pressure sensor, audio signal and photosensor. The gaps between each clip are only three seconds long, however. Sixth, the researchers who proposed the Kotani Thermal Facial recognition approach (KTFE) [116] analysed a visual and a thermal database together in order to enable the concurrent recording of both expression information and thermal information and thus better recognition of emotions.

2.5.4. Visible Facial Expression Databases

A summary of the existing visible image databases is given in Table 2-2; this shows the main reference, the number of samples, the age range, the collection environment, the expression distribution, the annotation method and additional information.

Table 2-2: Visible facial expression databases.

Database	Samples			Subject		Expression			Annotation Method
	Image	Seq	Env	NO	El	Ba	Com	Coll	
AR FACE [117]	4000		Lab	116	P	4		Acted	By subject
JAFFE [118]	219		Lab	92	P	7			Semantic ratings over 60 subjects
MMI [119]	1500	169	Lab	19	P	7			2 Coders FACS coded
GEMEP [106]		7K	Lab	10	P	18		Acted	28- 30
CK+ [120]		327	Lab	123	P & S	7*		Acted	2 Coders
RaFD [121]	8040		Lab	49	P	8		Acted	276 Percentage of agreement on emotion categorization
Multi-PIE [122, 123]	755370		Lab	337	P	6		Acted	
FER-2013 [124]	35887		Web		S	7		Search	Image
EmotionNet [125]	1M		Web		P & S	7	17		10% manually 90% automatically
AffectNet [10]	450K		Web		P & S	8			1 L/I
RAF-DB [11]	29672		Web		P & S	7			Distribution values from about labellers per image
iSAFE [126]	395		video	44	P	7			Professional and unprofessional annotator
SFEW 2.0 [12]	1635		Movie	330	P & S	7			2 L/I
AFEW [12]		1,426	Movie	330	P & S	7	-		

Seq=Sequences, Env=Environment, NO=number, El=Elicit, Ba=Basic, Com=Compound, Coll=Collection Method, P=posed and S=spontaneous. 7 basic expressions (6 basic expression + Neutral) --- 8 basic expressions (6 basic expression + Neutral + Contempt) *(6 basic expression + Contempt)

For the AR Face Database [117], the illumination conditions, and the distance from the camera to the subject, were strictly controlled throughout the whole image capturing process. The AR database has 4000 images; a total of 116 participants, 63 of whom were males and 53 females. Its images have a resolution of 768 × 576 pixels, and each pixel has 24 bits of depth.

One of the earliest static facial expressions datasets was the JAFFE Database [118], which has been extensively used in expression research. Each subject used for this was asked to tie her hair away from her face in order to expose all the expressive zones of the face. Moreover, the researchers positioned Tungsten lights so as to create even illumination on the face and to reduce back-reflection; a box enclosed the region

between the camera and a plastic sheet. The database contains 219 images of ten Japanese females. The subjects posed for six expressions each (anger, disgust, fear, happy, sad and surprise) and for the neutral expression.

The subjects used in the MMI Facial Expression Database [119] were instructed by an expert (a FACS coder) on how to display the required facial expressions. They were asked to display 79 series of expressions, which each included either a single AU or a combination of AUs. The subjects had to display the required expressions while minimising out-of-plane head motions. To allow easy access and ease of search, the MMI has a web-based direct-manipulation application. It contains more than 1500 samples, both static images and sequences of images, all of frontal or profile view faces, of 19 male and female subjects expressing various emotions facially.

The Geneva Multimodal Emotion Portrayal (GEMEP) [106] consists of more than 7,000 audio-video emotion portrayals, representing 18 emotions (including some rarely-studied subtle emotions). The subjects were professional theatre actors who were coached by a professional director. Ten actors were recruited for the scenarios and had the help of a director who supervised the acting during the recordings. Each emotion was represented through audio-only, video-only and audio-video portrayals, and 90 labellers were randomly assigned across these types of portrayal (31, 31, 28, respectively in the three categories).

The Extended Cohn-Kanade (CK+) Database [120, 127] includes 327 sequences, which were captured in a lab-controlled environment from 123 subjects. These were of varying ethnic backgrounds: 81%, Euro-American, 13% Afro-American and 6% other; their ages ranged from 18 to 50. They were instructed by an experimenter to perform a series of 23 different facial expressions, which included single AU expressions and

expressions which required combinations of AUs. The image sequences in CK+ vary in duration (i.e., from 10 to 60 frames), size (640 × 490 or 640 × 480 pixel arrays with 8-bit grayscale or 24-bit colour values) and views (frontal and 30-degree). These sequences started from the neutral expression and ended with a peak expression, which was one of the six basic expressions, plus contempt.

The Radboud Faces Database (RaFD) [121] contains portrait images of 49 subjects: 39 adults and ten children. The portrait images show eight facial expressions with three gaze directions; all of the portrayed facial expressions were based on prototypes from the FACS. Varying poses and illuminations were used, and the images were captured simultaneously from five different camera angles; three flashes provided the illumination. The images were captured in a highly controlled environment, and all were aligned, cropped and resized to 1024 × 681 pixels. For the validation, all images were rated by 315 labellers of nine randomly chosen subjects from the database. For each image, labellers rated the expression in terms of intensity, clarity, valence and the genuineness of the expression. Further, to select images from the dataset with a specific property, labellers were asked to rate the attractiveness of the neutral frontal gaze images for all nine models.

CMU Multi-PIE Datasets [123] contains both temporal and static samples recorded in the laboratory over five sessions. To address some of the limitations of the lab-controlled environment, the authors utilised a system of 15 cameras and 18 flashes connected to a set of PCs so as systematically to capture images of varying poses and illuminations across large numbers of subjects and samples. In total, the Multi-PIE database contains 755,370 images from 337 different subjects. The emotions portrayed were smiling, surprised, squinting, disgust, screaming and neutral.

Facial Expression Recognition FER-2013 [124] was introduced at ICML 2013 for the Facial Expression Recognition Challenge; it was designed to assist in the classification of the emotions expressed in photographs of the human face. FER-2013 contains 35,887 images collected and automatically labelled by the Google image search API. Using a set of 184 emotion-related keywords, such as “blissful”, “enraged,” etc, the API searched for images of faces that matched with these keywords. After the faces were automatically cropped from each image, labellers checked the image labels, corrected the cropping if required, and removed duplications. The approved images were then resized to 48 × 48 pixels and converted to grayscale. Unfortunately, the images are difficult to register well, and most facial landmark detectors are unable to extract facial landmarks from them at the provided resolution and quality. The images in FER-2013 represent six basic emotions plus neutral and a small number of images portraying disgust (547 images). It does not provide information about facial landmark locations, and only the categorical model of emotion is provided.

EmotioNet [125] is a large-scale facial emotion database employing a categorical model. It contains one million images that can be queried by specifying an AU, the AU’s intensity, the emotion category or an emotive keyword. The images were downloaded from the Internet by searching for images associated with texts which contained any of the words which can be derived from the word “feeling” via WordNet [128]. The authors presented a novel AU and AU intensity detection algorithm and used it to automatically construct and annotate 90% of the collected images of facial emotion. The remaining samples (100,000 images) were manually annotated with AUs by experienced coders, so as to assist in recognition of AUs and AU intensities in images of faces. EmotioNet uses six basic emotions plus neutral and 17 compound emotions. The accuracy of the automatically-annotated AUs has been estimated at

about 81% in relation to the manually-annotated selected group of 3000 images. EmotioNet is considered a novel source of FACS models found in the wild with a large number of samples and significant subject variation. The authors did not describe in detail the manual annotation procedure employed for the 10%, i.e., 100K, a subset of facial expression images, however, or the number of labellers/coders who were engaged in this. Also, the emotion categories are judged based on the AU label and not directly manually annotated; this reduces the reliability and validity of the emotion labels.

Another large-scale facial emotion database is AffectNet [10], which also applies categorical and dimensional models. This database contains more than one million images obtained from the Internet by querying three search engines (Google, Bing and Yahoo) and specifying emotion-related keywords in six different languages: English, Spanish, Portuguese, German, Arabic and Farsi. A total of 450,000 images were annotated by twelve expert annotators in order to label the face in the images using both discrete categorical and continuous dimensional (valence and arousal) models.

The rest of the images were automatically annotated by a software application developed for the purpose, again using both the categorical and dimensional models of effect. Furthermore, alongside the six basic emotions plus neutral, four other discrete categories were defined for the categorical model: Contempt, None, Uncertain and Non-face. The Non-category encompassed types of emotion not provided with a specific category, such as sleepy, bored, tired, seducing, confused, ashamed, focused, etc. The Non-face category was assigned to images which either do not contain a face, but instead a drawing, animation or painting, or which contain a face obscured by a watermark. An image was also tagged as Non-face if the face was distorted beyond

what would be considered a natural or normal shape for a face, or if the face detection algorithm otherwise failed to detect the face boundaries, even if an expression could be inferred from it. The Uncertain category was assigned to images where the annotators were completely uncertain about all of the facial expressions shown in it. Each image was labelled by only one annotator, however, due to time and budget constraints, and compound expressions were not included. Moreover, the class distribution of the training dataset was heavily imbalanced; that is, most of the images are in the majority classes (Happy 146198, Neutral 80276) and relatively few images are in minority classes (Disgust 5264, Fear 8191).

The Real-World Expression Database RAF-DB [11] contains 29,672 real-world images which were automatically downloaded from the Internet by an optimised algorithm and picked out using emotion-related keywords. To assure the reliability of the labelling of the collected images, sufficient well-trained labellers (315 annotators) were engaged to annotate each image independently about 40 times, utilising a website developed for the purpose. The images are divided into two different subsets: basic emotions and twelve classes of compound emotions and labelled with different expressions, age ranges, genders and posture features. Subjects in the RAF-DB database ranged in age from 0 to 70 years old, 52% of them were female, 43% male, and in 5% of cases, the gender was unclear. In terms of racial distribution, there were 77% Caucasian, 8% African American, and 15% Asian. The data are provided with precise locations and the size of the face region, as well as five manually-located landmark points on the face and 37 landmarks, the latter having been automatically annotated. Furthermore, the database includes four features of each image for training and testing sets: HOG [75], Gabor [129], base DCNN [11] and DLP-CNN [11] features. Even though the number of samples is fairly small compared to the number in the other

web-based facial expression databases, such as AffectNet and EmotioNet, with its valuable and reliable metadata, the RAF-DB database is considered to be a useful benchmark resource for facial emotion/expression analysis researchers.

The Indian Semi-Acted Facial Expression (iSAFE) [126] dataset contains 395 video clips of 44 subjects aged 17 to 22 years. These subjects were of two ethnic backgrounds; Indo-Aryan and Dravidian (Asian). The video clips were captured in a lab-controlled environment, but the subjects were not instructed to act. To capture the spontaneous facial expressions of the subjects, they were asked to watch a few stimulant videos and label their emotions with seven basic expressions plus uncertain. Then their video clips were annotated by a professional annotator, who was a psychologist trained in assessing the human emotions, as well as by an unprofessional annotator. iSAFE is very small in size, however, with a limit number of annotators.

Dhall et al. [12] released Acted Facial Expressions in the Wild (AFEW) and Static Facial Expression in the Wild (SFEW 2.0). AFEW is a dynamic, temporal facial-expression database comprising short video clips of facial expressions in close to real-life environments, whereas SFEW is a static subset which was created by selecting some of the frames from AFEW. Both databases were introduced for the EmotiW 2015 Challenge [111] and form the basis of two facial expression recognition sub-challenge solutions, focused on audio-video based and static image-based emotion recognition.

The source of these databases were 54 selected movies. To annotate these movies quickly, a video clip recommender system based on subtitle parsing was utilised to scan a full movie and recommend short clips, all of which were said to have a high probability of showing a subject manifesting a meaningful expression. Then the suggested video clips were annotated with six basic expressions plus neutral by two

independent labellers. The chosen movies covered a large number of actors, although many of these appear in multiple movies in the AFEW dataset — to enable researchers to study how their expressions changed over time. The clips contain varied scenes with indoor, night-time and natural outdoor illumination. Although movies filmed in studios have controlled illumination conditions, even for outdoor settings, they are still closer to real-life than videos made in lab-controlled conditions. AFEW/ SFEW contain 1,426 video clips/ 1635 images of 330 subjects aged 1-77 years. The databases cover unconstrained facial expressions, varied head poses, occlusions and differing resolutions of faces. The number of samples in the training dataset of AFEW/ SFEW is quite small, however, and also the class distribution is imbalanced.

2.6. Image-Based vs Video-Based Emotion Recognition

With the advances in GPUs and machine learning techniques, the robustness of image-based recognition systems has been significantly boosted. On the other hand, what such video-based recognition systems can achieve today is still greatly inferior to human perception. Human subjects are able to look at a video and instantly know what actions are being performed and/or what emotions are being portrayed in it and are able to detect the ways in which these emotions change. Due to their fast and accurate visual system, human subjects can perform complex tasks, such as interactions with many people at the same time or driving, with little conscious thought. Fast, accurate algorithms for recognition systems would allow computers to, for instance, be able to explain scene information in real-time to human users and so reduce the need for specialised sensors for complex tasks such as driving cars and would also unlock the potential for general purpose use.

Current studies treat video's temporal information by splitting a whole video either into groups [102, 103] or individual frames [42] and consequently process these portions multiple times. The next process utilises several models to aggregate the processed parts to implicitly infer the whole temporal information. Based on the number of frames processed at a time, the current state-of-the-art models for video-based recognition fall into four categories: Single-Frame, Set-of-Frames, All-Frames and Key-Frames, as illustrated in Figure 2-13.

***Single-Frame processing approaches*:** Here, the features of each frame are individually extracted and classified by utilizing handcrafted feature methods or/and 2D-CNNs models, then appropriate aggregation methods- decision fusion- are applied directly on the classifications of these frames. Otherwise, the extracted features of frames are combined by using features fusion methods [42], then classified, as illustrated in Figure 2-13 (a).

***Set-of-Frames processing approach*:** The video frames are divided into sets, then a suitable model, such as 3D-CNN, is used to extract, and then classified according to the features of each set. To predict the final classification, a decision fusion method is applied by employing a classification of the sequence of all sets. 3D-CNN models have been used widely for this [103]— to learn both the spatial and the temporal information simultaneously, as illustrated in Figure 2-13 (b).

***All-Frames processing approach*:** The sequence of all video frames are used directly in one model, such as 3D-CNN [130] or optical flow[131] to predict the final classification In this approach, no decision fusion method is needed, as illustrated in Figure 2-13 (c).

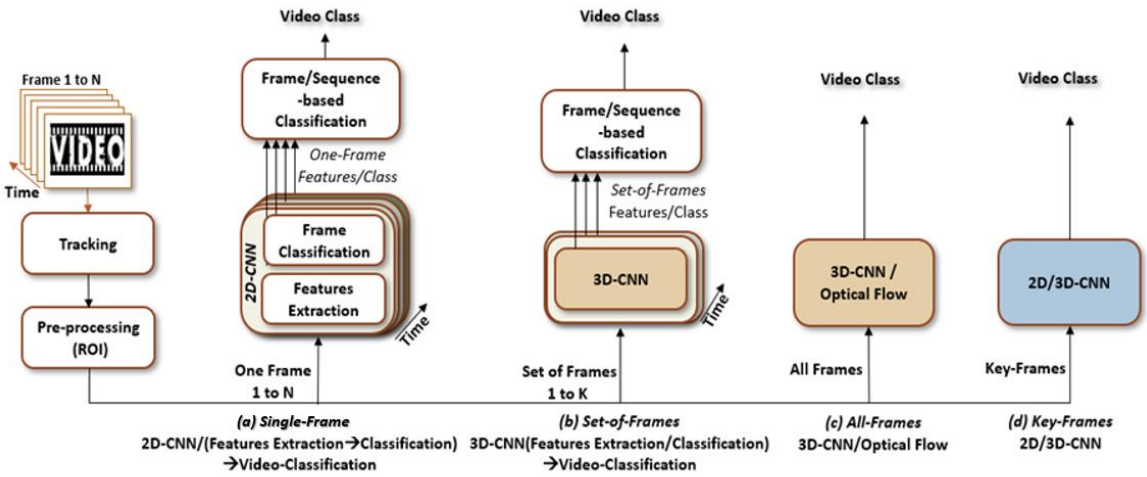


Figure 2-13: Categorization of video-based recognition approaches based on the number of frames processed at a time.

***Key-Frames processing approach*:** A small number of frames are selected to be key-frames and represent the video as one set. Then a suitable model, such as 3D-CNN or 2D-CNN [9], is used to obtain the final decision directly, as illustrated in Figure 2-13(d).

The fusion of spatio-temporal information is classified into three levels: *Feature-level fusion* combines the extracted features through specific approaches before classification, depending on whether the extracted feature is a feature vector or a descriptor. *Decision-level fusion* uses the scores generated from multiple classifiers to obtain the final decision by using multiple layers through a rule-based scheme such as AND, OR, MAX, and majority voting, or in a pattern classification sense, in which the scores are used as new features [179]. *Data-level fusion* combines the original data before using any feature extraction approaches.

The fusion level is selected dependent on the characteristics of the data and the requirements of the application problem. Single-Frame processing approaches use the feature-level and/or decision-level fusion models. Set-of-Frames processing approach may combine data-level fusion with feature-level and/or decision-level fusion models.

In contrast, All-Frames and Key-Frames processing approaches depend on data-level fusion to directly fuse and classify the spatio-temporal information by utilizing a single model such as 3D-CNN.

2.7. Emotion Recognition Algorithms

Shan et al. [132] proposed a Boosted-LBP feature extractor combined with an SVM for classification. Boosted-LBP features are the most discriminative LBP features obtained by adopting AdaBoost to learn from a large LBP feature pool. Liu et al. [77] proposed a Boosted Deep Belief Network (BDBN) to perform feature learning, feature selection and classifier construction for emotion recognition. Different DBN models for unsupervised learning of features in audio-visual emotion recognition have been compared to the work done by Kim et al. [133]. Li et al.[42] used CNNs on images collected from the web. They compared the performance of these CNNs on the CK+ dataset to the state-of-the-art methods to prove the effectiveness of CNNs. Many interesting approaches have also been proposed for audio-video based emotion classification challenges (EmotiW).

To solve the emotion recognition problem in video analysis, Kahou et al. [81] combined multiple deep neural network architectures, each based on a different data source. For frame-based classification, they used an architecture similar to the AlexNet model to classify aligned images of faces, and a shallow network to extract features of the mouth, which were then used as input to an SVM emotion classifier. For the audio information, they developed a deep belief net (DBN) to classify the audio signal available with the video clips. For the spatio-temporal information within the entire scene, they used a deep autoencoder-based classifier to model the spatio-temporal properties of human activity. They used an SVM for combination across frame

predictions, and a multilayer perceptron (MLP) trained on sequence-level features to provide the final video classification. Their architectures accuracy was 41.03% on the test set, the highest accuracy in the EmotiW 2013 [134] challenge.

Sikka et al. [72] followed a feature fusion approach based on MKL which was employed to find an optimal combination of audio and eleven visual features for input into a non-linear SVM classifier. They utilised four different models to extract ten handcrafted visual features: HOG with four and eight bins, PHOG with four bins, BoW on the extracted faces with dictionary sizes of 200, 400 and 600, BoW computed on the entire image, LPQ-TOP with block sizes of 5, 7, and 9. Their classification accuracy was 37.08 on the validation dataset.

Liu et al. [135] represented all frames of a video clip as an image set and modelled it as a linear subspace to be embedded in Grassmannian manifold. After the features were extracted from each video frame using CNN, Class-specific One-to-Rest Partial Least Squares (PLS) was employed to learn on video and audio features separately so as ultimately to distinguish between classes. Then, to find the optimal fusion of classifiers from both modalities (video and audio), a linear fusion was conducted at decision level by introducing a weighted term λ. The final accuracy achieved on the validation and test set was 35.85% and 34.61%, respectively.

The EmotiW 2014 challenge organisers provided aligned face images which were extracted from a selection of video clip frames. Liu et al. [136] applied three kinds of feature-extracting methods on these images: HOG, Dense SIFT and DCNN. They utilised a frame-feature fusion approach which combined the extracted features from successive image frames as a feature vector. Then, the combined feature vectors of the video sequences were represented by three types of image set models: linear subspace,

covariance matrix and Gaussian distribution. Then, different Riemannian kernels were utilised on these models correspondingly for similarity/ distance measurement. Three types of classifiers, kernel SVM, logistic regression and partial least squares, were investigated for comparison. Finally, a score-level fusion of classifiers learned based on different kernel methods, and different modalities (i.e. video and audio) was conducted in order to improve the performance further. Their pipeline achieved a 50.4% classification accuracy on AFEW 4.0 test dataset.

In order to improve training efficiency, Sun et al. [71] applied a PCA model to automatically select a better set of facial images which were extracted from the video's frames. Several visual features (SIFT, LBP-TOP, PHOG and LPQ-TOP) were extracted from the selected image set. The primer classifications of SVMs which were trained on the audio and visual features were hierarchically combined to provide the final classification. In their second submission, they applied the feature-level fusion MKL method with BoW and Audio features. Their methods achieved 47.17% on AFEW 4.0 test dataset.

Chen et al. [76] proposed a feature descriptor called (HOG_TOP) to extract the dynamic visual features from video sequences, and further adopted MKL to find an optimal combination of the visual and audio features. An SVM with multiple kernels was trained for classification. Their methods achieved overall classification accuracies on AFEW 4.0 validation and test datasets of 40.21% and 45.21%, respectively.

Yao et al. [137] proposed a novel pair-wise learning strategy to discriminate two particular emotion categories. The method automatically seeks a set of facial image patches using an undirected graph structure, which takes learnt facial patches as individual vertices to encode feature relations between any two learnt facial patches.

Then a robust emotion representation was constructed by concatenating all task-specific graph-structured facial feature relations sequentially. In their highest accuracy submission, they used the average of three linear SVMs trained with AU-aware facial feature relations (two face scales), audio model and CNN model. Without using an additional database, their method achieved competitive results on SFEW 2.0 and AFEW 5.0 test datasets for both sub-challenges: the image-based static facial recognition accuracy was 55.38%, compared to 53.80% for the audio-video based expression recognition. They did not specify the performance of their CNN and the accuracy of each model separately, however.

Fan et al. [103] won the EmotiW 2016 challenge by proposing a hybrid network that combines three models cascaded 2D-CNN with LSTMs, 3D-CNNs with RNN and an audio module. Their recognition accuracy is 59.02%, and the accuracy of the fused two visual models (CNN-RNN and C3D) can be reached 48.30%. Without audio information, the accuracy of the best single CNN-RNN model was 45.43%, while, the accuracy of the single C3D can reach only 39.69%.

To capture spatio-temporal information, Ouyang et al.[138] employed CNNs for feature extraction and directly contacted the nodes in lower CNN layers with LSTMs. They utilised three different models: VGG-LSTM, ResNet-LSTM and C3D Network. Deep neural network (DNN) was applied for emotion recognition of audio signals. Their overall accuracy was 57.2% on the test dataset and 54.2% on the validation dataset. The accuracy of their visual models, VGG-LSTM, ResNet-LSTM and C3D Network, were 47.4%, 46.7% and 35.2%, respectively, on AFEW 7.0 validation dataset. Likewise, Ouyang et al.[138], Vielzeuf et al.[139] proposed VGG-LSTM and C3D model, but, they utilized C3D model as features extractors for a set of consecutive

frames and LSTM was utilized to classify the combined features of all sets to gather. Then, they combined the scores for the visual models with the audio model to give the final classification. The reported accuracy for VGG-LSTM was 48.6% and C3D-LSTM was 43.2% on AFEW 7.0 validation set where the overall accuracy was 58.8 % on the test dataset.

2.8. Summary

This chapter has surveyed the important work and recent research on computer vision-based emotion recognition, encompassing both image-based and video-based classification design. Starting with the types of emotion explanation which the studies in computer vision use to describe the emotions, this chapter is organised by following the general design of image and video classification systems: Pre-Processing→(Feature Extracting and Deep learning and Convolutional neural network)→Classification. As the database is considered to be the most important component of any recognition system, an intensive survey was presented of the available visual and thermal facial expression databases, along with a comparison between the databases of spontaneous and posed facial expressions. The chapter also discussed the major disadvantages of a lab-controlled environment database which prevent them from representing the wide variety of real-world conditions.

An overview of the important differences between image-based and video-based recognition systems and the challenging problems faces these systems were also offered, followed by a review of the recent state-of-art algorithms and their performance in different databases. Some important challenges highlighted in this chapter will form the basis for the work in the following chapters of this thesis.

Chapter 3

Accuracy Enhancement of the Viola-Jones Algorithm for Thermal Face Detection

3.1. Introduction

Human facial analysis is an active research area due to its wide variety of potential applications, such as face recognition, emotion recognition and human-computer interaction. While standard visible spectrum cameras are regularly used as the sensors in these applications, there has recently been an increased interest in facial analysis applications in the thermal infrared spectrum, since these can combat some drawbacks of the visible spectrum and provide a higher level of liveness detection such as face temperature, emotions and health state [140] [3].

A necessary step towards an automated human facial analysis system, in any modality, is face detection. Considerable work has been done on developing face detection methods in the visible spectrum, whereas face detection in the thermal spectrum has received less attention. The current thermal face detection approaches do not suit real-time applications. To detect faces from thermal images, some related issues should be addressed. For example, detecting faces with eyeglasses is a challenging issue, and the presence of other heat-emitting objects may cause false positive face detections. Figure 3-1 shows Samples of the thermal facial images from the NVIE [8] database.

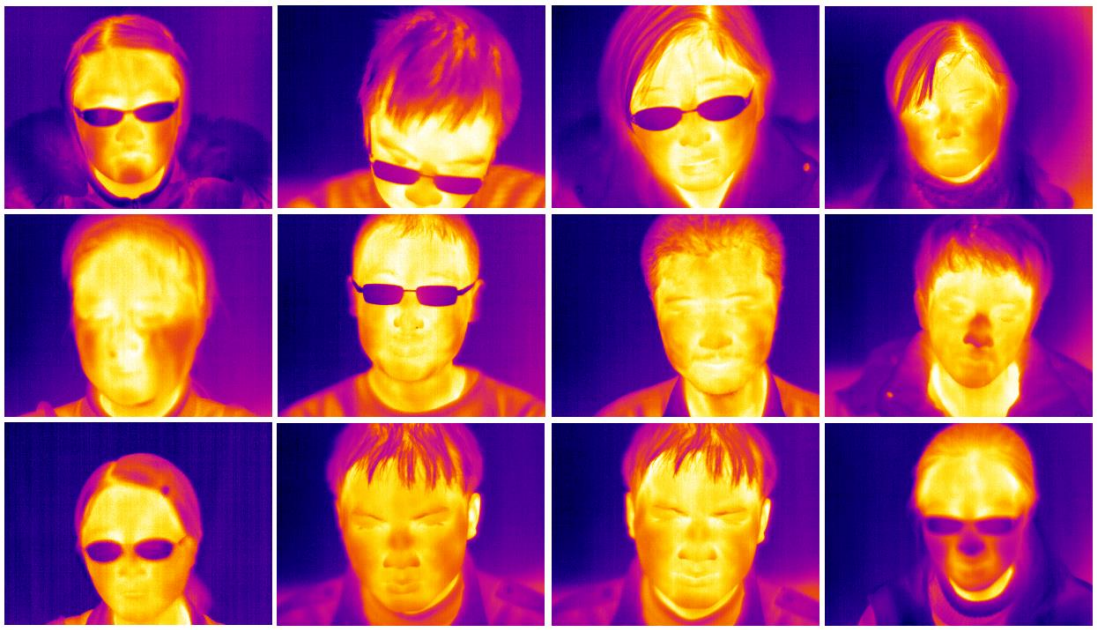


Figure 3-1: Samples of the thermal facial images from the NVIE database [8].

Reese et al. [140] compared the Viola-Jones [3] algorithm with Gabor feature extraction and classification [141] and the non-training method, Projection Profile Analysis [140] in respect to face detection from both thermal and visible images on Alcorn State University (ASU), the University of Notre Dame (UND), and the FERET databases [142]. Reese et al.'s study concluded that the Viola-Jones algorithm achieved better performance, with an average accuracy of 74.8% on thermal images from the ASU and UND databases (78.99% ASU, 70.62% UND) and an average speed of 0.03 seconds per image (0.05 ASU, 0.01 UND). Moreover, they suggested that more work could be done to improve the accuracy of the Viola-Jones algorithm for face detection from thermal infrared images.

In that context, we aim to create a thermal face detection that can meet the requirements of real-time applications described in Section 3.1. The main contribution of this chapter is an efficient and effective process to improve the performance of the Viola-Jones algorithm for face detection from thermal images, with or without eyeglasses. The enhancement process reduced the detection time of the Viola-Jones

algorithm by roughly a factor of two while retaining high detection accuracy [2]. The potential benefits of the process are also investigated. Firstly, the performance of two pre-processing methods, Otsu's method [143] and Gradient Magnitude, are compared. Secondly, the performance of LBP features [144] and HOG descriptors [145] was compared with the performance of the Haar-like features that were originally utilised by the Viola-Jones algorithm.

Some related work is reviewed in Section 3.2. The rest of this chapter is organised as follows. The proposed methods are introduced in Section 3.3. The experimental setup and databases used are described in Section 3.4. The experimental results are presented in Section 3.5 and followed by a discussion in Section 3.6. Section 3.7 draws conclusions.

3.2. Related Work

Since thermal cameras can capture facial skin temperature, many algorithms for thermal facial analysis use different threshold values to separate areas of interest from the background**.** Cheong et al. [144], Mekyska et al. [145], and Wang et al. [47] convert a thermal image into a grey scale image, and binarise it using Otsu's method [143], before identifying the location of the head region using the global minimum point from the horizontal projection of the image. Trujillo et al.[52] and Wong et al.[146], meanwhile, proposed a non-training face detection algorithm that utilises head curve geometry. The algorithm uses a threshold method on the red component extracted from the RGB image to generate a binary image and then performs a morphological closing on this. Reese et al. [140] used the projection profile analysis algorithm for thermal face detection. To separate areas of interest from the background, they used region growing segmentation, but this is slow and lowers the accuracy slightly.

Reese et al. [140] compared three face detection algorithms for thermal and visible images: the Viola-Jones algorithm [3], Gabor feature extraction and classification [141] and non-training Projection Profile Analysis [140]. Reese et al.'s experiments [140] showed that learning-based methods (Viola-Jones and Gabor) are able to detect faces from both thermal and visible images, with the Viola-Jones algorithm being the best for face detection in the thermal spectrum.

The current thermal face detection algorithms require critical three conditions to be fulfilled. For high performance, there should be only one person existing in the image, and no other heat-emitting objects should be captured by the camera, such as the subject's hands. Moreover, the face should be full frontal to the camera, similar to those in a passport [146 ,140].

3.2.1. Viola-Jones Method

Since Viola-Jones' face detector [3] is the most popular and state-of-the-art method for face detection, it is adopted as the baseline method to detect faces from thermal images. Their contribution is three-fold. First, in order to achieve fast calculation with high accuracy, they developed a simple and efficient classifier that used the AdaBoost learning algorithm to choose a small number of critical visual features from a large set of potential features.

Figure 3-2 shows the pseudocode for the AdaBoost algorithm adopted by [147]. Second, their method for combining classifiers in a cascade allows background regions of the image to be easily discarded while spending more computation time on promising face-like regions. Third, they introduced the concept of the "integral image", which allows Haar-like features to be computed very quickly.

- Given training images $(x_1, y_1), \ldots, (x_n, y_n)$ where x represents features and y represents class label. $y = 0,1$ for negative and positive examples respectively.
- Initialize weights $w_{1,i} = \frac{1}{2m}$ for $y_i = 0$, and $w_{1,i} = \frac{1}{2l}$ for $y_i = 1$, where m and l are the number of negative and positive respectively.
- For $t = 1, \ldots, T$:

1. Normalize the weights,

$$w_{t,i} \leftarrow \frac{w_{t,i}}{\sum_{j=1}^{n} w_{t,j}}$$

So that w_t is a probability distribution.

