Artificial Intelligence Tools for Facial Expression Analysis.

by Luma Akram Alharbawee

Department of Computer Science
University of Exeter

Submitted by Luma Akram Alharbawee, to the University of Exeter as a thesis for the degree of Doctor of Philosophy in Computer Science, November, 2019.

I certify that all material in this thesis which is not my own work has been identified and that no material has previously been submitted and approved for the award of a degree by this or any other University.

Signature

Luma Alharbawee

I am extremely grateful and my heartfelt thanks are reserved for both Professor Sinan Kalkan and Dr. Jacqueline Christmas - my viva examiners, for their suggestions, very useful comments, and invaluable insights.

Declaration

I, Luma hereby to declare that this thesis is written by me and the work presented in it are my own and has been generated by me as the result of my own original research, except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This thesis is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements.

by Luma Akram Alharbawee

Acknowledgements

I wish to express my deep thankfulness and appreciation towards my brilliant supervisor, Dr. Nicolas Pugeault for his up-to-date ideas, motivation, feedback, and valuable suggestions, and also to my second supervisor Prof. Richard Everson, for his expertise, wisdom, and advice concerning my research. Sincere thanks to my mentor Dr. Mustafa Aziz, for all the help and support throughout my work, and all the staff in my department, specially, Anastasios Roussos, and Luo Chunbo and everyone who helped me during my studies, including Andrea Donakey for her proofreading and Magdalena Katomeri for her wellbeing support. I am so indebted. Many thanks to Vina Flores for all the loving support, inspiration, encouragement, and unceasing guidance. I would like to send a greeting to St. David's Primary School, especially, a big thanks to Mrs. Fran Brinicombe for the excellent pastoral care and emotional support she has provided for both my children. Most of all, the biggest thank you goes to acknowledge the Iraqi cultural Attache for their financial support which made this project possible and especially over the last four years. I also extend my thanks to all my colleagues in the department, particularly George De Ath, Dmitry Kangin, ChengQiang Huang, and Yuan Zuo, who have all been on their own research journey and always been happy to help. I am immeasurably grateful for your support. Big thanks go to all the lovely friends and amazing people I met in England specially Bridget Sealey and all the residential Iraqi families for consistently pushing me to finish my studies. Back to my family in Iraq, who are all in my thoughts always, for the sustained love and enduring support they offer in my life. In particular, I would like to send a big warmest gratitude to my affectionate brother Ahmed for his dedicated support, as well as deepest thanks to my mother who always prays for me. This thesis is dedicated to my loving parents and to the loving memory of my late father, Akram Alharbawee, who was with me every step of this journey and whose everlasting support is with me every day of my life. To my husband, Mohammed, for his truly unconditional help and for having to take on the work-life balance, which is a really unenviable task. Without you, none of this journey would have ever happened. I would like to say sorry to my son Yousif for having to leave you when I go to university and to my other beloved children Noor and Yaman, I love you so much and I am always proud of you all. ***Shukran*** to all of you.

Abstract

Inner emotions show visibly upon the human face and are understood as a basic guide to an individual's inner world. It is, therefore, possible to determine a person's attitudes and the effects of others' behaviour on their deeper feelings through examining facial expressions. In real world applications, machines that interact with people need strong facial expression recognition. This recognition is seen to hold advantages for varied applications in affective computing, advanced human-computer interaction, security, stress and depression analysis, robotic systems, and machine learning. This thesis starts by proposing a benchmark of dynamic versus static methods for facial Action Unit (AU) detection. AU activation is a set of local individual facial muscle parts that occur in unison constituting a natural facial expression event. Detecting AUs automatically can provide explicit benefits since it considers both static and dynamic facial features. For this research, AU occurrence activation detection was conducted by extracting features (static and dynamic) of both nominal hand-crafted and deep learning representation from each static image of a video. This confirmed the superior ability of a pretrained model that leaps in performance. Next, temporal modelling was investigated to detect the underlying temporal variation phases using supervised and unsupervised methods from dynamic sequences. During these processes, the importance of stacking dynamic on top of static was discovered in encoding deep features for learning temporal information when combining the spatial and temporal schemes simultaneously. Also, this study found that fusing both temporal and temporal features will give more long term temporal pattern information. Moreover, we hypothesised that using an unsupervised method would enable the leaching of invariant information from dynamic textures. Recently, fresh cutting-edge developments have been created by approaches based on Generative Adversarial Networks (GANs). In the second section of this thesis, we propose a model based on the adoption of an unsupervised DCGAN for the facial features' extraction and classification to achieve the following: the creation of facial expression images under different arbitrary poses (frontal, multi-view, and in the wild), and the recognition of emotion categories and AUs, in an attempt to resolve the problem of recognising the static seven classes of emotion in the wild. Thorough experimentation with the proposed cross-database performance demonstrates that this approach can improve the generalization results. Additionally, we showed that the features learnt by the DCGAN process are poorly suited to encoding facial expressions when observed under multiple views, or when trained from a limited number of positive examples. Finally, this research focuses on disentangling identity from expression for facial expression recognition. A novel technique was implemented for emotion recognition from a single monocular image. A large-scale dataset (Face vid) was created from facial image videos which were rich in variations and distribution of facial dynamics, appearance, identities, expressions, and 3D poses. This dataset was used to train a DCNN (ResNet) to regress the expression parameters from a 3D Morphable Model jointly with a back-end classifier.

Table of contents

List of figures

List of tables

Chapter 1

Introduction

1.1 Research Context

As humans, we are particularly gifted at recognising other people and inferring mental states from even a cursory glance at their faces – in fact, even young children can recognise happiness and emulate smiles. If Artificial Intelligence and Robotic systems are to spread wider in society, they will need to be able to interact with people appropriately by recognising and taking into account their mood and state of mind. For example, the exact same words can carry very different meaning if spoken in anger, annoyance, amusement or anxiety through feelings such as rage, resentment, bitterness, discontent or irritation. One essential difficulty of this task is the large range of differences between people's faces and how they show emotion. There is considerable diversity in human facial expression, and as such, there is a lack of definitive scientific explanation for its many complexities [94]. In daily life, more than seven thousand facial Action Unit (AU) groups are observed even through the small numbers of AUs [535]. The human face can display an assortment of facial expressions and come in different shapes. In addition, the timing, speed, relative duration, and appearance of various facial action activation might differ between acted and spontaneous facial behaviour [424]. Previous work in computer vision has demonstrated that although it is possible to recognise emotions from images of known people's faces, it is difficult to do it on unknown persons. People can often show a mixture of emotional expressions. Furthermore, pure facial expressions are rarely elicited. Moreover, from a technical standpoint, detecting real-time facial expression presents a difficult challenge in computer vision due to the level and ambiguity of the variability, the subtlety, and the complexity in its appearance [230] and subjects can be extremely dynamic in their pose. Despite many recent successes in this field of study, there are several questions remain and not cover in this thesis: How do we accurately identify facial expressions? What does facial affective state mean? Are there

Fig. 1.1 Some applications of facial expression recognition systems.

facial expressions influenced by culture? How do people interpret, experience, deliver and express these emotions in the real world? What type of features is most important? Are transient or intransient features within faces, inner features such as eyes, eyebrows, facial lines, mouth, furrow, nose, and lips or outer features for example head shape, movement, and hairline used for an efficient facial expression recognition system? How do we read and understand facial actions when gathering the data? How do facial expressions effect communication? Is it always sufficient to use the known emotion categories? Can we do representational emotions? What is the most effective way of encoding facial expressions for computer-aided applications? How do we obtain reliable Ground Truth (GT) information? How do we best use temporal information? Could we get an improvement in the performance from the learning of temporal dependencies, and if so, how [92]? Does the detection of the occurrence of the target AUs need the modelling of entire sequences, or is a single frame sufficient [528]? How could we incorporate the facial expression recognition system with other models [536]? Does decreasing the dimensions of the feature space provide an improvement in computational capability? Do any approaches for reducing the dimensions produce improved recognising of facial expression? How much training data is required for the optimal automatic facial expression recognition (FER) system? Would providing more additional training samples yield diminishing returns results or counter-productive outcomes [190] [245]? For a video of facial expression, how can we detect per frame that AU or multiple AUs are activated [557]? How can we know that the activation is in the onset, the apex, or in offset phase, and what is the total dwell duration in the process of the activation [554]? What are the cues that are important to garner for modelling facial expressions and how can we encode them? Should we use spatial representations, latent temporality, spatio-temporal, or intensity? Would considering all the matters jointly in one context improve facial Action Units detection? Could we get a representation that generalises

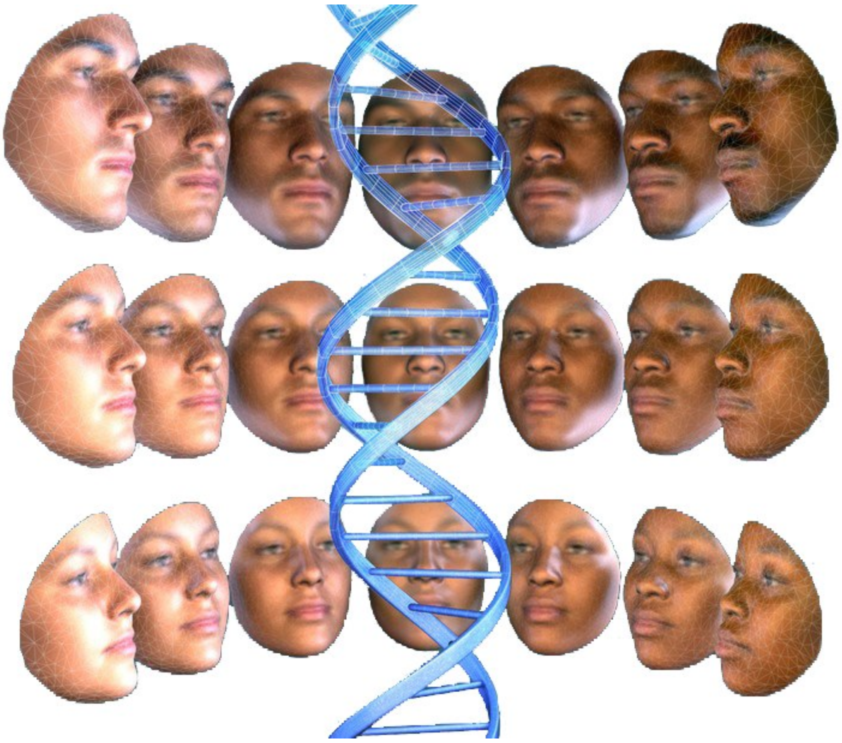


Fig. 1.2 A 3D model link between facial features and DNA; image adapted from [225].

well across datasets and subjects for better AU detection [92]? Does the use of unlabelled data affect the speed of the execution? Are there any improvements accomplished in the classification which consider the unlabelled data? For what conditions and for which datasets should the use of semi-supervised methods be employed? Would that improve the learner agent performance or the separator [33]? Are there any facial expressions which indicate action coordination, and if so, which expressions usually occur [40]? Can the use of the generated additional synthetic images by Generative Adversarial Networks improve the performance of the FER system [623]? Does using GANs such as DCGAN succeed in extracting good deep features representation for emotions/AUs? If yes, how? Which part is used as a feature extractor? And which features are better: frontal view, multi-view, or in the wild? How could we generate videos of diverse facial expression conditioned from only a single face image? How could we represent the face shape by learning a nonlinear 3DMM from a set of 2D face images in the wild [544]?

1.2 Motivations and Applications

Facial expression analysis refers to computer-aided applications that are designed to automatically analyse and recognise human emotions and characteristic facial feature changes

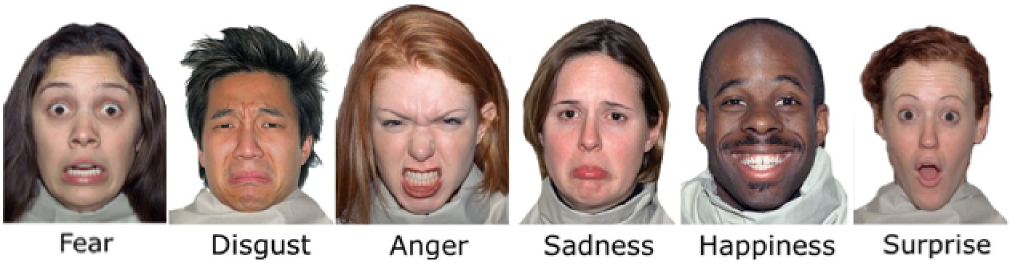


Fig. 1.3 Six universal basic emotions, from left to right: Fear, Disgust, Anger, Sadness, Happiness, and Surprise; image adapted from [226].

using visual information [45]. Modern science's ability to detect trauma through a facial presentation (i.e. facial expression affect analysis) has been viewed by scientists as noteworthy recently. Measurement of effective analysis facial expressions is involved in myriad applications such as Computer Vision and Machine Learning. It is necessary to have a look at all fields and seek to capitalise on the building and development of an automatic FER system [495] [364].

Some of this includes image understanding, the grading system of facial nerve function in medicine, compression of facial image, synthetic face animation [517], access to the security field, video surveillance, entertainment, reconstructions of a 3D face, biometric identification, content image retrieval, video conferencing, video coding (video indexing) [34], advanced intelligent human-computer interaction interfaces [393], applications of the low enforcement, behavior monitoring, behavioural science, psychoanalysis, and anthropology [495] [427], energy conservation, banking authentication, identity verification, systems of criminal justice, border control, interrogation, investigations of an image database, and applications of smart card [34] [364]. The rapid progression of facial expression recognition technology meant that it could be deployed in a multiplicity of fields such as: facial expression and emotion transfer, avatars with expressions, face replacement, face animation [206], analysis of user customer experience, the movie industry avatar animation and marketing (evaluation of customer affective feedback on services products) [618]; online security, control of fraud ID/passport protection, surveillance systems [69], affective robotics [409], access control [288], control of domestic appliance; virtual reality, safe driving, advanced driver fatigue monitoring assistant systems, mobile technology [206], job interviews, computing, call centre applications [288], real-time facial recognition system in mobile, emotional chat pots, digital entertainment, deceit and lie detection, online gaming and online education; assistance and investigation of medical treatment, diseases and depression detection, social robots, anxiety detection, healthcare and clinical monitoring [618], assisted living, evaluation of emotional impairment for those experiencing neuropsychiatric disorders, mental health conditions

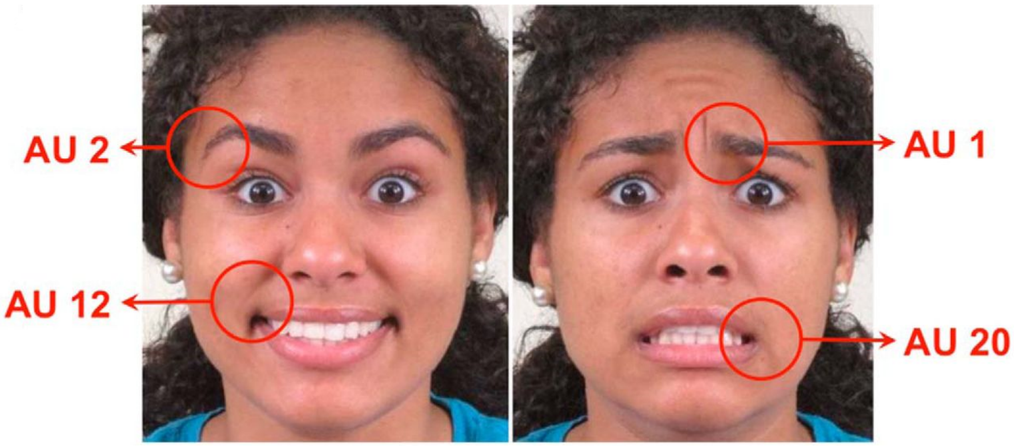


Fig. 1.4 Facial expression images have been annotated with the activation region according to the FACs; image adapted from [514].

detection and the improvement of a variety of expressions produced by autism spectrum disorder patients; engagement of student statistics estimation, social media, analytics of crowd behaviour, real-time attendance tracking among many other areas [125] [69]. Fig. 1.1 shows some applications of a facial expression recognition system. Scientists can determine an individual's sex and ancestry from their DNA; further genetic research is now exploring the links between genes and facial features. The research team started by examining both the genetic markers of face shape and the given physical face shape [94]. Fig. 1.2 represents a 3D based model linking facial features and DNA.

1.3 An Overview of the Current Systems

Knowing the different types of facial affect behaviour to quantify existing categories in the literature is an essential part of designing reliable perceptual integration models as well as gaining insights about human cognition [145]. Cross-cultural studies have shown that some emotions visible via facial expression are universal and some are reliant on one's cultural background. Ekman et al. [152] denoted a categorical model which is a set of discrete affective-related prototypical categories of facial expressions (e.g. happiness, fear, anger, surprise, disgust, and sadness) [157]. Fig. 1.3 shows the six basic emotion categories defined by Ekman.

In the past, all the proposed approaches to automatic expression analysis limit facial expression classification within the basic emotion categories [152] [496]. However, it is not certain whether all expressions able to be displayed on the face can be subsumed under the six basic emotion classes [495]. Nonetheless, to date, there has not been a clearly defined method of identifying Ekman's six basic emotion categories under every day, realistic situations; all of which remain unanswered. So far, the classification of facial expression tends to be subjective, so they are placed into several emotional categories [519]. Further comprehensive methodologies pursue the Facial Action Coding System (FACS) which atomically group facial muscle movements. This system described all the potential of facial movements in terms of facial Action Units (AU), and does not directly reflect the affective state. EMFACS (Emotional Facial Action Coding System) is one example method to transform AUs to affect space [398]. Every visible emotional response can be encoded as an amalgamation of facial Action Unit activations; see Fig. 1.4. This technique tries to either categorise which AU is initiated or to evaluate the level of their intensity.

The Facial Action Coding System (FACS) is the most comprehensive system that precisely describes the basic facial expression movements by encoding the configuration of AU or multiple AUs in terms of facial atomic activation muscle actions. In a muscle-based approach, FACS defines 46 Action Units assumed as the smallest fundamental measurement of visible discernible blocks of facial movements [268] [340] [556]. This system describes each specific AU and also their related four temporal segments by splitting the facial expression event into phases: Neutral, Onset, Apex and Offset. The Action Units can be described by two means: AU occurrence activations' presence or absence if noticed on the face, and the degree of AU intensity which spans from a six point (A-E) ordinal scale. Furthermore, this system supports mapping from facial appearance changes to emotion space. There are no other available systems that have the same descriptive power [384]. Facial changes can be identified by three levels of activities: the lowest level described by a facial component shape which is commonly represented by a key facial features landmarks. The midst level represented by the contraction of a particular set of facial muscles to define facial AUs. Finally, the highest level denoted the six basic prototypical emotional expressions [344]. Fig. 1.5 shows a close-up description of most of the facial AUs.

More automatic, less controlled or very refined facial movements can be detected as micro expressions (MEs). These expressions are rapid yet provide an accurate measure of a person's real and underlying emotions within a small time frame [644]. A micro expression can be characterized as brief, very short in duration length (as for example 1/25 seconds [441]), more spontaneous, sub-class from subtle expressions, rapid and involuntary which exposes repressed effect embedded in facial expression [437] [637] [484]. They can suppress an individual's underlying true hidden feelings, genuine emotions, and thoughts and it can be hard to forge micro expressions [599]. lying is one example of micro expressions where more contradictory attitudes in the verbal and non-verbal signs could be found [382]. Fig. 1.6 shows an example of the difference between micro expression and fake expression.

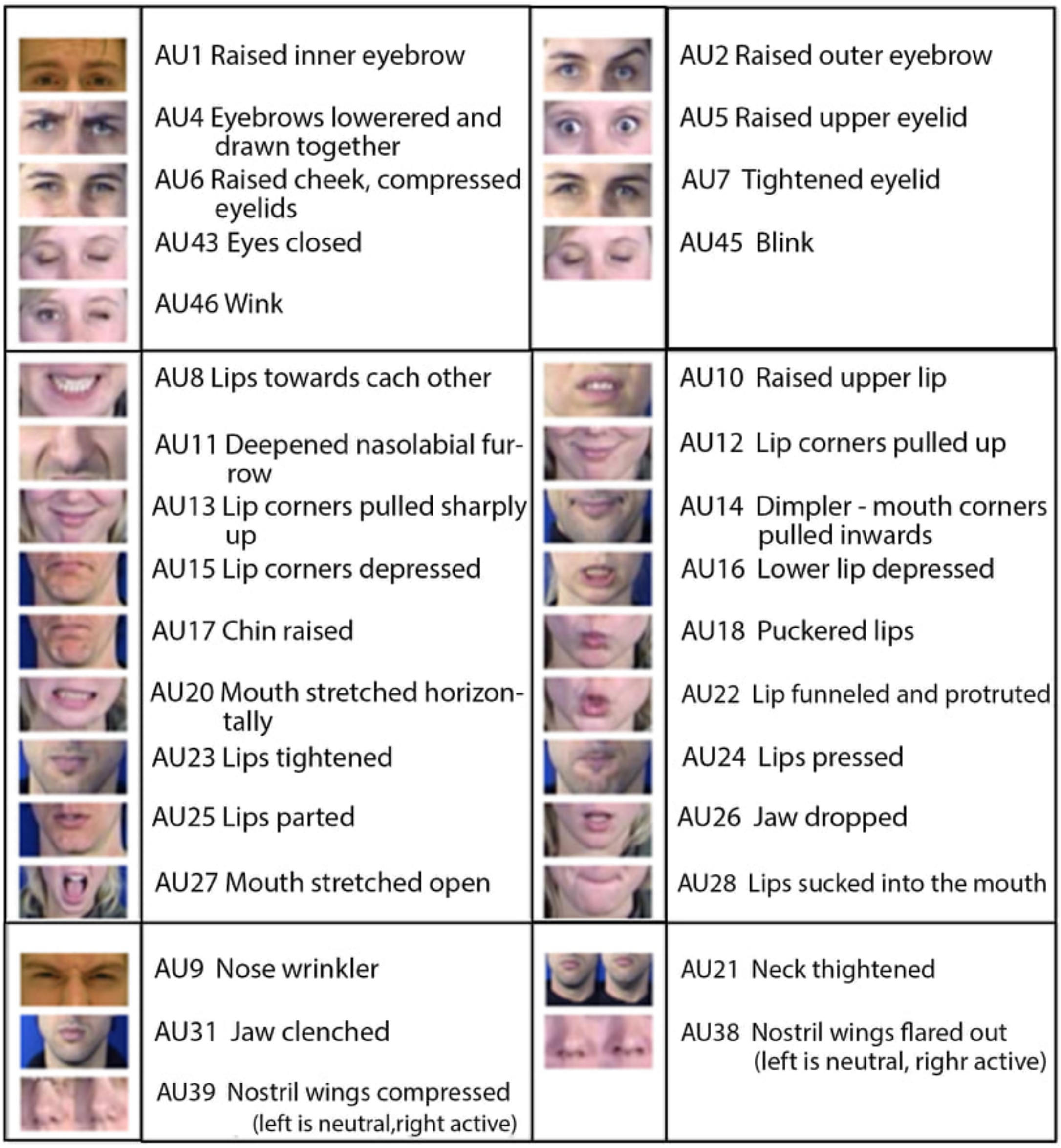

AU1 Raised inner eyebrow
AU4 Eyebrows lowerered and drawn together
AU6 Raised cheek, compressed eyelids
AU43 Eyes closed
AU46 Wink
AU2 Raised outer eyebrow
AU5 Raised upper eyelid
AU7 Tightened eyelid
AU45 Blink
AU8 Lips towards cach other
AU11 Deepened nasolabial furrow
AU13 Lip corners pulled sharply up
AU15 Lip corners depressed
AU17 Chin raised
AU20 Mouth stretched horizontally
AU23 Lips tightened
AU25 Lips parted
AU27 Mouth stretched open
AU10 Raised upper lip
AU12 Lip corners pulled up
AU14 Dimpler - mouth corners pulled inwards
AU16 Lower lip depressed
AU18 Puckered lips
AU22 Lip funneled and protruted
AU24 Lips pressed
AU26 Jaw dropped
AU28 Lips sucked into the mouth
AU9 Nose wrinkler
AU31 Jaw clenched
AU39 Nostril wings compressed (left is neutral,right active)
AU21 Neck thightened
AU38 Nostril wings flared out (left is neutral, righr active)

Fig. 1.5 Most of the available upper and lower facial AUs; image adapted from [555]

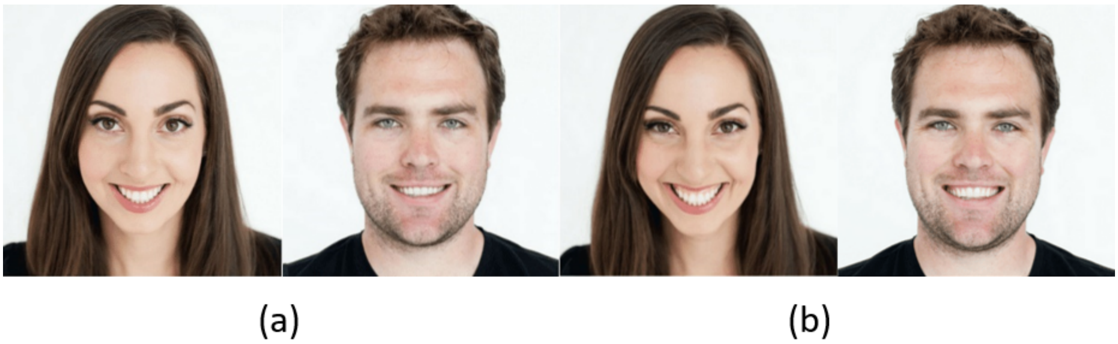


Fig. 1.6 Microexpressions can be used to distinguish between real and fake happiness, this is observed from the crinkle up lines of muscles around eyes and mouth which are more engaged in genuine smile expression: (a) the left two subjects show fake happiness, (b) the right two subjects display real happiness; image adapted from [149].