2. For each feature j, train a classifier h_j which is restricted to using a single feature. The error with respect to w_t is measured by $\epsilon_j = \sum_i w_i \left| h_j(x_i) - y_i \right|$.
3. Choose the classifier, h_t, with the lowest error ϵ_t.
4. Update the weights:

$$w_{t+1,i} = w_{t,i} \beta_t^{1-e_i}$$

where $e_i = 0$ if example x_i is classified correctly, $e_i = 1$ otherwise, and $\beta_t = \frac{\epsilon_t}{1-\epsilon_t}$.

- The final strong classifier is

$$h(x) = \begin{cases} 1 & \sum_{t=1}^{T} \alpha_t h_t(x) \geq \frac{1}{2} \sum_{t=1}^{T} \alpha_t \\ 0 & otherwise \end{cases}$$

where $\alpha_t = log \frac{1}{\beta_t}$

Figure 3-2: The pseudocode for the AdaBoost algorithm adopted by [3, 148].

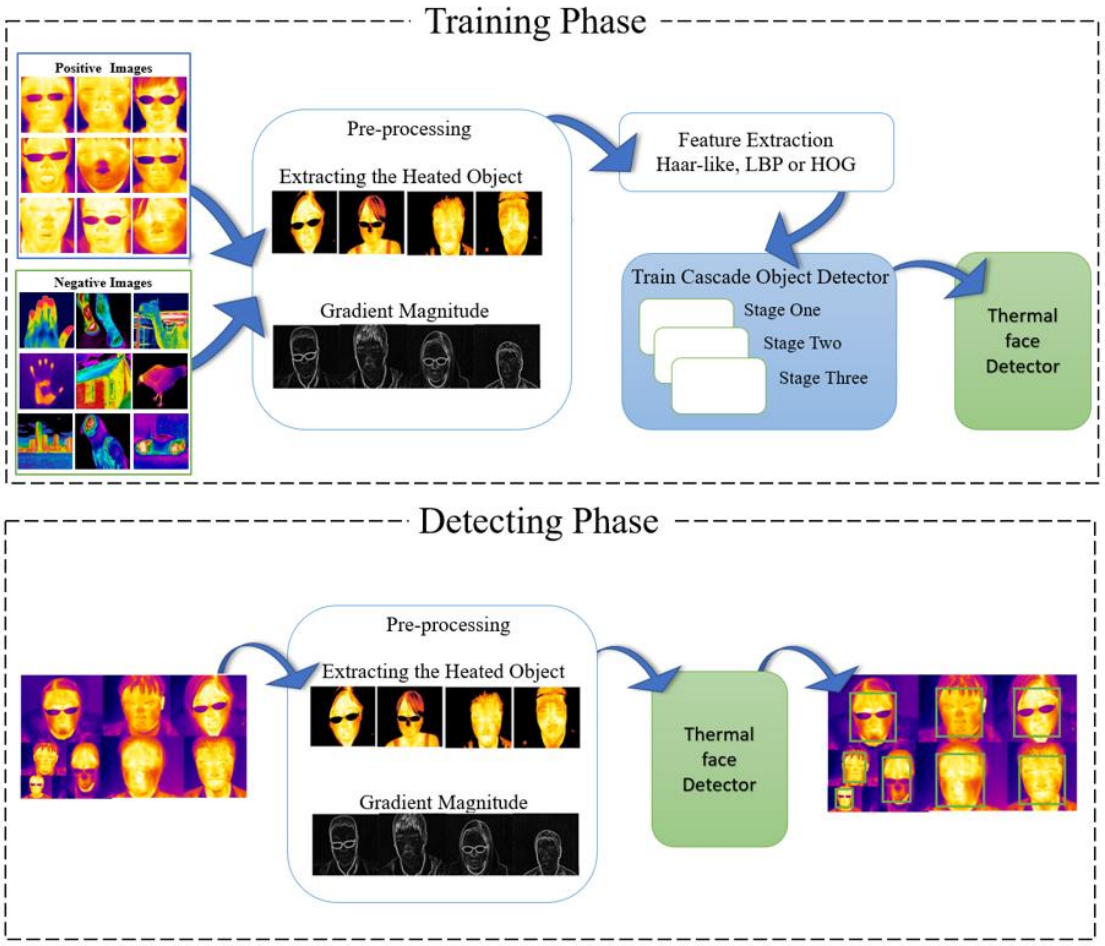


Figure 3-3: Training and detecting phases for thermal face detector.

3.3. Methods

This section describes the proposed process for face detection from thermal images. As shown in Figure 3-3, the proposed process consists of training and detecting phases, with each phase consisting of three stages. In the training phase, training images are divided into two groups to create positive and negative samples. A positive image contains a single face, whereas a negative image contains no face as shown in Figure 3-4. In the first stage, both positive and negative samples are pre-processed by using gradient magnitude or Otsu's methods [143]. The features are extracted from the pre-processed samples in the second stage. Both positive and negative features are used to train the cascade face detectors in the final stage.

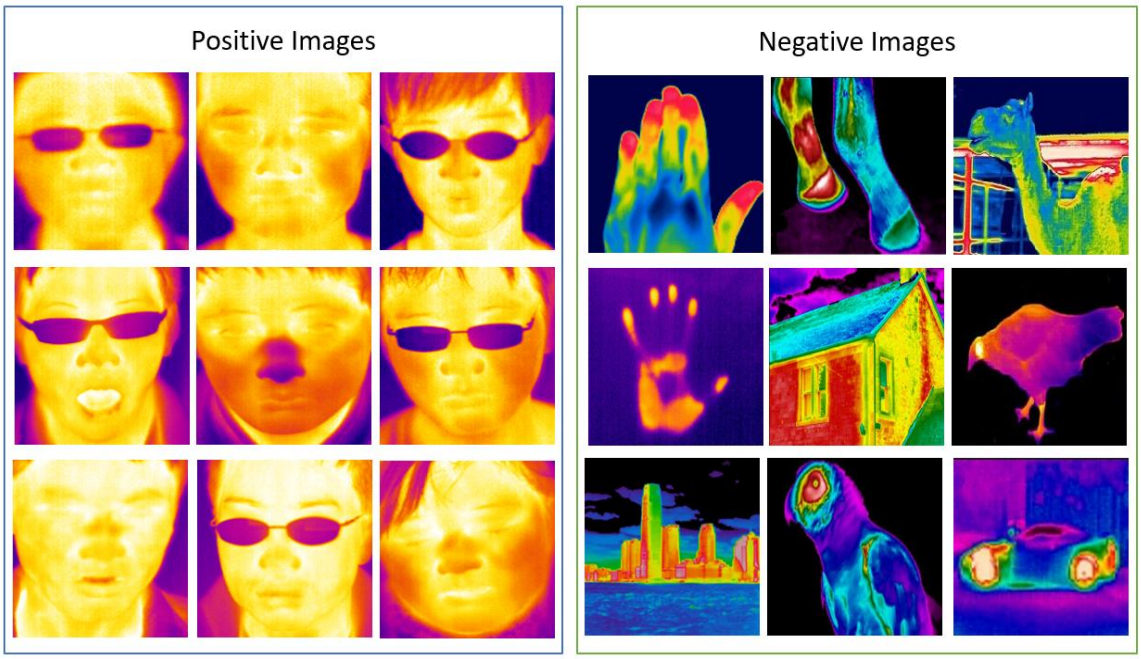


Figure 3-4: Sample of the positive and negative images used to train the thermal face detectors.

In the detecting phase, testing images are pre-processed by using the same pre-processing method as in the training phase. In the second stage, features are extracted from the testing image, which is scanned by creating sub-windows of different sizes to find the face. In the third stage, the features extracted from each scanned sub-window are evaluated by the cascade face detector. The detector rejects non-face sub-windows and detects sub-windows containing a face. If multiple sub-windows occur around each face in the scanned image, the detected sub-windows are combined to convert the overlapping detections into a single detection so as ultimately to return one detection per face. In the following, we explain the pre-processing, feature extraction and classification steps.

Figure 3-5: Samples of gradient magnitude images by using different colour maps.

3.3.1. Pre-processing

Due to the loss of facial feature properties and appearance features in thermal images, we propose using the gradient magnitude method to enhance the texture and the edges of the face in thermal images. We also suggest object extraction as a pre-processing step to increase detection accuracy.

3.3.1.1. Gradient Magnitude

Gradient magnitude is utilised to extract useful information from images, such as edges, by measuring the variation of intensity in a given direction. To compute the gradient magnitude of an image, the image is convolved with filters to identify gradients in the x and y directions. Then the gradient magnitude at each pixel is computed, using a variation of the distance equation to measure the steepness of the slope. The pixels with large gradient magnitude values become possible edge pixels, as shown in Figure 3-5. The Gradient magnitude can also be used for feature and texture matching.

Figure 3-6: Samples of extracted heated objects in thermal images by applying Otsu's method.

3.3.1.2. Object Extraction

It is essential in image processing to select an adequate threshold grey level to enable objects to be distinguished from their background. Otsu [143] proposed a method for selecting an optimal threshold by utilising the discriminant criterion to maximise the separability of the grey level classes. Due to its efficiency, most state-of-the-art methods for thermal image processing use Otsu's method for pre-processing. In order to extract heated objects from their background, the global threshold value is used to convert a thermal image into a binary image. Then, each pixel in the original image is multiplied by its correspondent pixel in the binary image, as shown in Figure 3-6.

3.3.2. Feature Extraction and Classification

Due to their efficiency, Haar-like features have been commonly used in face detection. Indeed, the Viola-Jones algorithm [3] utilises Haar-like features rather than the pixels directly. We propose using LBP features because they effectively describe the image texture features [149] and have an advantage on high-speed computation and rotation invariance, which facilitates broad usage in image segmentation, and retrieval, etc [150]. Additionally, LBP features show outstanding performance on several facial-related tasks such as face alignment [151], face recognition [152, 153], and gender recognition [154].

We also suggest using HOG features since they are useful for capturing the overall shape of an object. In the proposed process, Haar-like, LBP and HOG features are utilised in the training phase to form the cascade classifier and in the detecting phase to reject non-face areas. The Viola-Jones algorithm uses Adaboost training to select the most effective features. Since there are still a huge number of features to be calculated, the classifier is organised in the form of a cascade in order to avoid worthless calculation. As shown in Figure 3-3, the cascade classifier consists of several stages, in which multiple simple classifiers are used. If a sub-window fails on any weak classifier, it will be excluded as it contains no face. Only the sub-windows which have reached the final stage are considered to contain faces.

3.4. Experiments

In this section, we first describe the thermal facial image databases used in the training phase. Thereafter, we examine the effect of different parameters on the face detection performance on the testing dataset. Finally, we compare the performance of different face detectors.

3.4.1. Dataset

Two thermal facial image databases, NVIE [8] and I.Vi.T.E [114], were adopted in our experiments. The Natural Visible and Infrared facial Expression (NVIE) database contains both posed and spontaneous expressions of 215 subjects (157 males and 58 females) with three different illumination directions: left, right and front. The subjects were required to pose with seven different expressions, wearing eyeglasses and without them. However, the exact number of participating subjects is varied between sections of the database because several participants were limited to two or fewer facial expressions, and some thermal and visual video recordings were lost. The numbers of subjects contributed to the spontaneous database are 105, 111 and 112 under the front, left and right illumination, respectively, while 108 subjects presented to the posed database. To record infrared videos, an infrared camera capturing 25 frames per second, with resolution 320 × 240 and band wave 8–14 mm, was used. The camera was placed 0.75 m in front of the subject.

The naturalistic database of thermal emotion, which was named Italian Visible-Thermal Emotion (I.Vi.T.E.), consists of spontaneous expressions (one image per second in .PNG format) of 40 subjects (Italian undergraduate students aged 22 to 28 years). The thermal image resolution is 160 × 120 pixels. In order to create negative samples, 50 thermal videos were downloaded from YouTube and converted to 13,082 frames of thermal images, of which 10,000 images were used to train all detectors whilst 3082 images were used for testing. Depending on the length of the downloaded videos, the training and testing images are selected. To create a large number of negative samples, the frames of the longer videos which have a higher number of frames are used in the training set of the negative samples and the frames of the shorter

videos are used in the testing set. This is to avoid using similar frames in the training and testing sets.

3.4.2. Experiment Design

A dataset of 10,021 thermal infrared frontal face images from the NVIE database was constructed from posed images and the first frame of the spontaneous expression image sequences. The faces were manually extracted from images by using the Training Image Labeller app [155] in MATLAB. The app allows the user to interactively specify bounding boxes to define locations of ROIs which are used to train a classifier. To specify a face region, we placed the bounding box around each face just underneath the chin and about half-way between the hairline and the eyebrows. was used for extracting the faces from images. To help improve accuracy, we deliberately included extra visual information such as the contours of the chin and cheeks. No further alignment or resizing was done.

To make it subject-independent throughout our experiments, the NVIE database was split into three subsets: training, validation and testing according to the subjects in the database where each subset has different subjects. The training and validation subsets, consisting of 3530 thermal images each, were used to conduct two-fold cross-validation for parameter selection. The remaining 2961 thermal images from the NVIE database were used as the testing subset. To test the ability to detect more than one face in an image, a Twelve-In-One dataset was created from the NVIE database. This contained 180 images, which were selected from the thermal spontaneous expressions database. Each image comprised of twelve randomly selected faces. Samples of the Twelve-In-One dataset are shown in Figure 3-7.

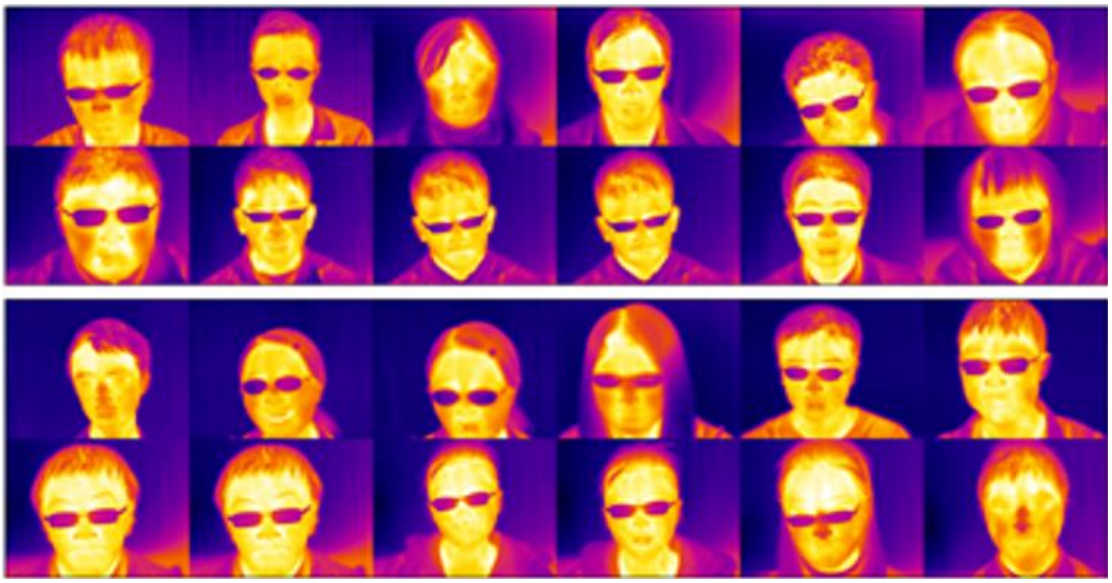


Figure 3-7: Two samples of the Twelve-In-One dataset, randomly selected from the NVIE database.

The third dataset was the I.Vi.T.E. database. Since this contains video frames with spontaneous expressions of each subject, some frames are not suitable for testing because a large part of the face is covered or the whole face is missing, as shown in Figure 3-8. The number of images were screened out on this basis was 3583. The number of images was screened out on this basis was 3583. The rectangular regions of interest were specified semi-manually for 31217 images. First, the initial regions of interest were determined by using three trained detectors. Then the three determined regions were combined to specify the bounding box. Finally, the regions of the box have been manually checked for all images. We have created a particular application, called ROI-Bounding Box, to assist us in setting and testing the region of interest for these images, whereas the Training Image Labeller app [155] in MATLAB was used for misdirecting faces in 624 images. Figure 3-9 shows the ROI-Bounding Box application.

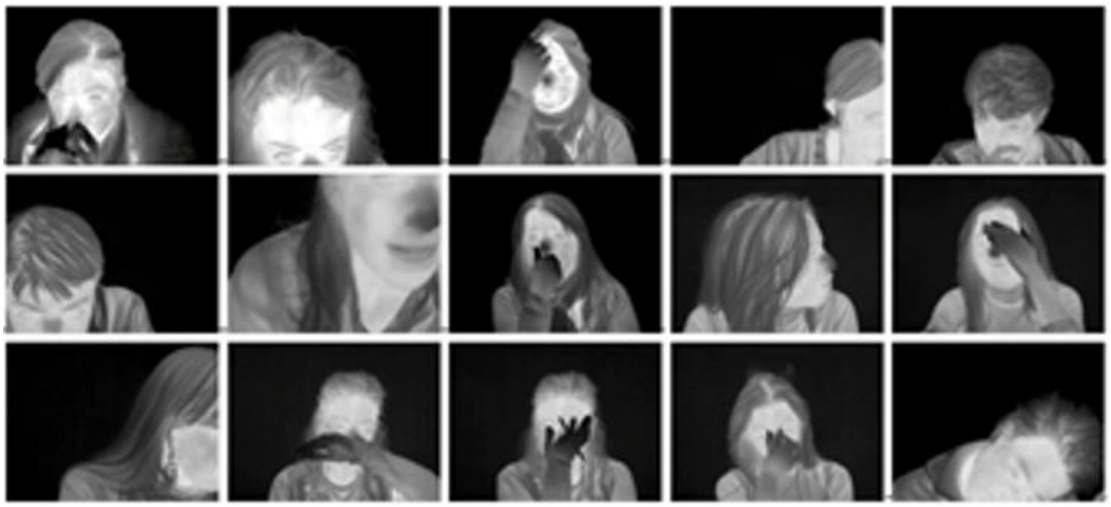


Figure 3-8: Samples of separated thermal images from the I.Vi.T.E. database where the face does not fully appear.

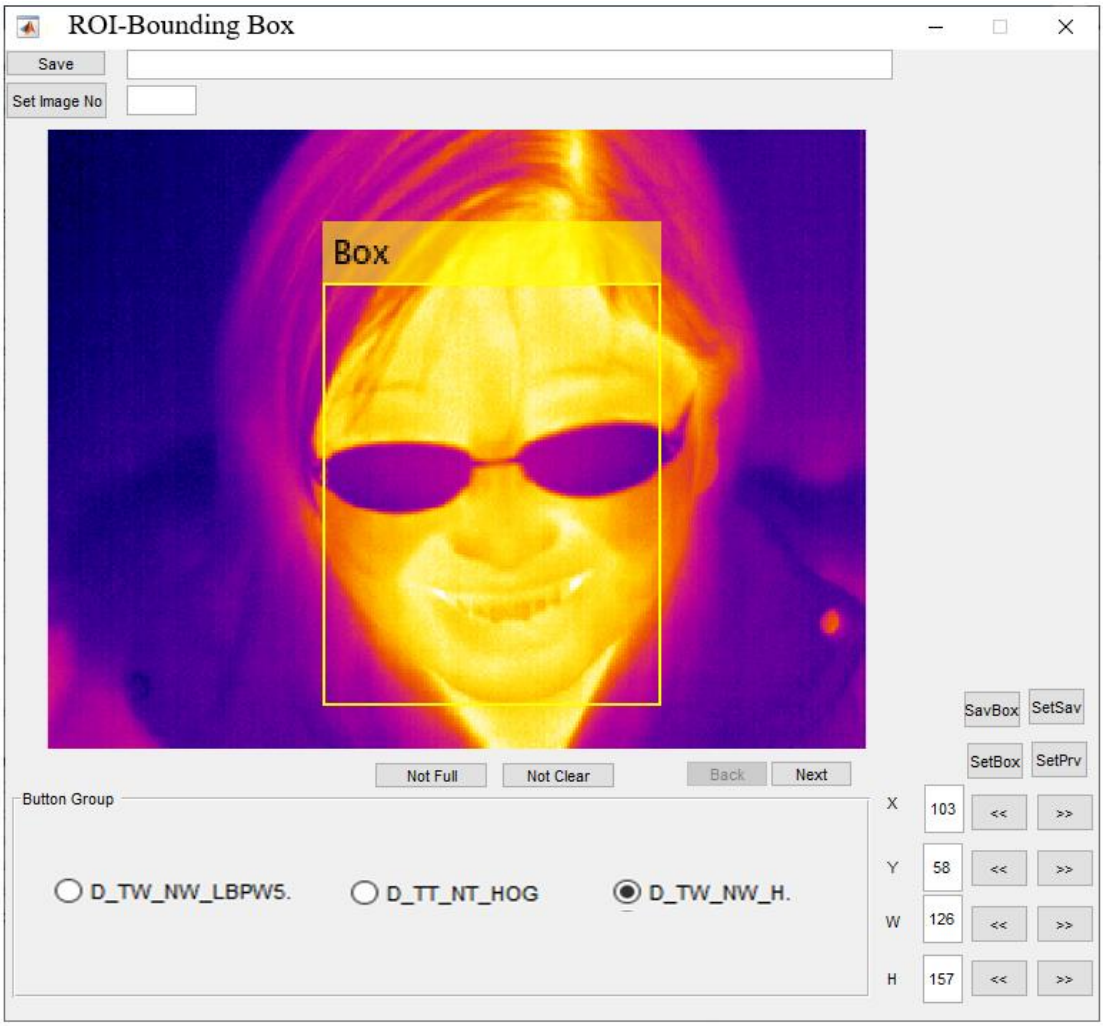


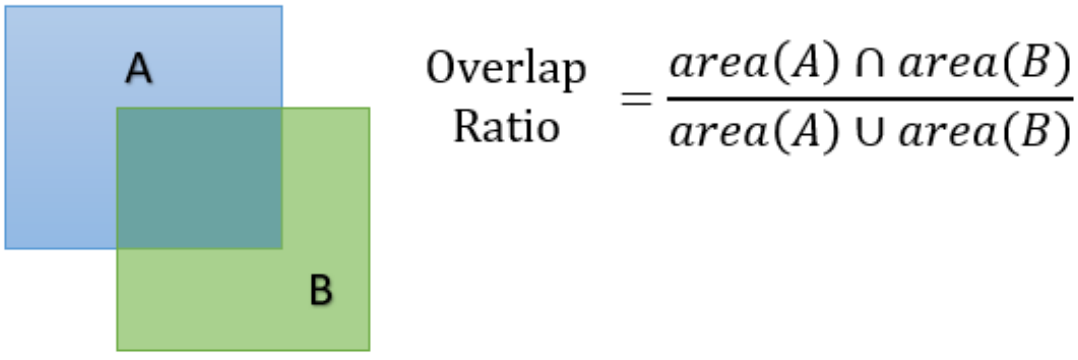
Figure 3-9: ROI-Bounding Box application created to set and test the region of interest Bounding Box for thermal images.

3.4.3. Criteria for Calculating True/False Positive Rate

Motivated by the work of Wang et al. [156], the face detection rate was used to measure the effectiveness of each detector by calculating the displacement from the automatically located rectangular of the target face from the true (manually annotated) rectangles, defined as the overlap ratio between them. To compute the ratio, the area of intersection between rectangle A and rectangle B was divided by the area of the union of the two, as shown in Figure 3-10. The value of the overlap ratio can be between 0 and 1, where 1 implies a perfect overlap and 0 implies no overlap.

As in other object detection challenges, the accepted degree of the overlap between the annotated and the detected regions is above 50% [157, 158]. Thus, for true positive cases, the minimum acceptance value of the overlap ratio was set to 0.5, whereas the cases were regarded as false positives if the overlap ratio was less than 0.5.

Figure 3-10: The overlap ratio between rectangle A and rectangle B.



3.5. Results

In order to demonstrate the performance of the proposed methods, Haar-like features were compared with LBP and HOG features for face detection. Before starting the comparisons, the holdout cross-validation was adopted to find the optimal values for the parameters of each pre-processing and feature selection method. There is a trade-off between the training parameters: the number of cascade stages, the false positive and true positive rates had to be set at each stage. In order to find optimal values for these parameters, several detectors were trained on different combinations of the parameters. The detectors with the highest mean accuracy on the validation datasets were selected for final testing.

Figure 3-11, Figure 3-12 and Figure 3-13 give the ROC curves comparing the performances of nine detectors tested on the NVIE testing dataset. To create the ROC curve, the threshold value for merging detected boxes around a face was adjusted from 7 to 1. This value was used to perform the merging operation where there were multiple detections around a face. In this case, the detections were grouped and then merged to produce one bounding box around the face, provided they met the merging threshold value. If the merging threshold were to be set to 0, all detections would be returned separately without performing thresholding or merging operation; this would serve to increase both detection and false positive rates. Adjusting the merging threshold value to ∞, meanwhile, would yield a detection rate of 0 and a false positive rate of 0. To compute the true positive and false positive rates, the number of true positive detections should be divided by the total number of all faces in all the images, while for calculating false positive rate the number of false positive detections should be divided by the total number of sub-windows scanned in all the images [3].

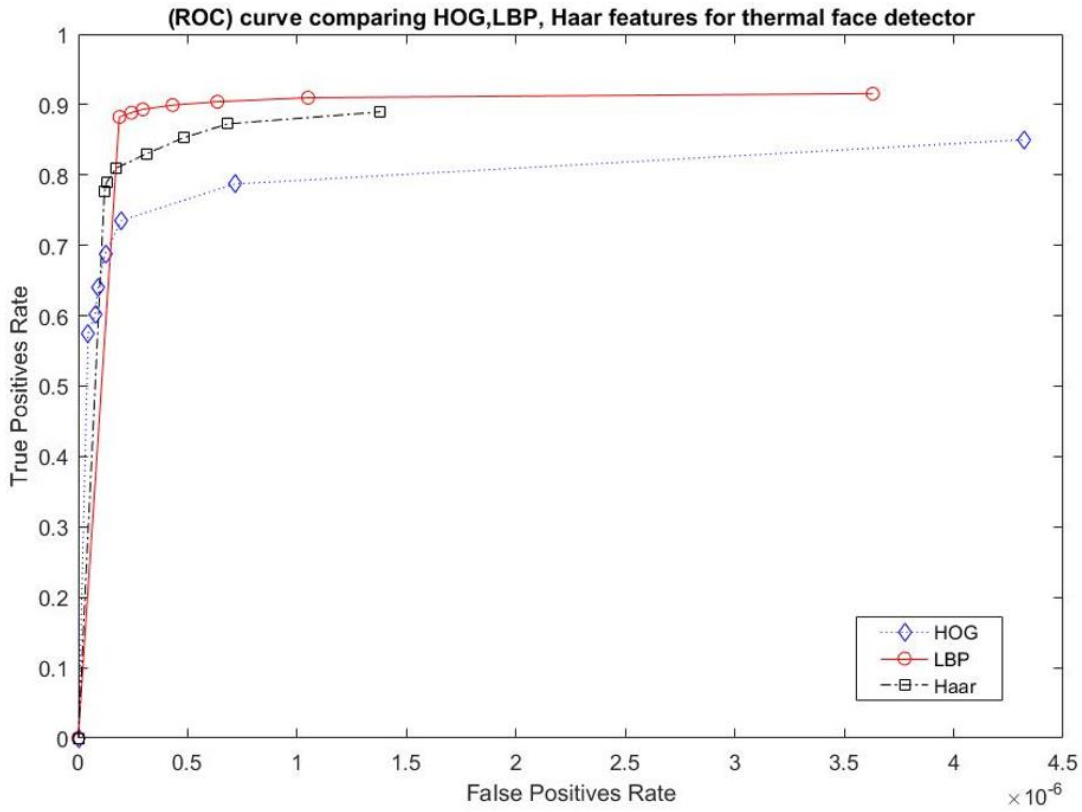


Figure 3-11: ROC curves comparing the performance of the Viola-Jones algorithm on thermal images when using different features (HOG, LBP and Haar-like), without pre-processing.

Figure 3-11 illustrates the Viola-Jones method's performance when using Haar-like, HOG and LBP features, respectively, without any pre-processing phase. Figure 3-12 and Figure 3-13, meanwhile, provide similar comparisons, but with the proposed pre-processing phase. Figure 3-12 and Figure 3-13 also show the performance of the three types of features when using the gradient magnitude method and Otsu's method respectively, in the pre-processing phase.

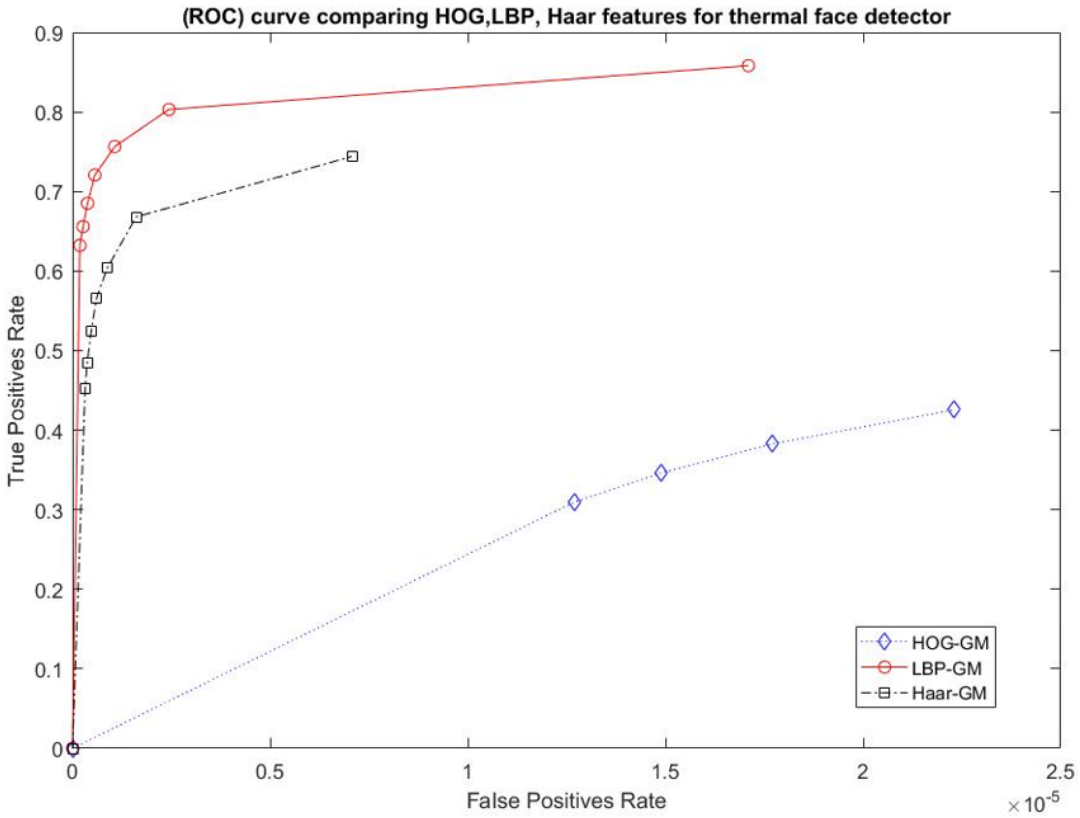


Figure 3-12: ROC curves comparing the performance of the Viola-Jones algorithm on thermal images when using different features (HOG, LBP and Haar-like) and applying the Gradient Magnitude (GM) method for pre-processing.