To the best of our knowledge, there are no present studies to detect spontaneous micro expressions [437]. Recently, the recognition of micro expressions has garnered growing attention from various disciplines because of its potential applications, for example, security systems, clinical diagnosis, and forensic investigation [417]. For recent ME works by [240] [241] have shown that the shape features are more discriminative for recognising certain MEs. Whereas the work by [626] and [160] illustrates that geometric features have been performed weakly than the other appearance and motion features because this is extremely dependent on the facial landmark components precision [417].

Fewer methods follow the dimensional approach, where a value is taken over a scale of continuous emotions. This is in relation to which facial expressions are considered a regression in the Arousal-Valence space. The small variations in the intensity of expressions can be covered by the dimensional model and can distinguish between different subtle facial displays on a continuous domain for each emotion category. Valence and Arousal is a 2D continuous scale where Valence reflects how optimistic, positive, pessimistic, or negative a person is when experiencing an expression event and Arousal represents how animated or composed the person is, and whether the event is confusing, thrilling, calming or soothing. An up-to-date assessment of facial expression analysis can be found in [483] [102] [380]. Nonetheless, there is comparatively fewer work in developing affective computing utilizing the continuous scale of a dimensional model. The reason is that the available annotated dataset which covers the whole dimensional scale of Arousal and Valence is still very limited, and the creation of a large dataset is expensive [398]. Fig. 1.7, represents a facial affect in the 2D domain of Valence and Arousal. It can be observed that there are several types of facial expression which are difficult to map in to the categorical model with ease.

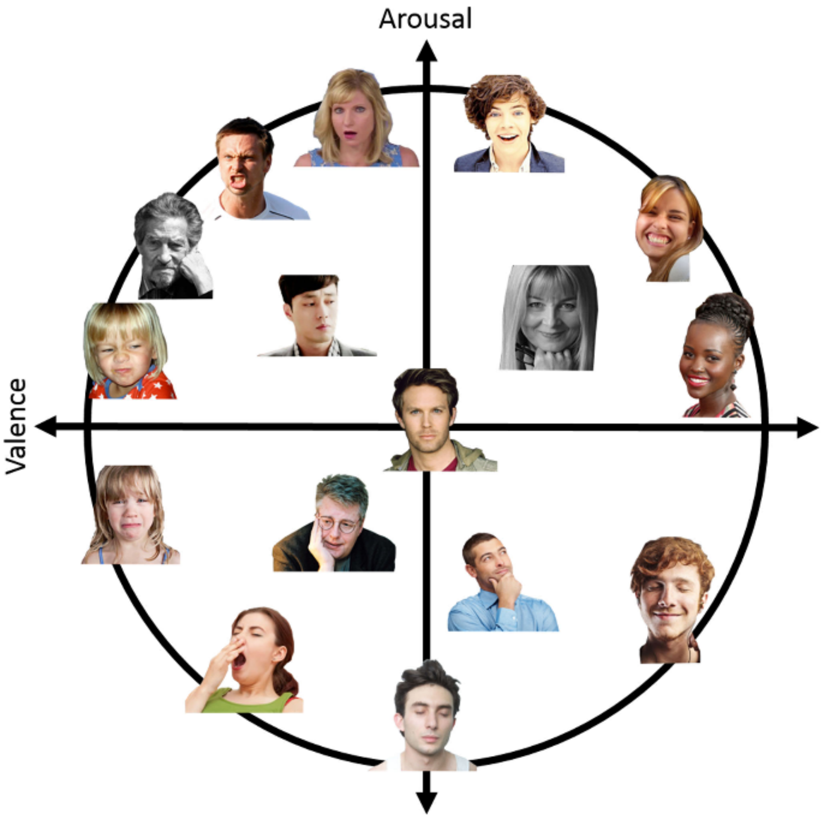


Fig. 1.7 Sample images of facial expressions were labelled to the 2D Valence and Arousal dimensional model; image adapted from [398].

Scientists report that humans can do deferentially a plethora of facial expressions which exist regularly every day. However, mixed emotions in the categorical model could not satisfactorily be transcribed to a restricted group of classes. Some researchers tried to realise multiple distinguished compound emotion categories to beat these limitations [398]. Compound Emotions (CEs) are constructed by uniting two basic emotion categories to construct new ones. The present work of [145] has compiled 21 distinct compound emotion categories. For example, sadly angry, sadly fearful, fearfully surprised, disgustedly surprised, happily surprised, fearfully angry, fearfully surprised, angrily surprised and angrily disgusted which are more distinct compound emotion categories [145] as shown in Fig. 1.8, including distress, contentment, guilt, coyness, sorrow, amusement, hate, boredom, confusion, triumph, desire, sympathy, anxiety, gratitude, repulsion, pride, grief, joyfulness and awe. The study found in the analysis of a FACS that although these 21 classes are different in production, they are strongly consistent with the related categories they represent and show how people move their facial muscles to differently express a range of emotions [144].

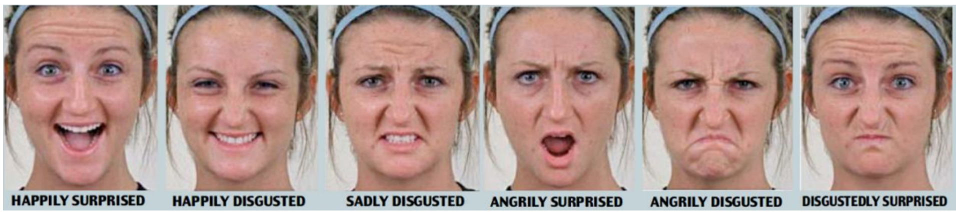


Fig. 1.8 Compound emotions, image adapted from [145].

A very recent work has shown that Deep Convolutional Neural Networks (DCNNs) have been trained straight from the image intensities to estimate the parameters directly of the 3D face Morphable Model (3DMM) and to regress exact and distinctive 3DMM representations, by using the foregoing points of facial landmark detection. These approaches were capable of estimating the shapes of the occluded faces in wild conditions [77]. The 3D Morphable Model (3DMM) is a powerful statistical model to date, of 3D facial shape and texture of the previously acknowledged characteristics of human faces in a space with the explicit correspondences [544]. The seminal work of Blanz and Vetter [51] was presented in the original construction of a traditional 3DMM learned from a group of 3D face scans from a single image or a collection of images [57] [460] [467] [487] and was used for the 2D well-controlled face recognition [50] [544]. 3D Morphable Models have various applications in many fields including computer vision, computer graphics, analysis of human behavior, analysis of medical images, model fitting, creative media, image synthesis, phenotyping of large-scale facial images, craniofacial surgery and biometrics [108] [516] [214] [9] [22]. 3DMM has shown an effective promising approach when applied to face recognition with pose and illumination variation being present, and also with invariant expression [399]. Fig. 1.9 represents 3DMM face reconstructions from a large number of human identities.

1.4 Challenges and Limitations

Recognising emotions and AUs automatically from videos or single images is undoubtedly a complex and challenging task. There are several obstacles associated with facial expression recognition which can be traced to many confounding factors which can significantly affect the system performance and the accuracy of the level of classification [495]. This includes the following: illumination is one of the biggest difficulties for automated facial expression recognition systems. Faces look different when various lighting conditions are used; for example, when lighting is put on one side of the face this side appears very bright whilst the other is very dark [222]. Illumination varies owing to different levels of skin reflection, lustre from eyes, teeth, and camera [222]. Fig. 1.10 illustrates the illumination difficulties, due to

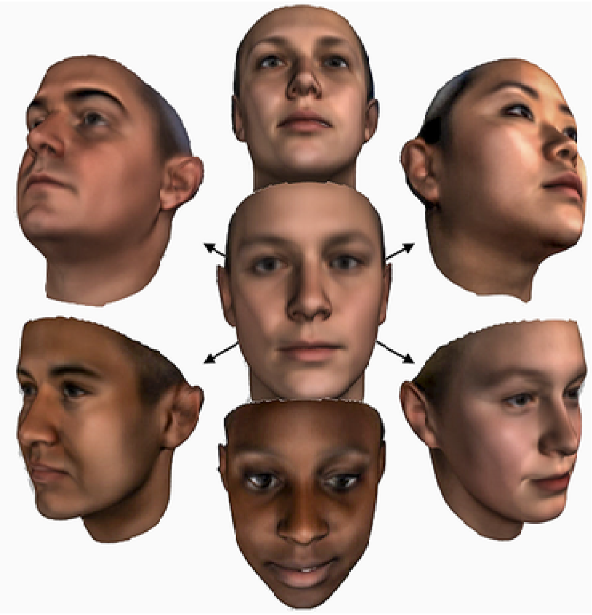


Fig. 1.9 3DMM face reconstruction using a large different human identity; image adapted from [55].

the variation in lighting, in which the same subject was manifested differently. Gabor wavelets can be used to filter the input image reducing the lighting variation [160] [593]. Second, users' non-frontal pose variation (in a plane, or out of plane rotation) and face misalignment in invariant head movement is a significant research problem found in unconstrained face recognition systems because of the 3D dynamic nature of a facial action [222]. In an unconstrained environment (real-life wild setting) this problem happens due to a subject's head pose (looking up, down, left or right) or camera's angle and distance at which a given face is being observed [222]. Most of the surveyed systems, however, are based on frontal view images of faces and many other subjects are well tested using frontal faces [426]. Fig. 1.11 represents the changes in pose views, due to the variation in viewing conditions, in which the same face subject appears differently. Facial occlusion is another research challenge in FER systems. Occlusion is frequently overlooked and supposed to be acquired from a controlled laboratory by true acquisition conditions. Lately, sparse representation was proposed as an efficient appropriate way to prevent occlusion whilst it is independent and uncorrelated to facial identity [242]. Next, the choice of evaluation method and the benchmarking of numerous algorithms is a crucial step in the facial expression recognition system.

A means to determine automatic recognition of emotions in the wild remain a difficult and complex problem [125] [662], as numerous factors come into play, such as providing an adequate simulacrum for real-world imaging conditions, and additional difficulties posed

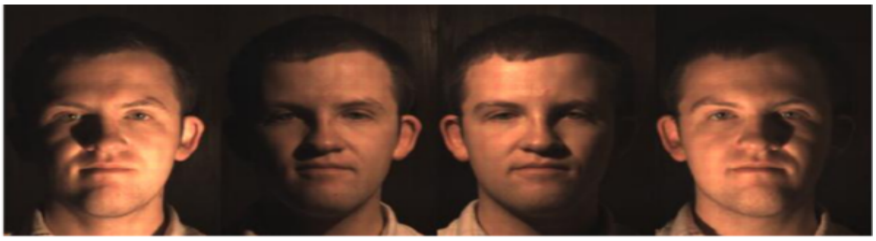


Fig. 1.10 An example of facial image with illumination variations; image adapted from [527].

by the substantial variations in the physiognomy of humans [206] such as anxiety, loss, deep distress, disappointment, misfortune or despondency a feeling suffered by a person's own self or others. Rigorous non-frontal face detection and alignment, precise facial points position and tracking, posed and spontaneous of facial expression displays, location specifics such as dim lighting in a scene, the level of intensity and description of AUs / emotions, subtle AUs recognition, AU correlations, temporal consistency, the availability of reliable dataset and ground truth information, low resolution imaging, extreme profile or out of plane rotations, eliciting conditions, scale and orientation, low intensity of facial expressions, the significant intra-class subject changeability of varying individual characteristics, such as age differences, different identities, face shapes, camera view, recording environment, level of expressiveness, unconstrained facial expressions and postures [630], orientation and the point of view [440], background noise and clutter, and self-occlusion [68] which rigorously damage current strategies [125]. Other contemporary factors are often considered such as registration errors, time delay, age progression, face-size, mood and behaviour, face identity, motion blur, gender, facial hair, permanent furrows, decorations, accessories and skin marks, make-up, glasses, piercings, tattoos, scarring, beards and scars which can either occlude or obscure the face [222] [426] [37] [380] [217] [284]. This includes also various identities across all subjects: babies, children, youngsters, adults, and elders. Subtle or large individual attribute differences between people's faces occur in key facial features such as intensity, appearance, shape, dynamics and conformation of the same facial expression [345]. Fig. 1.12 illustrates the variation and the problem space in recognising facial expression.

For instance, the reconstruction of a 3D face from a single RGB (monocular) facial capture or from commodity RGB-D camera view: when deployed in the wild, their 3D approximations are either unsound and vary between differing photos of the same subject or they become over-normalized and broad [548]. Imbalanced data with a scarce and limited AU image coded data annotation, according to the lack of adequately FACs coded dataset [159], represents a major issue impeding progress in the field. Furthermore, inadequate training data [630] and training classifiers were where relatively few examples are present for each class [505], and the procurement of labelled datasets is costly. Mistakes, or prejudice of human

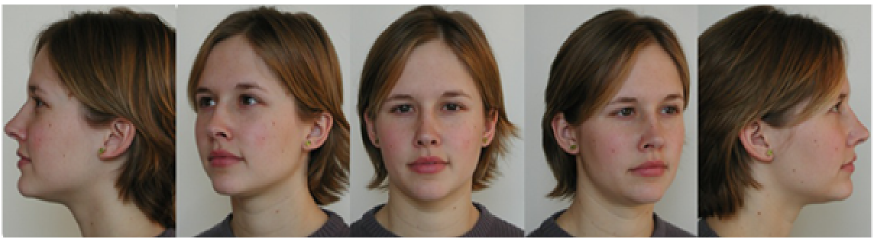


Fig. 1.11 Image source from MIT-CBCL database [583] shows face pose change; image adapted from [330].

annotations occur between different datasets [519]. The partiality of annotations is expected among different datasets because teams from a variety of backgrounds would have different principles in maintaining the standard of the released datasets [519]. Equally, if the size of the dataset is relatively small and current state-of-the-art deep learning algorithms cannot model all the previously stated difficulties, which constrain the ability to train large-scale models accordingly, this is likely to be prone to overfitting [157], therefore any results are likely to be diminished and restricted. To address the issue of training deep learning methods on small datasets, earlier investigations in this field have deployed transfer learning across tasks, where the weights of the CNN are initialised with those from a network trained for related tasks before making minor adjustments using the target dataset [409].

So, the preliminary stage in making an image classification system is assembling sufficient annotated data whereby each image is annotated with the correct category. Moreover, the current ubiquitous GPUs advance the training process of deep Neural Networks to tackle big data problems [157]. Modelling AU's temporal dynamics, which is highly variable, together with the information about location and duration within each facial expression, is difficult to determine in a sequence. Another challenge is that facial AU events can occur in very different time scales [395]. In real time, in most cases, certain positive examples of AUs are minimal, owing to the rarity of becoming activated due to natural facial expression (such as AU9 or AU20). This has to be taken into consideration to avoid "overfitting on the training data" [411].

Finally, as has been known, this would also not overlook the main problems of an FER system which are not yet fully solved: detection of an image segment as a face, extraction of information from the facial region, and classification of facial expressions [426]. Ideally, the typical structure of automatic facial expression recognition processes consists of multiple steps, in three main stages: detection of facial regions / alignment and tracking, facial feature extraction, and facial expression (emotions / AUs) classification [342] [364] [160], as shown in Fig. 1.13

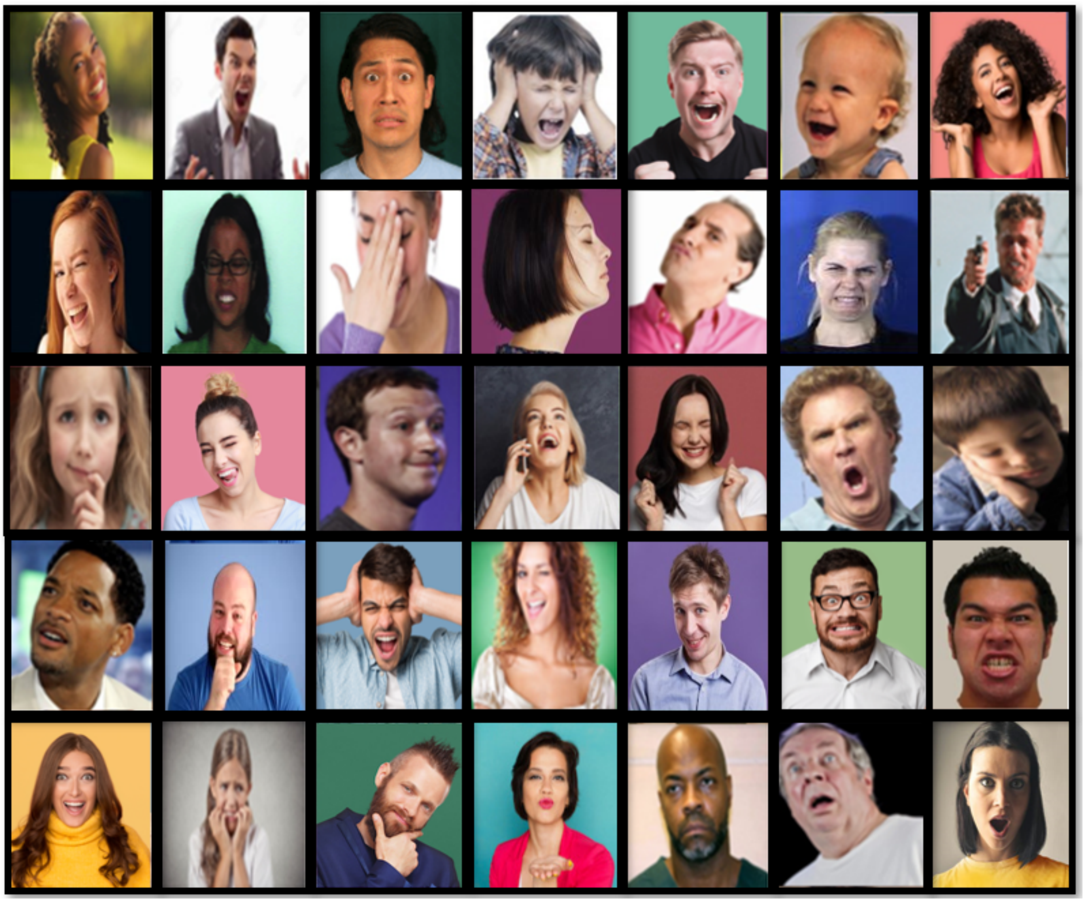


Fig. 1.12 Challenges and difficulties in recognising emotions.

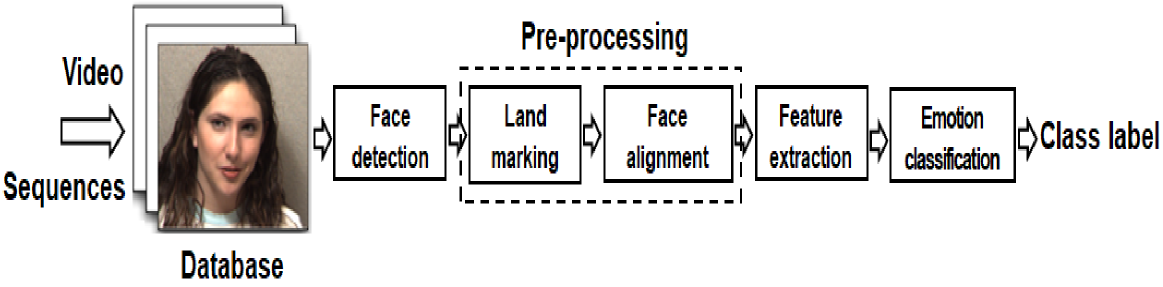


Fig. 1.13 Typical architecture of an FER system.

1.5 Thesis Aim and Objectives

The aim of this study involves the use of Artificial Intelligence tools (AI) and advanced Machine Learning techniques (ML) to design an automated system that is capable of recognising and estimating the emotions of different individual's feeling in real-time from live broadcast footage. This thesis will extend the state-of-the-art knowledge boundaries by looking at how emotional cues can be learnt and recognised by discovering temporal changes in facial appearance. How such patterns are learnt on test subjects can be generalised for applications to new individuals.

The Objectives:
The following objectives have been listed and are undergone entirely throughout the dissertation:

- Collect datasets for face and emotions/AUs recognition analysis and study the information provided within the dataset; evaluate the reliability of these datasets.
- Process images represented by face detection, facial feature extraction, emotion and AUs classification and recognition.
- Review facial emotion classification techniques and methods.
- Classify emotions based on Artificial Intelligence concepts.
- Evaluate emotion recognition algorithms on the collected dataset.
- Discover facial AUs temporal information and recognition to infer the subtle temporal emotional cues.
- Learn and develop how to extract features on 3D.

- Compare 2D & 3D features.
- Evaluate the models using feature extraction based on 3D face reconstruction.

1.6 Thesis Contributions

The thesis presents three main areas of research which are derived from the following list of contributions:

- First contribution: a benchmark of dynamic versus static methods for facial Action Unit detection.

  Chapter 3 elaborates on the recognition of Action Units, predicts the occurrence activation detection (presence or absence) of AUs at frame level basis and enhances the performance of the supervised proposed methods. Static and dynamic features were extracted using hand-crafted, engineered and deep learning-based representation methods. Temporal modelling was determined by knowing the hidden insights and investigating the underlying temporal facial features variations. This was done by hybrid multiple features of supervised and unsupervised methods, non temporal and temporal, temporal and temporal features from a video sequence. The importance of fusion features information in facial expression recognition was evaluated. Assessing and visualising the maximum expression of the desired target AU is also offered in this work.

- Second contribution: the usage of unsupervised Generative Adversarial Networks models as a feature extraction for the supervised tasks for facial expression recognition in the wild.

  The second contribution which is formulated in Chapter 4, in particular, considers the usage of Deep Convolutional Generative Adversarial Networks (DCGAN) for facial features extraction and for classifying the seven emotion classes in the wild together with Action Units. A constructive framework was proposed by using the Discriminator network as a feature extractor based on video frames and static images. More precisely, testifying was done to see whether the features learnt from the Discriminator's convolutional penultimate layer could provide information characterizing emotions and AUs. After that, the ability of DCGAN to generate arbitrary analogous images from a different perspective (predefined in front, multi-view settings and from real-life wild conditions) was discovered. The features trained on a large dataset potentially unlabelled can be experimentally transferred for the supervised task to a different one.

A manual re-annotation to the images of the Radboud dataset (emotions relabelled to AUs) was achieved. Higher quality discriminative representation features were derived from a large number of examples and from frontal face images. A generalisation across dataset evaluation performances was presented, using various pre-trained models which were obtained from the Discriminator of the DCGAN model to cope with the impact of the restricted number of the target dataset.

- Third contribution: a novel application of an existing disentangling 3D expression from identity method for real time facial expression recognition in the wild.

  Chapter 5 investigates the implementation of a novel framework to disentangle the identity from the expression for automatic emotion recognition. Accurate and real time facial expression recognition via Deep Neural Networks (DNNs) based mapping of 2D facial images to 3D morphable model. A large scale dataset was constructed and annotated from facial image videos termed as Face Vid. A deep CNN Network (ResNet) was well trained on the Face Vid dataset which was robust to illumination, occlusion and the changes of angle view. The approach was capable of estimating and regressing the facial expression coefficients independent of identity through 3D Morphable Models (3DMMs) from a single monocular RGB image in the wild. Though the 3D Morphable Model is not novel, its fitting using video sequences (temporal information) is somewhat different from the available literature, thus the approach is fairly novel. An experiment on stress analysis has been also accomplished. A multi Support Vector Machine classifier was added as a back-end stage to integrate the work for robust prediction of the emotion categories. An evaluation was made on the proposed method for both emotion recognition and stress detection. The experimental results on both applications achieved close to or outperformed state-of-the-art results. The proposed model works at 50 fps, thus providing a promising solution for real-time implementation.

1.6.1 Publications

- Journals
  Luma Alharbawee, Nicolas Pugeault, "A benchmark of dynamic versus static methods for facial Action Unit detection", it has been accepted by the IET Journal of Engineering on 05-05-2020.

- Conferences
  Mohammad Koujan, Luma Alharbawee, Giorgos Giannakakis, Nicolas Pugeault,

Anastasios Roussos, "Real-time Facial Expression Recognition "In The Wild" by Disentangling 3D Expression from Identity". It has been accepted at FG 2020 by the 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2020), IEEE, on 30/01/2020 .

1.7 Thesis Outline

This work is arranged in six chapters and the remaining chapters of this thesis are organised in the following order:

Chapter 2 briefly gives a review of state-of-the-art methods in facial expression analysis. It includes the data collection and the reliability of ground truth information. A methodology and overlap evaluation for a face detection algorithm was introduced and illustrations have been considered. Examples of supervised, unsupervised and semi-supervised learning feature extraction methods have been used as a baseline for the performance of emotion recognition methods. Also, this chapter contains the classification techniques that are used as an important part of this thesis and which are tested on different dataset features.

Chapter 3 deals with the modelling of AU target activation detection. Static and dynamic appearance features, which were extracted using both nominal hand-crafted features and deep learning representation, were compared for each frame of the video. This revealed the underlying temporal variation phases in a sequence using supervised and unsupervised methods which highlight and compare the exciting feature extraction representations on both static and dynamic data, and confer the importance of fusing more than one deep architecture. The proposed methods were evaluated by the third aspect: comparing the continuous scoring predictions by acquiring the best match between the predictions and the Ground Truths.

Chapter 4 investigates the idea of using Generative Adversarial Networks (GANs) for facial expression recognition in the wild. Specifically, adapting the ability of the unsupervised DCGAN model for the supervised tasks of classification and recognition of facial expression. This involves four main disparate experiments which were conducted for the aim of generating analogous images, indiscernible from their similar versions, in addition to the learning and extraction of high-level hierarchy representations of facial features. The fundamental aspects of pre-training models and transfer learning context were used. To evaluate the generalisation ability, the trained models were validated on more than five datasets. In addition to that, training the model on a specific amount of facial AU image samples and on one type of attribute is also included. The experimental setup of these methods was also presented;

then, the achieved results were analysed. Finally, conclusions and summaries from the contributions made in this chapter were drawn.

Chapter 5 introduces a novel framework for automatic recognition of emotions from a single monocular facial image, describing a proposed framework of emotion recognition from images. This accentuated the methodology in more detail, for instance producing 3DMM facial reconstruction from videos, plus the initialisation phase for the estimating of camera parameters, and the creation of ground truth annotations from a large-scale videos dataset, and providing the description of the dataset collection and the pre-processing, in addtion to training the model using a deep CNN expression Network (ResNet) which was performed on a large-scale dataset of videos termed as Face Vid which were collected from the internet. The estimation of expression parameters is included. The experimental results on the four selected available benchmarks are reported.

Finally, Chapter 6 provides a discussion about the main findings and contributions of this thesis, concludes the experimentation principle achievements, and identifies an outlook on the possibility of some recommendations and highlights the focus of future research directions.

Chapter 2

Related Work

2.1 Introduction

Automatic facial expression analysis has become an active research area for advance human-machine interaction. Facial expressions reflect not only intention, attention and emotions but other mental difficulties, social human interaction and physiological signals. This chapter briefly documents all the relevant aspects of the existing state-of-the-art literature for building a facial expression recognition system. Furthermore, a necessary critical analysis of the related background is given. It also includes the importance of data collection and the reliability of the Ground Truth information. The methodology of face detection algorithm was illustrated, together with the dataset used and the results. The adapted Viola Jones method to identify and locate the human face was used also for the detection of facial features such as eyes, nose, and mouth, which is an important measure in many subsequent facial analysis tasks carried out within this research. An overlap evaluation was done automatically using a clear and well-defined protocol. The performed evaluation regarding the influence of the overlap process (between Viola Jones method and Ground Truth annotations) on face detection error and accuracy analysed the improvements obtained and the reliability of this detector to address the problems which could be expected when applying this on new unlabelled data. Locating facial landmarks, aligning faces, and feature extraction stages were all presented. Examples of supervised, unsupervised and semi-supervised feature extraction methods have been introduced as a baseline for the performance of emotion recognition methods. Also, this chapter contains the classification techniques that have been used as an important part of this thesis and which were tested on different dataset features. Fig. 2.1 represents a flowchart that summarises the overall process and how the literature reviews are hanging together for our proposed work for facial expression recognition system.

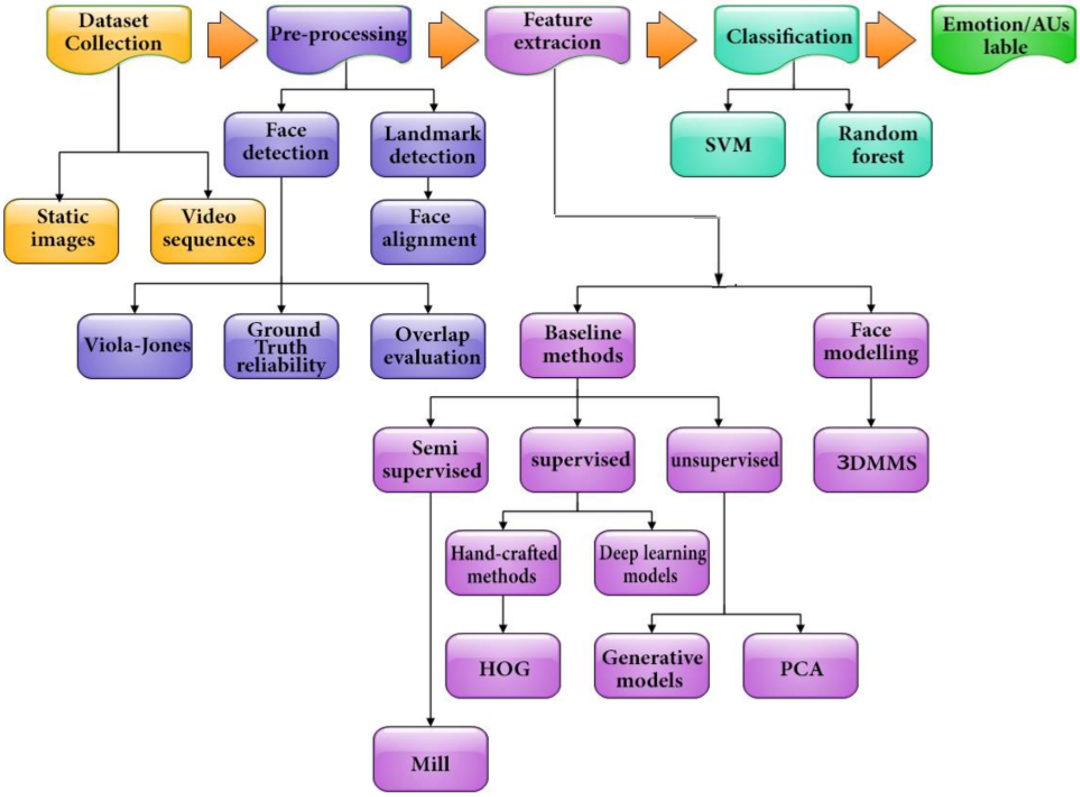


Fig. 2.1 Flow diagram simplifies the literature review and the related work of this study.

2.2 Collecting Datasets for Facial Expression Analysis

Datasets are considered a primary tool for evaluating the efficacy of facial expression recognition methods and play a big role in the progress of the field [490]. Most of the available datasets that have been introduced and used by the researchers are relatively limited, particularly, lacking a high-quality spontaneous facial expression display with a well-annotated large scale data, which represent a big hindrance in the research for training an effective facial expression recognition systems [644]. Therefore, the generalisability capabilities of different approaches to facial expression analysis was up to this time uncertain. In most datasets, the methods evaluation was limited for the classification of general facial expressions (within the six basic emotion categories) which have the ability to achieve a promising result in the controlled lab conditions. However, posed facial expressions do not reflect the complex real-world environmental challenges [492]. In addition to the subjects that have been few in number and homogeneous with respect to the age and ethnic background, recording and gathering conditions, apparent biases resulted from different cultures [336]. In humans, in reality, individuals' abilities from different cultural upbringings may display different or the same facial expressions for the same feelings [65] [41]. For example, the appropriate reaction response of anger expression becomes highly dependent on the cultural exhibit rules and refined with the integration of social interaction and the progression of age [318] [534]. Numerous approaches of facial expression analysis that have been elaborated in this way would inadequately transfer to related applications whereas the expressions, the subjects, the contexts, or the image attributes are more varying. Consequently, the relative strengths and shortages of different approaches are hard to determine and this was taken as evidence in the absence of comparative tests on the prevalent data [537]. As well as the recent development of face recognition algorithms, a large number of relative face datasets have been compiled of adequate size for evaluating the algorithms by controlling the affected variations. Therefore, much of the progress will be made by the availability of several important face databases and their associated evaluation methods [601]. In addition, there is a significant point which is the timing and appearance of the faces' articulations which might differ in spontaneously occurring behaviour. Collecting a dataset that represents all this space in a significative way is a difficult task. Although there are various datasets in current use, many of these datasets are tailored to the specific needs of the algorithm beneath the development, and especially the selection of a suitable dataset depending on the features to be tested for a given task [431]. It was recommendable for the researchers to utilise a standard dataset in order to compare the results while benchmarking a method. Every dataset available is defined with key features structure including the following properties: dataset characteristics, number of subjects, demographics, probe dimension, public accessibility and where it is available, gallery sets, recording conditions, image resolution, the total number of images, different statistical measures, input size, and the ground truth information [3] [200]. Subsequent studies should also be conducted on whether the training dataset size has an impact on the performance of the classifiers. Ultimately, the question is still open on how to the best choice of the favorable positive and negative training samples [190].

2.2.1 Labelled Faces in the Wild Dataset

Labelled Faces in the Wild [238] is a dataset of face images designed for the studying of the problem of uncontrolled face recognition. It was organized by Jain and Learned-Miller [253]. The dataset contained more than 133,233 images from 5,749 subject faces and was gathered from the web and downloaded from Yahoo! News. The Face Detection Database and the Benchmark (FDDB) includes the label annotations of 5,171 faces in a group of 2,845 originating from the Faces in the Wild dataset [238]. Every face in this dataset has been annotated with the name of the person pictured. There are 1,680 pictures of the people that have two or more distinguished photos in the dataset. The dataset was divided into 10-fold

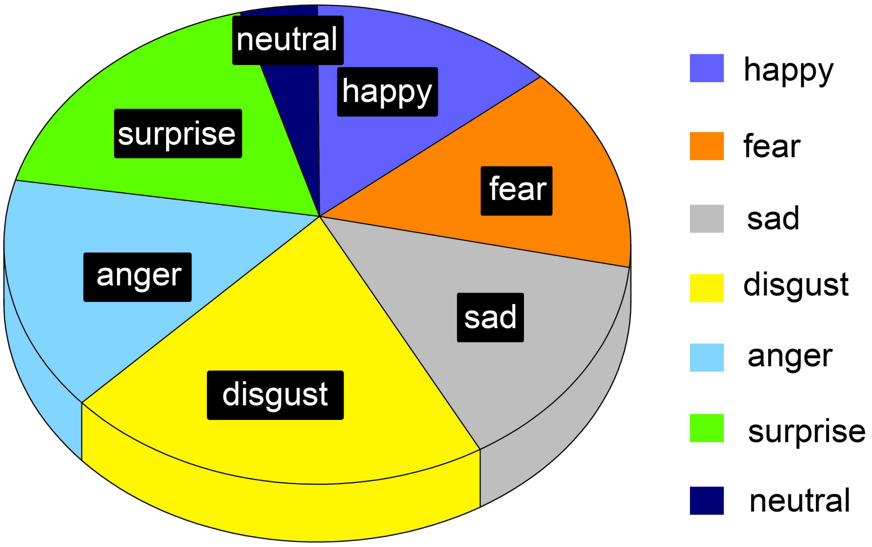


Fig. 2.2 MMI emotion classes distribution.

validation for the performance evaluation and it was provided with binaries and scripts [37]. The manually annotated bounding ellipses are given which serve as the ground truths for the detection experiments. The benchmark constituent of this challenge determines the performance by doing a matching between the outputs of a face detector with the ground truth ellipses for each image [253]. The lone constraint on this dataset is that the faces have already been detected by the Viola-Jones face detector [199].

2.2.2 YouTube Faces Dataset

The dataset consists of 3,425 video sequences from 1,595 different subjects. This dataset of face videos was designed for studying the problem of uncontrolled face recognition in the videos. All the videos were downloaded from YouTube. An average of 2.15 videos was available for each subject. The shortest clip duration was 48 frames, the longest clip was 6,070 frames, and the average length of a video clip is 181.3 frames. In the video collection of this dataset and the benchmarks they have followed the same instance of the Labelled Faces in the Wild dataset. The goal here was to create a large scale collection of videos along with the annotations determining the facial identities of a person who appeared in each video. Moreover, they published the benchmark tests, which were intended for measuring the performance of the video pair matching methods on these videos. Finally, they also provide descriptor encoding for all the faces visualised in these videos [585] [199].

2.2.3 Man Machine Interaction Facial Expression Dataset (MMI)

Man Machine Interaction Facial Expression dataset (MMI) was introduced by Pantic et al. [428]. It has been considered as a constantly growing resource and the most comprehensive free dataset for the detection of single and multiple AUs and the basic emotion categories as well as the facial expression temporal analysis [559]. It was conceived in 2002 as an expedient for constructing and evaluating facial expression recognition methods. The MMI dataset aims to provide a large volume of visible data of facial expressions to the scientific community of facial expression analysis. It consists of both side viewpoints (profile) along with frontal views. The dataset is composed of over 2,900 high-resolution video images from 75 different subjects, involves 30 subjects from both genders of research staff members and students (from 69 different faces), varying from mixed-aged between 19 to 62 years old, with different ethnicity backgrounds, for example, European, Caucasian (66%), African (4%), Carribean, South American and Asian (30%) [613]. Subjects were 48 % female and 11 children from 9-13 years old with 18 adults from 21-45 years old. It is completely FACS-coded for the presence of AUs activation in videos (entire event annotating of facial expression) by two FACS certified coders. A small part of this dataset was annotated for audio-visual laughter's recordings. In the MMI dataset, there are two parts of recordings, posed and spontaneous of facial expressions [491] [294]. In particular, the participant's face recordings start with the neutral face state to onset and then to the apex with one of the six basic emotions, and offset phase, ending in the neutral face phase again with a completed temporal dynamic segment onset-apex-offset pattern in the period separating [220] [397] [280]. A number of omissions were addressed by this dataset that do not exist in the other facial expression dataset. In contrast to the CK+ dataset [366], 68 facial landmark points annotations on the image are not included in the MMI dataset [640]. The facial images in each session were resized into 720 × 576 pixels with 24-bit true color values [640]. Subjects were instructed to display 79 series of facial expressions, six of which are prototypical emotions [220]. Expressions were recorded while the subjects watched TV programs or movies and listened to funny jokes which were told by an occupational comedian [286]. The recordings contain mostly facial expressions of different kinds of laughter, sleepiness, surprise, boredom, and disgust expressions, which were accompanied frequently by large head motions, and were made under variable lighting conditions [301]. It was developed entirely for a web-based interface application to allow integrated easy searching and scanning of the available images [554]. The online search engine will facilitate the research selection

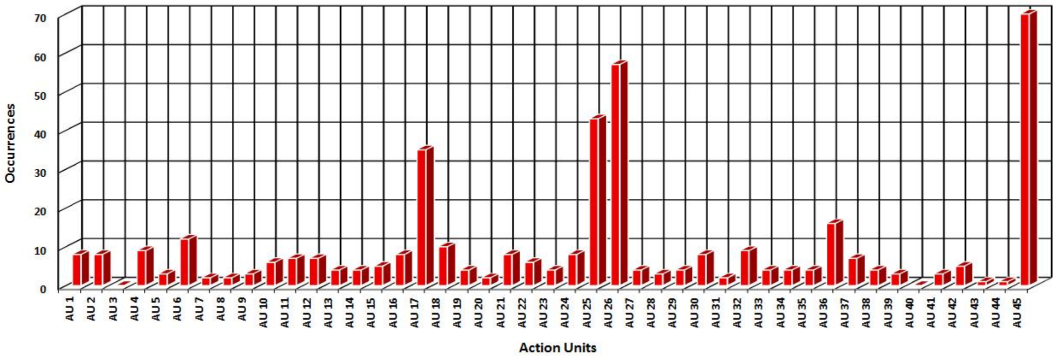


Fig. 2.3 MMI AUs activation occurrences.

of samples by setting different criteria. However, the database lacks potentially important metadata regarding the context in which the recordings were made, for example stimuli, environment, and presence of other people. Recently, recordings of naturalistic expressions have also been added. In this work, 400 video sessions were chosen from the dataset. Fig. 2.2 represents the MMI emotion classes distribution for the chosen 400 videos and Fig. 2.3 illustrates the MMI AUs occurrences also from the selected 400 videos. It can be observed that AU45 (Blink), AU25 (Lips part), AU26 (Jaw Drop), AU17 (Chin Raiser) were the most activated between the other AUs in this collection of videos.

2.3 Face and Facial Features Detection using Viola Jones Detector and Preprocessing

Detecting and locating faces typically serves as the initial basic stage, and the first step across all different facial analysis algorithms. Arguably, a popular strategy for finding a face bounding box uses the classic real time Viola-Jones method proposed by [566]. The difficulties encountered on uncontrolled datasets were elucidated by using this yet strong performing face detector. There are many openly available techniques used for face detection and alignment, and numerous new tools exist in the field, for example, DPM [161], Dlib [289], Seetaface [570], FaceReader [117], Av+EC2015 [463], Dlib3 [157], Emotient1, IntraFace [306], CNN [191], NVSIO3 [92] [224], multi channel features [140], MoPS4 [595] [661] and Intraface5 [595] [519]. One of the major difficulties of automated face detection is that the size and location of a specified face in an image are unknown [260]. Moreover, from a technical standpoint, detecting faces already presents a difficult challenge in computer vision due to the dynamic and challenging nature of a human face which has a high degree of variability in its appearance [230]. Human faces are detected using features derived from motion, contour geometry, colour, and facial analysis [230]. Given an arbitrary image, the aim of face detection is to determine whether or not there are any faces in the image and if present, return the image location of each face [230]. There are certain key issues to be considered if face detection is to be a success: what features to extract, the selection of the extracted features, and which new methods of classification and learning algorithms to apply on new facial data [34] [620]. The four most common approaches utilised to detect and track human faces are: Knowledge-based methods, Appearance-based methods, Feature invariant approaches and Template matching methods [606]. The recent advances of face detection techniques, implemented effectively in real time applications, which consist to delimit the face area with a rectangular [4], starting with the seminal Viola-Jones method suggested in 2001 [565], have made a breakthrough in face detection, and have made the detection process practically robust. Used as the basis of a large body of work on face detection, and the popularity of this algorithm is due to an efficient design that takes advantage of the asymmetries involved in the face detection problem [37]. Image detectors are usually composed of three main steps: rigid templates (the integral image); classifier learning with AdaBoost of boosting or by the application of Deep Convolutional Neural Networks (DCNNs); Deformable Parts-based Models (DPM) (the attentional Haar cascade structure) [620] [566]. Face tracking is another aspect of facial expression recognition which can often be a consequence of face detection. The aim of many systems is not only to detect a face, but to be able to locate it in real time. Tracking means realizing that the face in that frame of sequence is identical to the same face in the last frame of sequence. Tracking of common objects by detection has remained very challenging. Typically, a tracking system consists of three parts: image representation, appearance, and dynamic modelling [31]. Kalman filters or particle filters are statistical methods that are commonly used to track moving objects in images [426]. Feature segmentation is a simultaneous process which allows the isolation of transient and intransient features within faces, or it can be used to separate faces of interest from the background; see for example Fig. 2.4. Segmentation is considered to be easy and simple for many applications. However, further effort is devoted to the segmentation issue with the advancement of face recognition systems under complex backgrounds [364].

It is unfortunate that there still a lack of impartial real world face evaluation benchmarks as a result of the following three reasons: firstly, the present benchmarks of face detection do not support the fine-grained analysis of detection results. Secondly, the current benchmarks of face detection do not reflect the actual real world. Thirdly, recent benchmarks of face detection have not documented true state-of-the-art performances [601]. Benchmarks of the state of the art, carried out during the past several years, such as the FERET protocol [439], FRVT 2000 [438] [202], FRVT 2002 [438], FRVT 2006 [46], XM2VTS [391] and the FAT

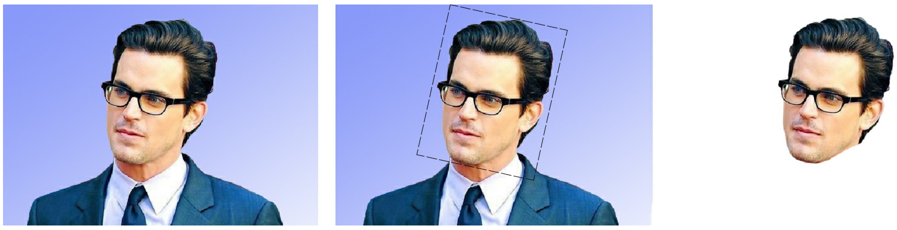


Fig. 2.4 Face detection and segmentation: (a) original image, (b) face detection, (c) face segmentation.

2004 evaluations [390], have shown that age, illumination and pose variations are the three crucial problems plaguing the current face recognition systems [137] [6].

2.3.1 Viola-Jones Detector

The Viola-Jones face detector uses a learning algorithm depending on a cascade of boosted classifier with Haar-like features to decide whether a region of an image is a face [456]. Viola-Jones uses a linear combination of weighted weak classifiers, known as a boosted classifier, which balances performance and high accuracy with high probability. There are three contributions of face detection framework [565] [566]:

- Integral image is a new intermediate representation of an image used to speed up the effective computation of the Haar-like features, as illustrated in Fig. 2.5 which is an example of how to use the integral image and Fig. 2.6 shows Haar-like features of a rectangular type [154], which allow these features to be resized arbitrarily and compared with the region of interest [566].

  The value at each pixel location (x,y) contains the sum of all pixel values within a rectangular region that has one corner at the top left of the image and the other at a location (x,y). To find the average pixel value in this rectangle, divide the value at (x,y) by the rectangle's area. Conveniently, $A+B+C+D$ is the Integral Image's value at location 4, $A+B$ is the value for location 2, $A+C$ is the value of location 3 and A is the value of location 1. So, with an Integral Image, the sum of pixel values for any rectangle in the original image with three integer operations is [422] [48]: $(x4,y4)-(x2,y2)-(x3,y3)+(x1,y1)$ or $D = A+B+C+D-(A+B)-(A+C)+A$. Another example of the integral image is in Fig. 2.7, where the blue matrices represent the original images, while the purple ones represent the images after the integral transformation. If the requirement is to compute the shaded area in the first image, it would have had to sum all the pixels individually [48].

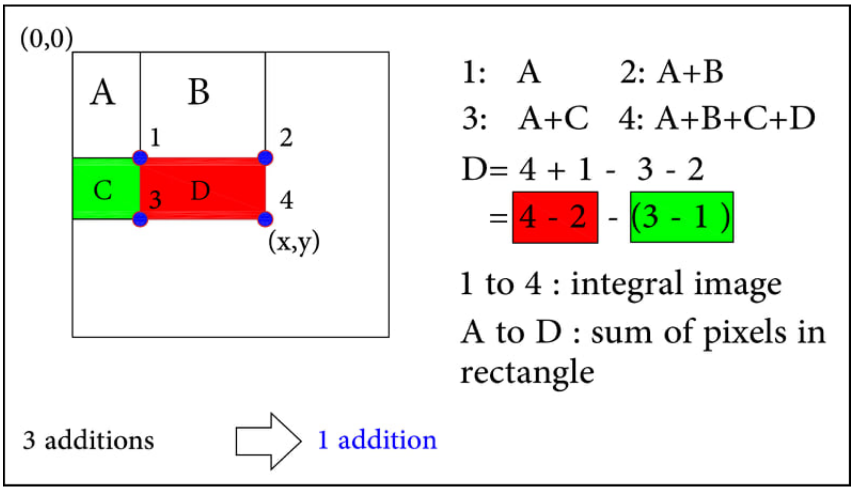


Fig. 2.5 An example of how to use the Integral Image, the sum of pixel values for any rectangle in the original image with three integer operations; image adapted from [422].

- Haar features: the presence of a Haar feature is determined by subtracting the average dark-region pixel value from the average light-region pixel value. If the difference is above a threshold (which is set during the learning), that feature is said to be present. In this regard, by working at this stage of scale and level of detail present in a set of data (region-level of granularity), Haar-like features are powerful in finding faces irrespective of variations in facial expressions [456].

- AdaBoost (Adaptive Boosting) is a machine-learning method for face detection, to train an extremely efficient classifier with a small number of critical visual features from a large set of potential features [565]. Boosted means that the classifiers at every stage of the cascade are complex themselves and they are constructed out of basic classifiers using one of the different boosting techniques (weighted voting) with minimum error. AdaBoost combines many weak classifiers to create one strong classifier [579] [154].