For a statistical evaluation of the Viola-Jones algorithm using LBP features for detecting faces from thermal images with or without eyeglasses, the McNemar's statistical test was used to determine the significance of the results, which is a statistically sound way of comparing the performance of two algorithms applied to the same dataset [159]. The null hypothesis assumes that there is no statistical difference in the performance of the Viola-Jones method when using LBP or Haar-like features. Based on the calculation of the McNemar's statistic using Equation (3-1), the null hypothesis can be rejected with an error probability of 0.05 if |Z|>1.96, which indicates that the differences in the performance are statistically significant.

$$Z = \frac{|N_{sf} - N_{fs}| - 1}{\sqrt{N_{sf} + N_{fs}}} \qquad \textbf{(3-1)}$$

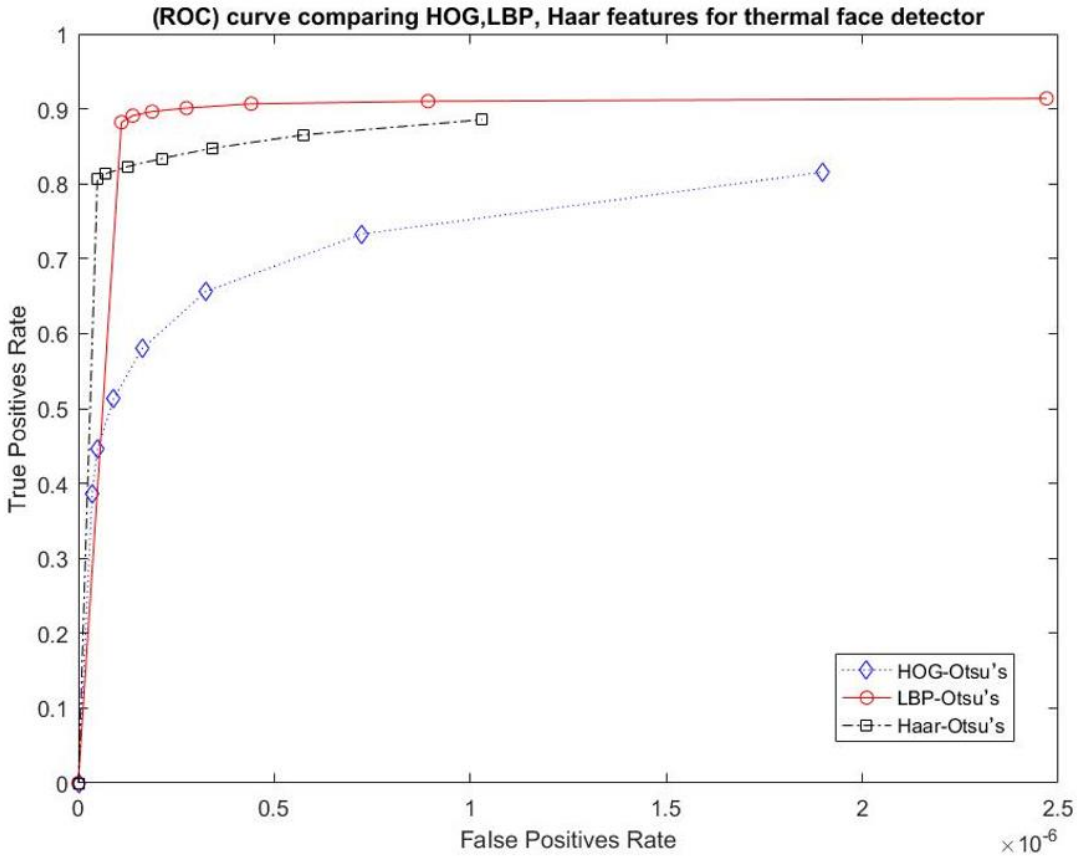


Figure 3-13: ROC curves comparing the performance of the Viola-Jones algorithm on thermal images when using different features (HOG, LBP, and Haar-like) and applying Otsu's method for pre-processing.

In Equation (3-1), N_{sf} is the number of occurrences when the first algorithm succeeds, and the second algorithm fails, while N_{fs} is the opposite. If $N_{sf} + N_{fs} > 20$, the statistic is reliable [159].

The Z-value and the related parameters (N_{ss}, N_{sf}, N_{fs}, N_{ff}) for the NVIE database as well as the I.Vi.T.E. database are shown in Table 3-1 and Table 3-2, respectively. The tables are split into three parts. The first part shows the results of McNemar's test for LBP features without a pre-processing phase versus HOG and Haar-like features, without a pre-processing phase in the first column and with pre-processing phases (Gradient Magnitude-GM, Otsu's method) in the second and third columns, respectively. Like the first part, the second and third parts show the results of

McNemar's test for LBP features with pre-processing phases (GM and Otsu) versus other features. The last column displays the comparison of McNemar's test results for the different pre-processing phases with LBP features. The Haar-like features without pre-processing phase represents the standard Viola-Jones algorithm trained to detect faces from thermal images.

Table 3-1: Comparison of the Z-value and the related parameters for LBP features with HOG and Haar-like features on the NVIE database.

LBP VS	**HOG**	**Haar-like**	**HOG GM**	**Haar-like GM**	**HOG Otsu's**	**Haar-like Otsu's**	**LBP-GM**
N_{ss}	2531	2636	2189	2210	2436	2621	2545
N_{sf}	200	95	542	521	295	110	186
N_{fs}	4	13	17	2	2	7	22
N_{ff}	226	217	213	228	228	223	208
Z-value	13.65	7.79	22.16	22.65	16.94	9.42	11.30

LBP-GM VS	**HOG**	**Haar-like**	**HOG GM**	**Haar-like GM**	**HOG Otsu's**	**Haar-like Otsu's**	**LBP Otsu's**
N_{ss}	2445	2527	2114	2193	2377	2464	2545
N_{sf}	122	40	453	374	190	103	22
N_{fs}	90	122	92	19	61	164	188
N_{ff}	304	272	302	375	333	230	206
Z-value	2.13	6.36	15.42	17.86	8.08	3.67	11.39

LBP-Otsu's VS	**HOG**	**Haar-like**	**HOG GM**	**Haar-like GM**	**HOG Otsu's**	**Haar-like Otsu's**	**LBP**
N_{ss}	2532	2639	2191	2208	2437	2623	2719
N_{sf}	201	94	542	525	296	110	14
N_{fs}	3	10	15	4	1	5	12
N_{ff}	225	218	213	224	227	223	216
Z-value	13.79	8.14	22.29	22.61	17.06	9.700	0.20

Table 3-2: Comparison of the Z-value and the related parameters for LBP features with HOG and Haar-like features on the I.Vi.T.E. database.

LBP VS	**HOG**	**Haar-like**	**HOG GM**	**Haar-like GM**	**HOG Otsu's**	**Haar-like Otsu's**	**LBP-GM**
N_{ss}	28306	27488	24244	23762	26023	26177	25509
N_{sf}	1567	2385	5629	6111	3850	3696	4364
N_{fs}	484	579	649	388	378	444	400
N_{ff}	860	765	695	956	966	900	944
Z-value	23.89	33.15	62.84	70.98	53.38	50.53	57.42

LBP-GM VS	**HOG**	**Haar-like**	**HOG GM**	**Haar-like GM**	**HOG Otsu's**	**Haar-like Otsu's**	**LBP Otsu's**
N_{ss}	25180	24146	21203	22260	23595	23630	25422
N_{sf}	729	1763	4706	3649	2314	2279	487
N_{fs}	3610	3921	3690	1890	2806	2991	4075
N_{ff}	1698	1387	1618	3418	2502	2317	1233
Z-value	43.72	28.61	11.08	23.62	6.86	9.79	53.11

LBP-Otsu's VS	**HOG**	**Haar-like**	**HOG GM**	**Haar-like GM**	**HOG-Otsu's**	**Haar-like Otsu's**	**LBP**
N_{ss}	28023	27204	23958	23682	25873	25999	29007
N_{sf}	1474	2293	5539	5815	3624	3498	490
N_{fs}	767	863	935	468	528	622	866
N_{ff}	953	857	785	1252	1192	1098	854
Z-value	14.91	25.44	57.21	67.44	48.03	44.79	10.18

3.6. Discussion

The results indicate that the Viola-Jones method using LBP features achieved higher accuracy than when Haar-like or HOG features were used to detect faces from thermal images. In addition, using Otsu's method in the pre-processing phase reduced the false positive rate. On the other hand, using HOG features reduced the accuracy with or without pre-processing.

Table 3-3: Comparisons of LBP, HOG and Haar-like features with the two pre-processing methods on the NVIE, I.Vi.T.E and Twelve-In-One databases.

	NVIE		I.Vi.T.E.		12-In-One		Speed
	TP rate	**FP rate**	**TP rate**	**FP rate**	**TP rate**	**FP rate**	**S/Image**
HOG	0.86	4.17E-06	0.92	4.42E-05	0.90	4.98E-07	0.0283
LBP	0.92	3.50E-06	0.96	7.59E-05	0.97	1.07E-06	0.0132
Haar-like	0.89	1.31E-06	0.90	9.13E-05	0.93	1.09E-06	0.0190
HOG-GM	0.75	6.64E-05	0.80	1.05E-03	0.70	2.55E-05	0.0313
LBP-GM	0.87	1.64E-05	0.83	1.94E-04	0.93	2.84E-06	0.0093
Haar-like-GM	0.75	6.87E-06	0.77	1.20E-04	0.87	8.26E-07	0.0101
HOG-Otsu's	0.82	1.80E-06	0.85	4.87E-05	0.80	7.39E-08	0.0271
LBP-Otsu's	0.92	2.34E-06	0.94	8.34E-05	0.98	5.50E-07	0.0088
Haar-like-Otsu's	0.89	9.90E-07	0.85	4.53E-05	0.92	3.39E-07	0.0096

Figure 3-11 and Figure 3-13 show that there is little difference in the accuracy of the Viola-Jones method when using LBP features, with or without a pre-processing phase. On the other hand, the pre-processing phase (Otsu's method) nearly halves the detection time. This is because the speed of the detector is related to the number of features evaluated per scanned sub-window and Otsu's method serves to exclude most of the features in non-heat-emitting regions. Thus, the first stage of the detection process discards a vast majority of the sub-windows, so that they are not evaluated in subsequent stages. This increases the detection speed and reduces the false positive rate at the same time.

According to McNemar’s test, LBP features outperformed other features in all cases except when using the Gradient Magnitude method in the pre-processing phase. Since some geometrical and appearance features are lost in thermal images, HOG features have lower accuracy than LBP and Haar-like features with respect to face detection from thermal images. Also, the differences between the grey values of pixels in the same heated area, such as the face, are quite small in the thermal image. Each

LBP feature represents the exact differences of each pixel with the eight neighbouring pixels, whereas each Haar-like feature represents differences in the value of white and grey rectangles which combine more than one pixel. This means that LBP features give a more accurate representation of the heated areas in the thermal image than Haar-like features. The performance of the Viola-Jones algorithm using LBP features is significantly better than when using Haar-like features for face detection in thermal images. The Z-value of LBP vs LBP-Otsu's in Table 3-1 (0.20) indicates that there is no significant difference between the two algorithms on the NVIE database, while the Z-value (10.18) of LBP vs LBP-Otsu's in Table 3-2 shows that the difference between them is significant on the I.Vi.T.E. database. Table 3-3 shows the speed of the detectors. It can be seen that the LBP-Otsu's detector has the highest speed in comparison with other detectors; it can process an image in about 0.0088 seconds on a 2.70 GHz Intel® i7 processor.

3.7. Conclusion

This chapter presents an efficient and effective process to improve the performance of the Viola-Jones algorithm for face detection from thermal images, with or without eyeglasses. McNemar's test was employed to test whether the difference in performance between the proposed process and the standard Viola-Jones algorithm is statistically significant. The results of the test demonstrate that the LBP features outperformed other features significantly in most cases and that applying Otsu's method in the pre-processing phase reduced the false positive rate in face detection from thermal images. The proposed enhancement process reduced the detection time of the Viola-Jones algorithm by roughly a factor of two while retaining high detection accuracy.

Chapter 4

Shallow Convolutional Neural Network for Eyeglasses Detection

4.1. Introduction

The challenges related to human facial analysis systems can be attributed to many appearance and technical factors. Appearance factors are related to the subject's face, such as pose and facial expression, whereas technical factors are related to the clarity and quality of images, such as variations in illumination, shadows, image resolution, and the presence of intervening components such as eyeglasses and hands. Eyeglasses are considered to be a particular confounding factor for human facial analysis systems due to reflection and frame occlusion which cover the most crucial part of the face over the ocular region. Moreover, the presence of eyeglasses may cause inaccurate classification, especially when the facial analysis system utilises convolutional neural networks in its models [160-162]. To increase accuracy in such circumstances, several human facial analysis systems have included an eyeglasses and non-eyeglasses image classification phase in their frameworks [6, 163], but this has the drawback of increased memory consumption and computation time. If these systems are to be robust enough to cope with real-world applications, a highly accurate and rapid eyeglasses detector is needed.

Most of the existing approaches for detecting eyeglasses utilise handcrafted feature extraction methods [160, 164]. Several pattern recognition projects [161, 165], however, have demonstrated that deep learning features may provide valuable

information about the relationships between raw data and learned features. The convolutional neural network (CNN) has become the most widely used approach in computer vision in recent years, and a number of recent studies indicate that the features extracted from the convolutional neural networks are compelling [32, 166]. This chapter presents an effective and efficient method for detection of eyeglasses in facial images based on extracting deep features from a well-designed shallow convolutional neural network. The proposed shallow CNN contains fewer complex structures and thus can be utilised in different facial analysis system frameworks without consuming their resources while still achieving high accuracy in real time. The proposed shallow CNN, called Shallow-GlassesNet, consists of just six layers: three convolution layers, two max-pooling layers, and one fully-connected layer.

The main contribution of this chapter is to address two essential aspects of CNN for eyeglasses detection: (1) the size of the training dataset required and (2) the depth of the network architecture. To this end, we initialise the learning parameters of the shallow CNN using the parameters of a deep CNN which is fine-tuned on a small dataset. The depth of the neural network is then decreased by removing some convolutional layers after testing its performance on the validation dataset [4].

The rest of this chapter is organised as follows. Some related work is reviewed in Section 4.2. The proposed framework and the structure of the Shallow-GlassesNet are introduced in Section 4.3. The fine-tuning and training process for Shallow-GlassesNet are explained in Section 4.4. The experimental setup and databases used are described in Section 4.5. Experimental results are presented in Section 4.6 and followed by a discussion in Section 4.7. Section 4.8 draws conclusions.

4.2. Related Work

The existing methods for detecting eyeglasses can be categorised into two approaches: handcrafted feature approach and deep learning approach. Jing [167] developed an eyeglasses detection and extraction algorithm in which detection is realised using edge information within a small area defined between the eyes, whilst extraction is achieved with a deformable contour, combining edge features and geometrical features. They obtained two false alarms in their test, by falsely detecting the presence of eyeglasses in facial images. With the facial database used in their experiment, 50% of eyeglasses were accurately extracted, 30% of eyeglasses were extracted with satisfactory results, and the remaining 20% were obtained with fair results. Their experimental investigation, which was conducted on their own images rather than a public database, showed an overall detection accuracy of 95.5%.

Bo et al. [168] presented a novel method for detecting eyeglasses in which detectors are trained by boosting simple wavelet feature based weak LUT (look-up-table) classifiers. They investigated the performance of their proposed method using Haar and Gabor features by utilising AdaBoost and SVM. Remarkable performance was achieved when Gabor features was combined with AdaBoost on the public FERET database [142], with a detection rate of 98.9% being reported. Experimental results show that the boosting methods have better performance than SVM.

Fernandez et al. [164] used Robust Local Binary Pattern and SVM for detection of eyeglasses. The proposed method was tested on the LFW database [5], demonstrating an accuracy rate of 98.65%. Du et al. [162] proposed a new set of Haar-like features to detect eyeglasses more robustly. Using the AdaBoost algorithm, their method achieved a detection rate of 95.11% on the face database CAS-PEAL [169].

Shao et al. [161] proposed a deep convolutional neural network called GlassesNet (GNet). They first pre-trained it for face identification and then fine-tuned it as an eyeglasses detection network. They evaluated their method on different databases, achieving accuracies ranging from 95% to 99.4 %. Their experiments on the Multi-PIE database show that the proposed method is highly robust to various challenging conditions.

Fernández et al. [163] proposed a real-time Big Data architecture in order to collect, maintain and analyse massive volumes of images related to the problem of automatic eyeglasses detection. This architecture can be used for automatic image tagging related to the detection of eyeglasses in facial images. Their innovative algorithm is based on Robust Local Binary Pattern and robust alignment. Experimental results demonstrate that a simple, yet efficient algorithm can obtain impressive classification accuracy, achieving 98.65% recognition rate on the LFW [5] database. This algorithm was also tested on the FERET database [142], achieving a 99.89% recognition rate. Experimental results also show that the proposed algorithm is robust under a wide range of lighting conditions and different poses, and can deal with occlusion, which is very common with sunglasses.

Mohammad et al. [160] proposed two schemes for the detection of prescription eyeglasses. The first proposed scheme is not learning based, and uses the Viola-Jones algorithm to detect regions of interest, followed by eyeglasses detection, yielding an overall accuracy of 99.0% on the FERET database and 97.9% on the VISOB database [170]. The second scheme is learning-based, which obtained a best overall accuracy of 99.3% on the FERET database and 100% on the VISOB database. Du et al. [162], meanwhile, proposed an accurate eyeglasses detection algorithm for in-plane rotated

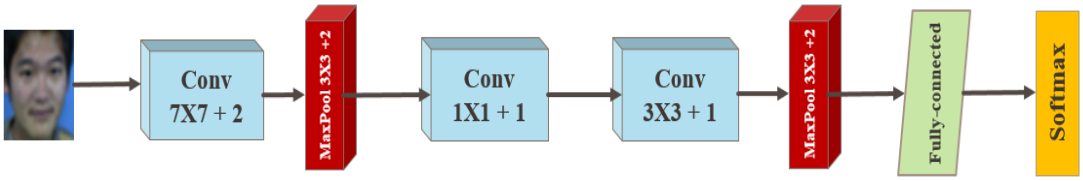


Figure 4-1: The shallow GlassesNet architecture.

faces by using a new set of Haar-like features which represent the features of rotated faces. Reese [140] proposed a face detection algorithm using projection profile analysis and an eyeglasses detection algorithm using block/region processing plus prior knowledge. Both algorithms were tested with the ASUIR database [171] (142 thermal face images from 71 subjects) in which the ground truths for both face region and eyeglasses region were established manually. All faces were successfully detected by their algorithms and the averaged overlapping ratio with the ground truths was 0.8998. The eyeglasses detection algorithm detected 22 of 23 eyeglasses, and the averaged overlapping ratio with the ground truths was 0.7986.

4.3. Methods

In this section, we describe in detail the architecture of the Shallow-GlassesNet and the proposed pipeline for eyeglasses and non-eyeglasses image classification. The pre-training and fine-tuning processes for the Shallow-GlassesNet are also introduced.

4.3.1. Shallow-GlassesNet Architecture

Inspired by GoogleNet [93] architectures, the proposed Shallow-GlassesNet, as shown in Figure 4-1, contains six layers: three convolutional layers and two max-pooling layers, followed by a fully-connected layer. The kernel size and stride of Conv1, Conv2 and Conv3 layers are set as 7×7 (2), 1×1 (1), and 3×3 (1), and their outputs are feature

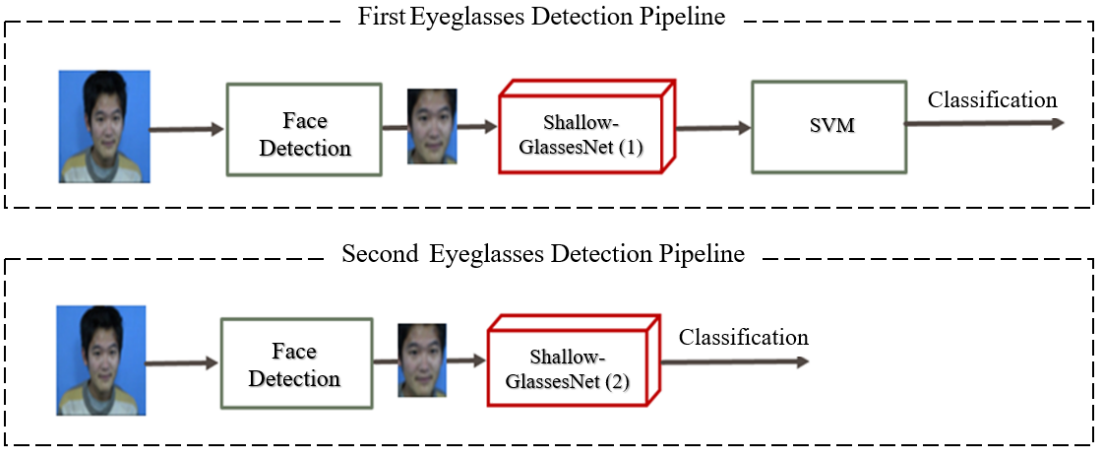


Figure 4-2: The proposed eyeglasses detection pipelines.

maps of sizes 64, 64 and 192 respectively. Each of these layers has similar corresponding layers in the GoogleNet. The three convolutional layers are followed by Rectified Linear Unit (ReLU), which is several times faster than other equivalents with tanh units [93]. Max-pooling is performed over a 3×3 pixel window, with stride 1.

4.3.2. Designing Shallow-GlassesNet

When designing a CNN, the initialisation of the network weights is critical since bad initialisation can cause gradient instability which could stop learning [94]. To avoid this problem, a GoogleNet was first fine-tuned with a small facial database. Then, each pooling layer was examined by training SVM on the features extracted from this layer. According to its performance, the layer achieving the highest accuracy was selected. Next, the convolutional layers of the Sallow-GlassesNet were initialised with the corresponding layers of the fine-tuned GoogleNet. Since the fully-connected layer is not used for feature extraction, it was initialised randomly. The initialised weights of Sallow-GlassesNet were kept fixed, which means no fine-tuning was performed for the convolutional layers. The following steps are followed to design and build the Shallow-GlassesNet (1) and (2):

Designing the Shallow-GlassesNet (1) and (2) steps:

(1) Split the database into three sets: training $T=\{t_1, t_2, t_3 \ldots t_N\}$, validation $V=\{v_1, v_2, v_3 \ldots v_M\}$, testing $S=\{s_1, s_2, \ldots s_Q\}$. Where $N= 70\%$, $M=15\%$ and $Q=15\%$ of the total number of images in the database.

(2) Fine-tuned the selected CNN model, CNN_{sm}, using T dataset.

Denote the set of pooling layers in the fine-tuned CNN_{sm} as $PL=\{pl_1, pl_2, \ldots. Pl_K\}$, where K number of pooling layers in the CNN_{sm}.

(3) For each layer in PL do:

(4) For each image in T do:

(5) Extracts the features vector $\hat{f}_i$ of ti from Pl_j

(6) Add the extracted features to the feature's vectors $\widehat{ft}_j = \widehat{ft}_j + \hat{f}_i$

(7) Go for step (3) for the next plj

Denote the set of features vectors as $\widehat{FT} = \{\hat{f}t_1, \hat{f}t_2, \hat{f}t_3 \ldots \hat{f}t_K\}$ of T dataset and the set of SVM as SVM = $\{SVM_1, SVM_2, SVM_3, \ldots SVM_K, \}$.

(8) Training SVM_j on $\widehat{ft}_j$

(9) Do steps (3) to (7) for all images in V

Denote the set of features vectors as $\widehat{FV} = \{\hat{f}v_1, \hat{f}v_2, \hat{f}v_3 \ldots \hat{f}v_K\}$ of V dataset.

(10) Testing SVMj by using $\widehat{fv}_j$

(11) Select the SVMHA which achieve the highest accuracy on V dataset and its corresponding pooling layer, plHA.

(12) Design the shallow CNN with a similar structure to the selected CNNsm model, starting from the input layer to the plHA.

(13) Add the fully connected layer to the designed shallow CNN.

(14) Initialised the convolutional layers of the sallow CNN model with the corresponding layers of the fine-tuned CNN model.

(15) To use the sallow CNN model for features extraction, ***Shallow-GlassesNet (1)***: Extract the features from the pl_{HA} in the sallow CNN model and use its SVM_{HA} for classification.

(16) To use the sallow CNN model for classification, ***Shallow-GlassesNet (2)***: Train the fully-connected layer of the sallow CNN model using T dataset. Hint: keep the initialised weights of the convolutional layers fixed during the training process.

The differences between the two pipelines are as follows:

- When utilising Sallow-GlassesNet (1) as feature extractor, the resulting image features were normalised and fed into a linear SVM classifier which was trained on another visible image database of face images with or without eyeglasses.
- When utilising Sallow-GlassesNet (2) as an end-to-end shallow CNN model, it is trained on another image database by freezing the learning weights of all kernels except the fully-connected layer.

4.3.3. Eyeglasses Detection Pipelines

As shown in Figure 4-2, the proposed pipelines for eyeglasses detection consist of three parts: face detection, features extraction by Shallow-GlassesNet, and classification /detection by SVM. To prepare training data, we used the face detection approach of Viola- Jones [3] and Joint-Face-Detection [172] face detectors. After obtaining and cropping the frontal face region from the visible image, the cropped face image is re-sized to match the input size of the Shallow-GlassesNet (224 × 224), while the mean RGB value, computed on the training set, is subtracted from each pixel.

In the first pipeline, the facial image features were extracted from the last max-pooling layer of the Shallow-GlassesNet (1). Then a Linear SVM classifier was trained on the extracted features to classify facial images with or without eyeglasses.

In the second pipeline, the facial image features were extracted and classified by the Shallow-GlassesNet (2) without any additional component. In other words, Shallow-GlassesNet (1) was utilised as a feature extractor in the first pipeline, Shallow-GlassesNet (2) was used to classify the facial images directly in the second pipeline.

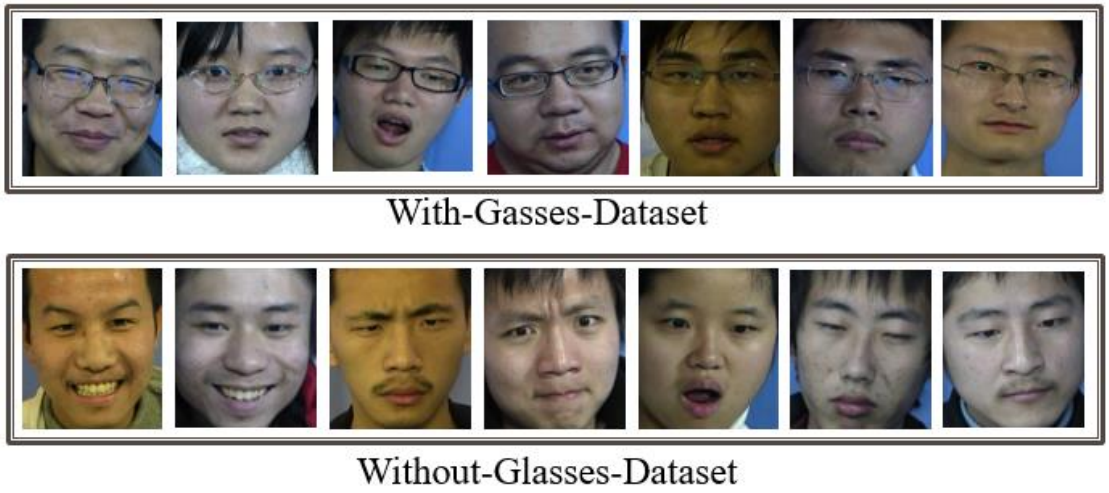


Figure 4-3: Samples from with-glasses-dataset and without-glasses-dataset from the NVIE posed database for the seven different facial expressions.

4.4. Fine-Tuning and Training Process

A CNN is an end-to-end model, a "black box", which receives the raw input data and gives the final classification results without any auxiliary process. Handling large training samples, the CNN automatically learns the features from the sample and classify these features by the neural networks. Conceptually, the complex CNN process can be divided into two sub-processes: a convolutional feature extractor and a neural network classifier. By separating these processes, we can construct a shallow CNN which can be utilised as a "black box" or as a feature extractor model. The performance of these models is equivalent or superior to the performance achieved by the original CNN and is accomplished with significantly lower network complexity.

When designing a CNN, the initialisation of the network weights is critical since bad initialisation can cause gradient instability which could stop learning [94]. To avoid this problem, a GoogleNet was first fine-tuned with a small facial database. Then, each pooling layer was examined by training SVM on the features extracted from this layer. According to its performance, the layer achieving the highest accuracy was selected.

Next, the convolutional layers of the Sallow-GlassesNet were initialised with the corresponding layers of the fine-tuned GoogleNet. Since the fully-connected layer is not used for feature extraction, it was initialised randomly. The initialised weights of Sallow-GlassesNet were kept fixed, which means no fine-tuning was performed for the convolutional layers.

4.5. Experiments

4.5.1. Experiment Design

The USTC-NVIE [8] (NVIE) database was adopted to fine-tune the GoogleNet, train the SVM, and test the proposed pipelines. It contains both posed and spontaneous facial expressions of 215 subjects, with illumination for three different directions. The posed database contains the apex expressional images with and without eyeglasses. As explained in Section 3.4.1, the number of participated subjects varied between the NVIE database sections. We used images of 101 subjects under the three illuminations in our experiments.

The database is divided into two parts: The With-Glasses-Dataset and the Without-Glasses-Dataset. Figure 4-3 shows some sample images from the posed database of seven different facial expressions. The database is small for training the SVM or the Shallow-GlassesNet (2). To deal with the over-fitting problem, we artificially enlarge the dataset by using techniques such as horizontal rotation. The original dataset was randomly partitioned into five almost equal-sized subsets, excluding overlapped subject images. Of the five subsets, one was retained for testing, and the remaining four were used for fine-tuning, training and validation. Table 4-1 illustrates the number of subjects and images in the training, validation, and testing datasets respectively.

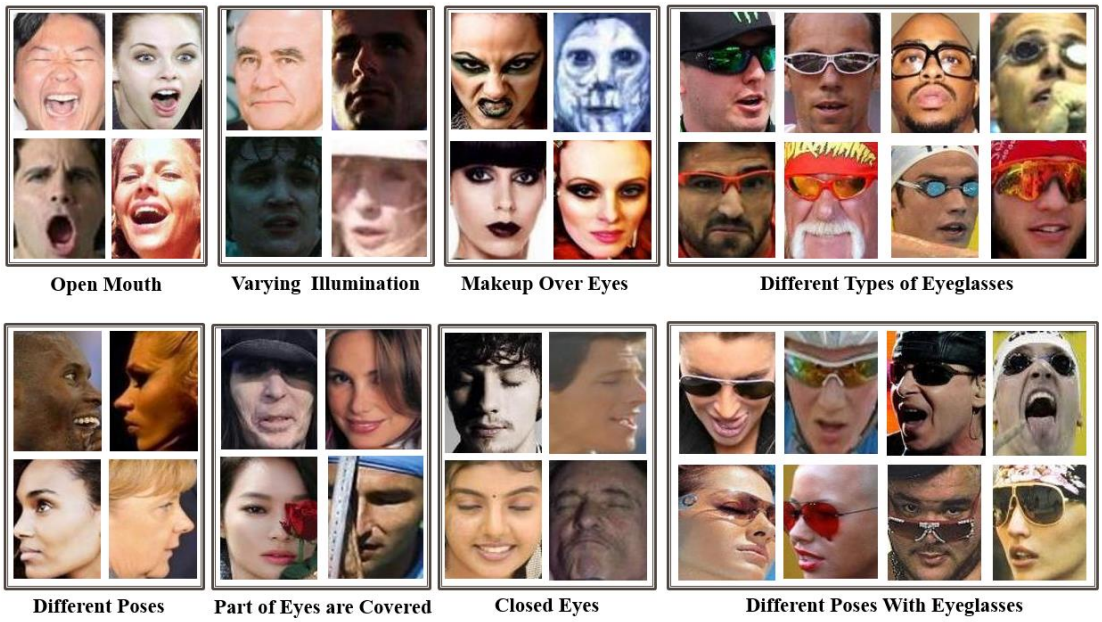


Figure 4-4: Samples from the Celebrity database (Celeba) [6].