Cascade classifier is a simple decision tree used to combine many features to allow background regions of the image to be quickly discarded while spending more computation on promising object-like regions [4] [154]; at each stage of the cascade the image in the current window is tested against a small number of features: those that fail are rejected quickly and those that pass progress to the next stage and this means it is capable of detecting faces with a high probability [349]. Training a conventional cascade requires finding a small set of weak classifiers that can achieve zero false negative rate (or almost zero) and a low false positive rate [565]. As pointed out by Sung and Poggio, “Both false positive and false negative detection errors can be easily corrected by further training with the wrongly

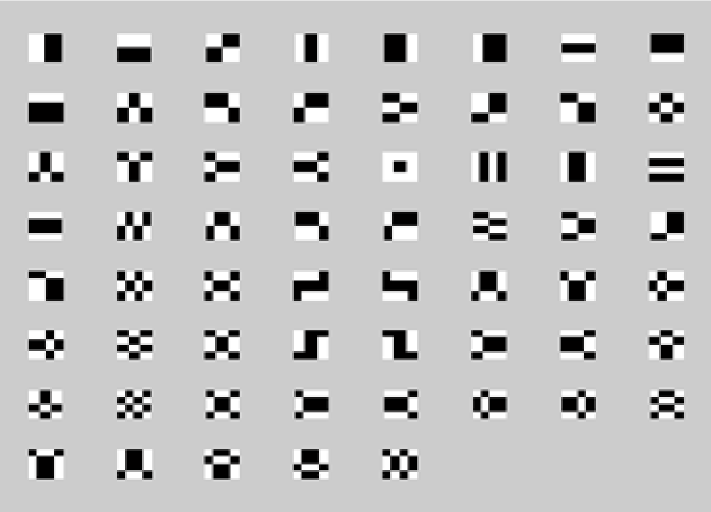


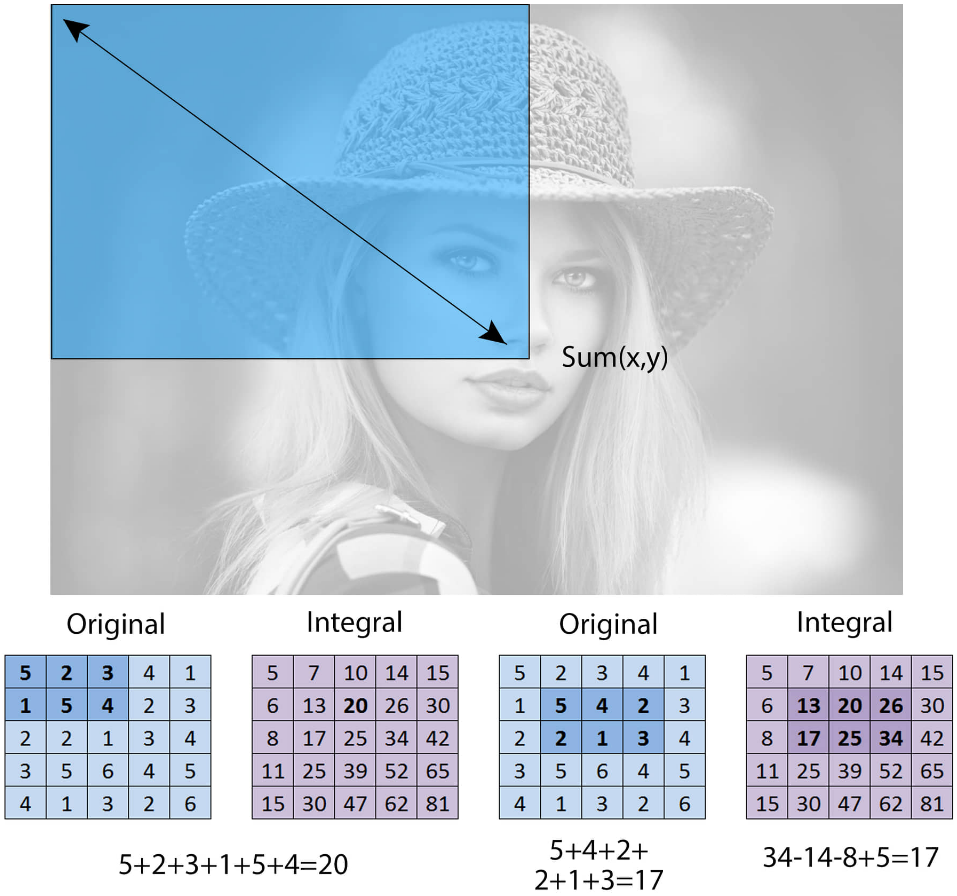
Fig. 2.6 Group of different patterns of Haar-like features; image adapted from [48].

classified patterns" [465]. Fig 2.8 shows a face and facial features (nose, eye, and mouth) localisation and detection using the VJ method.

2.3.2 Reliability of Ground Truth Information

When training and evaluating a system to recognise the facial expression, it is presumed that training and testing data have been accurately labelled. This supposition may be precisely correct or not. To make sure of internal validity, facial expression Ground Truth labels must be manually annotated, and verification is needed of the reliableness of the coding [537]. Ground Truth is a terminology used in various applications to denote the information provided by governing observation against the information provided by the inference. The term "Ground Truth" in Machine Learning refers to the supposed accurate labels of the training set used for supervised learning techniques. Within the past decade, significant effort has resulted in developing methods of facial feature tracking and analysis. The analysis includes both measurement of facial motion and recognition of expression. In the context of computer vision, Ground Truth data includes a set of images, and a set of corresponding labels of the images, defining a model for facial expression recognition, including the count, location, and relationships of key facial features. The collection of labels, such as interest points, corners, feature descriptors, shapes, and histograms, built a model [299]. The annotations have been made either by a certified person or automatically by image analysis, depending on the complexity of the problem.

To describe the problem space for facial expression analysis comprises numerous feature dimensions: level of description, temporal organization (transitions among expression),



Original

5	**2**	**3**	4	1
1	**5**	**4**	2	3
2	2	1	3	4
3	5	6	4	5
4	1	3	2	6

Integral

5	7	10	14	15
6	13	**20**	26	30
8	17	25	34	42
11	25	39	52	65
15	30	47	62	81

5+2+3+1+5+4=20

Original

5	2	3	4	1
1	**5**	**4**	**2**	3
2	**2**	**1**	**3**	4
3	5	6	4	5
4	1	3	2	6

5+4+2+
2+1+3=17

Integral

5	7	10	14	15
6	**13**	**20**	**26**	30
8	**17**	**25**	**34**	42
11	25	39	52	65
15	30	47	62	81

34-14-8+5=17

Fig. 2.7 Integral Image.

eliciting conditions, and reliability of manually coded expression (validity of training and testing data), individual differences in subjects, head orientation and scene complexity, image acquisition, image characteristics and the possibility to have non-facial behaviour. Most of the work to date has been restricted to a relatively limited region of this space [278]. An example of the Ground Truth information is of an image from the Labelled Faces in the Wild dataset, where the detection of every facial image is represented as an elliptical region in the Ground Truth information, and each row of the file contains the parameters of the Ground Truth information of the original image. The Ground Truth is provided in the following format:

2002 / 08 / 11 / big/ img _ 591
< major _ axis _ radius minor _ axis _ radius angle centre_ x centre_ y detection_ score >
1
Where:

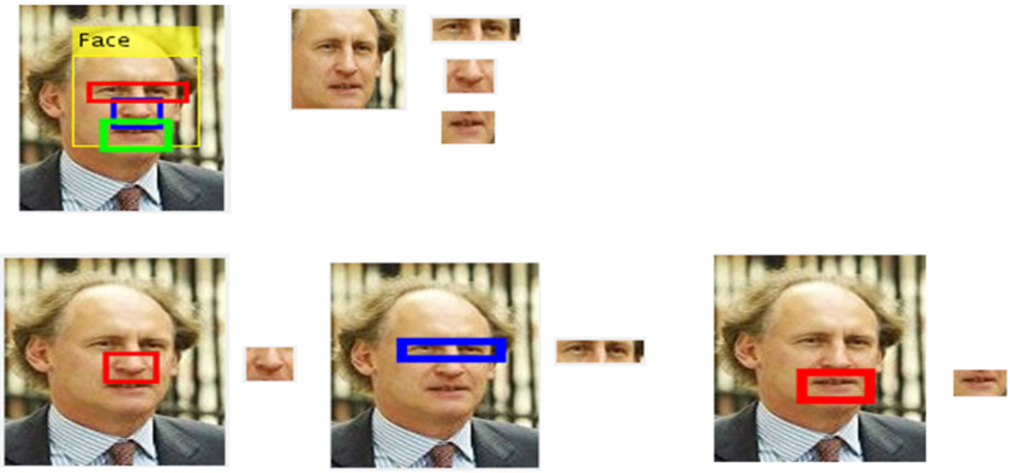


Fig. 2.8 Face and facial features localisation and detection using VJ method; nose detection and cropped, eye detection and one eye pair cropped, mouth detection.

- 2002 / 08 / 11 / big / img _ 591 is the path of the original image.
- 1: No of face annotations of the original image.
- Major _ axis_ radius: 123.583300
- Minor _ axis_ radius: 1.265839
- Angle: 85.549500
- centre _ x: 269.693400
- centre _ y: 161.781200
- detection _ score: 1

In order to find how reliable information available in this dataset the information in the Ground Truth dataset was used, and to specify a more accurate annotation for the image regions corresponding to human faces than is obtained with the specifications of ellipses regions. A rectangular region was defined around the pixels corresponding to these faces using the equations (2.1), (2.2), (2.3), (2.4) below for each face in the dataset. To find the coordinates of the rotated and shifted rectangle for each face region, the drawing of these points on the original image represents the heads of the rectangle: (x1, y1), (x2, y2), (x3, y3), (x4, y4) as shown in Fig. 2.9.

$$X_1 = (a \cdot \cos\theta - b \cdot \sin\theta) + X, \qquad Y_1 = (a \cdot \sin\theta + b \cdot \cos\theta) + Y \tag{2.1}$$

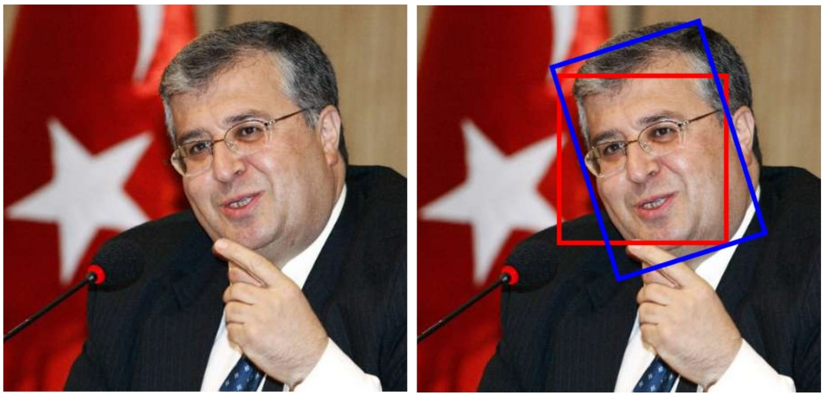


Fig. 2.9 Detection of face using Viola Jones method (red square) and using Ground Truth annotations (blue rectangle), on the image adapted from the Labelled Faces in the Wild dataset [314].

$$X_2 = (-a \cdot \cos\theta - b \cdot \sin\theta) + X, \qquad Y_2 = (-a \cdot \sin\theta + b \cdot \cos\theta) + Y \tag{2.2}$$

$$X_3 = (-a \cdot \cos\theta + b \cdot \sin\theta) + X, \qquad Y_3 = (-a \cdot \sin\theta - b \cdot \cos\theta) + Y \tag{2.3}$$

$$X_4 = (a \cdot \cos\theta + b \cdot \sin\theta) + X, \qquad Y_4 = (a \cdot \sin\theta - b \cdot \cos\theta) + Y \tag{2.4}$$

Where a, b are the major-axis-radius and minor-axis-radius were taken from the ellipse ground truth; θ is the angle of the ellipse; X, Y are the centre _ x, centre _ y of the ellipse respectively.

2.3.3 Overlap and Evaluation

There is still a significant question of how we can construct an equivalence among a group of detections and a group of annotations. This problem is easy for a good result in a given image, it could be more complicated for the increased numbers of false positives or the potentiality of multi overlapping detections, shown in Fig. 2.10 as an example. Blue rectangle signifies the true positives of the face detection output; the ellipses specify the face Ground Truth annotations. It can be noticed that the yellow rectangle shows the desired matching false positives. Also, the second face from the left has two detections overlapping with it, but the third face from the left has no detection overlapping with it, since no detection should be matched with this face [253].

An evaluation criterion was established for the detection algorithm, where some assumptions were first specified about the required output: each detection was corresponded to exactly one entire face, no more detections and no less. In other words, a detection cannot be considered to detect two faces at once, and two detections cannot be used together to detect a single face. A valid matching was required to take only one of these detections as the true

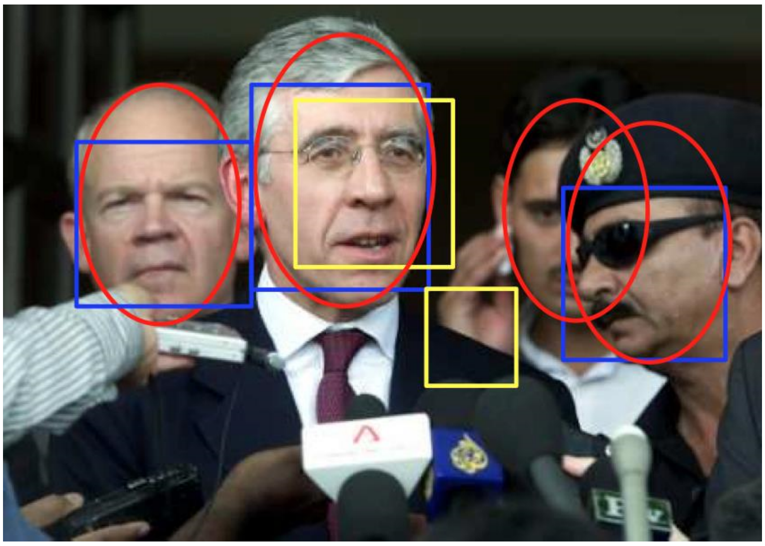


Fig. 2.10 Multiple face detection and overlapping; image adapted from [253].

match. Further debate was if an algorithm detects multiple disjoint parts of a face as separate detections, only one of them should contribute towards a positive detection and the remaining detections should be considered as a false positive. To represent the degree of match (the error in face detection) between a detection using the Viola-Jones method and an annotated region, the overlap between the two regions was calculated. The commonly used ratio of intersected areas to join the area between two bounding boxes was employed. The equation 2.5 illustrates how to find overlap between two regular geometrical shapes.

$$O(d_i, l_j) = (area(d_i \cap l_j))/(area(d_i \cup l_j)) \tag{2.5}$$

It was observed that not all of the annotated rectangular regions of the faces are in a regular form, so the Algorithm 1 below was used to find the overlap between the results of the square from the face detection method and the rotated rectangular resulting from the Ground Truth information. This was compared at every point of the original image to see whether it is located at the perimeter of the annotated rectangle or if it belongs to the Viola Jones's square or does not belong to either one of them, then the results were draw.

Where x_r is the $x-$coordinate of the rectangle; y_r is the $y-$coordinate of the rectangle; θ is an angle of the ellips; w_r and h_r are the width and height of the rectangle; x_p, y_p are the $x-$coordinate and $y-$coordinate, respectively, of the original image's pixel.

The matching score between a detection and a Ground Truth region is given by the ratio of the area of intersection to the combined area of both regions. Table 2.1 elucidates the overlap's results of the original image as show before in Fig 2.9. In order to assess the reliability and accuracy of the available Viola and Jones detector, it was applied on the widely

Algorithm 1 Calculate Overlap

```
Input: $x_r, y_r, \theta, x_p, y_p$
Output: 0 or 1.
initialization:
if $[(x_r + (y_p - y_r) \cdot \sin(\theta)) < x_p$ and $x_p < (x_r + w_r + (y_p - y_r) \cdot \sin(\theta))]$ and $[(yr + (x_p - x_r) \cdot sin(\theta)) < y_p$ and $y_p < (y_r + h_r + (x_p - x_r) \cdot \sin(\theta))]$ then
    $p(x_p, y_p) = 1$
else
    $p(x_p, y_p) = 0$
end if
```

Table 2.1 Overlap's results of the original image.

Rectangle's area	Viola Jones's area	Intersection	Union	Overlap
42292	34225	29598	46919	0.6308

used Faces in the Wild dataset. The code was implemented on a total number of 1,702 images. Fig 2.11 illustrates the final results of this implementation, in which the overlap between the Viola Jones (blue square) and Ground Truth annotation (green rectangle) for this image equals 0.6480. The results are summarized in Table 2.2. The perfect match between the detection and Ground Truth would require an overlap of 100%. As shown in Fig. 2.11c, the number of hits in the histogram is equal to 0 because there was no perfect matching. Fig. 2.11c also illustrates that what is considered to be a good match (True Positives) is equal to the sum of (overlap > 0.5) divided by the total number of images $= 1434/1702 = 0.8425\%$. A False alarm or False Positive is detecting a face when there is none: this is equal to the sum of (overlap > 0 and overlap $<= 0.5$) divided by the total number of images $= 152/1702 = 0.0893\%$. This seems obvious in Fig. 2.11 c, where the number of False Positive iterations is between $(0 - \ 0.5)$. Finally, False or missed Negatives, is not detecting the face that is present and is equal to the sum of (overlap $== 0$) divided by the total number of images $= 116/1702 = 0.0682\%$.

2.3.4 Locating Facial Landmarks and Aligning Faces

Face landmarking can be defined as the detection and localization of a set of certain characteristic salient points on the face [73]. These points are used to represent the essential information required to classify an individual [3]. Usually used landmarks are the corners of the eyes, nose tips, nostril and mouth corners, the end points of the eyebrow arcs, ear lobes, cheeks, chin and the facial component's midpoints [538]. These details are called *fiducial*

Table 2.2 Summary of results of the overlap between (VJ) and (GT) on faces in the wild data sets.

Test images	1702
Number of True positives	0.8425%
Number of False Positives	0.0893%
Number of False Negatives	0.0682%

points or *fiducial landmarks* in the face processing literature [73]. Fig. 2.12 shows a selection of faces with localized facial landmark positions.

Landmark localization represents an important intermediate stage for many subsequent facial expression recognition systems [73]. In general, facial key landmark extraction methods are divided into two types: discriminative methods such as cascaded regression models [282], and generative methods such as part-based and holistic models [52] [8]. Recent comprehensive studies have emerged for facial landmark points detection and tracking [88] [270] [64] [162] [53] [660] [297] [604] .

The purpose of face alignment is to locate semantic facial landmarks automatically and to rectify face images into the same canonical pose (typically, the front view) which is essential for some tasks. For instance, expression recognition [196], face swapping [79], head pose estimation [402], face tracking [285], face animation [74], blink detection [7] [510] and 3D face modelling [174]. Alignment is used to identify regions of similarity that may indicate functional, structural and/or evolutionary relationships between two images. Reliable facial landmarks and their associated detection and tracking algorithms can be widely used for representing the important visual features for face registration and expression recognition [130].

The objective of most of the conventional 2D facial alignment algorithms is to locate a sparse group of facial landmarks. Initial work was used the Active Appearance Model (AAM) [549] [481]. A recent breakthrough in locating the 2D facial landmarks has been made by the deep learning techniques even with the in-the-wild conditions datasets [434] [64]. However, 2D face alignment methods are incapable of addressing the big head poses and partial occlusion cases with environmental illumination because they can regress the visual properties only on the faces [53].

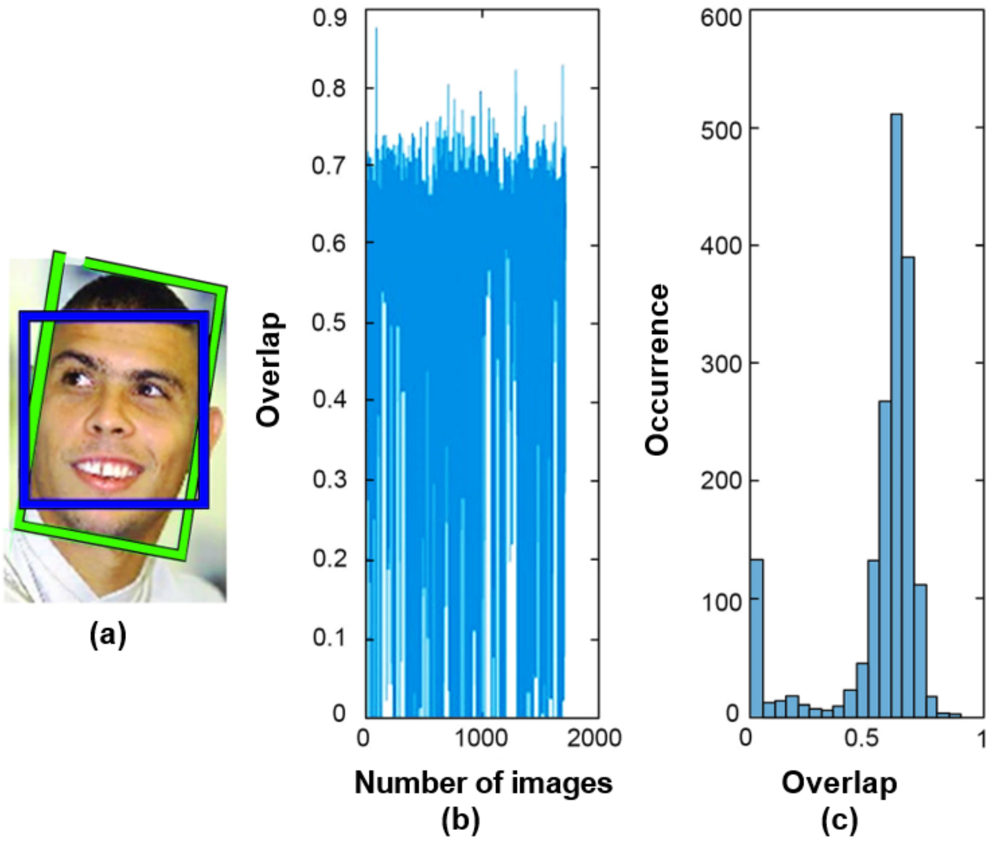


Fig. 2.11 The statistical results of the proposed study on test images representing faces in different imaging conditions. (a) The last sample from the data set. (b) The number of images with the values of the overlap corresponding to each image. (c) The histogram characterises the distribution of numerical data which is an approximation of the probability distribution of a continuous overlap.

2.4 Encoding Faces

Facial feature extraction and representation are an important step to any recognition system. Typically, the automated facial expression recognition system should have two stages: facial feature extraction and classification. After the face is detected, in this step, feature extraction methods are used to attain high level features that are fed into a classification system [426]. Feature extraction methods have the intention of achieving dimensionality reduction for the input data by minimising the variance among class variations from the effect of unwanted conditions, for example motion blur, sensitivity of colour variation in cameras, identity, lighting, alignment errors, illumination, and head pose. Feature extraction techniques can be divided according to whether the representation focuses on motion, texture, deformation, colour, the shape of faces or facial features [47]. In general, there are three types of

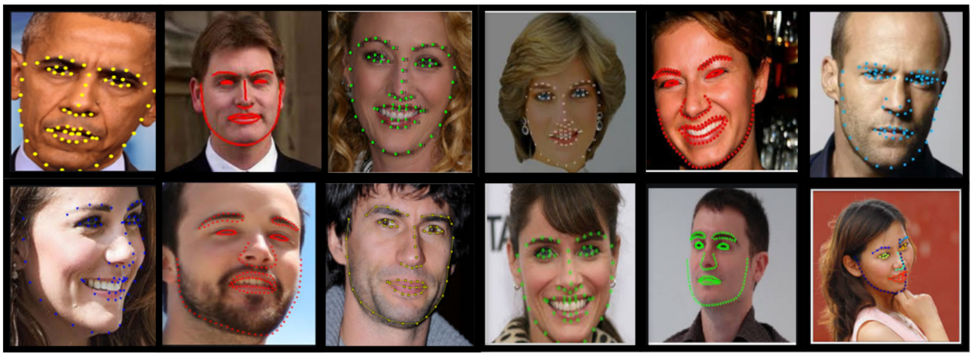


Fig. 2.12 An example of sample images which contain the predicted facial landmark points, where the points with different colours on the face (red, blue, green, yellow, white or a mixture of them) represent the detected facial landmarks depending on the method used, the number and the location of the landmarks.

feature extraction methods: geometric feature based, appearance based approaches, and hybrid methods [380]. The first one identifies the analysed shapes and the position of facial component changes even in small areas, and then finds the relation between them, relying fundamentally on key facial features' parts (mouth, nose, eyes, brows and chin) [47]. However, in the second approach (holistic), images are analysed as a whole and represent the appearance changes (skin texture) of the face or specific locations of the face [160] [364]. While local methods are suitable for subtle changes in small locations, holistic approaches are good at representing common facial expressions. More recently, methods based on the combinations of more than one type of features are more common to complementary information such as the hybrid of geometric and appearance features. Senechal et al. [489] won the FERA2011 challenge for AU detection with blended features. Finally, different Machine Learning techniques are adopted to classify the extracted feature into different categories.

2.4.1 Baseline Examples of Machine Learning Techniques: Supervised, Unsupervised, Semi-supervised

Computer vision (CV) is pertaining to automatic processing, acquiring, analysing, and a relative understanding of high-dimensional visual data from digital images and videos to output numeric information. It considers the advancement of theoretical underpinnings and an algorithmic basis to achieve automatic visual understanding. The ultimate objective of computer vision is to produce models and occasionally override the human vision by using computer software and hardware at different levels. From a scientific discipline, computer vision is regarded as a sub-field that aims to be the prelude for the theory and technology towards Artificial Intelligence systems (AI) for its dominance in a variety of fast-growing applications [164]. Computer vision methods encompass many various types of tasks which include: image processing, image retrieval, feature selection and description, image registration, depth calculation, object recognition, motion analysis, and heuristics [509]. Machine Learning (ML) is a class of Artificial Intelligence in that the system is built to learn from data through instructions and rules which the user provides with minimal intervention. It is via data analysis methods that the computer can learn new ways of developing, identifying patterns, and make decisions by the data provided. Machine Learning concentrates on the development of computer programs that teach themselves to act, grow and change when exposed to a new data problem. The process of Machine Learning involves seeking through data to find patterns as an alternative of extracting data and to modify program results accordingly. It requires a considerable effort from the developer and the programmer as is not an easy process to collect data and make a model automatically learn from them [443]. Algorithms of Machine Learning can often be categorized as being supervised or unsupervised. Supervised methods are usually used dataset sample labels provided by a human annotator. Unsupervised methods endeavor to learn implicitly the underlying structure of unlabelled data. PCA (unsupervised), HOG (supervised), and MIL (semi-supervised) feature extraction methods have been used in this chapter as a baseline for the performance of emotion recognition methods.

2.4.1.1 Principle Component Analysis (PCA) or Eigen Faces

Principle Component Analysis (PCA) is a distance metric learning method, known as a Karhunen-Loeve method, and is a statically extensively used method as a preprocessing step for feature selection and dimension reduction for facial expression recognition systems [403] [614]. It was originally presented by Pearson [433] in 1901, Hotelling [232], and the best recent reference is by Jolliffe [271] [272]. PCA is also known as Eigen faces [496]. PCA method was calculated using the Singular Value Decomposition (SVD). The idea behind the PCA method is to reduce the low-variance dimensionality of the dataset to a small number of principal components (identifying directions) while keeping a significant variability along with as much as possible of the related information [272]. It achieves a minimum reconstruction error. The key advantages of PCA algorithms are the minimising of their noise sensitivity, reducing the training time of the large computations, and the need for capacity and memory cost, capitalizing on thorough data structural ways of efficient visualisations. It highlights the similarities and dissimilarities in a multidimensional data set. It summarizes the underlying variance-covariance structure of a large set of variables. It manages the entire data without taking into consideration the fundamental class structure.

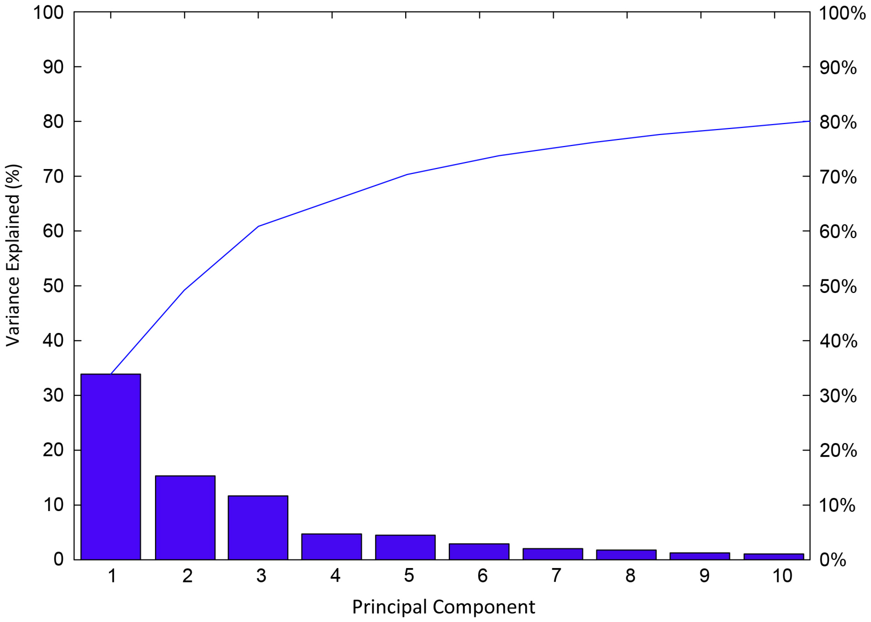
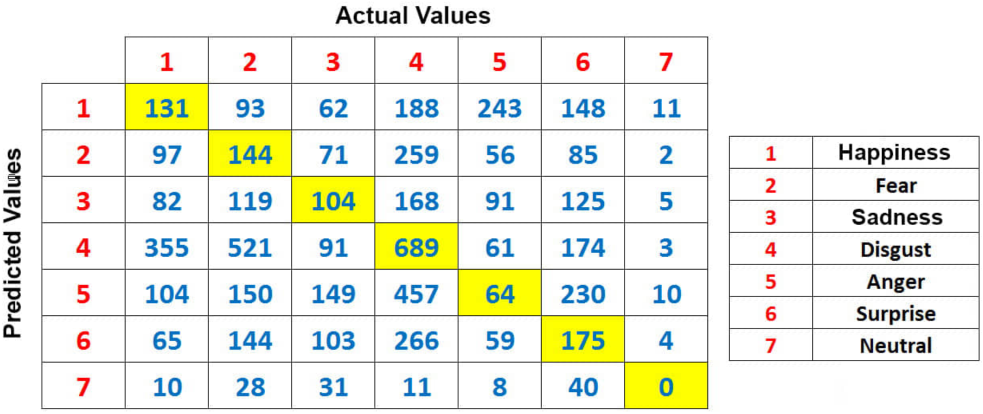


Fig. 2.13 A Pareto analysis chart of the Principal Component (PC) variances which show the percentage of variability explained by each principal component on the MMI dataset [428]. It consists of both a line graph and bars, the line explains the cumulative variance explained by the kept PCs, whereas the bars represent the variance explained by each principal component resulted from the PCA method.

A PCA algorithm is usually used for real-time applications. Nevertheless, it ignores the manifold structure of local data and does not scale to large databases. It is recommended not to use PCA if the number of dimensions (features) is bigger than the number of observations because of the curse of dimensionality as a result of overfitting, where data tends to move away from the centre in high dimension data. In Machine Learning terminologies, a familiar concept of the curse of dimensionality is well-known when the performance is degraded in which there are features which are not relevant for the predicting of the required output.

Fig. 2.13 clearly shows a screen plot of the first ten principle components (PCs) which retained collectively 80 % of the total proportion information variances contained in the given dataset (MMI). The only clear break in the amount of variance accounted for each component is still between the first, second and third components. However, the first component by itself explains more than 34 % of the variance, so more components might be needed. It can be observed that the first three principal components explain approximately two-thirds of the total variability in the standardized data so it might be a plausible mean to reduce the dimensions. After that, it was starting to produce diminishing returns for not continuing to have additional eigenvalues. Fig. 2.14 illustrates the confusion matrix which came out from



Actual Values

Predicted Values

	1	2	3	4	5	6	7
1	131	93	62	188	243	148	11
2	97	144	71	259	56	85	2
3	82	119	104	168	91	125	5
4	355	521	91	689	61	174	3
5	104	150	149	457	64	230	10
6	65	144	103	266	59	175	4
7	10	28	31	11	8	40	0

1	Happiness
2	Fear
3	Sadness
4	Disgust
5	Anger
6	Surprise
7	Neutral

Fig. 2.14 A confusion matrix was obtained from the SVM classifier after extracting the features using PCA approach on the MMI dataset.

the SVM classifier after extracting and reducing the features using the PCA method on the MMI dataset.

2.4.1.2 Histogram of Oriented Gradients (HOG) Based Features Descriptor

The Histogram of Oriented Gradients (HOG) is a local dense method which was first proposed by Dalal and Triggs [109] for the human description and recognition in videos as an object detection in a 2D image [113]. It represents the image's information by the orientations of the edges they have within. The paramount idea behind the HOG descriptor is to extract the local object shape and appearance features by the distribution of the intensity gradients across the image with the edge directions, weighted by its gradient magnitude, angle, and without a need for details about their locations [113].

A histogram was used for encoding the gradient information of the specified patch from the image; in particular, every image patch was partitioned into small spatial block regions called cells; within each cell a histogram of gradient directions was calculated for every pixel in the cell [380]. All the cells were gathered together to construct a big region of overlapping blocks, and the local histogram of a block was used to normalize the variance in all the cells in that block individually to 1, eliminating the influence of the local contrast variations of united illumination [380]. Consequently, these features became robust to illumination variations and misalignment. These normalized blocks are termed as the HOG descriptors [375] [384]. The output of the HOG descriptor was allocated every gradient in a block or patch as a histogram bin with a contribution proportional to its magnitude.

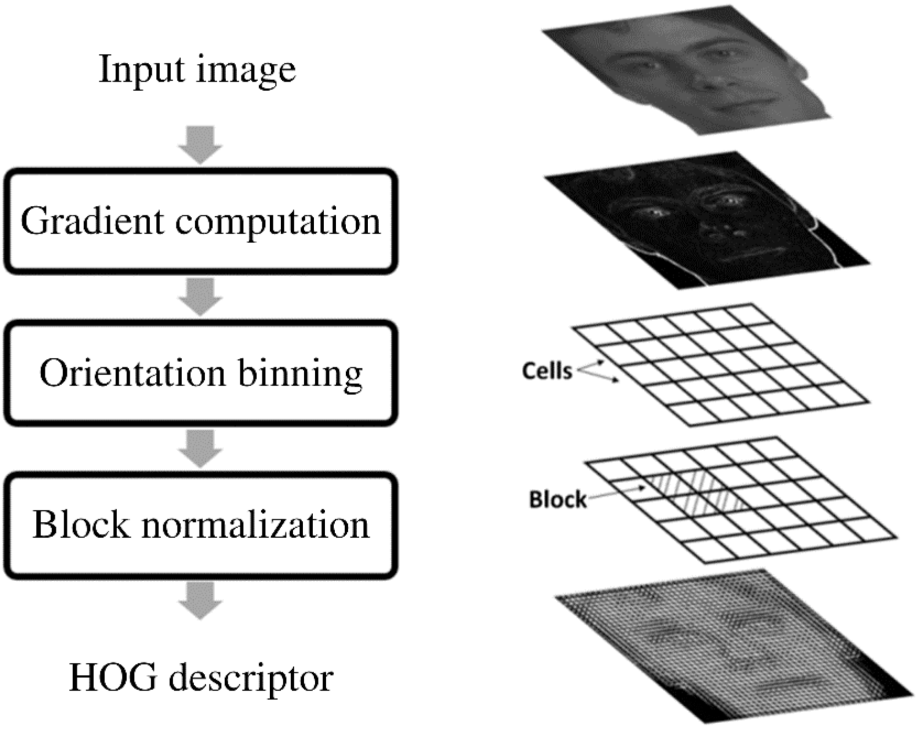


Fig. 2.15 The steps required to extract features using HOG descriptor for face and facial expression recognition; image sourced from [494].

The HOG descriptor representing the whole image face resulted in a matrix of feature vectors of length 3,894 dimension describing the face, and 265 × 265 pixel face images were used in this study. 8 × 8 pixel cells were defined with 8 orientations and blocks of 8 × 8 cells for the experiment in this work which show that it had the best result between the other choices for a given problem, and the method proposed by [109] was used. This required the representation of the features to be extracted topically to prevent the larger gradients to predominate the representation [380].

A well-known system by the FERA facial emotions challenge [105] has already been used by the HOG descriptor [482]. Therefore, it has become more suitable to represent various facial expressions [384]. Moreover, it has been used successfully for pedestrian detection [540], person detection, face recognition and object class classification. In general, generating a HOG descriptor consists of the following steps: Gradient calculation, Orientation binning, and Block normalization [493]. The whole process of the formulation the HOG descriptor is shown in Fig. 2.15. Also, Fig. 2.16 illustrates the visualisation of applying the HOG descriptor on some selected image samples from the MMI dataset. The HOG descriptor can be described by two primary parameters, the number of orientation bins and the cell size, in which the dimension of the block patch is involved in the individual histogram

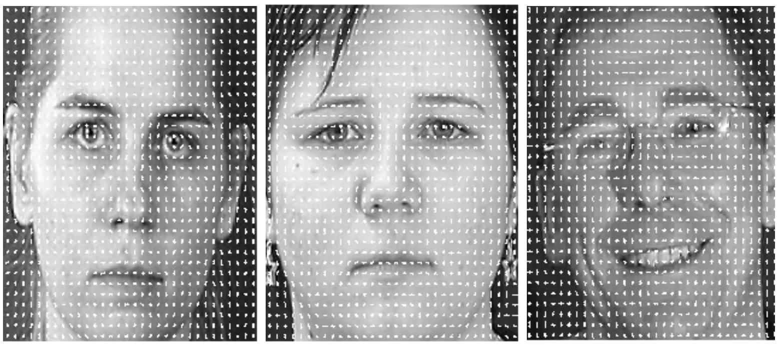


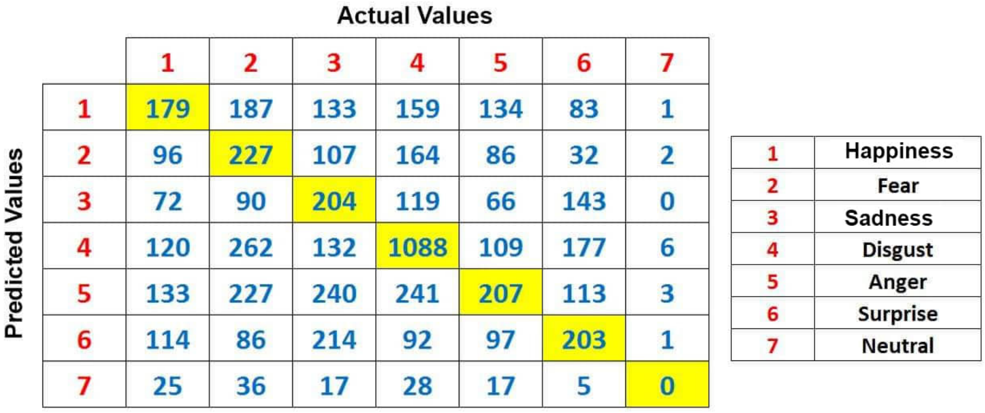
Fig. 2.16 The visualisation of the HOG descriptor on the selected images from the MMI dataset [428], and a local one dimension gradient histogram are calculated over all image cells with 8 × 8 pixels which are described as the blocks of the cells.

.

calculation and represented by the cell size. Furthermore, the quantization degrees of the gradient information can be denoted by the number of orientation bins. Using a small number of orientations may cause the lose of some of the information and consequently, the performance of the FER system would decrease. On the other hand, using a larger number of quantization levels may distribute the information over the bins, and this would also lower the performance of the FER system. Additionally, choosing a high number of cell sizes could lead to the loss of details of appearance information of a facial image because it is squeezed to a narrow or restricted space of an individual cell histogram. Also, useful and useless details are extracted by the use of a small number of cell sizes since the analysis of high resolution that has been accomplished [72]. Therefore, from Fig. 2.17 it can be deduced that the discriminative representation was given by applying this approach, and the choice of the parameters which are used in many different settings. A further study is required for the convenient use of the HOG descriptor for automatic facial expression recognition which is also needed for the achievement of a particular task and before any conclusion is drawn [380]. Fig. 2.18 shown 3D plot of two confusion matrices for the facial emotion category evaluation of the test set by the PCA and HOG methods.

2.4.1.3 Multiple Instance Learning (MIL)

MIL was proposed by Dietterich et al. [132] for solving drug activity prediction as an approach of semi-supervised learning (bag-level label models) where there is an uncertainty of complete knowledge on the labels of the training data examples (weakly labelled) [501]. The instances in the MIL are gathered into a group of bags (called as independent feature vectors in the MIL terminology) [24]. Furthermore, the labels of the bags are known and the labels of the instances which construct the bag are unknown. In the supervised learning



Actual Values

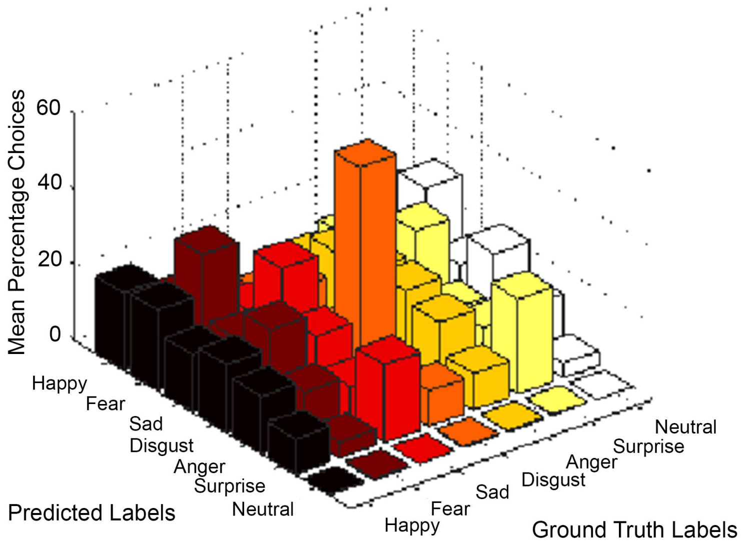
Predicted Values	1	2	3	4	5	6	7
1	179	187	133	159	134	83	1
2	96	227	107	164	86	32	2
3	72	90	204	119	66	143	0
4	120	262	132	1088	109	177	6
5	133	227	240	241	207	113	3
6	114	86	214	92	97	203	1
7	25	36	17	28	17	5	0

1	Happiness
2	Fear
3	Sadness
4	Disgust
5	Anger
6	Surprise
7	Neutral

Fig. 2.17 A confusion matrix was obtained from the SVM classifier after extracting the features from the HOG descriptor on the MMI dataset [428].

approaches, each of the training data instances are allocated a real value or discrete labels. However, most of these approaches entail a manual labelling of facial expression which is very time consuming [501]. In the typical MIL settings, the labels are assigned solely to the bags of instances, and there are no labels for the separate instances. For example, the whole video is associated with the label only with a bag that contains the instances (frames of the video). The purpose of MIL is to classify the invisible bags or instances and to predict whether the bag has the action of facial expression or not [13]. The MIL classification function has learned to predict the labels of bags in the testing data [603]. However, there are several important broad challenges available when learning from bags that are exclusive to the MIL and affect the learning by: the prediction level (bags or instance level), the structure of the composition bags (i.e. the attribution of instances for each class), the types and shape of data distribution (positive and negative), the instance labels' ambiguity such as not belonging to clearly defined categories as well as the label noise [71], the task to be carried out, and the ambiguity of the positive instances that are not given directly [264]. A positive bag in MIL learning is deemed positive if there is at least one instance in the bag that is positive, and a negative bag is labelled as negative if all the instances in the bag are negative [528], as illustrated in Fig. 2.19. However, this approach disregards the sequential type of data, and it remains for some applications to show a promising solution [528].

This approach has been further used in various useful application tasks including the famous molecule drug activity prediction [648], object tracking [13], object and event detection [167], text categorization model [633], classification of image or text [447], video classification [588] [321], document classification [598], modeling of protein family [457], image retrieval [603], speaker identification [24], retrieval of multimedia information [376]



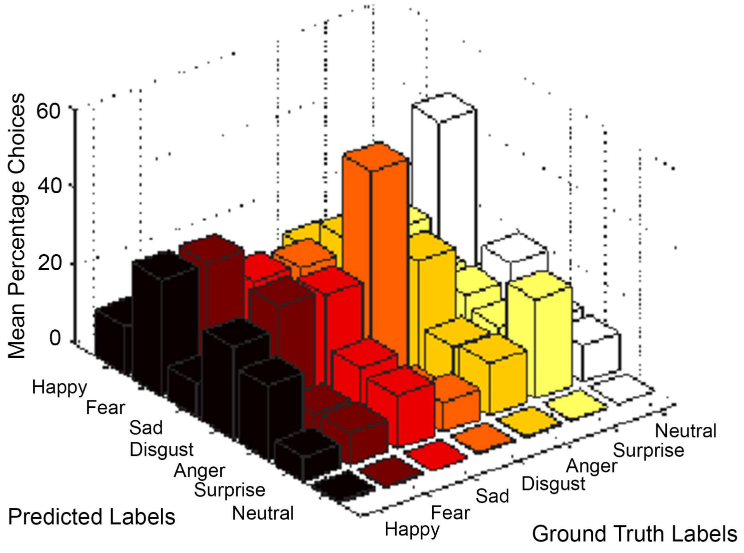


Fig. 2.18 A graph visualising a confusion matrix 3D bar performance for the seven emotions category validity obtained by top: PCA and bottom: HOG methods, respectively. The two x, y axes on the base of the graph can be represented by the actual target class and the output predicted class labels by the system that are stated from Happy to Neutral. The vertices are a mean of proportion scores for the required emotions as per rater selected emotion labels. The height bars show the percentages of wrong predictions and the number of successful and not successful recognition for the different classes. The better result is the lowest bars outer the main diagonal.

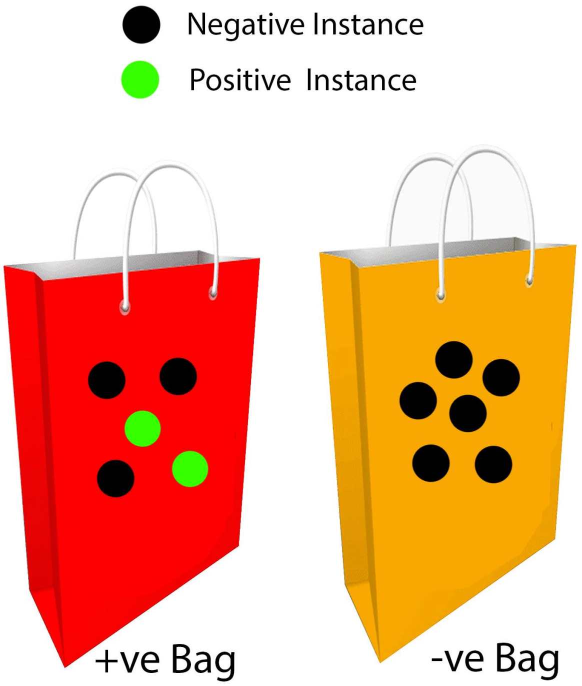


Fig. 2.19 Positive and negative bags in MIL; image adapted from [501].

[142], prediction of stock market [642], face recognition from videos [629], handling noise labelling of training data in video classification [321], detection of pain in videos [501], and learning of image concept [501]. These applications have been boosted by the availability of an open toolkit which is composed of many widely-employed MIL algorithms, and the possibility of comparative analysis on the performance of these methods on various popular applications [603].

With regards to the MIL case problems, it was noted experimentally that the classification performance at bags level [24] is comparatively worse or weak depending on the training data in comparison with the performance of the supervised methods which produce the better competitive results, though some methods systematically predominate over the others. Moreover, the performance of these various methods depended on the application [71]. It is unfortunate to say that there are few comprehensive studies for comparing the performance of a wide range of MIL family algorithms [30]. The weak performance was also shown in Fig. 2.20 which was obtained from applying MIL packages [603] [602] on the MMI dataset [428] in this work, where each facial expression sequence is represented by a bag.