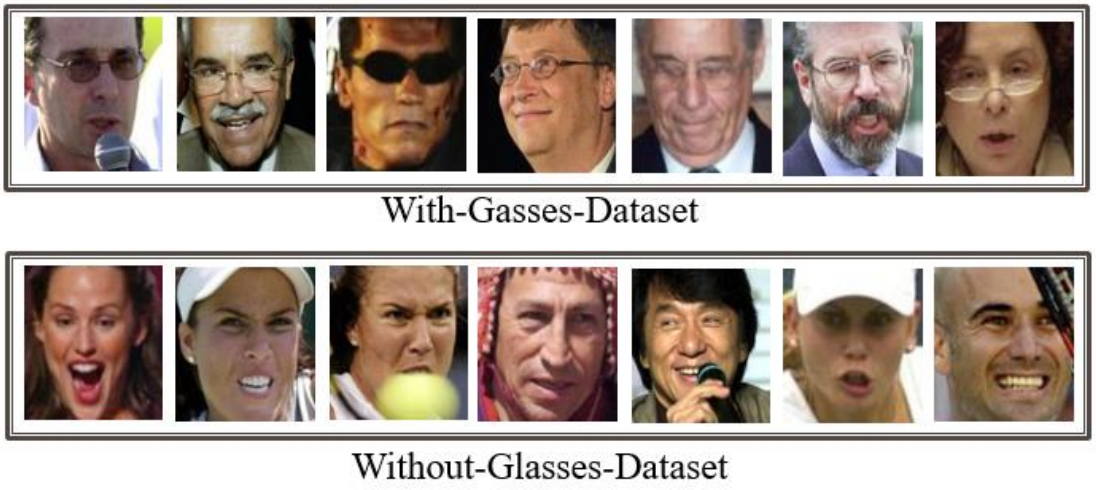


Figure 4-5: Samples from the Labelled Faces in-the-Wild (LFW) database [5].

To evaluate the performance of the proposed methods for eyeglasses detection, we adopted two large facial databases which were created for studying the unconstrained face recognition problem: (1) Labelled Faces in-the-Wild (LFW) [5], and (2) Celeb Faces (Celeba) [6]. The LFW database contains 13,233 face images of 5,749 different people collected from the web, with 1,244 images having eyeglasses. The Celeba database contains 202,599 face images of 10,177 different celebrities. Some samples

from the Celeba and LFW databases are shown in Figure 4-4 and Figure 4-5, respectively, from which it can be seen that eyeglasses detection is a challenging problem.

Table 4-1: The number of subjects and images in the training, validation and testing datasets.

NVIE Datasets	Without-Eyeglasses		With-Eyeglasses	
	Subject	Image	Subject	Image
Train	61	8946	61	8959
Val	20	641	20	675
Test	20	673	20	717

4.5.2. Neural Network Setup

We applied the Caffe toolkit [173] on NVIDIA GeForce GTX 980 GPU to fine-tune the pre-trained GoogleNet Deep CNN model [93] using the NVIE dataset. GoogleNet was fine-tuned using the stochastic gradient descent with a batch size of 50. The hyper-parameters of the applied training algorithm were as follows: momentum=0.9, weight decay=0.0002, initial learning rate=0.001.

4.6. Results

Before starting the testing, four-fold cross-validation was adopted to find the optimal extracted features which have the best average accuracy on the validation datasets. To analyse the efficiency of Shallow-GlassesNet (1) and Shallow-GlassesNet (2), we conducted two comparative evaluations to compare the accuracy and speed of the Shallow-GlassesNet and GoogleNet. Table 4-2 shows the accuracy with which eyeglasses were detected by GoogleNet, Shallow-GlassesNet (1), and Shallow-GlassesNet (2), respectively on the validation and testing datasets from the NVIE database. To calculate the accuracy of the proposed pipeline, we applied the following

Equation (4-1), which was defined by [159], where TP is the number of true positive detections, TN is the number of true negative detections, and N represents the number of face images tested.

$$(TP + TN)/N \quad \textbf{(4-1)}$$

Table 4-2: Comparison among GoogleNet, Shallow-GlassesNet (1) and (2) in terms of accuracy and speed on the validation and testing datasets from the NVIE database.

	Accuracy		Speed S/ Image	
	Val	**Test**	**CNN**	**Pipeline**
GoogleNet	89.97	92.09	0.1110	0.1560
Shallow-GlassesNet (1)	99.24	99.42	0.0297	0.0782
Shallow-GlassesNet (2)	96.96	97.63	0.0297	0.0550

Table 4-3: Confusion matrix and average accuracy of Shallow-GlassesNet (1) for eyeglasses detection on the LFW database and Celeba database.

	LFW		Celeba	
	Without	**With**	**Without**	**With**
Without	0.9869	0.0131	0.9604	0.0396
With	0.0273	0.9727	0.0235	0.9765
Accuracy	98%		97%	

Table 4-4: Confusion matrix and average accuracy of Shallow-GlassesNet (2) for eyeglasses detection on the LFW database and Celeba database.

	LFW		Celeba	
	Without	**With**	**Without**	**With**
Without	0.9553	0.0404	0.9412	0.0533
With	0.0456	0.9587	0.0623	0.9432
Accuracy	96%		94%	

To demonstrate the generalisation ability of the Shallow-GlassesNet models, cross-database validation experiments were conducted on the LFW [5] database and Celeba database[6]. Table 4-3 and Table 4-4 show the confusion matrix and average accuracy of the Shallow-GlassesNet models.

To conduct a statistical evaluation of Shallow-GlassesNet (1) and Shallow-GlassesNet (2) in comparison with GoogleNet, the McNemar's statistical test was used to determine the statistical significance of the results. The null hypothesis assumes that there is no statistical difference in the performance of Shallow-GlassesNet (1), Shallow-GlassesNet (2), and GoogleNet. The Z-value and the related parameters (N_{ss}, N_{sf}, N_{fs}, N_{ff}) on the NVIE, LFW and Celeba databases are shown in Table 4-5, respectively. The table is split into three parts where each part shows the results of the McNemar's test on each database. The first column shows the comparison between Shallow-GlassesNet (1) and Shallow-GlassesNet (2), the second column the comparison between Shallow-GlassesNet (1) and GoogleNet, and the last column the comparison between Shallow-GlassesNet (2) and GoogleNet.

For a fair comparison, our methods should be compared with the methods that used the same data sets we used for training and testing. However, the NVIE database, which we used to train our methods, is commonly used for emotion/expression recognition from visual and thermal images and has not been employed before to train eyeglasses detection models in the literature to the best of our knowledge. Therefore, we compared the generalization ability of our methods with the state-of-art methods that use the Celeba and/or LFW databases. Table 4-6 provides this comparison.

Table 4-5: Comparison of the Z-value and the related parameters for Shallow-GlassesNet (1) and Shallow-GlassesNet (2) with GoogleNet on the NVIE, LFW and Celeba databases.

VS		**Shallow-GlassesNet (1) Shallow-GlassesNet (2)**	**Shallow-GlassesNet (1) GoogleNet**	**Shallow-GlassesNet (2) GoogleNet**
NVIE	N_{ss}	1352	1280	1270
	N_{sf}	30	102	87
	N_{fs}	5	0	10
	N_{ff}	3	8	23
	Z-value	4.06	10.0	7.72
LFW	N_{ss}	12455	11455	11055
	N_{sf}	511	1511	1609
	N_{fs}	209	89	489
	N_{ff}	58	178	80
	Z-value	11.2	35.5	24.4
Celeba	N_{ss}	184586	169586	159576
	N_{sf}	11621	26621	31312
	N_{fs}	6302	1652	11662
	N_{ff}	90	4740	49
	Z-value	39.7	148.49	94.78

4.7. Discussion

Compared to the GoogleNet model, Shallow-GlassesNet (1) and Shallow-GlassesNet (2) performed better in the NVIE database and achieved accuracies of (99.24%, 99.42%) and (96.96%, 97.63%) in the validation and testing datasets, respectively. When SVM was used, the accuracy of Shallow-GlassesNet (1) increased slightly (by 2.28% - 1.79%). Table 4-2 also reports the speed of the Shallow-GlassesNet in comparison with GoogleNet, showing that Shallow-GlassesNet (1) and Shallow-GlassesNet (2) are much faster than GoogleNet. The results in Table 4-3 and Table 4-4 show that the proposed shallow CNNs have achieved very high accuracy. It can be clearly seen from the results that the proposed method is highly robust to various challenging conditions.

Table 4-6: Comparison of generalization ability between the state-of-the-art results on the LFW and Celeba databases.

Methods	Training Dataset	Testing Dataset	
		Celeba	LFW
FaceTracer: HOG & colour histograms + SVM [174]*	Celeba + LFW	98	90
PANDA-w: Multiple CNNs features + SVM [175]*		94	84
PANDA-l: Multiple CNNs features + SVM [175]*		98	89
Face detector[176] + ANet [177]		96	88
LNets + ANet(w/o) [177]		96	92
LNets + ANet [177]		99	95
Multi-Task CNN-AUX [178]	Celeba	99.63	---
PS-MCNN-LC [179]		99.85	92.78
LBP + SVM [164]	LFW	---	98.65
Multi-Task CNN-AUX [178]		---	91.3
CTS-CNN [180]	WebFace [181]	99	91
Shallow-GlassesNet (1) [4]	NVIE	97	98
Shallow-GlassesNet (2) [4]		94	96

* The method is trained and tested by Liu et al. [177] with the same data sets, and the results are reported in [177]. The underlined results show the cross-database testing performance.

According to the McNemar's test, Shallow-GlassesNet (1) outperformed other models in all databases. The Z-values of Shallow-GlassesNet (1) vs. Shallow-GlassesNet (2) in the first column (4.06, 11.2, 39.7) and those of Shallow-GlassesNet (1) vs. GoogleNet in the second column (10, 35.5, 148.49) indicate that the differences between them are significant on all databases. The performance of Shallow-GlassesNet (2) is significantly better than GoogleNet based on the Z-values in the last column (7.72, 24.4, 94.78). The statistic results are reliable because $N_{sf} + N_{fs} > 20$ in all cases. Therefore, the null hypothesis can be rejected as $|Z|>1.96$, which indicates that the differences in performance are statistically significant.

Note that Celeba, LFW and WebFace databases represent the real-world environment with significant variations in expressions, poses, races, illumination, background, etc. The WebFace database is a large-scale database containing about 10,000 subjects and 500,000 facial images collected from the Internet. However, unlike other methods reported in Table 4-6, our proposed mothers are trained on the lab-controlled environment NVIE database but tested on the real-world environment databases, Celeba and LFW. Shallow-GlassesNet (1) achieved 97%, which is 2% less accuracy, than those achieved by the methods applying the real-world environment databases in their training stage.

As shown in Table 4-6, the cross-database testing performance for Shallow-GlassesNet (1) and Shallow-GlassesNet (2) when tested on the LFW database achieved 98% and 96%, much superior to PS-MCNN-LC [179] and CTS-CNN [180], achieved 92.78% and 91% respectively. Furthermore, they lead to equivalent or superior performance to almost all approaches that use the LFW database for testing and training.

4.8. Conclusion

This chapter presents an efficient and effective eyeglasses detection framework based on a well-designed shallow CNN, called Shallow-GlassesNet. First, the pre-trained GoogleNet was fine-tuned with images containing eyeglasses and images that did not contain eyeglasses. Then the learned weights of the GoogleNet were copied to the corresponding layers in Shallow-GlassesNet (1) and Shallow-GlassesNet (2), which were used as a feature extractor. A linear SVM was trained on the extracted features to detect eyeglasses. The proposed Shallow-GlassesNet architecture reduced the detection time by roughly a factor of two while retaining high detection accuracy. The main

contribution of this work lies in designing a shallow CNN model for detection of eyeglasses. Unlike most CNN designs, this shallow network architecture design is characterised by its combination of high precision and high speed, making it ideal for use in real-time applications.

Chapter 5

A Neural Network Approach to Decision Score Fusion for Emotion Recognition

5.1. Introduction

Emotion recognition has been an interesting research topic due to its potential applications in behaviour science, education and human-computer interaction, among others, and various facial emotion recognition approaches have been proposed in recent years [8, 49, 182-185]. However, the variety of human expressions and emotions among different subjects continue to make automatic emotion recognition a very challenging problem.

Beside the challenges related to any facial detection or classification systems, as described in Section 4.1, facial emotion recognition system faces additional challenges which are related to the human emotion complexity. Whereas it is widely believed that emotions are universally recognised in facial expressions, there are different levels of intensity of expression which cause interference between different emotions that appear in facial expressions, thus compounding the challenges in developing emotion recognition systems in particular.

According to Wang et al. [47], deep learning models have been used to solve some challenges in computer vision tasks and could achieve better performance than other state-of-the-art representations. Due to the advances in deep learning, CNNs have recently become the most widely used approach in computer vision. Instead of using a

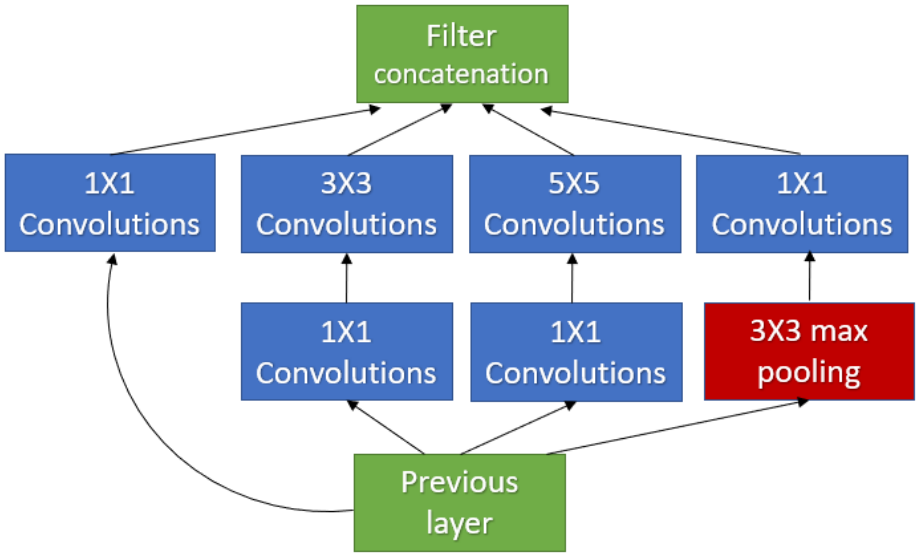


Figure 5-1: The inception layer in GoogleNet.

CNN to classify facial images directly, the system proposed in this chapter utilises a CNN as a feature extractor to extract facial features for other powerful classifiers. According to the GoogleNet design, the inception layer features may convey more useful information because the output of an inception layer is built up of the concatenated features from all the large convolutions. Thus, it could be considered as a feature fusion layer, as shown in Figure 5-1. We suggested utilising these layers as a feature provider in the proposed system.

Decision fusion strategies to combine multiple classifiers can produce more reliable and robust recognition than the application of a single classifier. Many studies have also proved that a single feature set with a single classifier which has a unique generalised classification approach often does not deal effectively with the high degree of variability and complexity encountered in many applications in the domain of computer vision. To deal with many complex applications such as emotion recognition, multiple classifiers can be utilised by acquiring information through multiple features extracted from multiple processes and then applying a decision fusion process to

combine them. The design of the multiple classifier system depends on the type of classifier outputs. Abstract classifiers produce only a label without any other information. Rank classifiers give an ensemble of possible classes ranked. Confidence or measurement classifiers give a vector of scores, each associated to a possible class [186].

The decision score fusion (score level fusion) refers to the combination of matching scores provided by the measurement classifier in the system. This is also known as fusion at the confidence level or measurement level. It is the most commonly used decision fusion approach, as evidenced by the experts in the field [187, 188]. To enhance the accuracy, multiple score fusion approaches have been developed by using the most common measurement classifier, i.e. SVM [189, 190]. The main contributions of this chapter [7] are as follows:

- An effective facial emotion recognition system is proposed to estimate the emotion from visible images by classifying them to one of the six universal emotions (Anger, Disgust, Fear, Happiness, Sad & Surprise) and Neutral. The proposed system designed special multi-layer perceptron neural network modes to fuse SVMs scores.
- A novel neural network model for decision fusion was proposed, which increase the classification accuracy by up to 10%.

The rest of this chapter is organised as follows. Related work is reviewed in Section 5.2. The proposed system is introduced in Section 5.3. The experimental setup and databases used are described in Section 5.4. Preliminary results and discussion are presented in Section 5.5 and Section 5.6 respectively. Section 5.7 draws conclusions.

5.2. Related Work

5.2.1. Facial Expression Recognition

Most existing facial emotion recognition methods can be generally categorised into two approaches: handcrafted feature approach and deep learning approach. With handcrafted feature approach expert knowledge is used to extract features from images according to a certain manually predefined algorithm. Traditional facial emotion recognition systems rely on many handcrafted features to recognise the features and parts of the face individually. On the other hand, Aleksic [191] proposed an automatic facial expression recognition system that uses multi-stream Hidden Markov Models (HMMs). They used Facial Animation Parameters (FAPs) which control the movement of the outer lips and eyebrows and used them for classification as visible features.

Contrary to the handcrafted features, deep learning features are derived from a training image dataset by using deep learning methods which effectively use the feedback information to investigate the suitability of the extracted features. In recent years, most existing facial emotion recognition methods rely on the deep learning approach [192-194]. CNNs are examples of deep neural networks which can be used to learn deep features.

5.2.2. Score Level Fusion

In a complex pattern recognition environment such as emotion recognition, multiple classifiers with decision combination can help alleviate many computer vision problems. Score-level fusion approaches can be broadly classified into two categories: (a) not-trainable (fixed fusion rules) and (b) trainable. The not-trainable method combines the outputs of the classifiers in simple ranking techniques such as voting,

sum, mean, median, product, min and max. The trainable method uses the scores as input features for a new pattern-classification problem such as Neural Networks, SVM [139].

Luis et al. [187] used a number of score fusion approaches (Neural Networks, SVM, Weighted Sum, and so on) with three independent monomodal biometric systems. They compared the behaviour of score normalisation techniques (z-norm, MAD, tanh, and so on), and proposed a new score normalisation procedure. They concluded that the improvement depends upon the normalisation algorithms as well as the case in consideration. R. Dzati et al. [195] enhanced their system performance by using SVM for score fusion. They created a score vector by appending the scores from multiple correlation filter outputs and used SVM to obtain the final decision. Nandakumar et al. [196] proposed a framework for the optimal combination of multimodal match scores which depend on the likelihood ratio test. He et al. [197] examined the performance of two score level fusion approaches: sum rule-based and SVM-based. They proposed a new robust normalisation scheme (Reduction of High-scores Effect normalisation) using three biometric traits: face, fingerprint and vein pattern, and demonstrated that their normalisation scheme with simple sum rule-based fusion could attain a better performance than ratio-based fusion [196].

5.3. Methods

A schematic diagram of the proposed score fusion based facial emotion recognition system is shown in Figure 5-2. The system can be divided into three modules: eyeglasses detection [4], deep feature extraction and classification, and decision fusion. The proposed system starts by detecting the face from the image; then, it is classified using the eyeglasses detector. Next, the fine-tuned CNNs are utilized to extract the

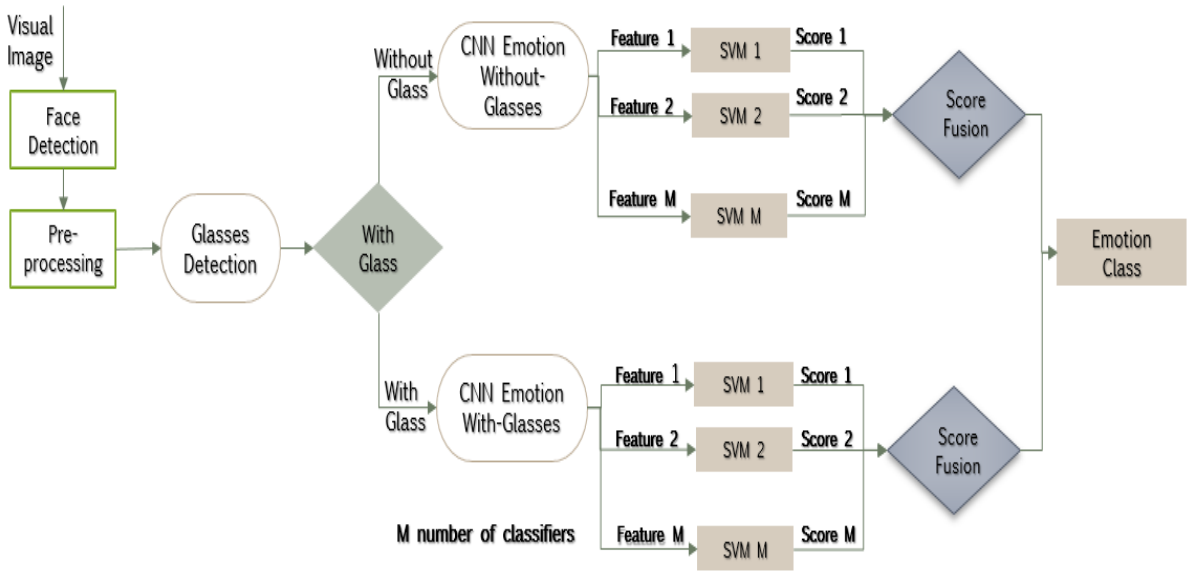


Figure 5-2: The proposed facial emotion recognition system.

features which SVMs use to predict the scores for the seven emotions. Finally, the proposed score fusion models classify the emotion of the image by using the collected scores from the SVMs. The details are provided in the following subsections.

5.3.1. Pre-processing and Eyeglasses Detection

We first used the Viola-Jones face detection approach [3]. After obtaining and cropping the frontal face region from the visible image, the cropped face image was re-sized to match the input size (224×224) of the CNN (GoogleNet). To enhance the system accuracy, the eyeglasses detector proposed in Chapter 4 was utilised to detect images of eyeglasses, which achieved a mean accuracy of 99.73% [4]. Appropriate CNNs were then selected to extract features, according to the eyeglasses detector result.

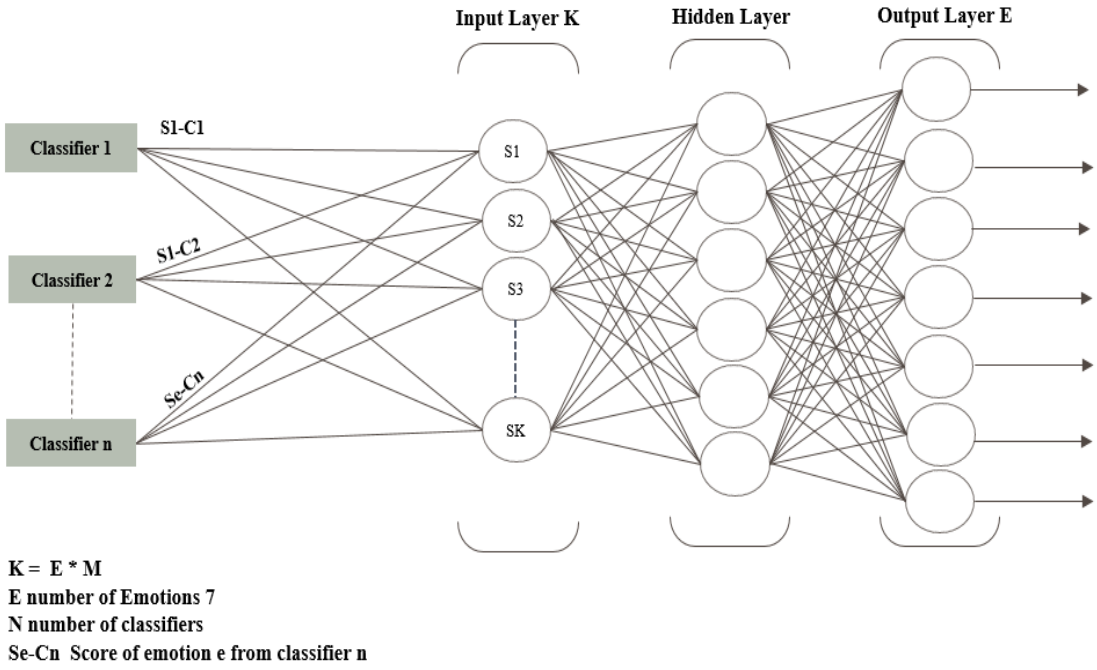


Figure 5-3: The multi-layer perceptron (MLP) neural network used for decision score fusion.

5.3.2. CNN-based Learning of Deep Features

Instead of using CNN to classify the images directly, CNNs were utilised as feature extractors to extract features for SVM classifiers. This was because we believed that the inception layer features may convey more useful information because the output of an inception layer is a big feature map which is built up from the concatenated features from all the large convolutions. Since GoogleNet CNN architectures consist of nine inception layers, we extracted a feature set from each inception layer. The nine feature sets which were extracted from the CNN trained on visible images were fed to SVM classifiers to predict scores for the seven emotions. The proposed framework consists of 18 classifiers: nine classifiers were used when eyeglasses were detected and the other nine when no eyeglasses were detected.

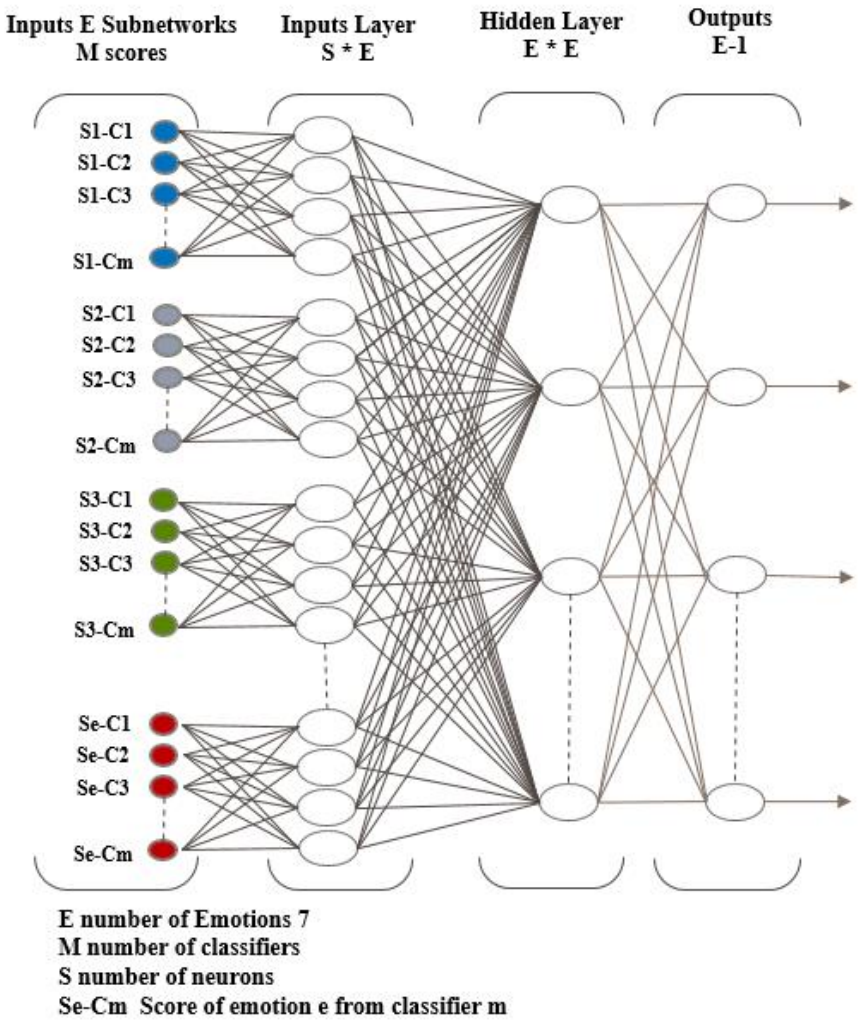


Figure 5-4: The neural network with Subnetworks in Input Layers (SIL) used for decision score fusion.

5.3.3. Decision Score Fusion

The output scores from nine SVM classifiers for the seven emotions were used by three not-trainable decision fusion methods and two trained neural network fusion models, as shown in Figure 5-3 and Figure 5-4, to evaluate the performance of the proposed methods. The first not-trainable fusion rule was majority voting, where each classifier gave its own vote and the class with more votes was chosen as the final result. The second not-trainable fusion rule was average voting, where the average score of each class was calculated and the final result given to the class with the highest average. The

last not-trainable fusion rule was maximum score, where the class with the maximum score was chosen.

The first trained fusion model, which is depicted in Figure 5-3, was a multi-layer perceptron (MLP) neural network. The input of this network came from the scores contributed by the nine SVMs. Each score was connected to all neurons in the input layer. The second trainable fusion model, as shown in Figure 5-4, was a neural network with Subnetworks in Input Layers (SIL), which consisted of an ensemble of seven subnetworks, each representing a single emotion. Each subnetwork had nine input scores contributed by the nine SVMs and nine hidden neurons. The input of each subnetwork were the scores for individual emotions collected from the nine SVMs. The SIL neural network had seven subnetworks to represent the seven emotions. The difference between the MLP and SIL neural networks lay in the input layer. In the MLP neural network, each score was connected to all neurons, whereas in the SIL neural network, scores of each emotion were connected to the subset of neurons which represented this specific emotion.

The process of training Subnetworks in Input Layers (SIL) network with one hidden layer.

The input scores are 63 =($E\ x\ M$), where E = 7 is the number of emotion classes and M=9 is the number of SVMs. Denote the set of scores $S = \{s_1, s_2, s_3, \dots s_E\}$ weights W= (0 - 1) is and the connections between the layers.

(1) The output value of each neuron in input layer is calculated by using the weighted summation of the 9 input scores. The following equation (5-1) is used to calculate:

$$\forall \boldsymbol{S} \in \{\mathbf{1}, \mathbf{2}, \dots, \boldsymbol{m}\}, \boldsymbol{h_l} = \sum_{\boldsymbol{i=1}}^{\boldsymbol{E}} \boldsymbol{W_{il}^H S_i} + \boldsymbol{\beta_l^H} \qquad \textbf{(5-1)}$$

where E is the total number of, W_{il}^H is the connection weight between i input neuron and the hidden neuron l, S_i is the input score i, and β_l^H is the bias of the l^{th} hidden neuron.

(2) The following sigmoid activation function (5-2) is used to map the weighted summation to the hidden layer

$$\forall \boldsymbol{l} \in \{\mathbf{1}, \mathbf{2}, \dots, \boldsymbol{j}\}, \boldsymbol{H_l} = \boldsymbol{sigmoid}(\boldsymbol{h_l}) = \frac{\mathbf{1}}{\mathbf{1}+ \boldsymbol{e^{-h_l}}} \qquad \textbf{(5-2)}$$

(3) The output of the network is calculated using the following equations (5-3), (5-4):

$$\forall \boldsymbol{p} \in \{\mathbf{1}, \mathbf{2}, \dots, \boldsymbol{n}\}, \boldsymbol{O_p} = \sum_{\boldsymbol{I=1}}^{\boldsymbol{j}} \boldsymbol{W_{lp}^O H_l} + \boldsymbol{\beta_p^O} \qquad \textbf{(5-3)}$$

$$\forall \boldsymbol{k} \in \{\mathbf{1}, \mathbf{2}, \dots, \boldsymbol{n}\}, \boldsymbol{O_p} = \boldsymbol{sigmoid}(\boldsymbol{O_k}) = \frac{\mathbf{1}}{\mathbf{1}+ \boldsymbol{e^{-op}}} \qquad \textbf{(5-4)}$$

Where W_{lp}^O is the weight between the l^{th} hidden neuron and the P^{th} output neuron which is equal to 7. β_p^O is the bias of the P^{th}output neuron.

5.4. Experiments

5.4.1. CNN Architecture Setup

The Caffe toolkit [173] on an NVIDIA GeForce GTX 980 GPU was applied to fine-tune the pre-trained GoogleNet Deep CNN model [93] using the NVIE dataset. The NVIE training dataset was used for fine-tuning by scaling the cropped facial image data to 224×224×3 so as to fit the CNN model input requirement. Four-fold cross-validation was adopted to find the optimal values for the parameters of each CNN. Two deep CNNs were trained using the stochastic gradient descent with a batch size of 50. The hyper-parameters of the applied training algorithm were as follows: momentum=0.9, weight decay=0.0002, initial learning rate=0.001.