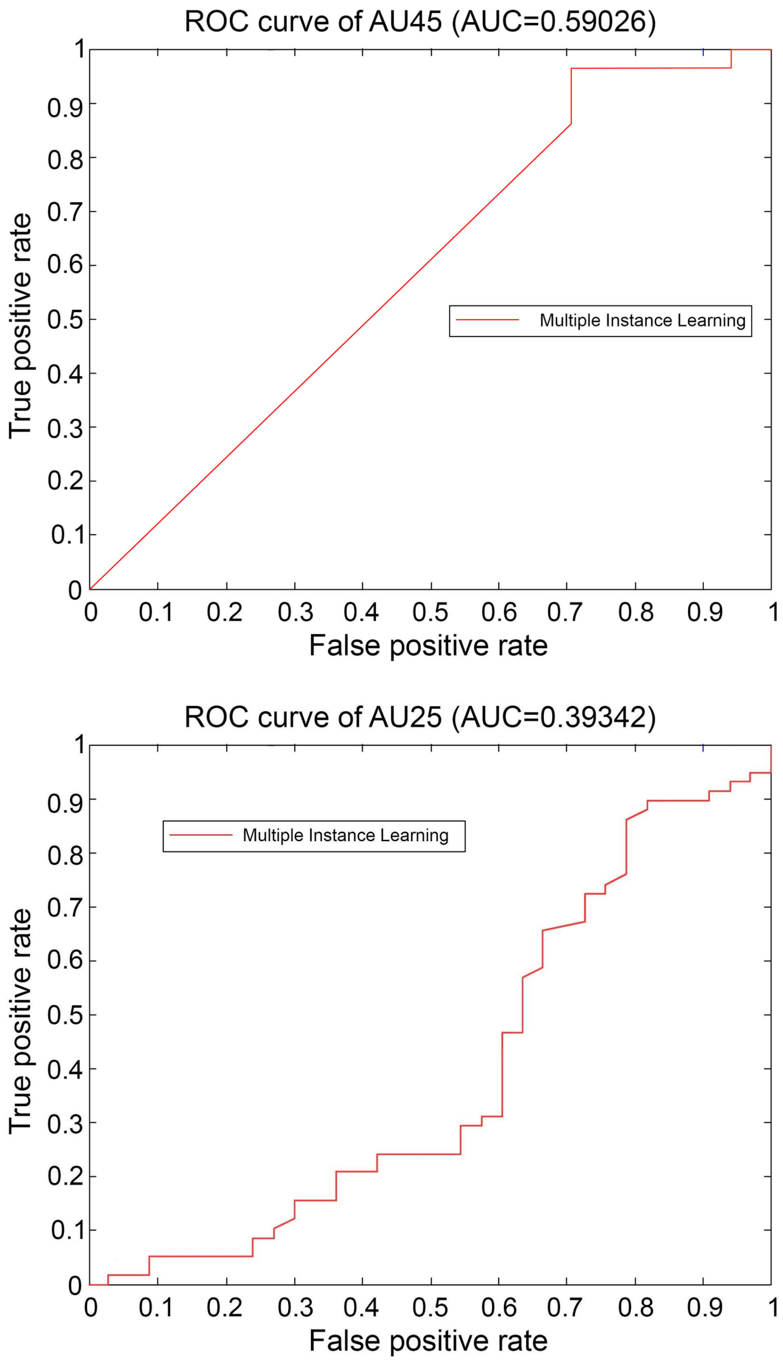


Fig. 2.20 Multiple Instance learning results using ROC curves on Top: AU45 and bottom: AU25 using the EM-DD algorithm [639] from the open-source MILL (Multiple Instance Learning Library) toolkit for MIL algorithms which were written in Matlab [603] [602].

2.4.2 Hand-crafted Features

Current facial expression recognition methods divide approximately into two groupings: traditional hand-crafted methods and deep learning representation models. Hand-crafted monomial features obtained from images and video for encoding emotions are available in an extensive array of literature. In the past, to date, many approaches adopted various conventional hand crafted feature representations for facial expression recognition, that can be broadly divided into appearance, geometric, dynamic, and fusion. Such cases include when facial landmarks' distance and angle relation are used, facial feature point tracking and the ability to extricate patches from specific key facial point locations [610]. Pantic [425] tracked a set of distinct facial qualities around eyebrows, eyes, nose, mouth and chin as the distinguishing markers to capture the geometric information of these facial features. In [78], 58 fiducial points were used to circumvent incorrect matching caused by non-linear image variations. Chang [78] trained an Active Shape Model (ASM) for feature representation [641] [10] [572]. Variations in either the whole face or specific face regions can be described by one of the first applications using Gabor wavelet coefficients analysis. Nevertheless, owing to the high calculation and memory complexity in extracting Gabor wavelet features [127] [212] [369], histograms of local binary pattern operators (LBP) were introduced as an effective appearance function for facial image analysis, as well as robust texture descriptor in many other applications [295] [319]. Local Binary Patterns (LBP) [492] and the family of descriptors of engineered representations: Local Binary Pattern histograms from Three Orthogonal Planes (LBP-TOP) [645], Local Gabor Binary Patterns from Three Orthogonal Planes (LGBPTOP) [16], Histograms of Local Phase Quantization (LPQ) [420], and their spatial / temporal extensions merits: Local Phase Quantization from Three Orthogonal Planes (LPQTOP)[102]. Two large benefits of LBP features are computational simplicity and tolerance against illumination variations across images. The outcomes of utilising LBP features for facial expression recognition were expansively studied by the authors [492] [572]. In their work, they attained better results in experiments compared to Gabor features [636]. LBP features and Discrete Wavelet Transform (DWT) have been used by Akram et al. [20] for feature extraction and classification using SVM which achieved from 72% to 100% accuracy. The authors of [645], suggested a temporal extension of the canonical LBP feature and called it LBP-TOP [16]. Spatial-temporal changes in texture are captured by LBP-TOP [127]. The dynamic appearance information between consecutive video frames can be represented in this application. Principle Component Analysis [29], Discrete Cosine transform (DCT) [396], Gabor motion energy [589], Non-negative matrix factorization [650] [333], Motion History Images (MHI) [397], Histogram of optical Flow [395] [484], Speed up Robust Features (SURF) [81], DAISY/Scale Invariant Feature Transform (SIFT) descriptors [590][386], Dense SIFT [100] [43], 3D SIFT [485], and D-SURF descriptors features were applied alongside a hierarchical classifier fusion method. The well-known appearance feature, a histogram of oriented gradients (HOG), was firstly proposed in [109] to tackle pedestrian detection. The occurrences of gradient orientations in localized portions of an image are calculated by the HOG feature and are intended to keep invariants to geometric transformations. HOG features were applied and revised by current works [346] [106] to capture appearance information and shape orientations of facial images for expression analysis [641]. Furthermore, Gaussian processes are used in more complex models [610]. PCA, LBP and HOG were used in a recent work by Ali et al. [12] to represent facial image feature vector, then the extracted feature vector was processed for dimensionality reduction using further principal component analysis. A weak classifier during boosting was trained utilising SURF features with the One-against-All logistic regression model which was used by Rao et al. [454]; afterwards, the weak classifier was selected to combine a fine multi-pose classifier. Their trials attained a rate of 90.64 % on the RaFD dataset and 74.05 % on the KDEF dataset. A Bayesian model was constructed by Mao et al. [378] by deploying multiple head poses to overcome the feature variation caused by head poses. Ionescu et al. [250] showed that a local learning approach improved the bag-of-words model for image-based facial expression recognition. Correspondingly, these hand-crafted applications have also been widely used in 3D FER; these are utilised to describe 3D facial shape data by coding different types of geometric maps [328] such as curvature-HOG [317], normal-LBP [327], and depth-SIFT [44]. Nevertheless, the hand-crafted features have shown their restrictions in practical applications. In general, facial expression (AUs and emotions) recognition methods can be divided into three categories. Frame level based approaches detect and evaluate emotions and AU occurrences (facial texture changes such as bulges and wrinkles) in each frame independently using appearance or geometric feature extraction methods, combined with binary or multi classification classifiers such as SVM or Adaboost [486]. While all the methods try to find landmarks, features location information or the geometry of the facial shape components signifies geometric features. Segment-level approaches use temporal dynamics in video sequences to detect AU from a set of temporally contiguous frames. Temporal phase modelling algorithms (transition detection) seek to discover constituent temporal segments: neutral, onset, apex, and offset in the event episode [136] [568] [612] [268] [395]. Intuitively, while facial actions express themselves over a time span, a dynamic pattern of information captures the trajectory changes of current state, and past state in a time space volume [552]. On the other hand, frame based methods are faster and easier to implement. However, static methods are very restricted in detecting affective expressive actions in real time, conveying less important information and neglecting to handle the latent

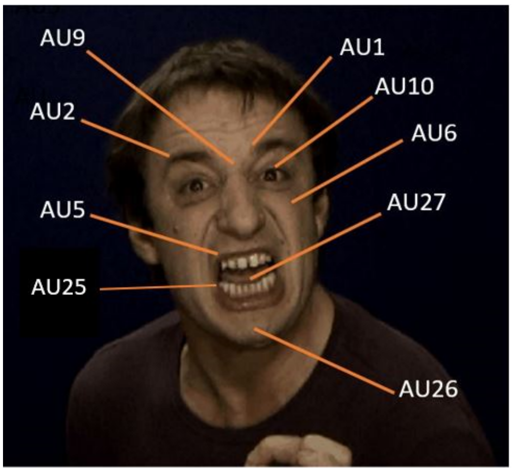


Fig. 2.21 The rules used to represent an uncontrollable rage expression (angry) by the activation of AU1, AU2, AU5, AU6, AU9, AU10, AU25, AU26, and AU27.

temporal variations among consecutive frames of the sequence [484] [528]. On the other hand, some AUs can be recognised using static features only, and also the remaining dynamic features are important; for example, the only lone difference between the AU43 (Eyes Closed) and the AU45 (Blink) lies in the area of temporal duration of eye closure. Nevertheless, a static image can often still provide enough beneficial information for AUs recognition [268]. The question is whether the detection of the occurrence of target AUs needs the modelling of the entire sequences, or whether a single frame is sufficient [528].

A plethora of published work on dynamic facial expression analysis has concentrated on incorporating the temporal relations of the frame order continuity in a sequence to improve the performance of video prediction. Previous studies which used a group of heuristic rules-based per AU with facial landmark positions [268], such as Fig 2.21, represented an uncontrollable rage expression from the GEMEP-FERA dataset [557] using some rules for mapping AUs to emotions by the activation of AU1, AU2, AU5, AU6, AU9, AU10, AU25, AU26, and AU27. Discriminative graph-based methods such as variants of Dynamic Bayesian Network (DBN) are probabilistic graphical models that can learn the full conditional joint probability of temporal cues for facial actions [486], such as Conditional Random Fields [219], Latent Dynamic Conditional Random Fields [568], the Kernel Conditional Ordinal Random Field [468], and Hidden Conditional Random Fields for action unit estimation [448]. Hidden Markov chain-shaped transition models are used to encode temporal persistence and the likelihood of label transitions throughout the sequence [380]. It was estimated the entire probability density with the observed data conditioned on the labels hidden state as an alternative of estimating the posterior category probabilities. Using the HMM approach, the concealed label sequence is not restricted; however, it can contain any value to represent the observed data. A normal distribution can often be used for modelling the conditional probability of the observation x_t with the given label state y_t for the real-valued observations [528]. Weakly-supervised learning such as Multiple Instance Learning are proposed to deal with incomplete labels. A semi-supervised learning approach can be effective in recognising all the positive samples of annotated data with potentially advantageous unlabelled data [488]. Segment-based classifiers use a bag of temporal words to represent the segments. For unsupervised approaches; sequence based clustering algorithms are used to group events of similar characteristics. Slow Feature Analysis describes a latent space time variation that correlates with the AU temporal segments [616]. An unsupervised Branch-and-Bound framework is used to force synchrony correlated facial actions in an unannotated sequence [92]. However, the aforementioned methods rely on specific problems under certain uses.

2.4.3 Deep Learning Representation

The convolutional neural network (CNN) is the best-received of the numerous deep learning models available; it has been proven to be particularly well suited for large-scale images recognition tasks. It has a robust visual representation capability due to the meticulous design of local to global feature learning with convolution, activation function, regularisation, loss function, pooling, design of layered architecture, optimization and fast computation [204]. Amongst the great benefits of CNNs are their ability to eliminate or greatly decrease the dependence on physics-based models and/or other pre-processing techniques by aiding end-to-end learning directly from input images [293]. On top of that, more recent work using Deep Convolution Neural Networks (DCNNs), involving robust accurate learning for more discriminative feature extraction from raw pixel image data, has triumphed over traditional methods. This is due to their exceptional ability of reporting improved results stemming from desired characteristic representations which result in high performance to expedite the process of training and testing at very low power consumption in many computer vision tasks. One of the major limitations of conventional CNN is that impartially extracted spatial relations of the facial components cannot consider the temporal variation relations [395] [293]. An alternative is to utilise deep neural networks, particularly CNN as a feature extraction way, and then implement an extra classifier, for example SVM or RF to get the optimal image representations [333]. Lately, deep learning methods have outstripped most state-of-the-art algorithms in the current literature for many vision tasks [17], for example, image classification [341], object detection [458], hand gesture recognition [383], scene understanding [340], segmentation [539], detection of visual saliency [204], action recognition [36], pose estimation [93], image generation [365], scene understanding [423], biometrics [307], face recognition [429], and facial expression recognition [87] [379]. A variety of techniques are deployed in up-to-date CNNs to decrease the training time and improve generalisation over input data, including data augmentation, dropout regularisation, ReLU activation functions, repetition balanced batches and GPU acceleration [308]. The input image is convolved through a filter collection in the convolution layers to produce a feature map, in CNN-based approaches. Every feature map is then pooled with fully connected networks and the facial expression is recognised as fitting a specific class-based output of the softmax algorithm [293]. Both low-level generic features and high-level semantic features are captured with a well-made CNN, as it is trained on millions of images which can parameterise a hierarchy of filters. Lu et al. [362] employed Convolutional Neural Networks (CNN) on facial appearance. Kahou et. al. [277] implemented convolutional neural networks (CNNs) to identify facial expressions and attained the 2013 Emotion Recognition in the Wild Challenge [407]. Deep learning was exacted on a geometric model from facial regions for facial expression analysis by Liu et al. [353]. Tang [610] recounted a deep CNN together learned with the output of a linear support vector machine (SVM) output. This process accomplished the first place on the FER-2013 Challenge [195]. Liu et al. [353] suggested a facial expression recognition framework based on 3D CNN and deformable action parts constraints to jointly localize specific facial action parts and discriminative learning part-based representations for expression recognition. Richardson et al. [460] extracted the face geometry from the image directly using a CNN based approach. A CNN model was deployed by Nezami et al. [408], to recognise facial expressions; a learned representation was used as an image captioning framework; this model implanted the identified facial expressions to produce additional human-like captions from images having human faces. Yu et al. [610] made use of a CNN model that was pre-trained on the FER-2013 dataset [195] with fine-tuning on the Static Facial Expression in the Wild (SFEW) dataset [126]. This method accomplished the first place on the FER-2013 Challenge [195]. Liu et al. [356] and Meng et al. [388] show identity-aware facial expression recognition models. CNNs were harnessed by Zhang et al. [634] to extract spatial information from video frames, where facial expressions were recognised by amalgamating the spatial information with temporal information. Peng et al. [435] focused on a synthesis CNN to produce a profile view from a static frontal face. Generalisability in recognising facial expressions was enhanced by Mollahosseini et al. [397] when CNN models were trained across various well-known FER datasets. Face registration procedures were applied to align faces to attain better levels of performance. Additionally, pre-trained Caffe CNN models were included by Liu et al. [355] to extract features of image-level. Both feature extraction and transfer learning can be implemented by CNNs, as shown by Kahou [277], where an example of work carried out by a team where training a CNN to extract audio features from video and by using a deep Restricted Boltzmann Machines (RBM) in a multiple modal data representation. Also, “Bag of mouth” features are extracted as well to enhance the performance. The effect of merging registered and unregistered face images on FER recognition systems were gauged by Kim et al. [287], which uses unregistered samples while the facial landmarks of the images were not detectable. Pramerdorfer [442] deployed a blend of recent deep architectures like VGGNet [506] into their CNN model to advance prediction accuracy using the FER-2013 dataset [195]. Duong et al. [147] suggested a new Deep Appearance Models (DAMs) method, as an improvement of the traditional AAM models by using Deep Boltzmann Machines (DBM) for robustly capturing the changes of the facial shapes and appearances [365]. Hu et al. [236] incorporated a new learning strategy into their CNN model referred to as Supervised Scoring Ensemble (SSE) in order to improve the accuracy of the prediction [157]. A CNN model was trained by Li et al. [338] via a modified back-propagation algorithm creating a locality preserving loss that aims to attract the local neighbouring faces jointly for the same class [177]. Ebrahimi et. al. [277], offered a deep learning-based technique for emotion recognition and the system was the challenge winner with a classification accuracy of 41.02% [128]. Sikka et al. [502] recommended a multi-kernel learning-based method, which had been taking second place in a technique throughout the EmotiW 2013 [128]. Mengyi [355]’s process centred on manifold learning and the convolutional neural networks which achieved the best throughout the EmotiW 2014. The latest models gain better effects centering on facial expression recognition in the wild using committees of different CNN classifiers [362]. The deep CNNs committee group was improved by adjusting the network architecture and the initialization of random weight [600]. While most of the preceding models concentrate on still images, facial expression analysis can be advantaged from temporal dynamic data [462] [139]. Cohen et al. [96] exploit Hidden Markov models (HMMs) for recognising facial expression from video sequences [95]. A recent breakthrough of deep hybrid approaches fusing a CNN and Long Short Term Memory [179] was developed for combining high level spatial features while preserving temporal dependencies simultaneously [293] [518]. Emotion perception modelling in video includes deep learning methods such as the use of CNNs for local feature extraction and LSTM for learning the temporal dynamics [610]. A network architecture suggested by Hasani and Mahoor [395] consists of 3D Inception-ResNet convolutional layers, followed by a Long Short Term Memory (LSTM) unit which has extracted jointly the spatial appearance information relations inside facial images alongside with the dynamic information relations among different frames in the sequence [176]. Connie et al. [100] provide an example of works which blend deep and hand-crafted features often employing an individual CNN model and several hand-crafted features, who use SIFT with dense SIFT, as well as Kaya et al. [281] who combine SIFT, Local Gabor Binary Patterns (LGBP), and HOG [177].

2.4.4 Generative Adversarial Networks (GANs)

Vast quality published methods were suggested and a large body of efforts present on modelling synthesis image in deep learning. Early work comprises Restricted Boltzmann Machines and include their variants such as Deep Belief Networks [591] [526] [228] [229] [412]. More recently, various successful models have developed in the area, including the Auto-Regressive models [311], and the Variational Auto Encoders (VAEs) [233] [292] which are a directed graphical model of latent variables to design complex generative models of data [291]. In reference [82], the authors state that training VAE is much easier than Generative Adversarial Networks (GANs) which use statistical inference [210]. In spite of the robust and stable training, VAEs have a tendency to produce blurry images [513], whereas combining a training of VAE and GANs jointly [312] (sharing the parameters of VAE encoder and GAN generator) enables use of the discriminator learned feature representations (gauge sample similarity) in the GAN for the objective of VAE reconstruction, and to learn an identity-invariant information representation [333]. Lately, great fresh cutting edge developments have been created by the approaches based on Generative Adversarial Networks (GAN) [68]. In 2014, Goodfellow et al. [194] offered the idea of Generative Adversarial Networks (GAN). However, the training process was not stable and the created images are predominantly noisy, and not intelligible, making it one of the main drawbacks of GAN. Realistic images have been made extensively by using the Generative Adversarial Networks, but it has been vaguely exploited for the classification capabilities [69]. Moreover, over the past three years, several methodologies [638] [61] [656] [630] [251] [594] have been recommended as ways to improve the existing GAN from different standpoints. Of late, Deep Convolutional Generative Adversarial Network (DCGAN) [449] have revealed a higher level of performance of the image generation. DCGAN merges GAN and CNN to provide techniques for enhancing the training stability [574]. Convolutional Neural Networks are adopted by the DCGAN to implement the Generator and the Discriminator, respectively [333]. Empirical instructions are also given on how to construct a stable GAN, for example, by taking the place of the pooling by strides convolution with utilising batch normalization [630]. Starting from the original network structure, Generative Adversarial Networks (GANs) have shown a great promising ability for synthesizing images. Extensions of several changes over the original design were proposed and exist recently, such as CycleGAN [363], InfoGAN [304], ExprGAN [133], Amgan [180], FF-GAN [609], and DR-GAN [545], SGAN [571] [632], TP-GAN [239], AC-GAN [416], CapsGAN [480], LR-GAN [605] model and Wasserstein GAN (W-GAN) [28]. More topical techniques concentrate on combining limitations on the input data of the generator or getting side information to improve synthesis. For instance, a conditional GAN is deployed conditional data to produce facial images from simple noise, the CGAN is an extension of the GAN [194], where G and D are given an extra variable Y as input [170] [392] [279], the generator was controlled by Y in this model. A more recent work by [334], using conditional GAN and added auxiliary constraints for augmenting the model using class labels and for governing the generator output and the discriminator, was exploited as a classifier to predict the classes. Additionally, Mirza and Osindero [392] depend on providing the class label to the G and D to formulate images which were conditioned on the class label. Luan et al. [545] and Springenberg [512] generalised GAN by learning a discriminative classifier where D was trained not just to differentiate between real and non-genuine subjects, but also to classify the processed images [630]. Info-GAN [82] learns explainable representations by using latent codes, and it utilises information regularisation for optimisation. WGAN [28] brings in Wasserstein distance to exchange the KL divergence in order to solve the problem of model collapse in GAN, and the samples were produced with greater diversity [587]. Odena [415] considers the samples of GANs to be a new class through semi-supervised training [547]. Zheng [649] also concentrates on semi-supervised training by allocating a unified label distribution upon all the current classes of GAN samples [574]. A method was proposed by [123] which uses a cascade of convolutional networks with a Laplacian hierarchical generation framework for generating higher quality images from a coarse to a fine pattern fashion; however, the objects' results were unsteady due to the releasing noise caused by enforcing multi models. GANs can also be utilised to reproduce an older version of the input image; this is seen in the reference of [26]. Even though the results were not validated, the procured images are largely realistic. The efficacy of GANs in image processing is demonstrated by the other cases of GAN which include: synthesizing front facial images from rotated images [239]; keeping the identity of the subject depicted in the images by altering images [347]; and eliminating additional illumination from facial images to confirm the appropriate conditions for face identification [652]. The variety of training techniques of GANs involve supervised learning, unsupervised, and semi-supervised, which also have provided various outputs for the classifiers [69].

2.4.5 3D Morphable Face Model Reconstruction (3DMM)

The 3D morphable face model was first founded by Blanz and Vetter [51]. It builds on the existing 2D active appearance model [101]. Blanz and Vetter showed how a face could be reconstructed using just one image through iteratively adding combined linear values of existing parameters such as registered scans and pose, camera, and lighting. Both the geometrical and textural features of the face scans were then decomposed via PCA and used to create separate and reduced-dimension geometry and texture spaces. Additional face scans were introduced in later work by [178], as well as more features in the model that enabled it to have expressions within a separate space. Factors such as the primary conditions and the complexity of the scenery heavily influence the convergence of the iterative fitting. Sela et al. [487] aim to address this deficiency through implementing a correspondence map. Although both of these approaches have impressive reconstruction results, the training data taken from the linear 3DMM model can be limiting. Consequently, out-of-subspace variations such as facial hair are not adequately dealt with. The goal of Booth et al. [55] is to optimise the texture of 3DMM through focusing on the in-the-wild feature based texture mode instead of controlled settings. 3DMMs have not gained popularity because they are difficult to construct. The process necessitates a 3D scanner that is both fast and exact, in addition to an output of several hundred scans and between these scans the computation of dense correspondence. It is widely acknowledged throughout various articles written on the topic of facial recognition that 3DMM based face image analysis is state of the art; however, it is simultaneously recognised that the main limitation is the complexity of its construction [27] [315]. Monocular case is another area in which 3DMM excels in facial reconstruction. The issue is that information about the surface is scarce, much like it is with a single image. Therefore, the process of 3D face reconstruction should use 3DMM as its starting point [544]. Additional work [50] [466] [532] [169] [323] and others have used various methods to enhance both the precision and reliability of the fitting. This has led to results that are accurate under conditions that are optimal. Throughout the literature, neutral expression is generally assumed in the case of 3D face recognition [477], whilst facial expression recognition is done using 2D imagery. Detailed analyses and surveys of 3D face recognition and facial expression recognition that used both 2D images and video can be seen in [62] [479]. Work that built on Blanz and Vetter's PCA based model was also done by Amberg et al. [22]. This work shifted to emotive facial shapes through adding a PCA modelling of the residuals taken from a pose that was neutral. The subsequent result of this was a single linear model that captured identity and expression variation in a facial shape that was 3D. Many methods have been suggested to address the limitations of ill-posed images in the process of using a single image for 3D face reconstruction. It has been highlighted by Vetter and Blanz [50] that it is possible to estimate the structure of the human face, both in terms of its texture and geometry, through a linear combination of orthogonal basis vectors that are taken from PCA of more than 100 male and female identities. Through this, the 3DMM is put forward as an ideal tool for representing the 3D face's shape and texture. Nonetheless, iterative methods remain unreliable, particularly in conditions that are in-the-wild, and so this results in more interest in methods that are regression-based [175]. Lately there has been more focus placed on this problem's most simple yet challenging manifestation, which is that of monocular face reconstruction. This infers that a 3D face mesh attained from one 2D photo poses difficulty, especially in terms of identity, camera and ill-posed images given that the process of image formation combines a series of facial components (shape, albedo) in addition to environment (lighting) into one colour per pixel.

The 3DMM comprises of a generative 3D shape that is parameterised, alongside a parameterised albedo (skin reflectant) model. These are combined with an associated probability density on the model coefficients. Describing a face is done through using a set of shape and albedo coefficients. When projection and illumination parameters are combined it is possible to render a face [432]. The construction of a linear 3DMM is done through performing a type of dimensionality reduction, on a training set of facial meshes (PCA). This dimensionality reduction is usually performed using a principal component analysis (PCA). Doing this necessitates each mesh firstly being re-parametrised into a consistent form. In this form, the amount of vertices, as well as the triangulation, and the anatomical meaning of each vertex, are made consistent throughout all of the meshes [57]. To illustrate this, an example can be that if a nose tip aligns with the vertex of index i in one mesh then it is necessary for all the remaining vertexes in each mesh to similarly correspond to the nose tip. One example of 3D face reconstruction is illustrated in Fig 2.22 and in this example every data vector entry aligns with the same point on the face, as well as the initial entry corresponding with the tip of the nose. Meshes that cater to the above-mentioned properties have been identified as densely corresponding with each other. Although it is easy to highlight the drawbacks of this, accurately and efficiently resolving amongst highly variable facial meshes can be a challenge. In addition, simply defining the anatomy of the human face, particularly in smoother areas of the face like the forehead and cheek, can be extremely difficult and so obtaining an objective measurement of correspondence quality poses a challenge. In order to register a common reference frame for the meshes, Patel and Smith [430] use a Thin Plate Splines (TPS) warp [544]. A collection of k-meshes $\{M_1, M_2, ..., M_k\}$ are inputted into a 3DMM construction algorithm. The individual input meshes have a unique amount of vertices and triangles, as well as a specific order in their topology. There are two stages to constructing a 3DMM. During the initial stage, it is necessary to create a state of dense correspondence from one training set mesh to the other. Next, there is a statistical analysis taken on the corresponding meshes and this results in linear models of both shape and texture [55]. The constructing of conventional linear 3DMM is discussed further in [544] [432] [529] [531], and non linear 3DMM construction is detailed in [530] [543] [544]. A regular type of optical flow is used in [51] to resolve the issue of correspondence being dense amongst different facial meshes. Nonetheless, there are limitations to this technique insofar as it works effectively only in more constrained settings, such as those where factors like age and ethnicity are the same between subjects [544]. Following this, a 3D morphable face model is made through taking a dataset of m exemplar faces. Once the registration is done, the faces are parameterised in the form of triangular meshes that have both a specific amount of vertices n and topology that is shared. Simply put, the face's geometrical structure can be shown with a shape vector S that has the (x,y,z) coordinates of the n vertices, while the face's texture can be shown using the texture vector T that is built from the colour values (r,g,b) for the n relevant vertices. The vertices $(x_j,y_j,z_j)^T \in R^3$ have an associated colour $(r_j,g_j,b_j)^T \in [0,1]^3$. This assumes that the amount of vertices are the same as the amount of texture values. Finally, it is possible to represent the face using these two dimensional vectors [432]:

$$S = (x_1,y_1,z_1,\dots x_n,y_n,z_n)^T \in R^{3n} \tag{2.6}$$

$$T = (r_1,g_1,b_1,\dots r_n,g_n,b_n)^T \in R^{3n} \tag{2.7}$$

The full range of the human appearance can be captured by a 3D morphable face model (3DMM) [51] due to the smooth, low-dimensional face that it spans over. Using one image to locate a person's co-ordinates in this space is a task that many applications undertake. Examples include: 3D avatar creation [248], facial animation transfer [153], virtual make-up [325], lip synchronisation [662], video editing [168] [49] [169] [532], faces animation [14], human skull geometry modelling [146], automatic face recognition [22], avatar puppeteering [59], transfer of facial expression among individuals [464], stimuli generation of experimental psychology [49] [169] [532] [320], image-to-image auto-encoder with a fixed, morphable -model-based decoder and an image-based loss [531] [175].

3D face reconstruction approaches traditionally use optimisation algorithms (optimisation-based) that search the space using inverse rendering [432] [169]. This optimises parameters such as shape, texture, pose and lighting to replicate the face from a photograph [323]. An example of this is iterative closest point [23], to regressing the 3DMM coefficients through finding the solution to the non-linear optimisation issue for establishing the correspondences of the points between a single face image and the canonical 3D face model [658], including facial landmarks [659] [316] [532] [66] [258] [198] and local features [198] [252] [466]. Following this, Bolkart and Wuhrer [54] demonstrate the way that it is possible to directly estimate a multilinear model using the 3D scans from a joint optimisation rather than the model parameters and group-wise registration of 3D scans [544]. Nonetheless, these techniques often take a great deal of time because of their high nonlinear optimisation complexity, the practical solving of which is arduous and also prone to being affected by a local optimal solution and initialisation that is bad and inefficient [542]. An alternative approach suggested by Paysan et al. [432] is employing a Nonrigid Iterative Closest Point (ICP) [23] for direct alignment with 3D scans. More recently, it has been shown by various works that quickly and

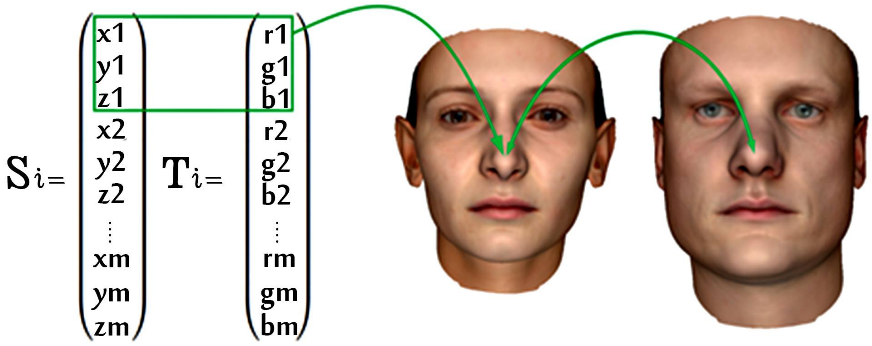


Fig. 2.22 Each point on the face is identical to the equivalent input data vectors ((x,y,z) coordinates and color values (r,g,b)). It can be noticed clearly from this example that the first input corresponds to the prominent pointed tip of the nose; image adapted from[432].

robustly fitting used regression from image pixels to 3DMM morphable model coordinates that use Convolutional Neural Network (CNNs) (learning based) [460][461] largely enhances both the efficiency and reconstruction quality [460][461][548][531]. Techniques that are CNN-based have also resulted in both successful 3D face reconstruction and dense face alignment [543][275][548][357]. [275] [658] [460] [661] show how regressing 3DMM coefficients has been done using cascaded CNN structures. As a result of the multi stage process, this can take a lot of time. Also, end-to-end methods [143] [548] [274] have been suggested to achieve a direct estimation of the 3DMM coefficients in a way that is holistic. Duong et al.'s earlier research [410] relies on presenting 2D Active Appearance Models using Deep Boltzmann Machines. Nonetheless, this is only successful when applied to 2D faces and sparse landmarks; therefore it is unable to optimally manage faces that have either large-pose variation or occlusion [544]. Variation Autoencoder (VAE) is employed by Bagautdinov et al. [32] as a means of learning how facial geometry can be modelled from 3D scans directly.

Tewari et al. [530] use multi-layer perception for embedding shape and albedo bases. Nonetheless, to achieve an accurate 3D face CNN regression model it is necessary to have a broad number of training faces that have 3D annotations. Collecting this is oftentimes both expensive and difficult to achieve [546]. The key problem with the regression method lies in its limited Ground Truth 3D face data for training purposes. It is often difficult to obtain scans of face geometry and texture as factors such as expense and privacy can create limitations. Earlier methods have focused on exploring how to synthesise training pairs of images and morphable model coordinates in a preprocess [460] [461] [548]. A method used by some works is that of employing large-scale synthetic data like [460]; however, these do not generalise at a high quality because there remains a domain gap with images that are real. Tewari et al. [531] identifies how 3DMM parameters can be regressed while unsupervised. Although the training is done on images that are in the-wild, there are still limitations due to the linear subspace. Due to this, the surface shrinks if the texture posed is more challenging [544]. This method is furthermore enhanced by Genova et al. [175] who compare both the reconstructed images and original input through focusing on higher-level features taken from a face recognition network that is pre-trained. Tran et al. [548] show how 3DMM representation can be regressed in a robust and discriminative fashion through using an assortment of images from one subject. Convolutional neural networks are used by Tran and Liu [533] to demonstrate how geometry and skin reflectance can appear in UV space. Although the weight of their representation is far stronger, it is still difficult for these models to accurately attain smaller details and nuances in the initial input images because of how their learning objectives have strong regularisations. Feng et al. [162] in more recent work stores the 3D facial geometry within a UV position map and trains an image to-image CNN so that it can regress the entire 3D facial structure directly, as well as the semantic information that can be acquired from one image. A refinement network is created by Richardson et al. [461] to add additional facial details on the existing geometry that is based on 3DMM. Recent work has suggested using CNN directly to acquire the reconstructed 3D face bypassing the 3DMM coefficients regression. Jackson et al. [252] suggest using a CNN-based regression for mapping the image pixels to a volumetric representation of the 3D facial geometry. Although their approach is no longer limited to the 3DMM space, predicting the voxel information requires both a complex network structure and a large amount of time.

2.5 Facial Expression classification

Classification and predictions of subject facial expression that appear in the image or video represent the output of the final stage of the facial expression recognition system. Facial changes can be identified as facial action units or prototypic emotional expressions depending on the type of information and whether the temporal information is used [426]. Facial expression classification studies are conducted for supervised learning algorithms to improve the recognition. In this regards, the extracted features were often utilised as emotion detectors and were then deployed to train a variety of classifiers; exemplars include spatial classification engaging Support Vector Machines (SVMs) [221], Random Forest classification method [348] [646] [581], Softmax [582] [329], Nearest Neighbor classifier [471], Linear Discriminant Analysis (LDA) [647] [156], Euclidean distance classifier [129], AdaBoost [576], Gaussian process [84], Fisher face [47], Discriminant analysis [500], Decision Trees [551], Hidden Markov Models (HMMs) [190] [453] [445], Naïve Bayesian learner [97], logistic regression [475] [452] and k-nearest neighbour (KNN) [293] [36]. Artificial Intelligence Approaches with tools such as Neural networks and Machine Learning techniques are able to automatically recognise the faces [403] [364]. In the work of Li Xia [592] a multi class-classification SVM classifier was used for facial emotion recognition and the results have shown an improvement in the performance for both training and testing stages in comparison with other conventional classification approaches. A fusion based method using CNN and SVM has been suggested by [567] where the results were promising on the CK dataset with about a 96.04% average accuracy rate. Omer et al. [511] proposed to use SVM for 3D facial expressions by using distance based features on the 3D Bosphorus Dataset. The experiments were executed on three emotions and achieved a 85% recognition rate [401]. Furthermore, probabilistic and dynamic classifiers are used such as Dynamic Bayesian Networks [36] and Restricted Boltzmann Machines (RBMs) [163].

2.5.1 Support Vector Machines

SVM [562] [234] is an optimal large margin supervised Machine Learning classifier for FER systems [38] [554] [560] [492], which has been proven to be an efficient, robust classifier used for regression and for both binaries as well as multi-classification problems. SVMs were first introduced by Boser and Vapnik and the recent pursuance of SVMs was employed by [103]. It is very powerful against outliers which must be contained in training data [190]. To separate two classes, SVMs attempt to construct a successful separation of a hyperplane that maximizes the margin while reducing the total distances of points among closest points which are positive and negative for specified observations. These are well known as support vectors [190], for example, splitting up two classes (facial Action Unit as such present or not) by a plane. Noticeably, regarding generalisation capacity, SVM proved to minimise the error estimation which is called the Structural Risk Function better than the popular Empirical Risk Function [459]. SVM is reliant on a non-linear mapping with finding a hyperplane to separate data classes. Therefore, SVM used a so-called kernel function trick to transform the input space to kernel space to increase the accuracy of the classification. Due to the SVM flexibility and the capability as a helpful classifier used for achieving predictions and projecting high dimensional data which emerged through feature extraction methods with less prone to overfitting. In the recent work, it has been shown that applying a linear kernel SVM classifier to learn the dependencies between different facial features and to distinguish between posed and spontaneous emotions is superior to other classifiers. A SVM classifier with different kernel functions was used to optimize the performance of numerous applications. For instance, Polynomial Kernel SVM, quadratic, RBF Kernel SVM, Linear Discriminant, Logistic Regression, k-Nearest Neighbour, Nave Bayes, and radial basis functions [374]. Therefore, SVM was commonly used for classification in our experiments because it generalises very well on unknown data. The Library of SVM was used for the training and testing of SVMs [75]. One drawback of conventional SVMs and many other classifiers is that they are not subsuming the temporal information structure knowledge of facial expressions [561]. For the task of facial expression recognition, for emotion detection, a multi-class decision has to be made. To accommodate this, a one-against-all-the-rest principle multi-class classification SVM was used such as Sad against not Sad, etc. For AU detection, a linear one-versus-all two class classification SVM was used, for example the interest of AU versus non-AU of the involved interest for the training of the linear SVM for each of the AU detectors. Moreover, all the features used for training SVM are linearly normalized into [0 1].

2.5.2 Random Forests (RFs)

A popular classification method was proposed by Breiman [63], who in his seminal work created the learning framework known as Random Forests. Random Forests was used widely in both computer vision and FER tasks. This is because it manages high-dimensional data like images or videos efficiently, in addition to being well matched to multi-class classification tasks [110]. It is a method that is useful as an ensemble approach, particularly because its predictions for classification, regression and numerous Machine Learning problems are usually very accurate [348]. The basis of the Random Forest classification approach is the decision-making tree with parameters that are random [646] [581]. They are a set of binary decision trees that are made on various samples from bootstrap. Bootstrapping is a randomised sampling technique that has replacement from the training set N [181]. Together with random subspace approaches for feature sampling and bagging, it results in a performance that resembles more well-known Machine Learning techniques; for example, SVM or Deep Neural Networks. Nonetheless, the probability for membership is better than the simple knowledge that the one is belonging to a part of a group. Currently both votes and the regression approach are probability estimation techniques made within Random Forests [324]. Constructing the decision tree was done via a Random Forest algorithm that took a group of random options and chose the option that was most suitable [218]. Of RF's characteristics, a key use was that of the Out-Of-Bag (OOB) error estimation, which is a sample set that did not exist in the current training tree. Therefore, its usage lies in error estimation. Internally estimating the generalisation error improves the level of accuracy in the tree classification results. In addition, it is a vital feature for importance quantification [181]. The popularity and usefulness of classification has been important as an application for Random Forests. Random Forests is seen as a more refined approach than traditional decision tree-based classifiers given that it has addressed many of their limitations. More specifically, the approach tackles the issue of overfitting. Random Forests maintains a high level of accuracy both for training data and unseen data. An additional advantage is that Random Forests is able to manage missing values in a more efficient manner [401]. The Random Forest classifier is not only strong but its recognition rate for images of higher resolution is also high [263]. The only disadvantage of the classifier is its limited computational complexity [298].

Chapter 3

A Benchmark of Dynamic Versus Static Methods for Facial Action Unit Detection

3.1 Introduction

Facial Action Unit (AU) is defined as the smallest observable localised movement units formed by a combination of facial muscles to formalise many facial expressions, which have been interpreted as emotions [5]. Facial action temporal cues are a reliable unique signal used to distinguish between different facial AUs, and these cues have an auto-correlation consistency at the time of the target signals. Fig. 3.1 shows how the importance of the temporal dynamic of AUs has a crucial impact on the real meaning of the facial expression and distinguishes between posed and spontaneously occurring expressions in the timing and appearance of facial muscle activations [21] [267] [144]. AU activation is a set of local individual facial muscle parts that occur at the same time constituting a natural facial expression event, which represents an essential part in predicting and formulating the task of facial expression recognition. AUs' occurrence activation detection can be inferred as temporally consecutive evolving movements of these parts, which may contain the four temporal segments of AU events: neutral, onset, apex, and offset together with a continuous variation in intensity [624][173].

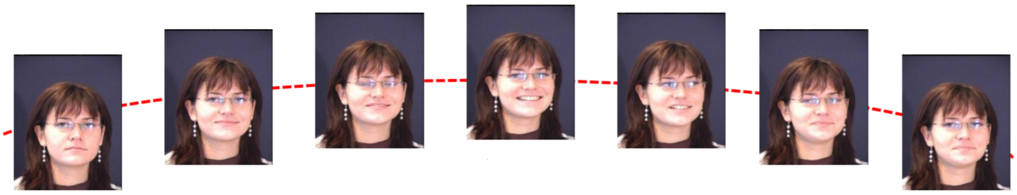


Fig. 3.1 Temporal dynamics of facial AUs; image adapted from [207].

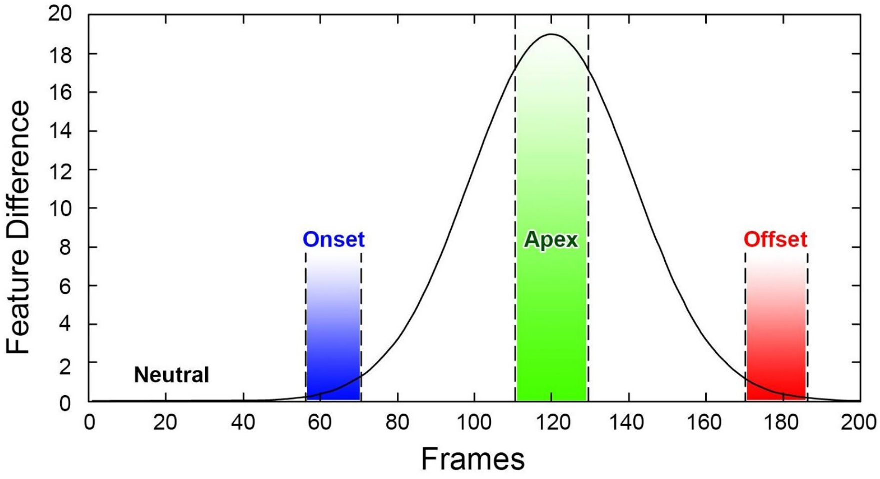


Fig. 3.2 The temporal segment phases of a facial expression: $Neutral \Rightarrow onset \Rightarrow apex \Rightarrow offset$; image adapted from [112].

To encode facial Action Units, the Facial Action Coding System (FACS), which was developed by Ekman and Friesen [165], is the most comprehensive system that accurately describes the basic facial expression movements by encoding the configuration of AU or multiple AUs in terms of facial atomic activation muscle actions. In a muscle-based approach, FACS defines 46 facial Action Units assumed as the smallest fundamental measurement of visible discernible blocks of facial movements [268] [340] [556]. Furthermore, this system supports mapping from facial appearance changes to emotion space [555]. Early work focused on extracting the features from still images or sequences separately and the combination of them was completely ignored. Detecting AUs automatically can provide explicit benefits since it considers both static and dynamic facial features. The discrimination between both posed and spontaneous facial expressions reveals that the ordinal temporal dynamics of certain facial muscles are particularly important for the classification of facial expressions; moreover, the difference between smiling and shouting is shown by the orientation of the parted lips and the area surrounding them. Thus, the important step in the analysis of facial expression is the detection of the AU activation themselves in a sequence [528].

Most people can not do the exaggerated act of the two signs of sadness: AU15 (Lip Corner Depressor) or AU1 (Inner Brow Raiser) with AU4 (Brow Lowerer). Even though some of these actions can be performed on instruction, the timing differed significantly from what occurs with the activated muscles of spontaneous facial actions [98]. It has been shown that for posed and spontaneous smile temporal characteristics (cues), such as frame rate, morphology, configuration, the speed of activation, time and total duration, co-occurrence, trajectory, and asymmetry [486], these are fundamental factors in distinguishing between

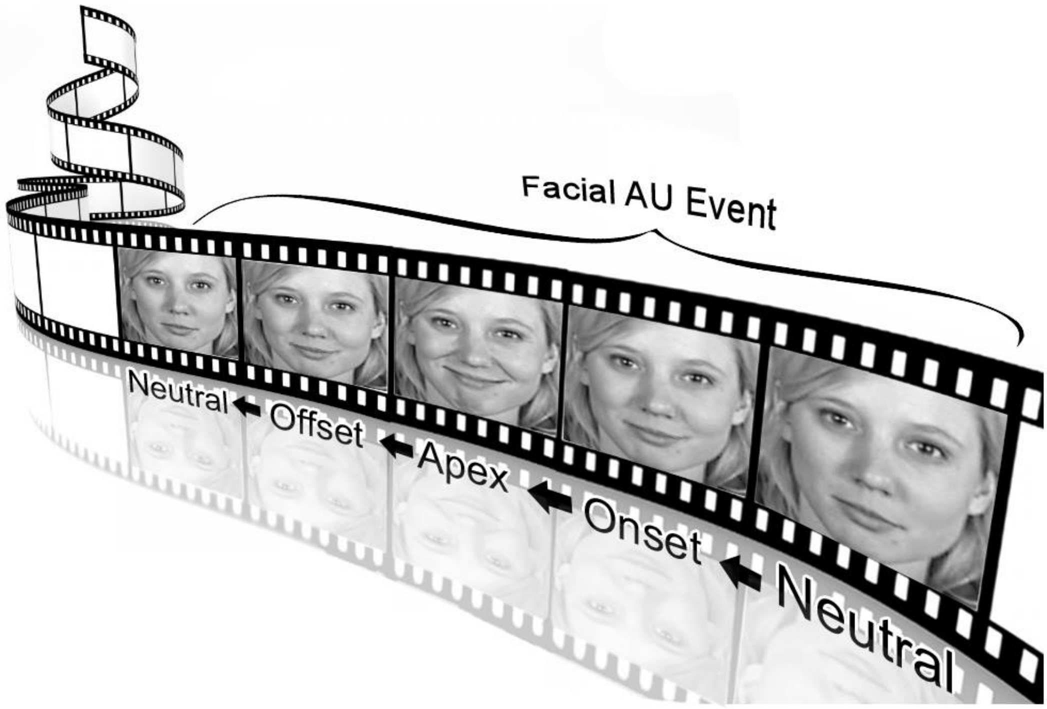


Fig. 3.3 Temporal facial dynamics of AU12 (Lip Corner Puller) transition detection; image adapted from [265].

the two types of classes. Furthermore, the difference between frustrated and delighted smiling faces showed the importance of analysing the temporal pattern code of AU12 in different affective states [411][173]. Frowning faces frequently present an embedded subtle appearance and dynamic consecutive contraction of muscles in one's brows of an AU4 episode [246]. Therefore, the temporal contexts determine the gradual relative duration changes of different AUs [342]. The temporal dynamics of AUs have a crucial influence on the nuanced meaning of facial expressions which should be considered when using dynamic features for the facial expression recognition system [342]. Moreover, it was reported in [213] that facial expressions can be recognized more efficiently from videos than from a still image. The question is whether the detection of the occurrence of target AUs needs the modelling of the entire sequences, or whether a single frame is sufficient [528].

In general, conventional AU activation detection methods are broadly divided into four techniques:

- Frame-level method has individually detected the presence or absence of the occurrence of one AU or other AUs from each frame of a video using appearance or geometric features [135], without a comparison with the proximal contiguous neighbouring frames [284]. Appearance features can be extracted from texture and motion changes in the face [192], such as wrinkles, bulges, and permanent furrows, and usually it is represented by SIFT, Gabor, and LBP descriptors. Geometric features contain information of facial feature shapes, including detecting landmark locations and the geometry of facial components. The first AU detection challenge (FERA) indicates that most approaches, including the winning one, were frame-based [135] [284]. Frame-based methods have shown to be more effective for the detection of subtle AU events because of the sensitivity to each frame. However, they are prone to noise due to the lack of temporal consistency.

- The sequence-level methods are based on learning the temporal related information of the sequence.

- Segment-based approaches use temporal dynamics in video sequences to detect AU from a set of temporally contiguous frames, for example, Hidden Markov Models, Bag of Words, and Rule-based methods. This approach is closer to human perception. However, these methods tend to link between the AU segments and the high level of AU intensity. Also, the detection of the temporal segments is largely impossible using a single frame [380]. Moreover, this method is more complicated to encode because the few available training images contrast with frame-based methods. Subsequently, it works inadequately in recognising subtle AU events.

- The transition-based method tries to detect the onset, apex, and offset segments for each AU event [136]. An event can be denoted as a maximum continuous interval of facial expression action [90]. The facial AU event activation can be divided into four constitute temporal segments: Neutral (when there is no facial expression), Onset (the beginning), Apex (the maximum), and Offset (the ending) of the activation of facial expression. Fig. 3.2 represents the four temporal segment phases ($Neutral \Rightarrow onset \Rightarrow apex \Rightarrow offset$) of a facial expression. Fig. 3.3 shows an example of the four temporal segments of the facial dynamics of AU12 (Lip Corner Puller) event which can be seen in [265].