5.4.2. Experiment Design

In this section, we first describe the image databases used in the training phase. Thereafter, we examine the effect of using the eyeglasses detector on the SVM's performance on the validation dataset. Finally, we compare the performance of common decision fusion methods against the proposed neural network model in respect to the validation and testing datasets.

The USTC-NVIE [8] (NVIE) database was adopted in the experiments. A detailed description of the NVIE's partitions and the enlargement techniques is explained in Section 4.5.1. Table 5-1 illustrates the numbers of samples for each emotion class in the training, validation and testing datasets.

Table 5-1: The number of visible samples for different emotions in the training, validation and testing sets on the NVIE database.

	Without-Eyeglass			*With-Eyeglass*		
	Train	***Val***	***Test***	***Train***	***Val***	***Test***
Anger	1281	57	63	1287	63	67
Disgust	1253	57	62	1267	63	67
Fear	1281	57	63	1288	63	67
Happy	1281	57	62	1288	63	67
Neutral	1300	299	298	1300	297	315
Sad	1279	57	63	1279	63	67
Surprise	1271	57	62	1271	63	67
No. Subjects	61	20	20	61	20	20
No. Images	8946	641	673	8959	675	717

To make it subject-independent throughout our experiments, the scores of classified images in the four-fold validation datasets were collected in order to create a new score dataset, which was then partitioned into two sets: training and validation score sets, on which the MLP and SIL neural networks were trained and validated. We tested these two models on the score testing dataset which was created from the collected scores of classified images in the four-fold testing datasets.

5.5. Results

To demonstrate the performance of the proposed approach, several SVMs were trained on features extracted from nine different inception layers, with and without using the eyeglasses detector. Before starting the comparisons between the trained SVMs, four-fold cross-validation was adopted to find the average accuracy among them.

Table 5-2: Average accuracy of four-fold cross-validation with GoogleNet inception layers, with and without the eyeglasses detector.

Classifiers	GoogleNet Layer-Name	Number of Features	Without Glasses Detection	With Glasses Detection	Accuracy Increase
1	Inception_3a	200704	41.7	48.5	6.8
2	Inception_3b	376320	38.3	45.6	7.3
3	Inception_4a	100352	38.4	45.2	6.8
4	Inception_4b	100352	35.2	40.2	5.0
5	Inception_4c	100352	35.0	43.7	8.7
6	Inception_4d	103488	30.4	40.4	10.1
7	Inception_4e	163072	31.4	37.6	6.3
8	Inception_5a	40768	30.3	35.1	4.8
9	Inception_5b	50176	24.8	33.0	8.2
Softmax		1024	26.5	29.5	3.0

The last column in Table 5-2 (Accuracy Increase) reports the difference between the previous two columns (With Glasses Detection and Without Glasses Detection) to demonstrate that the accuracy increased in all the layers when using the eyeglasses detector. To demonstrate the performance of the proposed neural network model for score fusion, comparisons of different decision score fusion approaches were done, as illustrated in Table 5-3. To conduct a statistical evaluation of the SIL and ML Neural Networks, the McNemar's statistical test was utilized to determine the significance of the results. The null hypothesis assumes that there is no statistical difference in the performance between the SIL and ML Neural Networks and the commonly used decision fusion strategies. The Z-value and the related parameters (Nss, Nsf, Nfs, Nff) for SIL and ML Neural Networks on the NVIE database are shown in Table 5-4 and Table 5-5, respectively.

Table 5-3: Comparison of the ML and SIL Neural Network with other decision fusion strategies (Average, Max, Majority voting and ML Neural Network) on the NVIE database.

Decision Fusion Methods	**With Glasses Detection**	
	Val	***Test***
Average	53.95	53.60
Max	53.04	50.36
Majority Voting	46.73	48.06
ML Neural Network	59.27	54.60
SIL Neural Network	64.13	61.15

Table 5-4: Comparison of the Z-value and the related parameters for the SIL Neural Network with other decision fusion strategies (Average, Max, Majority voting and ML Neural Network) on the NVIE database.

SIL Neural Network VS		***Average***	***Max***	***Majority Voting***	***ML Neural Network***
NVIE	N_{ss}	740	684	605	609
	N_{sf}	110	166	245	241
	N_{fs}	5	16	63	150
	N_{ff}	535	524	477	390
	Z-value	9.7	11.0	10.3	4.6

Table 5-5: Comparison of the Z-value and the related parameters for the ML Neural Network with commonly used decision fusion strategies (Average, Max and Majority voting) on the NVIE database.

ML Neural Network VS		***Average***	***Max***	***Majority Voting***
NVIE	N_{ss}	740	680	658
	N_{sf}	19	79	101
	N_{fs}	5	20	10
	N_{ff}	626	611	621
	Z-value	2.7	5.8	8.5

Among seven given expressions in the NVIE database, only three expressions, including happiness, disgust, and fear, were successfully induced by most subjects [1]. Thus, many state-of-the-art methods used these three expressions in their experiments, such as Wang et al. [8] and Lee et al. [198]. However, to develop a technique for recognising the facial expressions and emotions in real-time, it should classify most emotions and expressions that would typically occur in real life. Therefore, we conducted experiments for all seven expression classes involving the other expressions that most subjects did not successfully induce. This would significantly increase inter-class similarity cases and thus affect the classification accuracy of the proposed methods. This should be kept in mind when comparing against state-of-the-art methods.

Another critical factor that directly affects the performance of the comparison methods is the type of evaluation. Some state-of-the-art methods were evaluated by using cross-validation according to the images in the NVIE database. As a result, the training and testing datasets contain many similar images for the same subject performing the same expressions. Therefore, these state-of-the-art methods would achieve remarkably high accuracy. For example, the framework proposed in [199] achieved 93.63% of its average accuracy for recognizing five facial expressions on the NVIE database. Table 5-6 presents comparisons between the proposed methods and the reported results of recently state-of-the-art methods on the NVIE database. For a fair comparison, Table 5-6 includes only the state-of-the-art methods which evaluate their performance using cross-validation according to the subjects.

Table 5-6: Comparison with the state-of-the-art results on the NVIE dataset.

Methods		**No Emotions**	**Accuracy**
NVIE baseline [8]	PCA+LDA	3	58.47
	PCA+LDA+KNN	3	65.25
	AAM+KNN	3	67.80
	AAM+LDA+KNN	3	61.86
LBP + SRC [200]		3	59.5
LPQ + SRC [201]		3	62.17
Gabor + SRC [202]		3	65.00
LBP-TOP [69]+SRC		3	65.67
LPQ-TOP [107,203] + SRC		3	66.17
ML Neural Network [7]		**7**	54.60
SIL Neural Network [7]		**7**	61.15

5.6. Discussion

Table 5-2 also shows that higher accuracy was obtained from lower layers which compound a larger number of features. Since the number of features decreases in higher layers, Softmax has the lowest accuracy compared with when the SVM was trained on the deep features. In other words, the most discriminatory features could be lost as we go deeper through the CNN network. Thus, the proposed system uses CNN to extract the features instead of using CNN directly for recognition. The comparisons of different decision score fusion approaches indicates that the performance of the system using our neural network model was significantly better than the other score fusion approaches, as illustrated in Table 5-3.

According to the McNemar's test results shown in Table 5-4, the SIL Neural Network significantly outperformed the commonly used decision fusion strategies (Average, Max and Majority Voting) and the ML Neural Network because the Z values are greater than 1.96 in all cases. Table 5-5 shows that the ML Neural Network significantly outperformed the commonly used decision fusion strategies. All the statistic results are reliable because $N_{sf} + N_{fs} > 20$ in all cases. The null hypothesis is rejected with an error probability of 0.05, which indicates that the differences in performance are statistically significant.

The classification rates of the proposed system using the SIL Neural Network model achieves a recognition rate of 61.15%, which is higher than the recognition rates of the PCA+LDA baseline method. Moreover, it performs a competitive result of about 0.71–6% less than the other methods that classify three expressions compared to seven expressions to the proposed system.

5.7. Conclusion

This chapter shows that when the accuracy of individual classifiers is low, a neural network could be used as a useful decision fusion approach which learns from classifier mistakes and gives a more accurate decision. We proposed a novel neural network model for score fusion to improve the emotion recognition performance. Nine different sets of emotional features were extracted from faces by using inception layers in GoogleNet to train individual SVMs. According to the examination of the system accuracy for each individual feature set, classification rates increased by up to 7-15% when the eyeglasses detector was used. The classification rates of the system increased by about 10% when using the SIL neural network approach in the multiple classifier system for score fusion.

Chapter 6

One-Shot Only Real-Time Video Classification: A Case Study in Facial Emotion Recognition

6.1. Introduction

Nowadays, watching videos is considered to be a critical means for people to satisfy their entertainment and information needs. Due to the widespread use of videos, there is no doubt that analysing and recognising video content has become a staggeringly popular research area in the computer vision field. Indeed, the rapid advances in video technology has only served to increase the need for real-time applications for video analysis. Thus, video recognition systems have wide demand across many real-world applications, such as in visual surveillance, human-robot interaction and autonomous driving vehicles, etc.

This study designs two novel methods for real-time video classification and applies them to recognise emotion from videos. The proposed methods classify the video clips to one of the six universal emotions (Anger, Disgust, Fear, Happiness, Sad & Surprise). Inspired by the You Only Look Once (YOLO) system for real-time object detection [204], this chapter proposes a general model called One-Shot Only (OSO) [9] for video classification, which converts a video-based problem to an image-based one by using frame selection or clustering strategies to form a simple representative storyboard for spatio-temporal video information fusion. The work in this chapter is

different from that of Jing et al. [130] (Video You Only Look Once for Action Recognition), which uses complex 3D-CNNs to learn the appearance and temporal information from the whole video and classify the actions it contains in a single process by designing a total of eight types of 3D-CNN to handle different lengths of video clips. Using 2D-CNNs without losing the temporal information due to the use of the storyboard representation of videos, the OSO methods proposed in this chapter [9] can not only meet the requirements for real-time video analysis but also produce competitive video classification accuracy by combatting the overfitting problem existing in other commonly used 2D-CNN architectures for video classification. The main contributions of this chapter are as follows:

- A novel spatio-temporal data fusion approach to video representation is proposed, which speeds up video classification while delivering competitive accuracy to meet the requirements of real-time applications.
- Frame selection and clustering strategies are proposed to handle videos of different lengths, as well as the redundancy in consecutive video frames, leading to two OSO methods for effective video representation. The OSO methods reorganise video frames hierarchically as a single image, from which common 2D-CNN models can predict the video class probabilities.
- The OSO methods are evaluated using three sizes of storyboard for video representation and seven common 2D-CNN models for video classification. It is demonstrated that the OSO methods can be used to classify both images and videos and are able to improve the recognition accuracy when classifying emotion from images as well as videos.

The remainder of this chapter is organised as follows: Section 6.2 gives some insight into the challenges in developing video emotion recognition systems. Section

6.3 reviews the existing emotion recognition approaches and their advantages and disadvantages, according to their evaluation in both lab-controlled and real-world environments. Section 6.4 gives a review of the related work. Section 6.5 explains the methodology of the proposed systems. The experiments are described in Section 6.6. In Section 6.7, the results of the proposed approaches are analysed and discussed in Section 6.8. Finally, this study is concluded in Section 6.9.

6.2. Video Emotion Recognition Challenges

The state-of-the-art algorithms which currently implement video-based recognition still fall far short of the requirements of real-time applications due to a number of significant challenges such as:

Intra-class and inter-class variations: This kind of confusion is often met in relation to real-world image datasets, and is even more significant in relation to real-world video datasets — especially when a video sample is treated as a stack of images [205, 206]. A significant number of intra-class variation and inter-class similarity cases confuse many of the existing action recognition algorithms and thus affect their classification accuracy.

Emotions interference: In order to tackle this naturally-occurring problem, some real-world datasets have defined multiple compound emotion categories such as happily-surprised, fearfully-angry [11, 19]. For example, RAF-DB [11] has categorised the affective faces using two types of classes: basic and compound emotions. The basic type classes are surprise, fear, disgust, happiness, sadness and anger (the six basic emotions), plus neutral, whereas there are twelve compound emotion classes. Another way to address this problem is that there is an uncertain category, such as in AffectNet

[10] and RAF-DB [11]. The image is tagged as uncertain when the annotator is uncertain about any of the facial expressions exhibited in it. However, this is a challenging obstacle facing the video-based emotion recognition systems as the dynamic properties of emotional development of the subject in the video clip would increase the emotional interference problem in each frame, which significantly impacts recognition accuracy.

Data reliability: AffectNet [10] uses its *uncertain* category only where the annotators are completely uncertain about all of the facial expressions shown in the associated image. In their study, it was shown that the rate of agreement between two annotators in terms of their annotations of a randomly selected set of 36,000 images was 60.7%. Moreover, it is more difficult and less precise to label specific emotions from videos or reality shows, since the start and endpoints of the expression of an emotion is not always easily detectable. This may affect the performance of emotion recognition algorithms that do not have tolerance of inaccurate labelling. Due to budget and time constraints, however, the number of labellers used for most of the existing datasets can be considered too small, and often each image or video was classified by only one labeller (e.g., this is the case with AffectNet [10]).

Real-world conditions*:* Although the state-of-the-art results from a number of existing video-based recognition approaches have achieved impressive accuracies of around 96-97% [207, 208] on lab-controlled environment datasets (CK+ dataset [120]), the technical problems involved with transposing such approaches from the lab to real-world applications result in much lower accuracies, achieving 41~47% only on the AFEW dataset, for example [209]. It is a substantial challenge to generalise these approaches so that they can work adequately in the context of real-world applications

with extreme and "wild" environments, mostly due to the fact that, in relation to such environments, these approaches cannot be trained on large-scale datasets.

Insufficient annotated video data: The RAF-DB [11] and the large-scale facial emotion datasets in AffectNet [10] each contain more than a million images, whereas the existing video datasets, such as AFEW [12] and CK+ [120], contain only thousands of samples. To collect and annotate a large number of samples in AFEW, a video clip recommender system was utilized to automatically search in a movie for clips with a subject showing a meaningful expression. Then two annotators revised only the suggested video clips rather than manually scan the full movie. However, the samples in the video datasets are still insufficient for the purpose of training deep neural networks with millions of free parameters. Additionally, the number of annotators in this database is small, which affects the data reliability.

In general, recognising and knowing what is happening in a video is a very challenging task, and dealing with the temporal dimension in videos remains a vivid research issue as it has a direct impact on the efficiency of the system in terms of speed and accuracy.

6.3. Video Recognition Approaches

In order to explain what have been done in the literature to address these challenges, the frameworks of current video recognition systems are briefly described in this section and the main reasons for their weaknesses are clarified since these need to be clearly understood before applying them in reality.

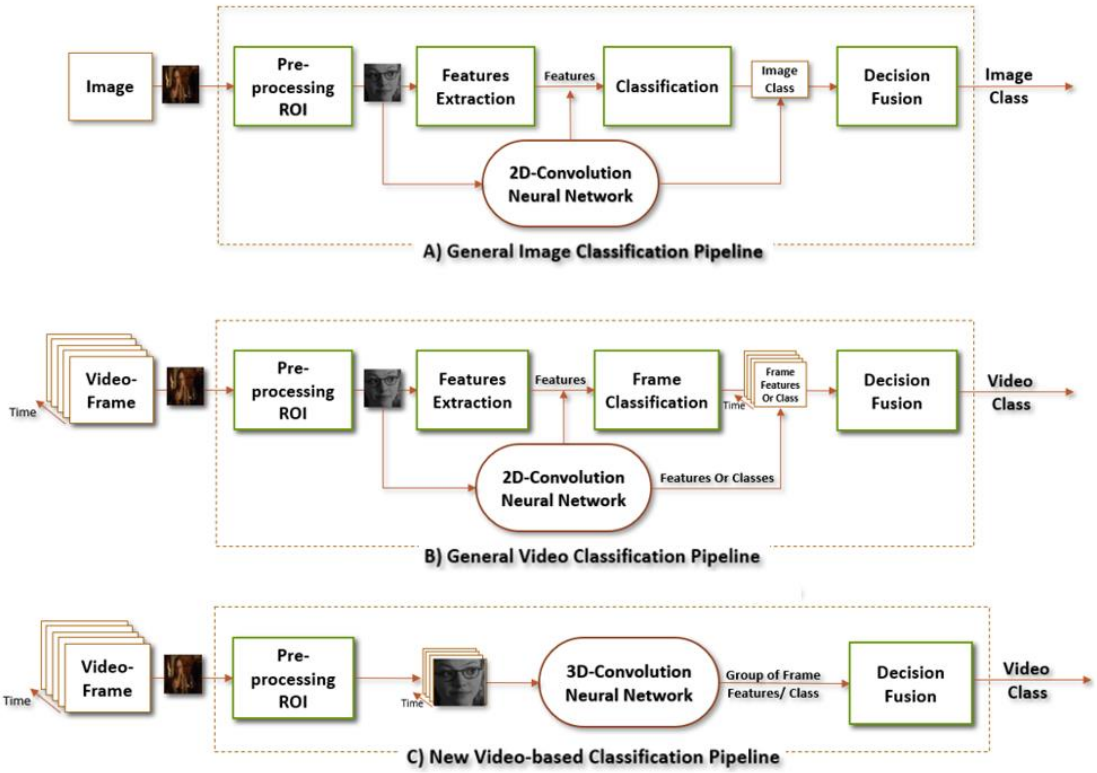


Figure 6-1: General pipelines of image-based and video-based classification systems.

As illustrated in Figure 6-1 (A), the general frameworks of image-based classification follow an old-fashioned pipeline which consists of four basic operations: pre-processing and region-of-interest extraction, features extraction, classification and decision fusion. The state-of-the-art methods regularly utilise CNNs either as an end-to-end model or as a feature extraction model. In both cases, CNNs achieve superior performance compared to handcrafted features and have had some notable successes at image-based recognition problems, elevating them to become an essential component in the current pipelines of image-based recognition frameworks [210].

Recently, numerous approaches have been proposed applying CNNs for video classification tasks such as action recognition and emotion recognition. Many of the state-of-the-art approaches follow a single-frame processing approach called the naïve approach, which treats video frames as still images and applies classifiers or CNNs to

classify each frame before fusing the predictions to get the final decision at the video level [42, 133, 211, 212], as illustrated in Figure 6-1 (B).

In the naïve approach, the classifiers, or a bunch of CNN models, are trained separately, before a decision rule, such as the sum rule or majority voting, combines their outputs. In this naïve approach, the order of frames is not very important and there are no relations between the successive frames so that the videos can be seen as unordered sets of frames. However, because a video is not just a stack of frames, but every single frame draws a small portion of the video's story, the single-frame approach without using temporal information has many drawbacks. To address this issue, it is natural to attempt to take advantage of temporal information in order to have a better performance. Indeed, some existing models [213-216] apply feature fusion approaches to aggregate the partial temporal information from individual frames or short clips.

6.3.1. Drawbacks of Single-Frame Processing Approaches

Misleading frames: A video might contain frames that are irrelevant or misleading in respect to the emotion or action of interest. This means that this approach could easily become confused and misclassify the video, especially when temporal components are not used and if there are many contaminant frames.

Overfitting: The training data is generally a set of images extracted from training videos frames. Most of the consecutive video frames are very similar, and some frames can be almost identical. This redundancy in the training data causes an overfitting problem, especially when training CNN models. Several frames are extracted from each video and the similarity of regions of interest in successive frames usually goes hand-in-hand with an overlapping problem, which reduces intra-class variation. This

in turn creates an overfitting problem in the training process. For example, in the facial-emotion recognition system, the region of interest is the subject's face, which is detected and extracted from each video frame. This inevitably results in numerous images containing the same subject's face with little variance. Thus, when the CNN is trained on these very similar images of a specific subject, it will learn the subject's face instead of his/her emotion and end up trying to find the similarities between the learned face and the tested faces. This considerably reduces the system accuracy. In addition, if the training dataset contains many videos of the same subject performing different emotions, this produces many similar images belonging to different classes which results in a reduced inter-class variation that makes the CNN hard to train.

Inconsistency: Another problem is the inconsistency between some of the extracted frames and the ground truth of the video clip, especially in respect to the emotion recognition system. As the complexity and intensity of the preformed emotion varies in video frames, not all extracted frames reflect the same category of the video. So even in a reliable database, these frames have a misleading ground truth which will add confusion in the learning and testing process.

To address the problems of the single-frame baseline approach for video-based classification, On the other hand, three-dimensional convolutional neural network models (3D-CNN) have made significant improvements in various video analysis tasks. Recently, various studies have utilised 3D-CNN models [103, 130, 217] to learn spatial-temporal features, as shown in Figure 6-1 (c). They are just like standard CNNs (2D-CNN), but applying additional spatio-temporal filters to represent spatio-temporal data. Since 3D-CNN models have many more parameters than 2D-CNN models, they are more complex and harder to train.

To summarise, compared to image-based methods, video-based methods are more complex and have three major problems in the process of constructing real-time recognition systems: overfitting, slow and hard to train or optimise.

6.4. Related Work

Video-based recognition is an essential branch among the studies of both computer vision systems and human perception. Numerous researchers have made intensive efforts to improve audio-visual emotion recognition based on images and videos [218]. This review of related work will focus on visual emotion recognition based on videos and what has been done to improve the classification performance by utilising spatio-temporal information.

Current studies deal with the temporal information in videos by splitting a whole video into either groups or individual frames and consequently processing these portions multiple times. Several models are usually utilised to aggregate the processed parts to implicitly infer the whole temporal information. Based on the number of frames processed at a time, the current state-of-the-art models for video-based recognition fall into four categories: *Single-Frame*, *Set-of-Frames*, *All-Frames* and *Key-Frames*, as illustrated in Figure 2-13.

6.4.1. Methods for Video Classification

Many methods for video classification for emotion recognition have been proposed in response to the Emotion Recognition in the Wild Challenge (EmotiW). Section 2.7 presents a comprehensive review of the video-based approaches proposed in EmotiW.

6.4.2. Key-Frame Selection Strategies

Since a short-length video clip of 2-3 seconds could contain 70-90 frames, processing each frame separately, even in a short-length video, is time-consuming and would usually affect the system accuracy. While many studies have proposed for key-frame selection strategies to handle this problem [219, 220], these strategies are either complicated and computationally expensive or do not work effectively. Some of the common key-frame selection strategies are as follows:

(1) Predefined-frames strategy. This takes specific frames depending on their position in the video, such as the middle and boundary frames [221, 222]. Although this strategy is straightforward and fast, it depends on some knowledge of the dataset [222]. In addition, predefined frames are usually not stable and do not detain most of the visual content.

(2) Motion-analysis strategy. This computes the optical flow for each frame in order to evaluate the changes in the facial expression [19, 223, 224], using specific points such as the left and inner eyebrow or the corners of the mouth.

(3) Visual-content strategy. This computes similarities between frames represented by colour histograms or other features [225]. The first frame is chosen as the first key-frame, then the similarity between adjacent frames is computed and the frame which has a significant change in content is selected as the next key-frame.

(4) Clustering strategy. This clusters similar frames by assigning each frame to a corresponding group and then selects the centroid frame of each group as key-frames [225, 226]. Although clustering methods can achieve good results in general, noise and motion can easily affect their performance. In addition, the key-frames selected may be only from the dominant clusters.

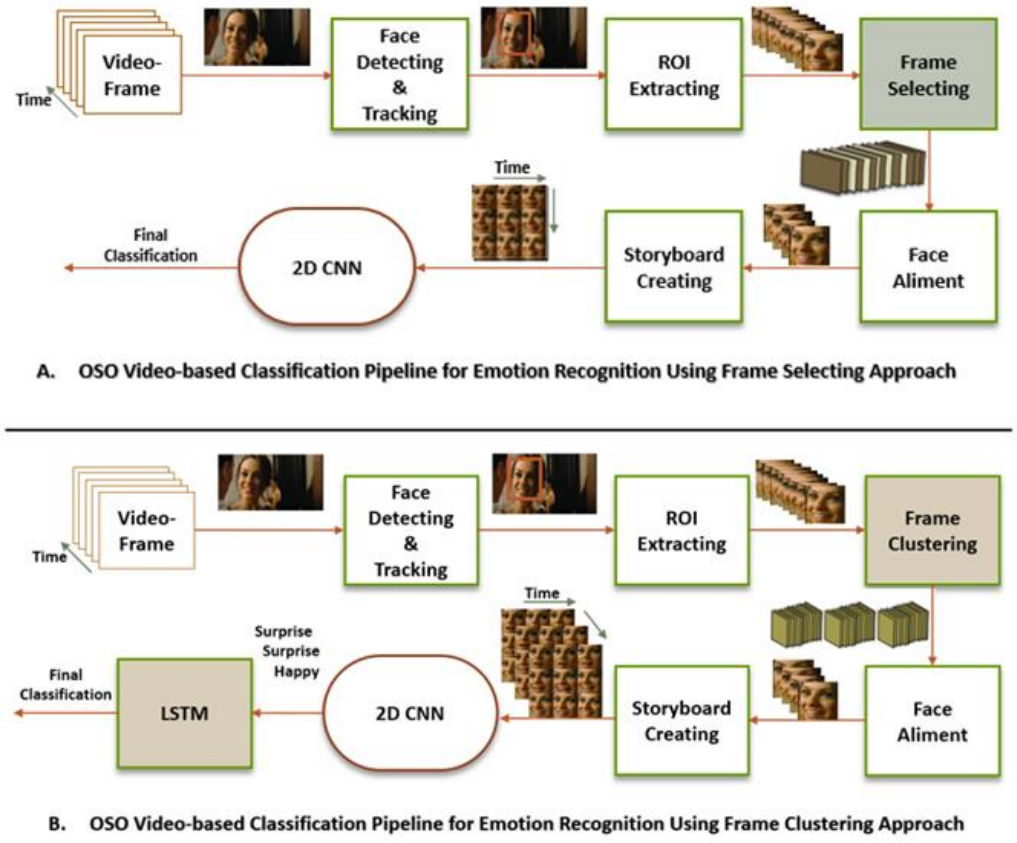


Figure 6-2: OSO video-based classification pipelines for emotion recognition.

6.5. Proposed Methods for One-Shot Only Real-Time Video Classification

Two OSO methods for facial emotion recognition based on video classification are proposed in this chapter, named *Frame Selecting Approach* and *Frame Clustering Approach*, which benefit from the hierarchical representation of spatio-temporal information in video frames. The structures of the proposed OSO approaches are shown in Figure 6-2 (A, B). Both approaches apply three pre-processing steps that detect and track faces across the video frames, and then extract the ROI of the detected faces. Then the facial landmark points are used to align the faces of the frames chosen by frame selection or clustering strategies. The pre-processed facial images are combined to create a storyboard in the form of a single image, in which spatio-temporal information

fusion is conducted at the raw data level, i.e., at ROIs of the selected video frames.

In the frame selecting approach, as shown in Figure 6-2 (A), the storyboard is created from selected frames and is used as the input to a 2D-CNN which predicts the emotion class of the video directly. Video clips have different lengths or different number of frames. Also, the period of the same emotion may vary when performed by different subjects or by the same subject at different times. When selecting only a small number of frames from a video clip showing the emotion, it is critical to select the frames which are most different from the average frame of the whole original video.

In the frame clustering approach, as shown in Figure 6-2 (B), the video frames are clustered into groups of frames with certain similarity, and a storyboard is then created for each group respectively. In a video clip, the subject might start by showing one emotion and end up by presenting quite another. In other words, the clip may contain several consecutive emotions: pre-emotion, post-emotion and the main one. For example, the “surprised” emotion may be followed by one of the post-emotions, perhaps “happy” or “fear”. By modelling the temporal relationships between consecutive emotions, we can distinguish between the compound and the individual ones. Based on this idea, we propose to produce pre-prediction of class for each storyboard using 2D CNNs, and the sequence of these class pre-predictions are sent to a LSTM network to obtain the final class prediction of the whole video.

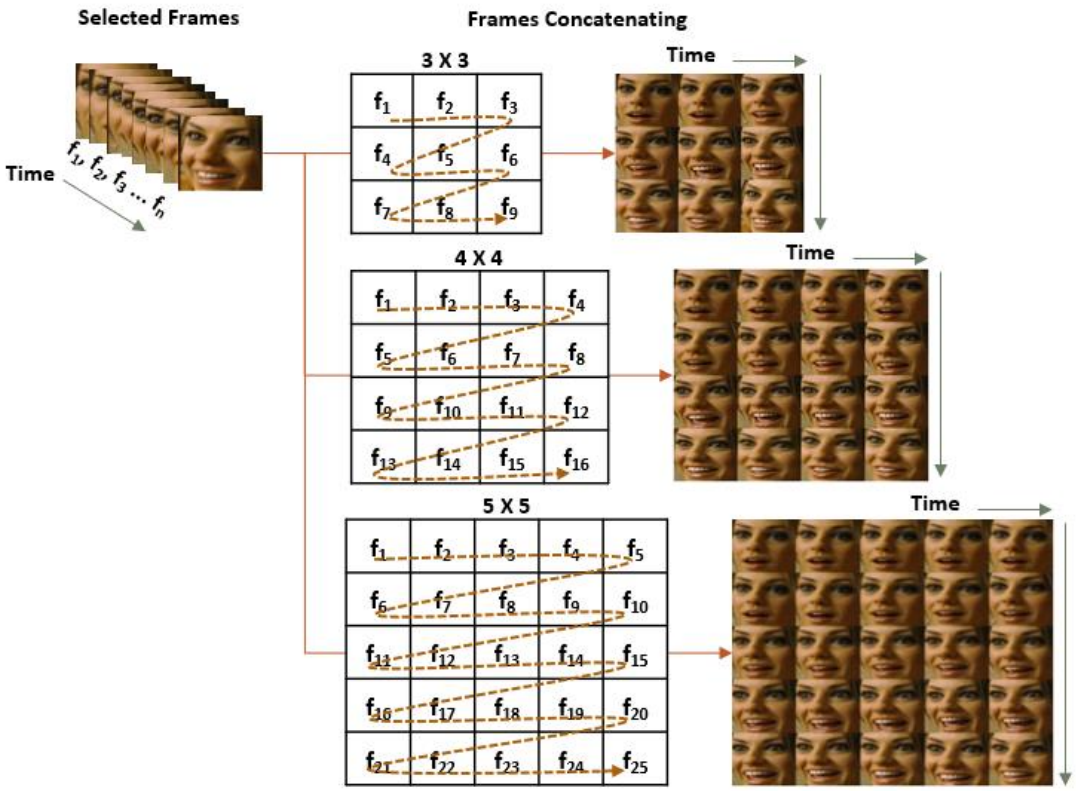


Figure 6-3: The three dimensions used in presenting the storyboard for video-based classification.

6.5.1. Spatio-Temporal Information Fusion (Storyboard Creating)

The facial emotion of a subject in a video generates space-time images within a 3Dspace, which encode both spatial and temporal information related to the subject's emotion. Instead of creating a 3D volume for the space-time information, this chapter proposes a storyboard creation technique that conflates video frames into one image based on keyframe selection or clustering. Figure 6-3 shows the three dimensions used in presenting the storyboard for video-based classification. Before constructing the storyboard, the selected frames (ROIs) are resized to a fixed size of 224 × 224 in order to reduce the interference caused by the images' boundaries. Then these frames are concatenated to build one image, as illustrated in Figure 6-3 . After that, the constructed storyboard is resized to 224 × 224 pixels because this size fits most 2D-CNN models.

6.5.2. Key-frame Selection and Clustering Strategies

The purpose of keyframe selection is to find a set of representative frames from an image sequence, while the purpose of frame clustering is to segment the set of sequential frames into subsets based on similarity matching.

In this chapter, a clustering-based strategy is used to achieve automatic keyframe selection and frame clustering. The proposed clustering-based strategy works, fundamentally, by measuring the dissimilarity or distance between frames d_t using the Euclidean distance:

$$d_t = \sqrt{\sum_{j=1}^{L} \left(\hat{f}_t(j) - \hat{f}_{t+1}(j)\right)^2} \quad \textbf{(6-1)}$$

where $\hat{f}_t$ denotes a frame feature vector at a specific time and L is the length of the vector. The following steps are followed to assign frames to the most similar cluster:

(1) Normalise the ROIs which are extracted from successive frames by resizing them to 224 × 224 and converting them to grey images.

(2) Represent every frame as a feature vector $\hat{f}$.

(3) Compare adjacent frames with each other using Equation (6-1) to determine how dissimilar they are. Denote the set of frames vectors as $\hat{F} = \{\hat{f}_1, \hat{f}_2, \hat{f}_3 \ldots \hat{f}_N\}$, where N is the number of frames in the video clip, and the difference between these frames as $D_{if} = \{d_1, d_2, d_3 \ldots d_{N-1}\}$.

(4) Determine a boundary-threshold value Ψ by calculating the mean value of the D_{if} set.

(5) Use the threshold Ψ to determine the borders of each cluster — where a dissimilarity value higher than Ψ indicates the start or end of a frame cluster.

Denote the set of clusters as $C = \{c_1, c_2, c_3 \ldots c_M\}$, where M is the number of clusters that consist of similar frames.

(6) For the keyframe selection strategy:

a) If the number of clusters M is smaller than the preset storyboard size (9, 16 or 25), decrease the value of Ψ. Alternatively, increase it when M is larger. Then go back to step 5.

b) If M is equal to the preset storyboard size (9, 16 or 25), select the mid-frame of each as a keyframe.

Or for the frame clustering strategy, preset a cluster-threshold value γ as the maximum number of clusters to be generated (it is set to 3 in this work):

a) If the number of clusters M is larger than γ, decrease the value for Ψ, Alternatively, increase it when M is smaller. Then go back to step 5.

b) If M is equal to γ, use all the frames in each cluster to build a storyboard.

(i). If the number of frames in a cluster Q is larger than the preset storyboard size (9, 16 or 25), choose the middle frames of the cluster.

(ii). If Q is smaller than the preset storyboard size (9, 16 or 25), duplicate the middle frames of the cluster to compensate.

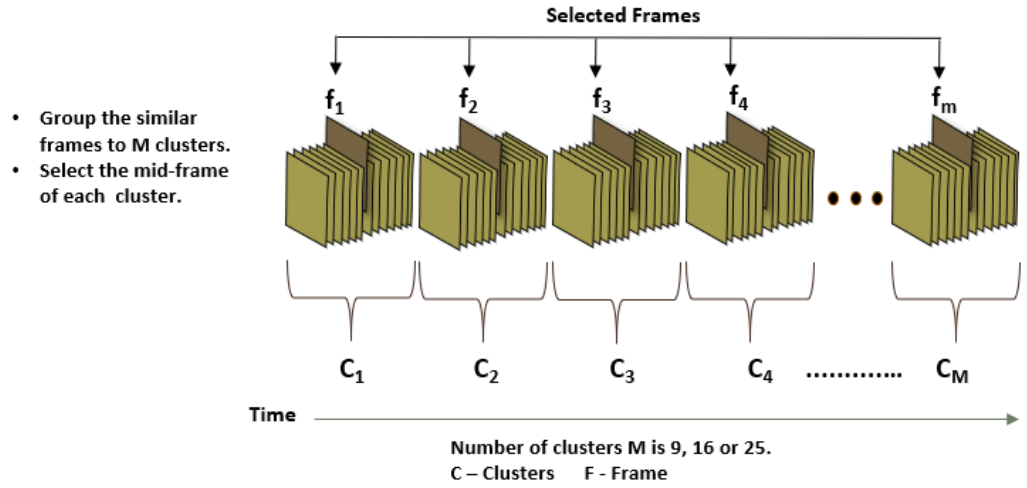


Figure 6-4: Frame selection strategy.

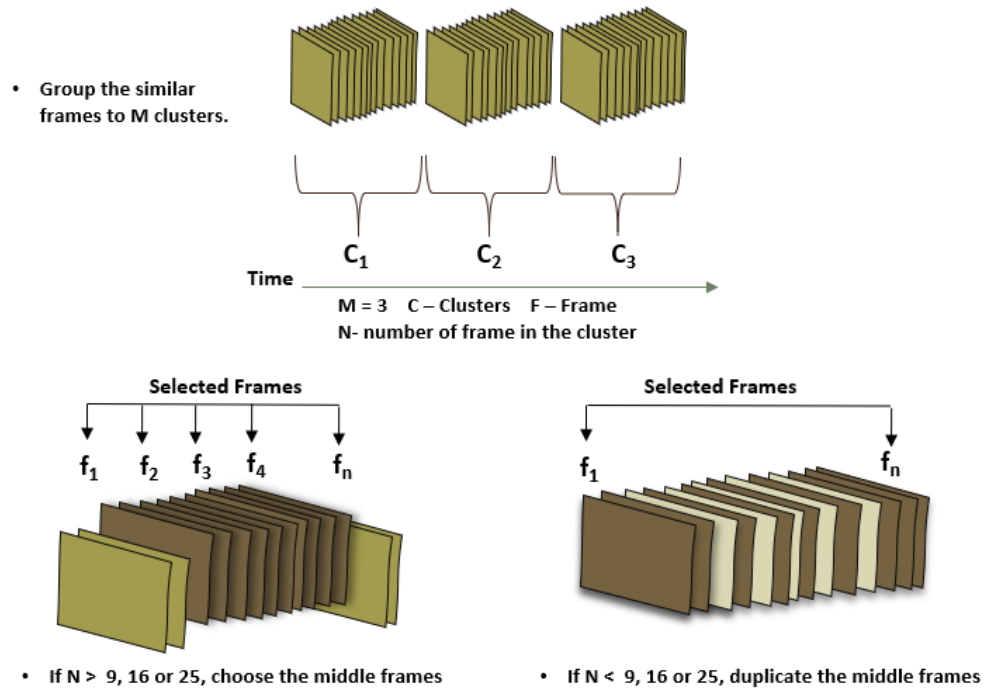


Figure 6-5: Frame cluster strategy.

In addition, since the video clips in the training database are short, varying from 3 to 5.4 seconds [111], we set γ, the cluster-threshold value (representing the maximum number of clusters), to 3. Figure 6-4 illustrates the frame selection strategy and Figure 6-5 illustrates the frame cluster strategy.

Figure 6-6: Samples from the AFEW [12], CK+ [120], AffectNet [10] and RAF-DB [11] databases.

6.6. Experiments

In this section, the facial emotion datasets used in the experiments are briefly described first, and the details about the implementation of the proposed methods are explained, including the pre-processing and the settings of the CNNs.

6.6.1. Databases

In order to evaluate our approaches, we conducted experiments on four facial emotion databases: AFEW [12], CK+ [120], AffectNet [10] and RAF-DB [11]. The first two are videos and the last two are still images. Some samples from these databases are shown in Figure 6-6.

Acted Facial Expressions in the Wild (AFEW) is a dynamic, temporal facial-expression dataset consisting of short video clips of facial expressions in close to real-

life environments. The number of samples in the training dataset of AFEW 7.0 is quite small, (training 773, validation 373 and test 653), and the class distribution is imbalanced. As the ground truth of the test dataset was kept hidden from the competitors, we tested our models on the validation dataset instead of the test dataset.

The Extended Cohn-Kanade (CK+) Database [120, 127] is a commonly used dataset, which includes 327 image sequences captured in a lab-controlled environment. These sequences start from the neutral expression and end with a peak expression, which is one of the six basic expressions, plus neutral and contempt.

The large-scale facial emotion dataset, AffectNet, was used for training the seven CNN models. In order to test the generalisation ability of our models, the RAF-DB [11] was used for testing. Section 2.5.4 provides a comprehensive description for AFEW, CK+, AffectNet and RAF-DB databases.

6.6.2. Implementation Details

6.6.2.1. Pre-processing

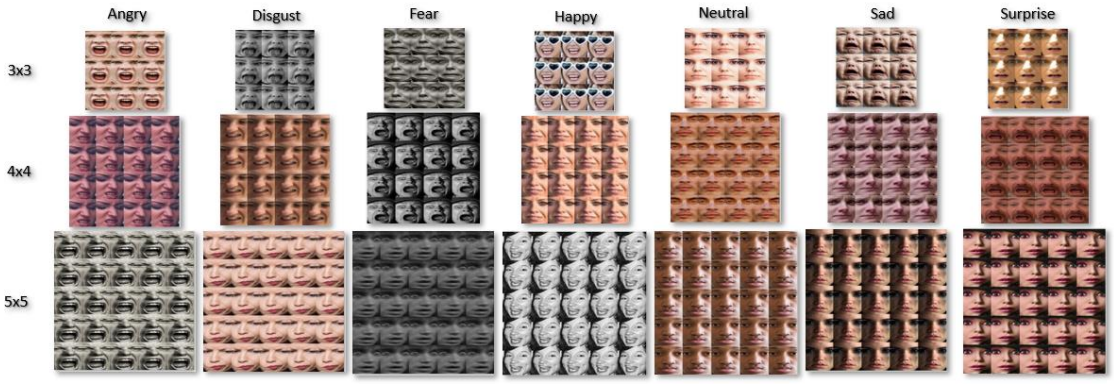
The chosen databases were produced in close to real-life environments and have variations in pose, illumination, occlusions and background. This high level of variation makes face detection and alignment challenging. The most common real-time face detection algorithm proposed by Viola-Jones [3] is useful for front views of faces but it is not robust enough to deal with the faces in videos where the subject moves without restrictions.

As the videos in the AFEW 7.0 dataset might be taken from more than one subject, we used the MATLAB “Face Detection and Tracking (FDT)” to track the main subject’s face in the videos automatically. The FDT uses the Kanade-Lucas-Tomasi

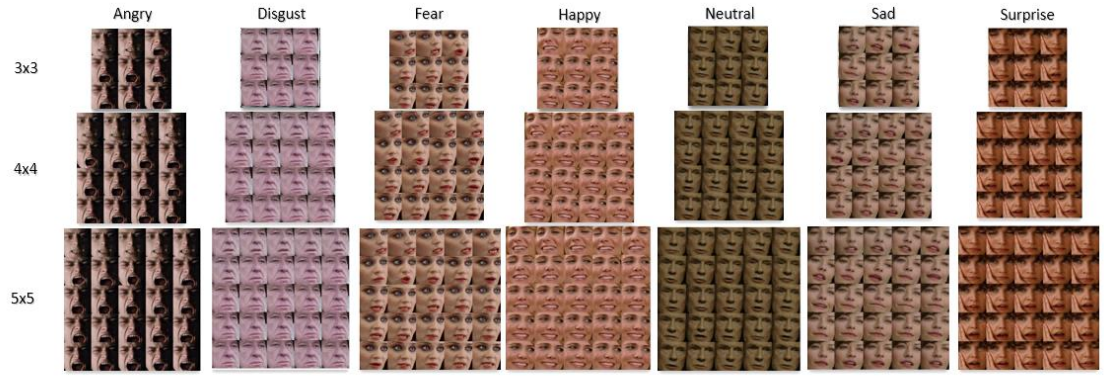
(KLT) algorithm to keep track of the face, even when the subject freely moves his or her face. Instead of applying Viola-Jones to detect the face across the video frames, which is a computationally expensive process, the FDT detects the face from the first frame. Then a set of feature points in the detected facial region were identified by using the standard of "good features to track" process [227] and tracked using the KLT algorithm across the video frames.

For facial landmark detection and face alignment, we utilised the algorithm proposed by Zhang et al. [172]. The generated landmarks (two eyes, nose and mouth corners) were then used to determine the inner area of the face as a ROI. For normalisation, the face is cropped to a 224 × 224 RGB image. The pre-processing steps were applied on all images in two databases AffectNet and RAF-DB. A preset number (9, 16, or 25) of cropped facial images were used to create the storyboard image by concatenating them. The storyboard image was then resized to 224 × 224 to fit the CNN input size. For training and testing purposes, three storyboard datasets of three different sizes (3 ×3, 4 ×4, 5 ×5) were created from these two still image datasets (AffectNet, RAF-DB).

After extracting the tracked faces from each video clip in the AFEW training, validation and test datasets, we applied the pre-processing steps on them and then followed the frame selection and cluster strategies to build the storyboard images for each strategy with the three storyboard sizes. As a result, six datasets were created in total, with different sizes and frame selection strategies. Some examples of the storyboard images built from the RAF-DB and AFEW training datasets are shown in Figure 6-7.



A. Samples of storyboard images built from RAF-DB training dataset



B. Samples of storyboard images built from AFEW training dataset

Figure 6-7: Samples of storyboard images of three sizes (3×3, 4×4, 5×5), built from RAF-DB and AFEW training dataset.

6.6.2.2. Training the 2D-CNN and LSTM Models

In order to increase the generalisation ability of the CNN models and to tackle the overfitting problem, seven well-known pre-trained 2D-CNNs (GoogleNet [93], VGG16, VGG19 [94], ResNet18, ResNet50, ResNet101[95], Inceptionv3 [228]) were utilised in our experiments and a large number of still images in the AffectNet dataset were used to fine-tune these models via two-fold cross-validation. To find the appropriate size for the storyboard, these models were fine-tuned on three different sizes, 3×3, 4×4 and 5×5. This resulted in 28 fine-tuned models that were able to classify emotional images.

We used the Caffe [173] toolkit on NVIDIA GPU to fine-tune the seven CNN models. These models were fine-tuned with a batch size of 16 on two GPUs (TitanX) and GeForce GTX 1080 using the stochastic gradient descent. The other parameters of the applied training algorithm were as follows: momentum=0.9, weight decay=0.0002.

To train the frame clustering OSO model for video classification, each video in the AFEW training dataset was clustered into three groups of frames and classified by the fine-tuned 2D CNNs, producing three emotion-words. Then the produced series of emotion-words were used to train the LSTM whose output is the final classification for each video. The AFEW validation dataset was used to evaluate the trained models.

6.7. Results

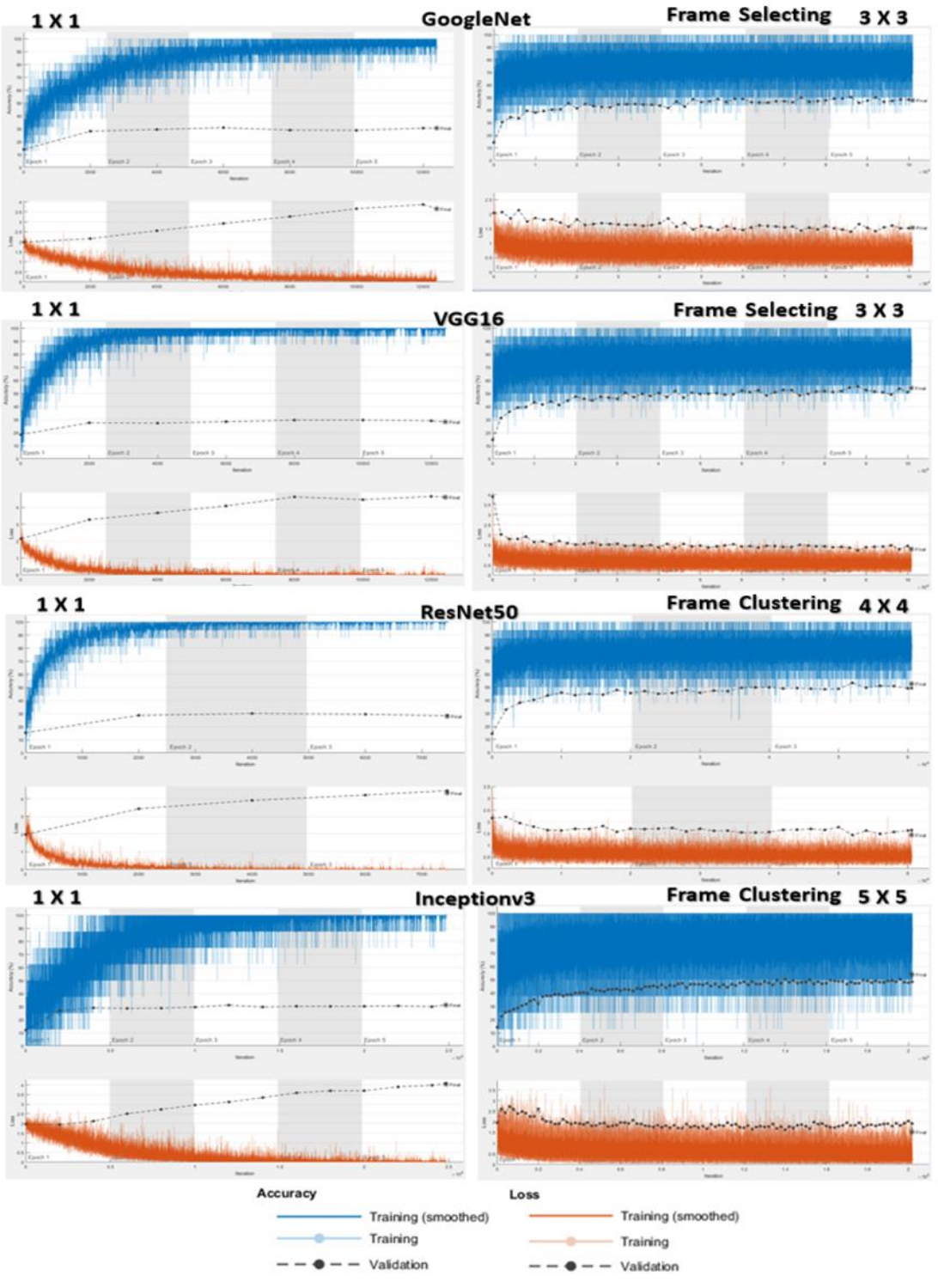
The major challenge in training a CNN model is to improve its generalisation ability and prevent overfitting, especially when training it on the extracted frames from the video clips since redundant identical frames reduce the intra-class variation. The proposed OSO methods address this challenge by training the CNN models on the storyboard images instead of the video frames. Figure 6-8 shows the training processes of the different 2D-CNN models when training them on video frames (1×1) and on the storyboard images with different sizes taken from the training and validation datasets of AFEW. As shown in Figure 6-8 (A), the four CNN models, GoogleNet, VGG16, ResNet50 and Inceptionv3, have experienced the overfitting problem when training on images extracted from video frames. The extremely low training loss and high validation loss indicate that these models perform well on the training dataset but fail to do so on the hold out samples in the validation dataset. Due to the high similarity between the consecutive frames, the CNN models memorise the training samples, which make it difficult to recognise the new one.

On the other hand, as shown in Figure 6-8 (B), the same four CNN models were trained on the storyboard images which were created by utilising frame selecting or frame clustering approaches. When comparing the training processes of these models, we can immediately notice that the gaps between training loss and validation loss are small in all four models, indicating that these models can overcome the overfitting problem and improve generalisation by using the storyboard.

Table 6-1: Accuracy of the seven common 2D-CNN models when training using single frame images and storyboard images of different sizes using the validation datasets and testing datasets of AffectNet and RAF-DB.

	Accuracy	**AffectNet**				**RAF-DB**			
		1x1	**3x3**	**4x4**	**5x5**	**1x1**	**3x3**	**4x4**	**5x5**
CNN	GoogleNet	50.1	40.5	46.7	43.3	71.5	67.4	69.4	68.2
	VGG16	53.7	53.6	48.6	48.0	73.3	75.3	71.3	71.4
	VGG19	53.5	53.8	53.7	49.0	72.8	**75.9**	74.1	69.0
	ResNet50	54.1	50.7	51.5	50.3	72.5	74.3	72.0	70.5
	ResNet101	54.3	52.9	53.8	50.7	74.2	73.8	70.7	71.3
	ResNet18	53.2	52.7	52.5	50.9	74.9	75.2	72.2	71.4
	Inceptionv3	56.1	53.3	53.1	53.5	70.6	71.6	72.7	68.7
	Average	53.6	51.1	51.4	49.4	72.8	73.4	71.8	70.1

In order to demonstrate the performance of the storyboard method for data fusion for image classification, cross-database validation experiments were conducted on the RAF-DB database. Table 6-1 shows the comparison among the seven CNN models when classifying emotions by using individual frame images and storyboard images with different sizes. The average validation results on AffectNet dataset are reported in the left section of the table and the results on RAF-DB database are illustrated in the right section of the table.



(A) 2D CNN models trained and tested on frames extracted from the video clips of training and validation datasets of *AFEW*

(B) 2D CNN models trained and tested on storyboard images created from frames of the video clips of training and validation datasets of AFEW

Figure 6-8: A comparison of the training process among four different 2D-CNN models using frames extracted from videos and the storyboard images created from frames from the video clips of the training and validation datasets of AFEW.

In our experiments, the proposed OSO video classification pipelines for emotion recognition using the frame selecting approach and frame clustering approach respectively (as shown in Figure 6-2), were evaluated based on three storyboard sizes (3×3, 4×4, 5×5), with their performance compared to those of the two single-frame processing (1×1) methods. In the first baseline method, a decision level fusion method based on majority voting was used to combine the CNN emotion predictions of all the frames in a video. The second baseline method followed a feature fusion approach, where an LSTM model was trained to classify videos using the fused features of all the frames extracted by the 2D-CNNs. Table 6-2 and Table 6-3 show the results on the AFEW dataset in terms of validation accuracy and the runtime of the OSO methods using seven 2D CNNs respectively, in comparison with the baseline methods.

Table 6-2: Validation accuracy of the OSO methods using 2D CNNs for video classification on the AFEW dataset, in comparison with single frame baseline (1×1) approaches.

		1x1		3x3		4x4		5x5	
	Accuracy	**Decision Fusion**	**Feature Fusion**	**Selecting**	**Clustering**	**Selecting**	**Clustering**	**Selecting**	**Clustering**
CNN	GoogleNet	33.8	35.1	46.4	48.7	49.0	46.7	41.9	45.0
	VGG16	39.3	34.6	**51.1**	55.3	47.7	**55.6**	47.7	50.1
	VGG19	37.7	30.1	49.8	53.8	47.7	51.0	46.7	47.8
	ResNet50	33.8	41.1	50.0	54.9	50.6	53.8	43.0	47.3
	ResNet101	36.9	37.2	48.7	52.8	46.7	51.1	44.6	48.0
	ResNet18	34.5	30.1	46.9	51.1	45.1	46.3	46.9	40.9
	Inceptionv3	36.1	34.8	50.8	54.6	48.0	51.3	41.7	45.7
	Average	36.0	34.71	49.1	53.03	47.83	50.83	44.64	46.4

Table 6-3: Comparison of validation time on the AFEW dataset between the OSO methods using 2D-CNNs and the single frame baseline (1×1) approaches for video classification.

	Speed S/Video	**1x1**		**3x3**		**4x4**		**5x5**	
		Decision Fusion	**Feature Fusion**	**Selecting**	**Clustering**	**Selecting**	**Clustering**	**Selecting**	**Clustering**
CNN	GoogleNet	0.31	0.33	0.038	0.047	0.055	0.061	0.074	0.091
	VGG16	0.60	0.63	0.037	0.046	0.056	0.059	0.073	0.090
	VGG19	0.69	0.71	0.039	0.047	0.056	0.061	0.075	0.092
	ResNet50	0.40	0.43	0.038	0.047	0.056	0.061	0.075	0.091
	ResNet101	0.52	0.56	0.046	0.050	0.062	0.064	0.078	0.099
	ResNet18	0.23	0.27	0.035	0.047	0.053	0.061	0.075	0.088
	Inceptionv3	0.59	0.62	0.045	0.052	0.063	0.064	0.077	0.099
	Average	0.48	0.51	0.040	0.048	0.057	0.062	0.075	0.093

Most winners of the EmotiW Challenge (2017-2019) utilised both audio and visual information in their approaches in order to increase the overall accuracy. To further evaluate the proposed OSO methods, we compared their performance with the EmotiW baseline performance as well as with those of the winning methods reported in the literature that used visual information only. Table 6-4 shows the comparison results on the AFEW dataset. To demonstrate the generalisation ability of the OSO methods, cross-database validation experiments were conducted on one of the lab-controlled environment databases, the CK+ dataset. We tested the models on all the image sequences in the CK+ dataset, and the results are shown in Table 6-5.

Table 6-4: Comparison with the state-of-the-art results on the AFEW 7.0 dataset.

		Methods	Val	Test
EmotiW baseline [111]		LBP-TOP-SVM	36.08	39.33
3D CNN	Ouyang et al. [138]	3D CNN	35.20	
	Lu et al. [214]	3D CNN	39.36	
	Fan et al. [103]	3D CNN	39.69	
	Vielzeuf et al. [139]	3D CNN+LSTM	43.20	
Decision / Feature Fusion	Yan et al. [41]	Trajectory + SVM	37.37	
	Fan et al. [229]	MRE-CNN (AlexNet)	40.11	
	Yan et al. [41]	VGG-BRNN	44.46	
	Ding et al. [230]	AlexNet	44.47	
	Fan et al. [103]	VGG16 + LSTM	45.43	
	Ouyang et al. [138]	ResNet + LSTM	46.70	
	Ouyang et al. [138]	VGG + LSTM	47.40	
	Fan et al. [229]	MRE-CNN (VGG16)	47.43	
	Vielzeuf et al. [139]	VGG16 + LSTM	48.60	
	Proposed OSO 1 [9]	FrameSelecting-VGG16-3x3	**51.10**	**51.15**
	Proposed OSO 2 [9]	FrameClustering-VGG16-4x4+LSTM	**55.60**	**52.37**

Since each image sequence in the CK+ dataset starts from the onset, neutral, expression in the first frame, to peak expression in the last frame, various state-of-the-art methods have utilised predefined frame/s to classify the image sequences, for example, by selecting the last frame [67, 231], the last three frames [232-237], etc. For a fair comparison, the OSO methods were therefore compared with the state-of-the-art methods which do not determine certain frames for training and testing.

Table 6-5: Accuracy of the OSO methods using 2D-CNNs for video classification on the CK+ dataset, in comparison with single frame (1×1) baseline approaches.

	Accuracy	1x1		3x3		4x4		5x5	
		Decision Fusion	Feature Fusion	Selecting	Clustering	Selecting	Clustering	Selecting	Clustering
CNN	GoogleNet	68.6	68.9	67.7	69.6	69.9	68.4	65.8	48.4
	VGG16	74.1	71.2	**95.9**	85.0	82.8	69.9	80.5	80.5
	VGG19	73.8	68.0	86.9	83.2	80.9	80.1	76.0	70.3
	ResNet50	80.6	76.7	90.0	81.3	73.3	74.8	60.5	69.2
	ResNet101	82.8	77.0	94.5	81.3	73.7	75.6	66.5	55.2
	ResNet18	81.2	65.7	90.3	**87.3**	80.5	69.2	77.5	67.3
	Inceptionv3	81.9	77.3	80.5	82.8	71.8	70.3	61.6	67.3
	Average	77.6	72.1	86.5	81.5	77.4	74.4	69.8	65.5

Table 6-6: Comparison with the state-of-the-art results on the CK+ dataset.

	Methods	
CK+ baseline [120]	Active Appearance Models (AAMs)+SVM	82.3
Liu et al. [238]	3DCNN	85.9
Jung et al. [239]	Deep Temporal Appearance Network (DTAN)	91.4
Sanin et al.[240]	Cov3D	92.3
Liu et al. [238]	3DCNN-DAP (Deformable Action Parts)	92.4
Liu et al. [241]	Spatio-Temporal Manifold (STM)-ExpLet	94.2
Sikka et al. [242]	Latent ordinal model (LOMo)	95.1
Zhang et al.[104]	Spatial–Temporal Recurrent Neural Network (STRNN)	95.4
Proposed OSO 1	FrameSelecting-VGG16-3×3	**95.9**
Proposed OSO 2	FrameClustering-VGG16-3×3+LSTM	**87.3**

6.8. Discussion

Compared to using single frame images, the storyboard method performs better in most models, achieving an accuracy of 75.9% when the size of the storyboard is (3×3), and 74.1% when the size of the storyboard is (4×4) by using VGG19. The accuracy decreases to 69.0% when using the (5×5) storyboard, however. This implies that increasing the number of the combined images in the storyboard will reduce the classification accuracy. Overall, the results show that our proposed data fusion methods (storyboard) can work well compared with the image classification method used in most of the 2D-CNN models.

We found that storyboard sizes do not relate consistently to performance across different datasets. For example, the average accuracies of different storyboard sizes (3×3, 4×4 and 5×5) are 73.4%, 71.8% and 70.1%, respectively, on the RAF-DB test dataset, which is much higher than the average accuracies achieved by the same storyboard sizes on the AffectNet dataset (51.1%, 51.4% and 49.4%). This is due to the different database categories. RAF-DB, for example, has defined multiple compound emotion categories, which reduces the emotion interference contained in each dataset.

As shown in Table 6-2, the OSO approaches outperformed both baseline methods in terms of validation accuracy by 10% to 17%. The frame clustering approach outperformed the frame selecting approach in almost all cases by 1.1% (VGG19, 5×5) to 7.9% (VGG16, 4×4) and on average by 3.93%, 3.0% and 1.76%, corresponding to storyboard sizes 3×3, 4×4 and 5×5, respectively. The highest accuracy was achieved by the OSO method using frame clustering and VGG16 with a storyboard size of 4×4. On average, the OSO method using frame clustering with a storyboard size of 3×3 achieved the highest accuracy of 53.03%. It can be observed that, among the seven 2D

CNNs, VGG16 achieved the highest accuracy in almost all cases.

One key advantage of the OSO approaches is their efficiency. To show this, we compared the runtime of the OSO approaches with that of the two baseline methods on the AFEW validation dataset, using a single NVIDIA TITAN X GPU. As Table 6-3 clearly demonstrates, the OSO approaches are about ten times faster than the single frame baseline methods. The proposed OSO methods using frame selecting and frame clustering approaches with the best 2D-CNN and storyboard size achieved validation accuracies of 51.10% and 55.60%, respectively, and test accuracies of 51.15% and 52.37%, respectively, much superior to the competition baseline performance as well as those other methods reported in the literature that used 2D- or 3D-CNNs for video-based emotion recognition without using audio information.

Table 6-5 reveals that the OSO approaches outperformed both baseline methods again, by 13.1% to 18.6%. Furthermore, the frame selecting approach with VGG16 and a storyboard size of 3×3 achieved the highest accuracy of 95.9%. It can be noticed that the frame selecting strategy with the storyboard size of 3×3 achieved better results than other storyboard sizes and the frame cluster strategy. Unlike the results in Table 6-2, the frame selecting approach outperformed the frame clustering approach in almost all cases by 0.8% (VGG19, 4×4) to 17.4% (GoogleNet, 5×5), and on average by 5.0 %, 3.0% and 4.3%, corresponding to storyboard sizes 3×3, 4×4 and 5×5, respectively. That is because this database consists of image sequences instead of video clips. Due to the small number of images in a sequence the storyboard of size 3×3 fits perfectly. Moreover, the differences between these images are high; thus, it does not suit the clustering strategy, which groups similar frames (images).

Table 6-6 shows the comparison of our methods and other state-of-the-arts on the

CK+ dataset. The proposed OSO methods using frame selecting approach with the best 2D-CNN and storyboard size achieved 95.9% accuracy, much superior to the CK+ baseline performance and those achieved by the methods applying 3D-CNNs reported in the literature.

6.9. Conclusion

This chapter proposes fast OSO methods for video-based facial emotion recognition to meet the requirements of real-time applications. In contrast to other approaches that aggregate temporal information from video frames, the proposed methods take advantage of spatio-temporal data fusion based on novel frame selection and clustering strategies and use 2D-CNN models to predict emotional categories from videos with facial expressions. The experimental results show that the proposed OSO methods are not only fast but also capable of achieving competitive accuracy in video classification.

Chapter 7

Conclusions and Future Work

7.1. Conclusions

Affective computing aims to bridge the gap between computational technology and humans in emotion recognition by developing new ways to understand, interpret, communicate and respond to human emotion. Emotion recognition is a broad and growing research area with tremendously significant applications across many areas, such as healthcare, education, the understanding of social interaction and behavioural science, etc., particularly in relation to developing reliable and intelligent models for real-time affective computing applications.

The distinguishability of human emotions within still images or videos captured from real-world situations is generally poor, due to the variations in the environment. In addition to the major difficulties that any facial recognition system may face, such as variable illumination, head pose, and facial occlusion, there are many other significant challenges that arise specifically from the nature of human emotions, and these have a considerable impact on recognition accuracy. These challenges include the variety of expressions and emotions as expressed by differing subjects (humans can express differing emotions with differing strengths), and interference between competing emotions. With these in mind, many researchers have attempted to combine several complex models to improve emotion recognition accuracy. This complexity, however, leads to the requirement for massive computing and storage resources, and

brings with it new problems relating to excessive resource consumption for data processing.

The methods developed in this thesis are the result of a radical rethink and are designed to minimise computational complexity while retaining high accuracy. Furthermore, these methods not only possess the capability to address the common challenges in developing facial emotion recognition systems but are also the basis for proposing several techniques for tackling other challenges, such as real-time video classification. This section summarises the contributions of the work presented in this thesis. Summary of the contributions will be discussed in Section 7.2, followed by limitations and some potential future research directions in Section 7.3.

7.2. Summary of the Contributions

This research has suggested that the requirements of real-time applications should be considered when designing each element of the framework of an emotion recognition system. The major contributions of this thesis work can be summarised as follows.

Firstly, an efficient and effective algorithm was developed to improve the performance of the Viola-Jones algorithm, widely used for face detection, in respect to the recognition of emotions from thermal images. Experimental results showed that the proposed method achieved significantly higher detection accuracy (95%) than the standard Viola-Jones method (90%) in face detection from thermal images, while also doubling the detection speed, as explained in detail in Chapter 3.

Secondly, to improve the robustness of the face detection method and to cope with real-world applications, an automatic subsystem for detecting eyeglasses, Shallow-GlassNet, was proposed to address the facial occlusion problem in face detection. This was done by designing a shallow convolutional neural network capable of detecting

eyeglasses rapidly and accurately. By training a convolutional neural network on a small dataset and decreasing its depth to just three convolutional layers, Shallow-GlassNet significantly reduced the computational complexity whilst retaining a high level of accuracy, as explained in detail in Chapter 4.

Thirdly, a novel neural network model for decision fusion was proposed in order to make use of multiple classifier systems, which can increase the classification accuracy by up to 10%. In this approach, convolutional neural networks were used to extract nine different sets of emotional features from faces, which were then classified by SVM. The proposed neural network model for decision fusion was then trained based on the output scores from the nine SVM classifiers, as explained in detail in Chapter 5.

Finally, in order to reduce the computational cost without sacrificing recognition accuracy, a high-speed approach to emotion recognition from videos, called One-Shot Only (OSO), was developed based on a novel spatio-temporal data fusion method for representing video frames, as explained in detail in Chapter 6. To cope with real-time applications, the OSO method tackled video classification as a single image classification problem, which not only made it extremely fast but also reduced the overfitting problem known to occur when training deep neural networks. More specifically, the contributions related to the proposed OSO methods are as follows:

- A novel spatio-temporal data fusion approach is proposed for video representation. A strong advantage of this approach is that a single convolution neural network can be used to predict the whole video's class probabilities directly. This speed up the classification rate to the extent that the system is competent to perform in real-time. Another significant advantage of this approach is that it prevents the occurrence of the overfitting problem, which is often considered to be the most

difficult challenge in the use of CNNs when they are trained on nearly identical frames extracted from a single video.

- Frame selection and clustering strategies are proposed to handle videos of different lengths, as well as the redundancy in consecutive video frames, leading to two OSO methods for effective video representation. The OSO methods reorganise video frames hierarchically as a single image, from which common 2D-CNN models can predict the video class probabilities. The OSO methods are evaluated using three sizes of storyboard for video representation and seven common 2D-CNN models for video classification. It is demonstrated that the OSO methods can be used to classify both images and videos and are able to improve the recognition accuracy when classifying emotion from images as well as videos.
- In order to handle compound emotions, with transitions from one emotion to another, the OSO method with clustering-based video representation transposes the video classification task from the vision domain to the multi-class text classification domain to some extent, hence mimicking some operations of the human mind. In this approach, each video is translated into a sentence of a few words, in which form it can be addressed using a LSTM network. Using the clustering strategy, similar consecutive video frames are combined into one storyboard image which is classified to a single emotion using a CNN model.

7.3. Limitations and Future Work

Whilst several achievements have been made with regard to the theoretical and practical aspects of emotion recognition in this research, some interesting new issues relating to these have also been encountered. Some of these relate to improving the capabilities of emotion recognition techniques, such as the use of video-based

emotional databases, but mostly represent subsidiary topics which could not be considered or investigated here due to the limited time available. This section discusses these limitations and related issues.

The performance of an emotion recognition system depends to quite a large extent on the amount and diversity of the annotated samples provided for training purposes. Although several large-scale facial emotion focused databases, such as AffectNet and RAF-DB, were used in this research, existing video-based databases are still quite small and were created using a limited number of annotators, affecting their reliability. Therefore, a large-scale and reliable emotion database with a large number of variations extracted from 'the wild' is needed. Moreover, as described in Chapter 4 and Chapter 5, the presence of eyeglasses may cause inaccurate classification, this database should also have a large number of video clips with and without eyeglasses.

Human emotion recognition systems have mostly attempted to sense emotions by using facial expressions as representing the inner state of the human. While significant progress has been made in the field of facial expression recognition with respect to the visible spectrum [243], the performance of existing methods is vulnerable to illumination changes, darkness or excessive light. This is because illumination changes can significantly influence the appearance of visible images. On the other hand, thermal infrared images record temperature distributions and are not usually affected by illumination conditions. In addition, the detection of skin temperature changes can be helpful to classify emotions [13] and facial expressions. In recent years, facial expression recognition via the thermal infrared spectrum, which can represent the internal state of the human's emotions, has attracted greater attention [4]. In this thesis, we have tried to employ thermal images by first detecting faces (and this is where our contribution to this specific area mainly lies). This was achieved by

improving the Viola-Jones algorithm as used to detect the face from thermal images, and by utilising two thermal facial image databases, NVIE [8] and I.Vi.T.E [114]. These databases were collected in laboratory-controlled environments and thus do not adequately represent the real-world environment. This prevented us from using these databases in the remaining experiments undertaken in this thesis. To the best of our knowledge, currently, there is no real-world environment emotion database which combines both visible and thermal spectrum. There is a significant need for such a database.

The One-Shot Only (OSO) methods developed in this thesis for real-time emotion recognition have the potential to be extended and applied to other video-based classification tasks, such as human action recognition, video activity recognition, vehicle detection, etc. The proposed OSO framework provides starting points for several such relevant models and approaches.

Spatio-temporal data fusion is currently an active research topic, and techniques for achieving this are useful for integrating appearance and motion information for video-based classification processing. The storyboard which is utilised in the OSO models has shown good performance when used in a number of different scenarios. The selection strategy experiments, however, revealed that the number of frames in the storyboard has a significant influence on the classification performance. When the width and height of the storyboard image were fixed, but the number of frames was varied from 9 to 25, it was discovered that the accuracy decreased as the number of frames increased. Hence, how to choose the appropriate number of frames per storyboard when considering the length of a video has not been answered well in this thesis.

Since the proposed keyframe selection and keyframe clustering methods depend on measuring the similarity of consecutive frames, it should be taken into consideration that the candidate frames are not always the best among the frames available, when considering issues such as clarity, illumination, head pose and facial occlusion. Thus, these strategies should take into consideration such essential criteria when choosing the frames to be employed in creating the storyboard.

New fundamental components and possible improvements to the current methods would shed light on future works. Some suggested points are listed in the following:

- Creating a large-scale and reliable emotion database that contains a large number of video clip samples and reflects the characteristics of the real world with variations extracted from 'the wild' is needed. This database should be annotated to the dimensional and categorical models that include multiple compound categories of emotion, such as happily-surprised, fearfully-angry.
- Another large-scale and reliable emotion database is necessary to meet the need for studies of internal manifestations, including changes in the face and body temperature. This database should be created in a real-world environment and combine both visible and thermal spectrums by simultaneously capturing visual and thermal images and video clips.
- Resolving the above problems and finding answers to the related questions will significantly improve not only the OSO methods presented in this thesis, but also most video-based classification models and facial emotion recognition systems. These issues should be given adequate attention in future research.

REFERENCES

[1] P. Ekman and W. V. Friesen, "Constants across cultures in the face and emotion," *Journal of Personality and Social Psychology,* vol. 17, no. 2, pp. 124–129, 1971.

[2] A. M. Basbrain, J. Q. Gan, and A. Clark, "Accuracy enhancement of the viola-jones algorithm for thermal face detection," in *International Conference on Intelligent Computing*, 2017: Springer, pp. 71-82.

[3] P. Viola and M. J. Jones, "Robust real-time face detection," *International Journal of Computer Vision,* vol. 57, no. 2, pp. 137-154, 2004.

[4] A. M. Basbrain, I. Al-Taie, N. Azeez, J. Q. Gan, and A. Clark, "Shallow convolutional neural network for eyeglasses detection in facial images," in *9th Computer Science and Electronic Engineering (CEEC)*, 27-29 Sept. 2017 2017, pp. 157-161, doi: 10.1109/CEEC.2017.8101617.

[5] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," Technical Report 07-49, University of Massachusetts, Amherst, 2008.

[6] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *IEEE International Conference on Computer Vision*, 2015, pp. 3730-3738.

[7] A. M. Basbrain, J. Q. Gan, A. Sugimoto, and A. Clark, "A neural network approach to score fusion for emotion recognition," in *10th Computer Science and Electronic Engineering (CEEC)*, 2018: IEEE, pp. 180-185.

[8] W. Shangfei *et al.*, "A natural visible and infrared facial expression database for expression recognition and emotion inference," *IEEE Transactions on Multimedia,* vol. 12, no. 7, pp. 682-691, 2010, doi: 10.1109/tmm.2010.2060716.

[9] A. Basbrain and J. Q. Gan, "One-Shot Only real-time video classification: A case study in facial emotion recognition," in *Intelligent Data Engineering and Automated Learning – IDEAL*, Cham, 2020: Springer International Publishing, pp. 197-208.

[10] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "Affectnet: A database for facial expression, valence, and arousal computing in the wild," *IEEE Transactions on Affective Computing,* vol. 10, no. 1, pp. 18-31, 2019, doi: 10.1109/TAFFC.2017.2740923.

[11] S. Li and W. Deng, "Reliable Crowdsourcing and Deep Locality-Preserving Learning for Unconstrained Facial Expression Recognition," *IEEE Transactions on Image Processing,* vol. 28, no. 1, pp. 356-370, 2019, doi: 10.1109/TIP.2018.2868382.

[12] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon, "Collecting large, richly annotated facial-expression databases from movies," *IEEE Multimedia,* vol. 19, no. 3, pp. 34-41, 2012.

[13] A. Konar and A. Chakraborty, *Emotion recognition: A pattern analysis approach*, First ed. John Wiley & Sons, 2015.

[14] V. C. Appel, V. L. Belini, D. H. Jong, D. V. Magalhães, and G. A. Caurin, "Classifying emotions in rehabilitation robotics based on facial skin temperature," in *5th IEEE International Conference on Biomedical Robotics and Biomechatronics RAS/EMBS*, 2014, pp. 276-280.

[15] W. Shangfei, L. Siliang, and W. Xufa, "Infrared Facial Expression Recognition Using Wavelet Transform," in *International Symposium on Computer Science and Computational Technology ISCSCT '08. ,* 2008, vol. 2, pp. 327-330, doi: 10.1109/iscsct.2008.356.

[16] P. C. Petrantonakis and L. J. Hadjileontiadis, "A novel emotion elicitation index using frontal brain asymmetry for enhanced EEG-based emotion recognition," *IEEE Transactions on information technology in biomedicine,* vol. 15, no. 5, pp. 737-746, 2011.

[17] Y. P. Lin *et al.*, "EEG-based emotion recognition in music listening," *IEEE Transactions on Biomedical Engineering*, vol. 57, no. 7, pp. 1798-1806, 2010.

[18] R. Plutchik and H. Kellerman, *Theories of emotion*. Academic Press, 2013.

[19] S. Du, Y. Tao, and A. M. Martinez, "Compound facial expressions of emotion," *Proceedings of the National Academy of Sciences,* vol. 111, no. 15, pp. E1454-E1462, 2014, doi: 10.1073/pnas.1322355111.

[20] W. M. Wundt, *An introduction to psychology*. Macmillan, 1912.

[21] H. Schlosberg, "Three dimensions of emotion," *Psychological review*, vol. 61, no. 2, p. 81, 1954.

[22] D. C. Rubin and J. M. Talarico, "A comparison of dimensional models of emotion: evidence from emotions, prototypical events, autobiographical memories, and words," *Memory (Hove, England),* vol. 17, no. 8, pp. 802-808, 2009, doi: 10.1080/09658210903130764.

[23] P. Ekman and W. V. Friesen, *Facial action coding system: Investigator's guide*. Consulting Psychologists Press, 1978.

[24] J. Hamm, C. G. Kohler, R. C. Gur, and R. Verma, "Automated facial action coding system for dynamic analysis of facial expressions in neuropsychiatric disorders," *Journal of neuroscience methods*, vol. 200, no. 2, pp. 237-256, 2011.

[25] M. Del Giudice and L. Colle, "Differences between children and adults in the recognition of enjoyment smiles," *Developmental Psychology*, vol. 43, no. 3, pp. 796-803, 2007, doi: 10.1037/0012-1649.43.3.796.

[26] M.-H. Yang, D. J. Kriegman, and N. Ahuja, "Detecting faces in images: A survey," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 24, no. 1, pp. 34-58, 2002.

[27] R. Lienhart and J. Maydt, "An extended set of haar-like features for rapid object detection," in *International Conference on Image Processing*, 2002, vol. 1, pp. I-I.

[28] S. S. Farfade, M. J. Saberian, and L.-J. Li, "Multi-view face detection using deep convolutional neural networks," *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*, Shanghai, China, 2015. [Online]. Available: https://doi.org/10.1145/2671188.2749408.

[29] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097-1105.

[30] J. Vana, S. Mracek, M. Drahansky, A. Poursaberi, and S. Yanushkevich, "Applying fusion in thermal face recognition," in *International Conference of the Biometrics Special Interest Group (BIOSIG)* 2012, pp. 1-7.

[31] Z.-H. Feng and J. Kittler, "Advances in facial landmark detection," *Biometric Technology Today*, vol. 2018, no. 3, pp. 8-11, 2018, doi: 10.1016/s0969-4765(18)30038-9.

[32] N. Wang, X. Gao, D. Tao, H. Yang, and X. Li, "Facial feature point detection: A comprehensive survey," *Neurocomputing*, vol. 275, pp. 50-65, 2018, doi: 10.1016/j.neucom.2017.05.013.

[33] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham, "Active shape models-their training and application," *Computer Vision and Image Understanding*, vol. 61, no. 1, pp. 38-59, 1995.

[34] S. Milborrow and F. Nicolls, "Locating facial features with an extended active shape model," in *European Conference on Computer Vision*, 2008: Springer, pp. 504-513.

[35] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, no. 6, pp. 681-685, 2001.

[36] D. Cristinacce and T. Cootes, "Automatic feature localisation with constrained local models," *Pattern Recognition*, vol. 41, no. 10, pp. 3054-3067, 2008.

[37] E. Zhou, H. Fan, Z. Cao, Y. Jiang, and Q. Yin, "Extensive facial landmark localization with coarse-to-fine convolutional network cascade," in *IEEE International Conference on Computer Vision Workshops*, 2013, pp. 386-391.

[38] Y. Sun, X. Wang, and X. Tang, "Deep convolutional network cascade for facial point detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 3476-3483.

[39] B. Scholkopf and A. J. Smola, *Learning with kernels: Support Vector Machines, regularization, optimization, and beyond*. MIT press, 2001.

[40] R. Salakhutdinov, A. Mnih, and G. Hinton, "Restricted Boltzmann machines for collaborative filtering," in *24th International Conference on Machine Learning*, 2007: ACM, pp. 791-798.

[41] J. Yan *et al.*, "Multi-clue fusion for emotion recognition in the wild," *ACM International Conference on Multimodal Interaction*, Tokyo, Japan, 2016.

[42] C. Liu, T. Tang, K. Lv, and M. Wang, "Multi-feature based emotion recognition for video clips," *ACM International Conference on Multimodal Interaction*, Boulder, CO, USA, 2018.

[43] I. Kotsia, N. Nikolaidis, and I. Pitas, "Facial expression recognition in videos using a novel multi-class support vector machines variant," in *International Conference on Acoustics, Speech and Signal Processing - ICASSP '07*, 2007, vol. 2: IEEE pp. II-585-II-588, doi: 10.1109/ICASSP.2007.366303.

[44] I. Kotsia and I. Pitas, "Real time facial expression recognition from image sequences using support vector machines," in *International Conference on Image Processing* 2005, vol. 2: IEEE pp. II-966, doi: 10.1109/ICIP.2005.1530218.

[45] Y. Wu, Z. Wang, and Q. Ji, "Facial feature tracking under varying facial expressions and face poses based on restricted boltzmann machines," in *IEEE*

Conference on Computer Vision and Pattern Recognition, 2013, pp. 3452-3459.

[46] P. Shen, S. Wang, and Z. Liu, "Facial expression recognition from infrared thermal videos," in *Intelligent Autonomous Systems 12*, vol. 194, S. Lee, H. Cho, K.-J. Yoon, and J. Lee Eds., (Advances in Intelligent Systems and Computing: Springer Berlin Heidelberg, 2013, ch. 31, pp. 323-333.

[47] S. Wang, M. He, Z. Gao, S. He, and Q. Ji, "Emotion recognition from thermal infrared images using deep Boltzmann machine," *Front. Comput. Sci.*, vol. 8, no. 4, pp. 609-618, 2014, doi: 10.1007/s11704-014-3295-3.

[48] B. Hernández, G. Olague, R. Hammoud, L. Trujillo, and E. Romero, "Visual learning of texture descriptors for facial expression recognition in thermal imagery," *Computer Vision and Image Understanding*, vol. 106, no. 2–3, pp. 258-269, 2007, doi: http://dx.doi.org/10.1016/j.cviu.2006.08.012.

[49] Z. Liu and S. Wang, "Emotion recognition using hidden markov models from facial temperature sequence," in *Affective Computing and Intelligent Interaction*, vol. 6975, S. D'Mello, A. Graesser, B. Schuller, and J.-C. Martin Eds., (Lecture Notes in Computer Science: Springer Berlin Heidelberg, 2011, ch. 26, pp. 240-247.

[50] W. Shangfei, S. Peijia, and L. Zhilei, "Facial expression recognition from infrared thermal images using temperature difference by voting," in *2nd International Conference on Cloud Computing and Intelligent Systems (CCIS)* 2012, vol. 01: IEEE pp. 94-98, doi: 10.1109/ccis.2012.6664375.

[51] S. Wold, K. Esbensen, and P. Geladi, "Principal component analysis," *Chemometrics and intelligent laboratory systems*, vol. 2, no. 1-3, pp. 37-52, 1987.

[52] L. Trujillo, G. Olague, R. Hammoud, and B. Hernandez, "Automatic feature localization in thermal images for facial expression recognition," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2005, pp. 14-14, doi: 10.1109/CVPR.2005.415.

[53] S. Jarlier *et al.*, "Thermal analysis of facial muscles contractions," *IEEE Transactions on Affective Computing*, vol. 2, no. 1, pp. 2-9, 2011, doi: 10.1109/t-affc.2011.3.

[54] Y. Yoshitomi, K. Sung-Ill, T. Kawano, and T. Kilazoe, "Effect of sensor fusion for recognition of emotional states using voice, face image and thermal image of face," in *9th IEEE International Workshop on Robot and Human Interactive Communication*, 2000, pp. 178-183, doi: 10.1109/roman.2000.892491.

[55] H. Nguyen, F. Chen, K. Kotani, and B. Le, "Human emotion estimation using wavelet transform and t-ROIs for fusion of visible images and thermal image sequences," in *Computational Science and Its Applications – ICCSA*, vol. 8584, B. Murgante *et al.* Eds., (Lecture Notes in Computer Science: Springer International Publishing, 2014, ch. 17, pp. 224-235.

[56] Y. Nakanishi, Y. Yoshitomi, T. Asada, and M. Tabuse, "Facial expression recognition of a speaker using thermal image processing and reject criteria in feature vector space," *Artif Life Robotics,* vol. 19, no. 1, pp. 76-88, 2014, doi: 10.1007/s10015-013-0136-7.

[57] S. Wang, S. He, Y. Wu, M. He, and Q. Ji, "Fusion of visible and thermal images for facial expression recognition," *Front. Comput. Sci.*, vol. 8, no. 2, pp. 232-242, 2014, doi: 10.1007/s11704-014-2345-1.

[58] T. Asada, Y. Yoshitomi, and M. Tabuse, "A system for facial expression recognition of a speaker using front-view face judgment, vowel judgment, and thermal image processing," *Artif Life Robotics,* vol. 17, no. 2, pp. 263-269, 2012, doi: 10.1007/s10015-012-0052-2.

[59] Y. Yoshitomi, T. Asada, K. Shimada, and M. Tabuse, "Facial expression recognition of a speaker using vowel judgment and thermal image processing," *Artif Life Robotics,* vol. 16, no. 3, pp. 318-323, 2011, doi: 10.1007/s10015-011-0939-3.

[60] Y. Yoshitomi, N. Miyawaki, S. Tomita, and S. Kimura, "Facial expression recognition using thermal image processing and neural network," in *6th IEEE International Workshop on Robot and Human Communication*, 1997, pp. 380-385, doi: 10.1109/roman.1997.647016.

[61] Y. Sugimoto, Y. Yoshitomi, and S. Tomita, "A method for detecting transitions of emotional states using a thermal facial image based on a synthesis of facial expressions," *Robotics and Autonomous Systems,* vol. 31, no. 3, pp. 147-160, 2000, doi: http://dx.doi.org/10.1016/S0921-8890(99)00104-9.

[62] M. M. Khan, R. D. Ward, and M. Ingleby, "Automated classification and recognition of facial expressions using infrared thermal imaging," in *IEEE Conference on Cybernetics and Intelligent Systems*, 2004, vol. 1, pp. 202-206 doi: 10.1109/iccis.2004.1460412.

[63] W. Ni, "Facial image registration," Ph.D. dissertation, Dept. Elect. Eng., Université de Grenoble., Grenoble, France, 2012.

[64] O. Rudovic, M. Pantic, and I. Patras, "Coupled Gaussian processes for pose-invariant facial expression recognition," *IEEE Transactions on Pattern*

Analysis and Machine Intelligence, vol. 35, no. 6, pp. 1357-1369, 2013, doi: 10.1109/tpami.2012.233.

[65] M. K. Bhowmik, B. K. De, D. Bhattacharjee, D. K. Basu, and M. Nasipuri, "Multisensor fusion of visual and thermal images for human face identification using different SVM kernels," in *Long Island Systems, Applications and Technology Conference (LISAT)*, 2012: IEEE, pp. 1-7, doi: 10.1109/lisat.2012.6223195.

[66] M. Shin, M. Kim, and D. Kwon, "Baseline CNN structure analysis for facial expression recognition," in *25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, 2016, pp. 724-729, doi: 10.1109/ROMAN.2016.7745199.

[67] J. Li and E. Y. Lam, "Facial expression recognition using deep neural networks," in *IEEE International Conference on Imaging Systems and Techniques (IST)*, 2015, pp. 1-6, doi: 10.1109/IST.2015.7294547.

[68] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 7, pp. 971-987, 2002, doi: 10.1109/TPAMI.2002.1017623.

[69] G. Zhao and M. Pietikainen, "Dynamic texture recognition using local binary patterns with an application to facial expressions," *IEEE Transactions on Pattern Analysis & Machine Intelligence,* no. 6, pp. 915-928, 2007.

[70] S. Sahoo and A. Routray, "Emotion recognition from audio-visual data using rule based decision level fusion," in *IEEE Students' Technology Symposium (TechSym)*, 2016, pp. 7-12, doi: 10.1109/TechSym.2016.7872646.

[71] B. Sun, L. Li, T. Zuo, Y. Chen, G. Zhou, and X. Wu, "Combining multimodal features with hierarchical classifier fusion for emotion recognition in the wild," in *16th International Conference on Multimodal Interaction*, 2014: ACM, pp. 481-486.

[72] K. Sikka, K. Dykstra, S. Sathyanarayana, G. Littlewort, and M. Bartlett, "Multiple kernel learning for emotion recognition in the wild," in *15th ACM on International conference on multimodal interaction*, 2013: ACM, pp. 517-524.

[73] V. Ojansivu and J. Heikkilä, "Blur insensitive texture classification using local phase quantization," in *International Conference on Image and Signal Processing,* 2008: Springer, pp. 236-243.

[74] J. Päivärinta, E. Rahtu, and J. Heikkilä, "Volume Local Phase Quantization for Blur-Insensitive Dynamic Texture Classification," in *Scandinavian Conference on Image Analysis*, 2011: Springer, pp. 360-369.

[75] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)* 2005, vol. 1, pp. 886-893, doi: 10.1109/CVPR.2005.177.

[76] J. Chen, Z. Chen, Z. Chi, and H. Fu, "Emotion recognition in the wild with feature fusion and multiple kernel learning," in *16th International Conference on Multimodal Interaction*, 2014: ACM, pp. 508-513.

[77] Z. Liu, H. Wang, Y. Yan, and G. Guo, "Effective facial expression recognition via the boosted convolutional neural network," in *CCF Chinese Conference on Computer Vision*, 2015: Springer, pp. 179-188.

[78] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray, "Visual categorization with bags of keypoints," in *Workshop on Statistical Learning in Computer Vision, ECCV*, 2004, vol. 1, no. 1-22: Prague, pp. 1-2.

[79] J. Zhang, M. Marszałek, S. Lazebnik, and C. Schmid, "Local features and kernels for classification of texture and object categories: A comprehensive study," *International journal of computer vision*, vol. 73, no. 2, pp. 213-238, 2007.

[80] K. Sikka, T. Wu, J. Susskind, and M. Bartlett, "Exploring bag of words architectures in the facial expression domain," in *European Conference on Computer Vision*, 2012: Springer, pp. 250-259.

[81] S. E. Kahou *et al.*, "Combining modality specific deep neural networks for emotion recognition in video," in *15th ACM on International conference on multimodal interaction*, 2013, pp. 543-550, doi: 10.1145/2522848.2531745.

[82] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong, "Locality-constrained linear coding for image classification," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2010: Citeseer, pp. 3360-3367.

[83] M. Ranzato, J. Susskind, V. Mnih, and G. Hinton, "On deep generative models with applications to recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 2011: IEEE, pp. 2857-2864.

[84] Y. Bengio, "Learning deep architectures for AI," *Foundations and trends in Machine Learning*, vol. 2, no. 1, pp. 1-127, 2009.

[85] Y. Guo, Y. Liu, A. Oerlemans, S. Lao, S. Wu, and M. S. Lew, "Deep learning for visual understanding: A review," *Neurocomputing,* vol. 187, pp. 27-48, 2016, doi: 10.1016/j.neucom.2015.09.116.

[86] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning* (The Adaptive Computation and Machine Learning). Cambridge, MA: The MIT Press, 2016.

[87] A. Voulodimos, N. Doulamis, A. Doulamis, and E. Protopapadakis, "Deep learning for computer vision: A brief review," *Computational Intelligence and Neuroscience,* vol. 2018, p. 7068349, 2018, doi: 10.1155/2018/7068349.

[88] R. Yamashita, M. Nishio, R. K. G. Do, and K. Togashi, "Convolutional neural networks: an overview and application in radiology," *Insights into Imaging,* vol. 9, no. 4, pp. 611-629, 2018, doi: 10.1007/s13244-018-0639-9.

[89] A. Khan, A. Sohail, U. Zahoora, and A. S. Qureshi, "A survey of the recent architectures of deep convolutional neural networks," *Artificial Intelligence Review,* 2020, doi: 10.1007/s10462-020-09825-6.

[90] C. Nwankpa, W. Ijomah, A. Gachagan, and S. Marshall, "Activation functions: Comparison of trends in practice and research for deep learning," *arXiv preprint arXiv:1811.03378,* 2018.

[91] J. Gu *et al.*, "Recent advances in convolutional neural networks," *Pattern Recognition,* vol. 77, pp. 354-377, 2018.

[92] J. Deng, W. Dong, R. Socher, L. Li, L. Kai, and F.-F. Li, "ImageNet: A large-scale hierarchical image database," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248-255, doi: 10.1109/CVPR.2009.5206848.

[93] C. Szegedy *et al.*, "Going deeper with convolutions," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1-9.

[94] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *The International Conference on Learning Representations*, Banff, Canada, 2014.

[95] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770-778.

[96] A. S. Vyas, H. B. Prajapati, and V. K. Dabhi, "Survey on face expression recognition using CNN," in *5th International Conference on Advanced Computing & Communication Systems (ICACCS)*, 2019, pp. 102-106, doi: 10.1109/ICACCS.2019.8728330.

[97] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273-297, 1995.

[98] M. M. Khan, R. D. Ward, and M. Ingleby, "Infrared thermal sensing of positive and negative affective states," in *IEEE Conference on Robotics, Automation and Mechatronics*, 2006, pp. 1-6, doi: 10.1109/ramech.2006.252608.

[99] R. J. Williams and D. Zipser, "A learning algorithm for continually running fully recurrent neural networks," *Neural Computation*, vol. 1, no. 2, pp. 270-280, 1989.

[100] F. A. Gers, J. Schmidhuber, and F. Cummins, "Learning to forget: Continual prediction with LSTM," in *9th International Conference on Artificial Neural Networks: ICANN*, Edinburgh, UK, 1999.

[101] A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional LSTM and other neural network architectures," *Neural Networks*, vol. 18, no. 5, pp. 602-610, 2005, doi: https://doi.org/10.1016/j.neunet.2005.06.042.

[102] L. Yang, I. O. Ertugrul, J. F. Cohn, Z. Hammal, D. Jiang, and H. Sahli, "FACS3D-Net: 3D convolution based spatiotemporal representation for action unit detection," *International Conference on Affective Computing and Intelligent Interaction (ACII)*, 2019.

[103] Y. Fan, X. Lu, D. Li, and Y. Liu, "Video-based emotion recognition using CNN-RNN and C3D hybrid networks," in *18th ACM International Conference on Multimodal Interaction*, 2016: ACM, 2016, pp. 445-450.

[104] T. Zhang, W. Zheng, Z. Cui, Y. Zong, and Y. Li, "Spatial–temporal recurrent neural network for emotion recognition," *IEEE Transactions on Cybernetics*, vol. 49, no. 3, pp. 839-847, 2019, doi: 10.1109/TCYB.2017.2788081.

[105] K. Zhang, Y. Huang, Y. Du, and L. Wang, "Facial expression recognition based on deep evolutional spatial-temporal networks," *IEEE Transactions on Image Processing*, vol. 26, no. 9, pp. 4193-4203, 2017, doi: 10.1109/TIP.2017.2689999.

[106] T. Bänziger and K. R. Scherer, "Introducing the geneva multimodal emotion portrayal (gemep) corpus," *Blueprint for affective computing: A sourcebook*, pp. 271-294, 2010.

[107] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A Survey of affect recognition methods: Audio, visual, and spontaneous expressions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 1, pp. 39-58, 2009, doi: 10.1109/tpami.2008.52.

[108] F. Weninger, M. Wöllmer, and B. Schuller, "Emotion recognition in naturalistic speech and language—A Survey," in *Emotion Recognition*: John Wiley & Sons, Inc., 2015, pp. 237-267.

[109] Z. Liu and S. Wang, "Posed and spontaneous expression distinguishment from infrared thermal images," in *21st International Conference on Pattern Recognition (ICPR)* 2012: IEEE, pp. 1108-1111.

[110] C. P. Sumathi, T. Santhanam, and M. Mahadevi, "Automatic facial expression analysis a survey," *International Journal of Computer Science and Engineering Survey*, vol. 3, no. 6, pp. 47-59, 2012.

[111] A. Dhall, O. V. R. Murthy, R. Goecke, J. Joshi, and T. Gedeon, "Video and image based emotion recognition challenges in the wild: EmotiW 2015," *ACM on International Conference on Multimodal Interaction*, Seattle, Washington, USA, 2015.

[112] C. J. Powell and A. Jablonski, "The NIST electron effective-attenuation-length database," *Journal of Surface Analysis*, vol. 9, no. 3, pp. 322-325, 2002, doi: 10.1384/jsa.9.322.

[113] *Dataset 02: IRIS thermal/visible face database*. [Online]. Available: http://vcipl-okstate.org/pbvs/ bench/

[114] A. Esposito, V. Capuano, J. Mekyska, and M. Faundez-Zanuy, "A naturalistic database of thermal emotional facial expressions and effects of induced emotions on memory," in *Cognitive Behavioural Systems*, vol. 7403, A. Esposito, A. Esposito, A. Vinciarelli, R. Hoffmann, and V. Müller Eds., (Lecture Notes in Computer Science: Springer Berlin Heidelberg, 2012, ch. 12, pp. 158-173.

[115] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic, "A multimodal database for affect recognition and implicit tagging," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 42-55, 2012, doi: 10.1109/t-affc.2011.25.

[116] H. Nguyen, K. Kotani, F. Chen, and B. Le, "A thermal facial emotion database and Its analysis," in *Image and Video Technology*, vol. 8333, R. Klette, M. Rivera, and S. i. Satoh Eds., (Lecture Notes in Computer Science: Springer Berlin Heidelberg, 2014, ch. 34, pp. 397-408.

[117] A. Martinez and R. Benavente, "The AR face database (Tech. Rep. 24)," *Barcelona, Spain: Computer Vision Center, Universitat Autònoma de Barcelona*, 1998.

[118] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba, "Coding facial expressions with gabor wavelets," in *Third IEEE international conference on automatic face and gesture recognition*, 1998, pp. 200-205.

[119] M. Pantic, M. Valstar, R. Rademaker, and L. Maat, "Web-based database for facial expression analysis," in *IEEE International Conference on Multimedia and Expo*, 2005, p. 5

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

[121] O. Langner, R. Dotsch, G. Bijlstra, D. H. Wigboldus, S. T. Hawk, and A. Van Knippenberg," Presentation and validation of the Radboud Faces Database," *Cognition and Emotion*, vol. 24, no. 8, pp. 1377-1388, 2010.

[122] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker, "The CMU multi-pose, illumination, and expression (Multi-PIE) face database," *Robotics Inst., Carnegie Mellon Univ., Pittsburgh, PA, USA, Tech. Rep. TR-07-08*, 2007.

[123] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker, "Multi-pie," *Image and Vision Computing*, vol. 28, no. 5, pp. 807-813, 2010.

[124] I. J. Goodfellow *et al.*, "Challenges in representation learning: A report on three machine learning contests," in *International Conference on Neural Information Processing*, 2013: Springer, pp. 117-124.

[125] C. Fabian Benitez-Quiroz, R. Srinivasan, and A. M. Martinez, "Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 5562-5570.

[126] S. Singh and S. Benedict, "Indian Semi-Acted Facial Expression (iSAFE) Dataset for human emotions recognition," in *5th International Symposium, SIRS*, Singapore, 2020: Springer Singapore, in Advances in Signal Processing and Intelligent Recognition Systems, pp. 150-162.

[127] T. Kanade, J. F. Cohn, and T. Yingli, "Comprehensive database for facial expression analysis," in *Fourth IEEE International Conference on Automatic Face and Gesture Recognition* 2000, pp. 46-53, doi: 10.1109/AFGR.2000.840611.

[128] G. A. Miller, "WordNet: a lexical database for English," *Communications of the ACM*, vol. 38, no. 11, pp. 39-41, 1995, doi: 10.1145/219717.219748.

[129] C. Liu and H. Wechsler, "Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition," *IEEE Transactions on Image processing*, vol. 11, no. 4, pp. 467-476, 2002.

[130] L. Jing, X. Yang, and Y. Tian, "Video you only look once: Overall temporal convolutions for action recognition," *Journal of Visual Communication and Image Representation*, vol. 52, pp. 58-65, 2018, doi: https://doi.org/10.1016/j.jvcir.2018.01.016.

[131] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," *Advances in neural information processing systems*, vol. 27, pp. 568-576, 2014.

[132] C. Shan, S. Gong, and P. W. McOwan, "Facial expression recognition based on Local Binary Patterns: A comprehensive study," *Image and Vision Computing*, vol. 27, no. 6, pp. 803-816, 2009, doi: 10.1016/j.imavis.2008.08.005.

[133] B.-K. Kim, J. Roh, S.-Y. Dong, and S.-Y. Lee, "Hierarchical committee of deep convolutional neural networks for robust facial expression recognition," *Journal on Multimodal User Interfaces*, journal article vol. 10, no. 2, pp. 173-189, 2016, doi: 10.1007/s12193-015-0209-0.

[134] A. Dhall, R. Goecke, J. Joshi, M. Wagner, and T. Gedeon, "Emotion recognition in the wild challenge 2013," in *15th ACM on International Conference on Multimodal Interaction*, 2013: ACM, pp. 509-516.

[135] M. Liu, R. Wang, Z. Huang, S. Shan, and X. Chen, "Partial least squares regression on grassmannian manifold for emotion recognition," in *15th ACM on International Conference on Multimodal Interaction*, 2013: ACM, pp. 525-530.

[136] M. Liu, R. Wang, S. Li, S. Shan, Z. Huang, and X. Chen, "Combining multiple kernel methods on riemannian manifold for emotion recognition in the wild," *International Conference on Multimodal Interaction*, Istanbul, Turkey, 2014.

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

[138] X. Ouyang *et al.*, "Audio-visual emotion recognition using deep transfer learning and multiple temporal models," in *19th ACM International Conference on Multimodal Interaction*, 2017, 2017, pp. 577-582.

[139] V. Vielzeuf, S. Pateux, and F. Jurie, "Temporal multimodal fusion for video emotion classification in the wild," *ACM International Conference on Multimodal Interaction*, Glasgow, UK, 2017.

[140] K. Reese, Y. Zheng, and A. Elmaghraby, "A comparison of face detection algorithms in visible and thermal spectrums," in *International Conference on Advances in Computer Science and Application*, 2012.

[141] S. G. B. Gupta, and A. Tiwari, "Face detection using gabor feature extraction and artificial neural networks," *ISCET*, 2010.

[142] P. J. Phillips, H. Wechsler, J. Huang, and P. J. Rauss, "The FERET database and evaluation procedure for face-recognition algorithms," *Image and Vision Computing*, vol. 16, no. 5, pp. 295-306, 1998, doi: http://dx.doi.org/10.1016/S0262-8856(97)00070-X.

[143] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 9, no. 1, pp. 62-66, 1979, doi: 10.1109/tsmc.1979.4310076.

[144] Y. K. Cheong, V. V. Yap, and H. Nisar, "A novel face detection algorithm using thermal imaging," in *IEEE Symposium on Computer Applications and Industrial Electronics (ISCAIE)*, 2014, pp. 208-213, doi: 10.1109/iscaie.2014.7010239.

[145] J. Mekyska, V. Espinosa-Duro, and M. Faundez-Zanuy, "Face segmentation: A comparison between visible and thermal images," in *IEEE International Carnahan Conference on Security Technology (ICCST)*, 2010, pp. 185-189, doi: 10.1109/CCST.2010.5678709.

[146] W. K. Wong, J. H. Hui, J. B. M. Desa, N. I. N. B. Ishak, A. B. Sulaiman, and N. Yante Binti Mohd, "Face detection in thermal imaging using head curve geometry," in *5th International Congress on Image and Signal Processing (CISP)*, 16-18 Oct. 2012 2012, pp. 881-884, doi: 10.1109/CISP.2012.6469684.

[147] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," *CVPR (1)*, vol. 1, pp. I-511-I-518, 2001, doi: 10.1109/cvpr.2001.990517.

[148] Z. Jie and C. Zhiqian, "Real Time Face Detection System Using Adaboost and Haar-like Features," in *2nd International Conference on Information Science and Control Engineering (ICISCE)*, 24-26 April 2015, pp. 404-407, doi: 10.1109/ICISCE.2015.95.

[149] T. Ahonen, A. Hadid, and M. Pietikäinen, "Face recognition with local binary patterns," in *European conference on computer vision*, 2004: Springer, pp. 469-481.

[150] A. Kumar, A. Kaur, and M. Kumar, "Face detection techniques: a review," *Artificial Intelligence Review*, vol. 52, no. 2, pp. 927-948, 2019.

[151] S. Lin and X. X. Dai, "Sketch-based face alignment for thermal face recognition," in *21st International Conference on Pattern Recognition (ICPR)*, 11-15 Nov. 2012, pp. 2347-2350.

[152] N. Zaeri, "Component-based Thermal Face Recognition," *British Journal of Applied Science & Technology*, vol. 4, no. 6, pp. 945-966, 2014.

[153] L. Guan, "Face recognition with visible and thermal IR images," *M.S.E.E.*, Temple University, Ann Arbor, 2010.

[154] S. Wang, Z. Gao, S. He, M. He, and Q. Ji, "Gender recognition from visible and thermal infrared facial images," *Multimed Tools Appl*, pp. 1-24, 2015/06/30 2015, doi: 10.1007/s11042-015-2756-5.

[155] "Training Image Labeler." Mathworks. http://uk.mathworks.com/help/vision/ref/trainingimagelabeler-app.html (accessed) 1-12-2020.

[156] S. Wang, Z. Liu, P. Shen, and Q. Ji, "Eye localization from thermal infrared images," *Pattern Recognition*, vol. 46, no. 10, pp. 2613-2621, 2013, doi: http://dx.doi.org/10.1016/j.patcog.2013.03.001.

[157] S. Zafeiriou, C. Zhang, and Z. Zhang, "A survey on face detection in the wild: Past, present and future," *Computer Vision and Image Understanding*, vol. 138, pp. 1-24, 2015/09/01/ 2015, doi: https://doi.org/10.1016/j.cviu.2015.03.015.

[158] H. Li, Z. Lin, J. Brandt, X. Shen, and G. Hua, "Efficient boosted exemplar-based face detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1843-1850.

[159] N. Kanwal, E. Bostanci, and A. F. Clark, "Evaluation method, dataset size or dataset content: How to evaluate algorithms for image matching?," *Journal of Mathematical Imaging and Vision*, journal article vol. 55, no. 3, pp. 378-400, 2016, doi: 10.1007/s10851-015-0626-4.

[160] A. S. Mohammad, A. Rattani, and R. Derahkshani, "Eyeglasses detection based on learning and non-learning based classification schemes," in *IEEE International Symposium on Technologies for Homeland Security (HST)*, 25-26 April 2017 2017, pp. 1-5, doi: 10.1109/THS.2017.7943484.

[161] L. Shao, R. Zhu, and Q. Zhao, "Glasses detection using convolutional neural networks," in *11th Chinese Conference Biometric Recognition (CCBR)* Chengdu, China,, Z. You *et al.*, Eds., 2016: Springer International Publishing, pp. 711-719, doi: 10.1007/978-3-319-46654-5_78. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-46654-5_78

[162] S. Du, J. Liu, Y. Liu, X. Zhang, and J. Xue, "Precise glasses detection algorithm for face with in-plane rotation," *Multimedia Systems,* journal article vol. 23, no. 3, pp. 293-302, 2017, doi: 10.1007/s00530-015-0483-4.

[163] A. Fernández, R. Casado, and R. Usamentiaga, "A real-time big data architecture for glasses detection using computer vision techniques," in *3rd International Conference on Future Internet of Things and Cloud*, 24-26 Aug. 2015 2015, pp. 591-596, doi: 10.1109/FiCloud.2015.78.

[164] A. Fernández, R. García, R. Usamentiaga, and R. Casado, "Glasses detection on real images based on robust alignment," *Machine Vision and Applications,* journal article vol. 26, no. 4, pp. 519-531, 2015, doi: 10.1007/s00138-015-0674-1.

[165] S. Bekhet and H. Alahmer, "A robust deep learning approach for glasses detection in non-standard facial images," *IET Biometrics,* vol. n/a, no. n/a, 2020, doi: https://doi.org/10.1049/bme2.12004.

[166] E. Hjelmås and B. K. Low, "Face Detection: A Survey," *Computer Vision and Image Understanding,* vol. 83, no. 3, pp. 236-274, 2001, doi: http://dx.doi.org/10.1006/cviu.2001.0921.

[167] Z. Jing and R. Mariani, "Glasses detection and extraction by deformable contour," in *15th International Conference on Pattern Recognition*, 2000, vol. 2: IEEE, pp. 933-936.

[168] W. Bo, A. Haizhou, and L. Ran, "Glasses detection by boosting simple wavelet features," in *17th International Conference on Pattern Recognition (ICPR)*, 2004, vol. 1: IEEE, pp. 292-295 doi: 10.1109/ICPR.2004.1334110.

[169] W. Gao *et al.*, "The CAS-PEAL large-scale chinese face database and baseline evaluations," *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans,* vol. 38, no. 1, pp. 149-161, 2008, doi: 10.1109/TSMCA.2007.909557.

[170] A. Rattani, R. Derakhshani, S. K. Saripalle, and V. Gottemukkula, "Competition on mobile ocular biometric recognition," in *IEEE International Conference on Image Processing (ICIP)*, 2016, pp. 320-324, doi: 10.1109/ICIP.2016.7532371.

[171] Y. Zheng, "Orientation-based face recognition using multispectral imagery and score fusion," *Optical Engineering,* vol. 50, no. 11, p. 117202, 2011.

[172] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters,* vol. 23, no. 10, pp. 1499-1503, 2016.

[173] Y. Jia *et al.*, "Caffe: Convolutional architecture for fast feature embedding," in *22nd ACM international conference on Multimedia*, 2014: ACM, pp. 675-678.

[174] N. Kumar, P. Belhumeur, and S. Nayar, "Facetracer: A search engine for large collections of images with faces," in *European conference on computer vision*, 2008: Springer, pp. 340-353.

[175] N. Zhang, M. Paluri, M. A. Ranzato, T. Darrell, and L. Bourdev, "Panda: Pose aligned networks for deep attribute modeling," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 1637-1644.

[176] J. Li and Y. Zhang, "Learning surf cascade for fast and accurate object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2013, pp. 3468-3475.

[177] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *IEEE International Conference on Computer Vision*, 2015, pp. 3730-3738.

[178] E. M. Hand and R. Chellappa, "Attributes for improved attributes: A multi-task network utilizing implicit and explicit relationships for facial attribute classification," in *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[179] J. Cao, Y. Li, and Z. Zhang, "Partially shared multi-task convolutional neural network with local constraint for face attribute learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4290-4299.

[180] Z. Yang, J. Sullivan, and L. Haibo, "Face attribute prediction using off-the-shelf CNN features," in *International Conference on Biometrics (ICB)*, 13-16 June 2016, pp. 1-7, doi: 10.1109/ICB.2016.7550092.

[181] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Learning face representation from scratch," *arXiv preprint arXiv:1411.7923*, 2014.

[182] T. E. Haley-Hermiz, "Emotion capture: Emotion mimicry using facial motion capture," *M.S.*, Purdue University, Ann Arbor, 2014.

[183] E. M. Schmidt, J. J. Scott, and Y. E. Kim, "Feature learning in dynamic environments: modeling the acoustic structure of musical emotion," in *13th International Society for Music Information Retrieval Conference ISMIR*, 2012: Citeseer, pp. 325-330.

[184] Y. Du and X. Lin, "Emotional facial expression model building," *Pattern Recognition Letters*, vol. 24, no. 16, pp. 2923-2934, 2003, doi: http://dx.doi.org/10.1016/S0167-8655(03)00153-3.

[185] L. Zhou, H. Pang, and H. Liu, "Emotion recognition from physiological signals based on ASAGA," in *International Conference on Communication, Electronics and Automation Engineering*, G. Yang, Ed., 2013, vol. 181: Springer Berlin Heidelberg, in Advances in Intelligent Systems and Computing, pp. 735-740, doi: 10.1007/978-3-642-31698-2_103. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-31698-2_103

[186] A. F. R. Rahman and M. C. Fairhurst, "Multiple classifier decision combination strategies for character recognition: A review," *Document Analysis and Recognition,* vol. 5, no. 4, pp. 166-194, 2003.

[187] P. Luis, M. J. Poza, J. M. Gómez, and D. Carrero, "Multimodal biometrics: topics in score fusion," in *Computational Intelligence in Security for Information Systems*: Springer, 2009, pp. 155-162.

[188] A. A. Ross, K. Nandakumar, and A. K. Jain, *Handbook of multibiometrics*. Springer Science & Business Media, 2006.

[189] K. Tiwari and P. Gupta, "An adaptive score level fusion scheme for multimodal biometric systems," in *Adaptive Biometric Systems: Recent Advances and Challenges*, A. Rattani, F. Roli, and E. Granger Eds. Cham: Springer International Publishing, 2015, pp. 119-131.

[190] S. C. Dass, K. Nandakumar, and A. K. Jain, "A principled approach to score level fusion in multimodal biometric systems," in *International Conference on Audio-and Video-based Biometric Person Authentication*, Berlin, Heidelberg, 2005: Springer Berlin Heidelberg, in Audio- and Video-Based Biometric Person Authentication, pp. 1049-1058.

[191] P. S. Aleksic and A. K. Katsaggelos, "Automatic facial expression recognition using facial animation parameters and multistream HMMs," *IEEE Transactions on Information Forensics and Security,* vol. 1, no. 1, pp. 3-11, 2006.

[192] J. Yan *et al.*, "Multi-clue fusion for emotion recognition in the wild," in *18th ACM International Conference on Multimodal Interaction*, Tokyo, Japan, 2016.

[193] A. Yao, D. Cai, P. Hu, S. Wang, L. Sha, and Y. Chen, "HoloNet: towards robust emotion recognition in the wild," in *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, 2016: ACM, pp. 472-478.

[194] T. Zhang, W. Zheng, Z. Cui, Y. Zong, J. Yan, and K. Yan, "A Deep Neural Network-Driven Feature Learning Method for Multi-view Facial Expression Recognition," *IEEE Transactions on Multimedia*, vol. 18, no. 12, pp. 2528-2536, 2016, doi: 10.1109/TMM.2016.2598092.

[195] R. Dzati Athiar, S. A. Samad, and A. Hussain, "Score information decision fusion using support vector machine for a correlation filter based speaker authentication system," in *International Workshop on Computational Intelligence in Security for Information Systems CISIS'08*, 2009: Springer, pp. 235-242.

[196] K. Nandakumar, Y. Chen, S. C. Dass, and A. Jain, "Likelihood ratio-based biometric score fusion," *IEEE transactions on pattern analysis and machine intelligence*, vol. 30, no. 2, pp. 342-347, 2008.

[197] M. He *et al.*, "Performance evaluation of score level fusion in multimodal biometric systems," *Pattern Recognition*, vol. 43, no. 5, pp. 1789-1800, 2010.

[198] S. H. Lee, W. J. Baddar, and Y. M. Ro, "Collaborative expression representation using peak expression and intra class variation face images for practical subject-independent emotion recognition in videos," *Pattern Recognition*, vol. 54, pp. 52-67, 2016, doi: https://doi.org/10.1016/j.patcog.2015.12.016.

[199] A. Sun, Y. Li, Y.-M. Huang, and Q. Li, "The Exploration of Facial Expression Recognition in Distance Education Learning System," in *International Conference on Innovative Technologies and Learning*, 2018: Springer, pp. 111-121.

[200] M.-W. Huang, Z.-w. Wang, and Z.-L. Ying, "A new method for facial expression recognition based on sparse representation plus LBP," in *3rd International Congress on Image and Signal Processing*, 2010, vol. 4: IEEE, pp. 1750-1754.

[201] Z. Wang and Z. Ying, "Facial Expression Recognition Based on Local Phase Quantization and Sparse Representation," in *8th International Conference on Natural Computation*, 29-31 May 2012 2012, pp. 222-225, doi: 10.1109/ICNC.2012.6234551.

[202] S. Zhang, X. Zhao, and B. Lei, "Robust Facial Expression Recognition via Compressive Sensing," *Sensors*, vol. 12, no. 3, pp. 3747-3761, 2012. [Online]. Available: https://www.mdpi.com/1424-8220/12/3/3747.

[203] B. Jiang, M. F. Valstar, and M. Pantic, "Action unit detection using sparse appearance descriptors in space-time video volumes," in *IEEE International*

Conference on Automatic Face & Gesture Recognition (FG), 2011: IEEE, pp. 314-321.

[204] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look Once: unified, real-time object detection," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 779-788, doi: 10.1109/CVPR.2016.91.

[205] Y. Kong and Y. Fu, "Human action recognition and prediction: A survey," *arXiv preprint arXiv:1806.11230* 2018.

[206] Y. Song *et al.*, "Large margin local estimate with applications to medical image classification," *IEEE Transactions on Medical Imaging*, vol. 34, no. 6, pp. 1362-1377, 2015, doi: 10.1109/TMI.2015.2393954.

[207] J. Chen, Z. Chen, Z. Chi, and H. Fu, "Facial expression recognition in video with multiple feature fusion," *IEEE Transactions on Affective Computing*, vol. 9, no. 1, pp. 38-50, 2018, doi: 10.1109/TAFFC.2016.2593719.

[208] H. Yan, "Collaborative discriminative multi-metric learning for facial expression recognition in video," *Pattern Recognition*, vol. 75, pp. 33-40, 2018, doi: https://doi.org/10.1016/j.patcog.2017.02.031.

[209] A. Dhall, R. Goecke, J. Joshi, K. Sikka, and T. Gedeon, "Emotion recognition in the wild challenge 2014: baseline, data and protocol," in *16th International Conference on Multimodal Interaction*, Istanbul, Turkey, 2014, 2666275: ACM, pp. 461-466, doi: 10.1145/2663204.2666275.

[210] J. Yue-Hei Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici, "Beyond short snippets: Deep networks for video classification," in *IEEE conference on computer vision and pattern recognition*, 2015, pp. 4694-4702.

[211] B.-K. Kim, H. Lee, J. Roh, and S.-Y. Lee, "Hierarchical committee of deep CNNs with exponentially-weighted decision fusion for static facial expression recognition," in *15th International Conference on Multimodal Interaction*, 2015, pp. 427-434, doi: 10.1145/2818346.2830590

[212] W. Li, F. Abtahi, and Z. Zhu, "A deep feature based multi-kernel learning approach for video emotion recognition," in *ACM international conference on multimodal interaction*, 2015, pp. 483-490, doi: 10.1145/2818346.2830583.

[213] V. Vielzeuf, C. Kervadec, S. Pateux, A. Lechervy, and F. Jurie, "An occam's razor view on learning audiovisual emotion recognition with small training sets," in *ACM International Conference on Multimodal Interaction*, Boulder, CO, USA, 2018.

[214] C. Lu *et al.*, "Multiple spatio-temporal feature learning for video-based emotion recognition in the wild," in *ACM International Conference on Multimodal Interaction*, Boulder, CO, USA, 2018.

[215] B. Knyazev, R. Shvetsov, N. Efremova, and A. Kuharenko, "Convolutional neural networks pretrained on large face recognition datasets for emotion classification from video," *arXiv preprint arXiv:1711.04598*. [Online]. Available: https://ui.adsabs.harvard.edu/abs/2017arXiv171104598K

[216] S. A. Bargal, E. Barsoum, C. C. Ferrer, and C. Zhang, "Emotion recognition in the wild from videos using images," in *18th ACM International Conference on Multimodal Interaction*, Tokyo, Japan, 2016, pp. 433-436, doi: 10.1145/2993148.2997627.

[217] D. M. Truong, H.-G. Doan, T.-H. Tran, H. Vu, and T.-L. Le, "Robustness analysis of 3D convolutional neural network for human hand gesture recognition," *International Journal of Machine Learning and Computing*, vol. 9, no. 2, pp. 135-142, 2019.

[218] N. Samadiani *et al.*, "A review on automatic facial expression recognition systems assisted by multimodal sensor data," *Sensors*, vol. 19, no. 8, p. 1863, 2019. [Online]. Available: https://www.mdpi.com/1424-8220/19/8/1863.

[219] A. R. Doherty, D. Byrne, A. F. Smeaton, G. J. Jones, and M. Hughes, "Investigating keyframe selection methods in the novel domain of passively captured visual lifelogs," in *International Conference on Content-based Image and Video Retrieval*, 2008: ACM, pp. 259-268, doi: 10.1145/1386352.1386389.

[220] F. Noroozi, M. Marjanovic, A. Njegus, S. Escalera, and G. Anbarjafari, "Audio-visual emotion recognition in video clips," *IEEE Transactions on Affective Computing*, vol. 10, no. 1, pp. 60-75, 2017.

[221] A. F. Smeaton and P. Browne, "A usage study of retrieval modalities for video shot retrieval," *Information Processing & Management*, vol. 42, no. 5, pp. 1330-1344, 2006, doi: https://doi.org/10.1016/j.ipm.2005.11.003.

[222] B. Yang, J. Cao, R. Ni, and Y. Zhang, "Facial expression recognition using weighted mixture deep neural network based on double-channel facial images," *IEEE Access*, vol. 6, pp. 4630-4640, 2018, doi: 10.1109/ACCESS.2017.2784096.

[223] K. Anderson and P. W. McOwan, "A real-time automated system for the recognition of human facial expressions," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 36, no. 1, pp. 96-105, 2006, doi: 10.1109/TSMCB.2005.854502.

[224] S. M. Guo, Y. A. Pan, Y. C. Liao, C. Y. Hsu, J. S. H. Tsai, and C. I. Chang, "A key frame selection-based facial expression recognition system," in *First International Conference on Innovative Computing, Information and Control (ICICIC'06)*, 2006, vol. 3, pp. 341-344, doi: 10.1109/ICICIC.2006.383.

[225] Q. Zhang, S.-P. Yu, D.-S. Zhou, and X.-P. Wei, "An efficient method of key-frame extraction based on a cluster algorithm," *Journal of Human Kinetics*, vol. 39, no. 1, pp. 5-14, 2013.

[226] S. Hasebe, M. Nagumo, S. Muramatsu, and H. Kikuchi, "Video key frame selection by clustering wavelet coefficients," in *12th European Signal Processing Conference*, 2004, pp. 2303-2306.

[227] J. Shi, "Good features to track," in *IEEE conference on computer vision and pattern recognition*, 1994, pp. 593-600.

[228] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2818-2826.

[229] Y. Fan, J. C. K. Lam, and V. O. K. Li, "Multi-region ensemble convolutional neural network for facial expression recognition," in *International Conference on Artificial Neural Networks*, Cham, 2018.

[230] W. Ding *et al.*, "Audio and face video emotion recognition in the wild using deep neural networks and small datasets," in *18th ACM International Conference on Multimodal Interaction*, 2016, pp. 506-513.

[231] S. Ouellet, "Real-time emotion recognition for gaming using deep convolutional network features," *arXiv preprint arXiv:1408.3750*, 2014.

[232] X. Liu, B. Vijaya Kumar, J. You, and P. Jia, "Adaptive deep metric learning for identity-aware facial expression recognition," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017, pp. 20-29.

[233] H. Yang, U. Ciftci, and L. Yin, "Facial expression recognition by de-expression residue learning," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2168-2177.

[234] Z. Zhang, P. Luo, C. C. Loy, and X. Tang, "From facial expression recognition to interpersonal relation prediction," *International Journal of Computer Vision*, vol. 126, no. 5, pp. 550-569, 2018, doi: 10.1007/s11263-017-1055-1.

[235] Z. Meng, P. Liu, J. Cai, S. Han, and Y. Tong, "Identity-aware convolutional neural network for facial expression recognition," in *12th IEEE International*

Conference on Automatic Face & Gesture Recognition (FG), 2017, pp. 558-565, doi: 10.1109/FG.2017.140.

[236] J. Cai, Z. Meng, A. S. Khan, Z. Li, J. O. Reilly, and Y. Tong, "Island loss for learning discriminative features in facial expression recognition," in *13th IEEE International Conference on Automatic Face & Gesture Recognition (FG)*, 2018, pp. 302-309, doi: 10.1109/FG.2018.00051.

[237] Y. Fan, V. Li, and J. C. K. Lam, "Facial expression recognition with deeply-supervised attention network," *IEEE Transactions on Affective Computing*, pp. 1-1, 2020, doi: 10.1109/TAFFC.2020.2988264.

[238] M. Liu, S. Li, S. Shan, R. Wang, and X. Chen, "Deeply learning deformable facial action parts model for dynamic expression analysis," in *Asian Conference on Computer Vision*, Cham, 2015: Springer International Publishing, in Computer Vision -- ACCV 2014, pp. 143-157.

[239] H. Jung, S. Lee, J. Yim, S. Park, and J. Kim, "Joint fine-tuning in deep neural networks for facial expression recognition," in *IEEE International Conference on Computer Vision*, 2015, pp. 2983-2991.

[240] A. Sanin, C. Sanderson, M. T. Harandi, and B. C. Lovell, "Spatio-temporal covariance descriptors for action and gesture recognition," in *IEEE Workshop on Applications of Computer Vision (WACV)*, 2013, pp. 103-110, doi: 10.1109/WACV.2013.6475006.

[241] M. Liu, S. Shan, R. Wang, and X. Chen, "Learning expressionlets on spatio-temporal manifold for dynamic facial expression recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1749-1756.

[242] K. Sikka, G. Sharma, and M. Bartlett, "Lomo: Latent ordinal model for facial analysis in videos," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 5580-5589.

[243] S. Li and W. Deng, "Deep facial expression recognition: A Survey," *IEEE Transactions on Affective Computing*, pp. 1-1, 2020, doi: 10.1109/TAFFC.2020.2981446.

APPENDIX: PAPERS DELIVERED

- A. M. Basbrain, J. Q. Gan, and A. Clark, "Accuracy enhancement of the viola-jones algorithm for thermal face detection," in International Conference on Intelligent Computing, 2017: Springer, pp. 71-82.
- A. M. Basbrain, I. Al-Taie, N. Azeez, J. Q. Gan, and A. Clark, "Shallow convolutional neural network for eyeglasses detection in facial images," in 9th Computer Science and Electronic Engineering (CEEC), 2017, pp. 157-161.
- A. M. Basbrain, J. Q. Gan, A. Sugimoto, and A. Clark, "A neural network approach to score fusion for emotion recognition," in 10th Computer Science and Electronic Engineering (CEEC), 2018, pp. 180-185.
- A. Basbrain and J. Q. Gan, "One-Shot Only real-time video classification: A case study in facial emotion recognition," Cham, 2020: Springer International Publishing, in Intelligent Data Engineering and Automated Learning – IDEAL 2020, pp. 197-208.
- I. Al-Taie, N. Azeez, A. Basbrain, and A. Clark, "The effect of distance similarity measures on the performance of face, ear and palm biometric systems," in International conference on digital image computing: Techniques and applications (DICTA), 2017: IEEE, pp. 1-7.
- I. Al-Taie, N. Azeez, W. Yahya, A. Basbrain, and A. Clark, "Biometric Recognition Systems Based on SVM pca and SVM pca, lda Techniques," in Third World Conference on Smart Trends in Systems Security and Sustainability (WorldS4), 2019: IEEE, pp. 158-161.
- N. Azeez, I. Al-Taie, W. Yahya, A. Basbrain, and A. Clark, "Regional Agricultural Land Texture Classification Based on GLCMs, SVM and Decision Tree Induction Techniques," in 10th Computer Science and Electronic

Engineering (CEEC), 2018: IEEE, pp. 131-135.

- N. Azeez, W. Yahya, I. Al-Taie, A. Basbrain, and A. Clark, "Regional Agricultural Land Classification Based on Random Forest (RF), Decision Tree, and SVMs Techniques," in Fourth International Congress on Information and Communication Technology, 2020: Springer, pp. 73-81.