The automatic AU event detection (temporal segments) and intensity levels is a general challenging problem [267] owing to the following reasons: First, the facial AUs event might occur in a different time scale [115], for example, very short activations or long ones or none [411]. Second, the expression can originate from a different affective state [173]. For instance, facial action units can reveal that deliberate expression is more intense than spontaneous expression, since the subjects have more control of their emotions when expressing a stayed an exaggerated expression. In real life, humans may show deceitful or unfelt expressions,

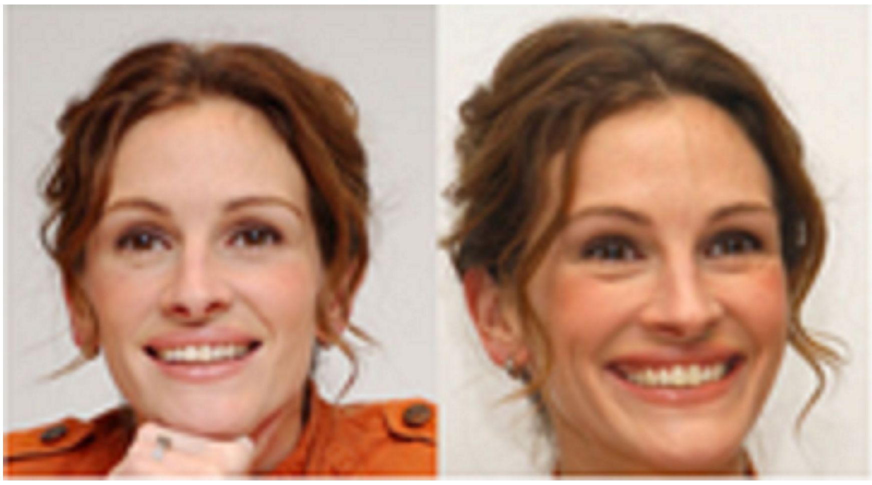


Fig. 3.4 The differences can be noticed between: the left image displays an unfelt or deceitful expression and on the right is a genuine expression; image sourced from [302].

and may show difficulties in expressing real emotions and display different emotions than are appropriate for the given situation, or may have an ability to hide their internal true affective states from others and occasionally to mislead observers, or these emotions may not correspond with the emotional state experienced. This is shown in Fig. 3.4. Moreover, it is observed by Liong et al. [352] that the ordinal temporal structure of eye blinking movement is more intense than a micro expression movement [418]. Third, the co-occurring of different AUs can be additive and is not isolated; AUs may occur separately (non-additive) or in a combination to realize nearly all of the possible facial expressions [135] [380]. The example AU12 + AU6 is a more common combination of the latent signal of the happiness emotion than others. People can show more than 7,000 AU combinations every day [192], and each AU is associated with an intensity score [484]. This is given by a five-point scale levels (A, B, C, D, and E) which expresses the intense degree of activation, starting from 0 (not present) and A (a Trace of the action) to E (refers to Maximum intensity evidence) [380], where C and D scales are an indication of the large appearance of changes between the other levels [246]. Fig. 3.5 represents the facial appearance graded changes of AU4 (Brow Lowerer) and AU12 with the corresponding intensity variation levels, according to the FACS system [246].

Some sequences may include temporal inconsistencies; for example, training and testing sequences may contain all the segment phases of expression whereas test sequences may contain only some of them [484]. Facial expressions may start and end with a neutral face [503]. More importantly, the transition of AU points from inactivation to activation is difficult to detect [136]. Consequently, a reliable facial AU recognition system that develops affective computing systems significantly lacks the training and evaluating on a sufficient annotated data which can be unitized in building such models [398]. Unfortunately, existing datasets

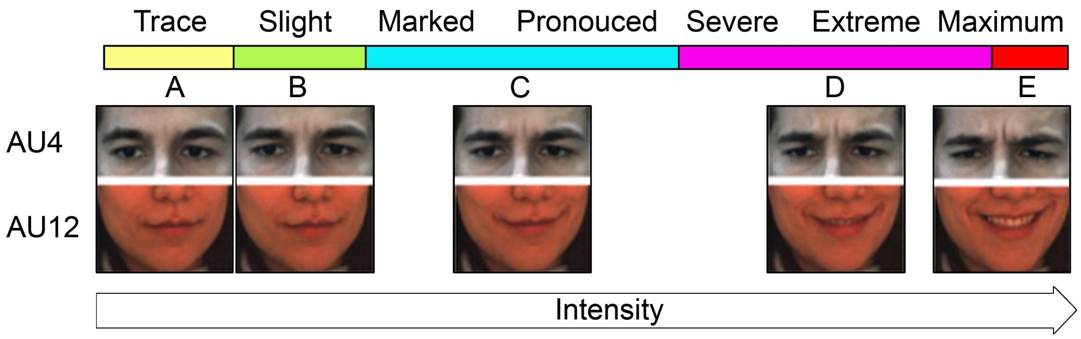


Fig. 3.5 The AU intensity level scores: A. Trace, B. Slight, C. Markedly pronounced, D. Severe, Extreme, and E. Maximum; image adapted from [246].

do not meet the demands; there are no comprehensive, spontaneous, and dynamic databases available annotated in terms of AU transition detection and this entails inadequate features because of the small changes in intensity if we deemed it also [417], where only video level annotations are available and there are no annotations for individual video frames which may drive a poor classification performance, especially if the training samples are less. This occurs because the annotation process is time-consuming, prone to errors, tends to be tedious and can only be performed by a highly certified FACS annotators [15]. The known challenges of automatic facial expression recognition systems consist of big individual variations across people such as face scale, shape, appearance and morphology of facial features, all the complexity with the human effect [283], registration errors are related to complicated landmarks. As well as the availability of irrelevant facial movements, such as open and closed mouth with eye-blinking at the same time [418]. Lastly, individual AUs may need to be detected separately which makes the AU predictions more accurate from static images and one image may include multiple AUs [340].

The CNN-based approaches have outperformed the state of the art performance and surpassed other methods by a big margin, thus confirming their beneficial learning as a source of data representation [334][223]. The four main advantages that have been contributed to using CNNs for facial expression recognition are the avoiding of doing manual extraction of features, and the input to the network was not a set of hand-coded features but instead, it is a raw image [361]. Furthermore, the ability to capture a high level of spatial information due to the use of various filters. It enables "end-to-end" learning straight from input images, which completely removing the use of pre-processing methods. In addition to the cutting-edge recognition reported results, and the retraining of existing networks from other available recognition tasks [414]. In order not to build and train the networks from scratch, for deep spatial representation, we depended on fine-tuning from the available networks that is trained

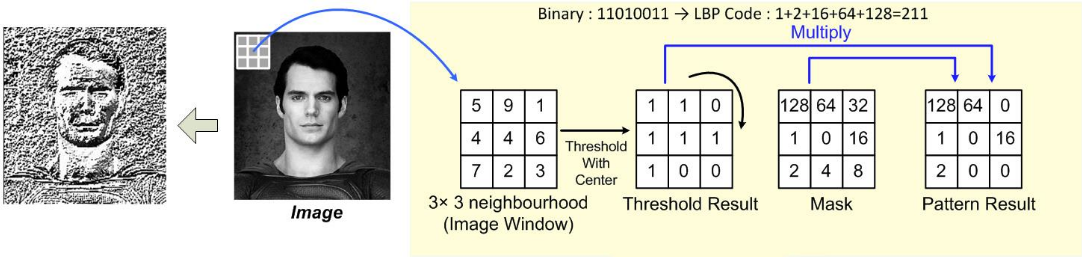


Fig. 3.6 The steps required for computing LBP pattern response from each block of the image, image sourced from [305].

on a large scale dataset. A new trend line is integrating numerous features and classifiers, in which more accurate AU predictions can be gained and the fusion of different features can further benefit the performance [340].

Our aim in this chapter is to address three main complementary aspects: the problem of modelling AU target activation detection, and then, to discover the underlying temporal variation phases in a sequence using supervised and unsupervised methods which highlight and compare the exciting feature extraction representations on both static and dynamic data, which confer the importance of fusing more than one deep architecture. The proposed methods were evaluated by the third aspect: comparing the continuous scoring predictions by acquiring the best match between the predictions and the Ground Truths. We demonstrated that both methods (static and dynamic) can compete with the state-of-the-art available methods and the results were promising when tested on the available enhanced Cohn-Kanade dataset; the achieved results illustrate the effectiveness of the proposed methods.

This chapter is organized as follows: after this introduction in Section 3.1, the methodology of the feature extraction methods proposed in both categories, static and dynamic, are presented together with the proposed hybrid recognition architecture, the used dataset, all detailed in Section 3.2, which also discusses the experimental settings and gives the results in Section 3.3. The conclusions are provided with possible future directions in Section 3.4.

3.2 Methodology

3.2.1 Local Binary Patterns (LBP)

Local Binary Patterns and their extensions were originally proposed for grey scale invariant local image texture analysis. Since then, it has proved to be a very efficient feature descriptor used in many applications because of its computational simplicity and discriminating power for texture classification in real world complex settings. It also remains robust to monotonic

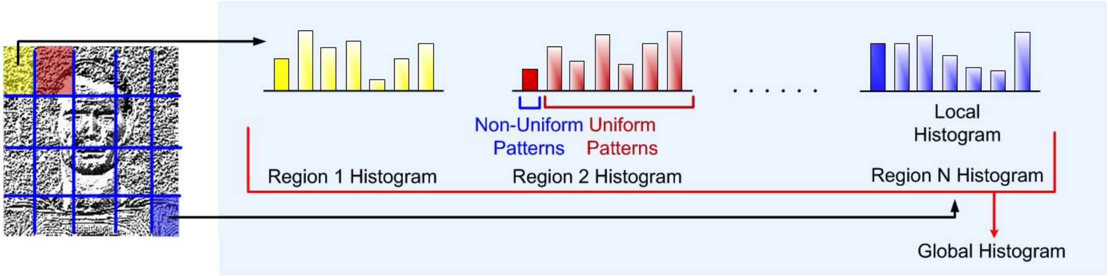


Fig. 3.7 An example of histogram concatenation of LBP descriptor; image sourced from [305].

grayscale changes, in addition to its sensitivity to local structure tolerance to variations in face alignment [15], though it is not robust to rotations and is prone to noise. In practice, an 8-bits binary pattern (LBP code) response of a pixel is computed. In other words, the image labels are made by comparing and thresholding the value of a central pixel intensity with the intensity of all the local pixels in the neighbourhood [419]. If the intensity of the central pixel is larger or equal to its neighbour's, it is encoded by one, or otherwise zero [273]. Fig. 3.6 represents an example of the steps required to calculate the LBP code using a rectangle of 3x3 grey pixel's local surrounding neighbourhood. Later on in the aforementioned process, each bin will correspond to one of the different possible binary patterns and produce a flow of binary numbers with eight surrounding pixels which will result in 256 possible combinations of the LBP dimensional descriptor [303]. Fig. 3.7 illustrates this spatial representation by dividing the image into small local regions; texture descriptors are extracted from each location separately, and concatenated to get the LBP histogram generation of the facial image [400]. A dynamic texture extension approach of the LBP descriptor to the temporal domain was adopted in [269] [645]. A further description review of LBP descriptor can be found in [268] [541] [404] [237] [492]. The LBP operator can be calculated using the thresholding function from the following equation [42] [394]:

$$LBP-code=\sum_{p=0}^{n-1} U(I_p-I_t)\times 2^p, U(x)=\begin{cases} 1 & \text{if } x\geq 0 \\ 0 & \text{if } x<0 \end{cases} \tag{3.1}$$

where: I_t is the central selected threshold grey value. I_p are the grey intensity values of the neighbours surrounding window pixels for $n=0,1,....,7$, where n count the number of pixel neighbors [116].

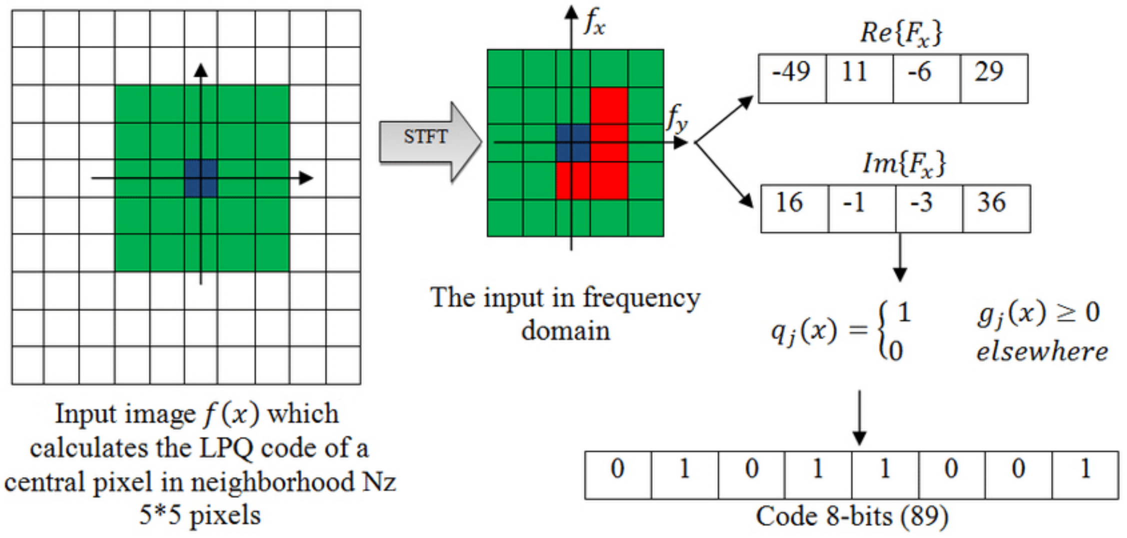

STFT
f_x
f_y
$Re\{F_x\}$
-49 11 -6 29
$Im\{F_x\}$
16 -1 -3 36
The input in frequency domain
$q_j(x) = \begin{cases} 1 & g_j(x) \geq 0 \\ 0 & elsewhere \end{cases}$
Input image $f(x)$ which calculates the LPQ code of a central pixel in neighborhood Nz 5*5 pixels
0 1 0 1 1 0 0 1
Code 8-bits (89)

Fig. 3.8 Steps required to calculate Local Phase Quantisation descriptor; image adapted from [42].

3.2.2 Local Phase Quantisation (LPQ)

The LPQ operator is a static local appearance texture descriptor using the 2D Short-Term Fourier Transform Phase (STFT) on local image window neighborhoods to transform the input image to frequency domain [455], and was first suggested as a texture descriptor by Ojansivu and Heikkila [420]. It has been initially fully used for the classification of the blurry invariant texture property [42] as a particular blurring method [653]. Both LBP and LPQ have been applied successfully for AU recognition and are resistant to image blur. LPQ depends on the blur invariance possession of the Fourier phase spectrum. In LPQ descriptor only four complex coefficients were used related to 2D frequencies [405]. The real and the imaginary part for each pixel position x in the Fourier coefficient is calculated from the input image $f(x)$ through a local rectangular M-by-M neighbourhoods N_z; the local phase frequency spectra is computed by the 2D STFT using the following equation:

$$F(u,x) = \sum_{y \in N_x} f(x-y)e^{-j2\pi u^T y} = W_u^T F(x) \tag{3.2}$$

W_u represents the basis vector of the 2-D DFT at frequency u and $F(x)$ represents a vector containing all M^2 samples from N_x. Where $W_{u,y} = e^{-j2\pi u^T y}$, and $x \in \{x_1, x_2, ..., x_N\}$ is a 1-D convolution [653]. In the LPQ process, the local Fourier coefficients $F(u,x)$ are computed at four angles $[0, \pi/2, \pi, 3\pi/2]$. In 2D frequency points, the angles were indicated as ($u1$ to $u4$), in that $u1 = [a,0]^T$, $u2 = [0,a]^T$, $u3 = [a,a]^T$, $u4 = [a,-a]^T$ where a=1/window size is a small value, and the STFT window size is set to 7. For each pixel position, only the first four frequency coefficients are extracted; this produced a vector of:

$$F_x^c = [F(u_1,x), F(u_2,x), F(u_3,x), F(u_4,x)] \tag{3.3}$$

Next, the real and the imaginary parts of F were separated, the result is obtained from the following [394]:

$$W = [Re\{F(x)\}, Im\{F(x)\}] \tag{3.4}$$

where, $Re\{.\}$ and $Im\{.\}$ are real and imaginary parts of the complex number. The resulting vectors are quantized using a criteria of simple scalar quantizer [380] [394]. The quantization approach can be represented by [657]:

$$q_j = \begin{cases} 1 & \text{if } W_j \geq 0 \\ 0 & \text{otherwise} \end{cases} \tag{3.5}$$

Where q_j is the quantization for the elements in W, and $q_j(x)$ is the j^{th} component of the vector $W(x) = [Re\{F(x)\}, Im\{F(x)\}]$. The four low frequency values of the phase information are encoded and mapped to an embedding histogram of codes of features for the classification. The resulting eight bit binary coefficients coding $q_j(x)$ are represented as integers using the equation of the image label $f_{LPQ(x)}$ as:

$$f_{LPQ(x)} = \sum_{j=1}^{8} q_j^{2^{j-1}} \tag{3.6}$$

As a result, a 256-dimensional feature vector was reached [124]. Fig. 3.8 describes an example of computing LPQ descriptor and Fig. 3.9 illustrates the description of encoding a face after applying an LPQ descriptor on the image.

3.2.3 LPQTOP

The Local Phase Quantisation from Three Orthogonal Planes (LPQTOP) descriptor [102] is an extension of the basic LPQ operator to the time domain where the LPQ features are extracted autonomously from three orthogonal slices, denoted by x-y, x-t and y-t respectively [224] [645]. The main advantages of the LPQTOP descriptor are robustness against image transformations such as rotation, insensitivity to illumination variations, computational simplicity, and multi-resolution analysis [645]. The LPQTOP dynamic texture descriptor was originally introduced to extract the latent temporal information clues (to learn feature representation from video volume), demonstrating facial appearance changes occurring in

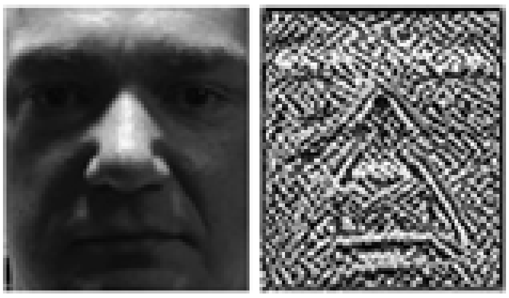


Fig. 3.9 An encoded face using LPQ descriptor; image adapted from [653].

facial AUs over time, in terms of expressing temporal segments of facial AUs [268]. On the other hand, LPQTOP encompasses texture analysis and combines static local appearance with shape attribute features (x-y plane provides texture spatial domain) and motion change features (x-t, and y-t planes provide the temporal information domain), in three directions (x-y, x-t, y-t) to encode the phase transition information per image position for each space and time volume, exhibited in facial expressions [224]. For more details about the LPQTOP discriptor, refer to [268]. The consequence resulting from binary patterns is stacked for the three orthogonal planes and is concatenated in a single histogram [224] as shown in Fig. 3.10. In the end, we got 768 bins = (256 $\times$ 3) LPQTOP features extracted per spatial-temporal volume containing 3, 5, or 7 second window frames. In our experiment, all the images of Cohn-Kanade are in frontal view and therefore it is not necessary to consider in plane head movement. The cropped face region of the input frame of size 256 $\times$ 256 pixels was split in to 10 $\times$ 10, 5 $\times$ 5, 7 $\times$ 7 blocks separately with a different frame rate each sequence. Lastly, the Support Vector Machine and Random Forests were used as binary classifiers for predicting the occurrence of AUs. Besides, the optimal size of temporal windows was investigated in dynamic descriptors as Fig. 3.11, Table 3.1, and Table 3.2 explain the Area Under the ROC Curve for AUs activation detection using LPQTOP descriptor with two classifiers (SVM and RF) based on different parameters: grid 10 $\times$ 10 volume 3-3-3, grid 10 $\times$ 10 volume 3-3-5, grid 10 $\times$ 10 volume 3-3-7, grid 5 $\times$ 5 volume 3-3-3, grid 5 $\times$ 5 volume 5-5-3, grid 7 $\times$ 7 volume 3-3-3. Also, Fig. 3.12 gives an example of extracting the temporal features using LPQTOP descriptor by using two different datasets.

3.2.4 Non linear-Slow Feature Analysis

Facial AU temporal dynamics analysis can be modelled using the nonlinear Slow Feature Analysis method. The SFA was first investigated by [584] as an unsupervised learning approach for describing the most slowly time-varying facial sequence latent space features of rapidly temporal varying signals that grasp time dependencies, ranked by their continuous

Table 3.1 AUC values for each Action Unit resulted from the Support Vector Machine classifier using features that were extracted from the LPQTOP descriptor with different parameters.

AUs	**Support Vector Machines**					
	Grid 10×10	Grid 10×10	Grid 10×10	Grid 5×5	Grid 5×5	Grid 7×7
	Vol 3-3-3	Vol 3-3-5	Vol 3-3-7	Vol 3-3-3	Vol 5-5-3	Vol 3-3-3
AU1	0.9884	**0.9909**	0.9887	0.9707	0.974	**0.9909**
AU 2	**0.9648**	0.969	0.964	0.9211	0.9375	0.9639
AU4	0.8314	**0.84952**	0.83634	0.75216	0.76149	0.75264
AU 5	0.9503	0.9622	0.9807	0.9871	**0.9913**	0.9891
AU 6	**0.8875**	0.8709	0.8493	0.7812	0.7648	0.8174
AU 7	0.99851	0.99887	0.99899	0.99988	**1**	**1**
AU 9	0.98493	0.97917	0.98384	0.97566	**0.98972**	0.97729
AU 12	0.97063	0.95287	0.9584	0.95517	0.94657	**0.98262**
AU15	0.95797	0.95662	**0.96971**	0.95868	0.94202	0.93009
AU 17	0.73664	0.72542	0.72099	0.74139	**0.79407**	0.69756
AU 23	0.8598	0.8385	0.8449	0.8687	0.8892	**0.9313**
AU 24	0.9527	0.933	**0.9551**	0.9313	0.9422	0.953
AU 25	0.9858	0.9877	0.9803	**0.9913**	0.9861	0.9793
AU 27	0.6888	0.6852	**0.6968**	0.6416	0.6206	0.6707
Average	**0.927831**	0.923383	0.924202	0.911818	0.917992	0.921985

Table 3.2 AUC values for each Action Unit resulted from Random Forest classifier using features that were extracted from the LPQTOP descriptor with different parameters.

AUs	**Random Forests**					
	Grid 10×10	Grid 10×10	Grid 10×10	Grid 5×5	Grid 5×5	Grid 7×7
	Vol 3-3-3	Vol 3-3-5	Vol 3-3-7	Vol 3-3-3	Vol 5-5-3	Vol 3-3-3
AU 1	0.9894	0.9941	0.9795	0.9927	0.9821	**0.9974**
AU 2	**0.9071**	0.8661	0.8742	0.8616	0.8878	0.9041
AU 4	**0.96658**	0.93453	0.93282	0.92864	0.93872	0.95613
AU 5	**0.9913**	0.9843	0.9727	0.9738	0.975	0.9752
AU 6	**0.7797**	0.7874	0.74543	0.66224	0.7146	0.70045
AU 7	0.98007	0.98407	0.9337	0.99667	**0.9978**	0.96488
AU 9	0.99063	0.99136	0.98554	**0.99323**	0.98627	0.9904
AU 12	0.90767	**0.92295**	0.8849	0.90059	0.85041	0.89587
AU 15	0.92358	0.92389	0.88936	0.92363	0.9251	0.92938
AU 17	**0.87415**	0.81938	0.82539	0.77865	0.79199	0.75594
AU 23	0.847	**0.8634**	0.8487	0.8359	0.8446	0.8262
AU 24	**0.919**	0.8825	0.9118	0.9148	0.9144	0.9162
AU 25	**0.9838**	0.9759	0.9543	0.9251	0.9599	0.9517
AU 27	0.7172	**0.763**	0.7333	0.7304	0.6872	0.7315
Average	**0.92163**	0.9127	0.899165	0.899033	0.896288	0.891299

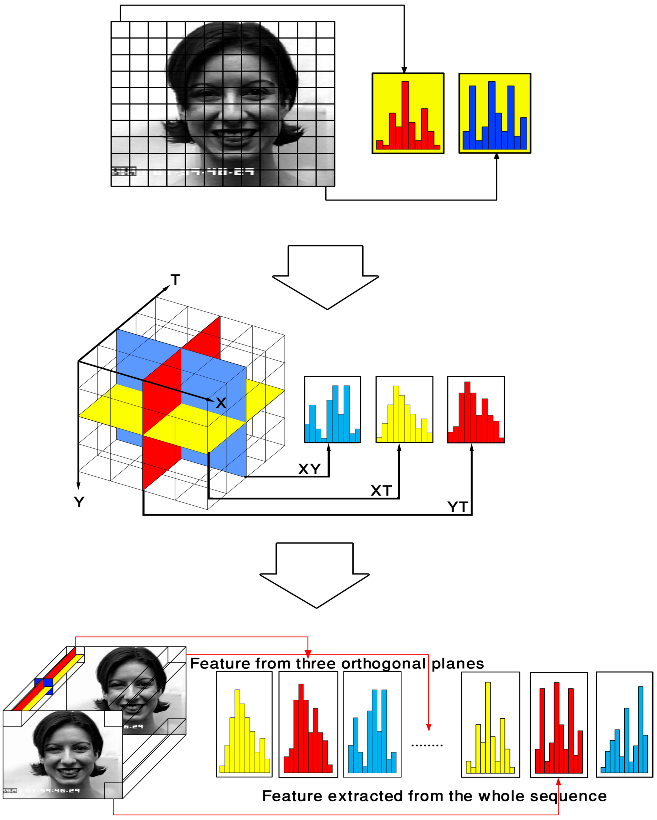
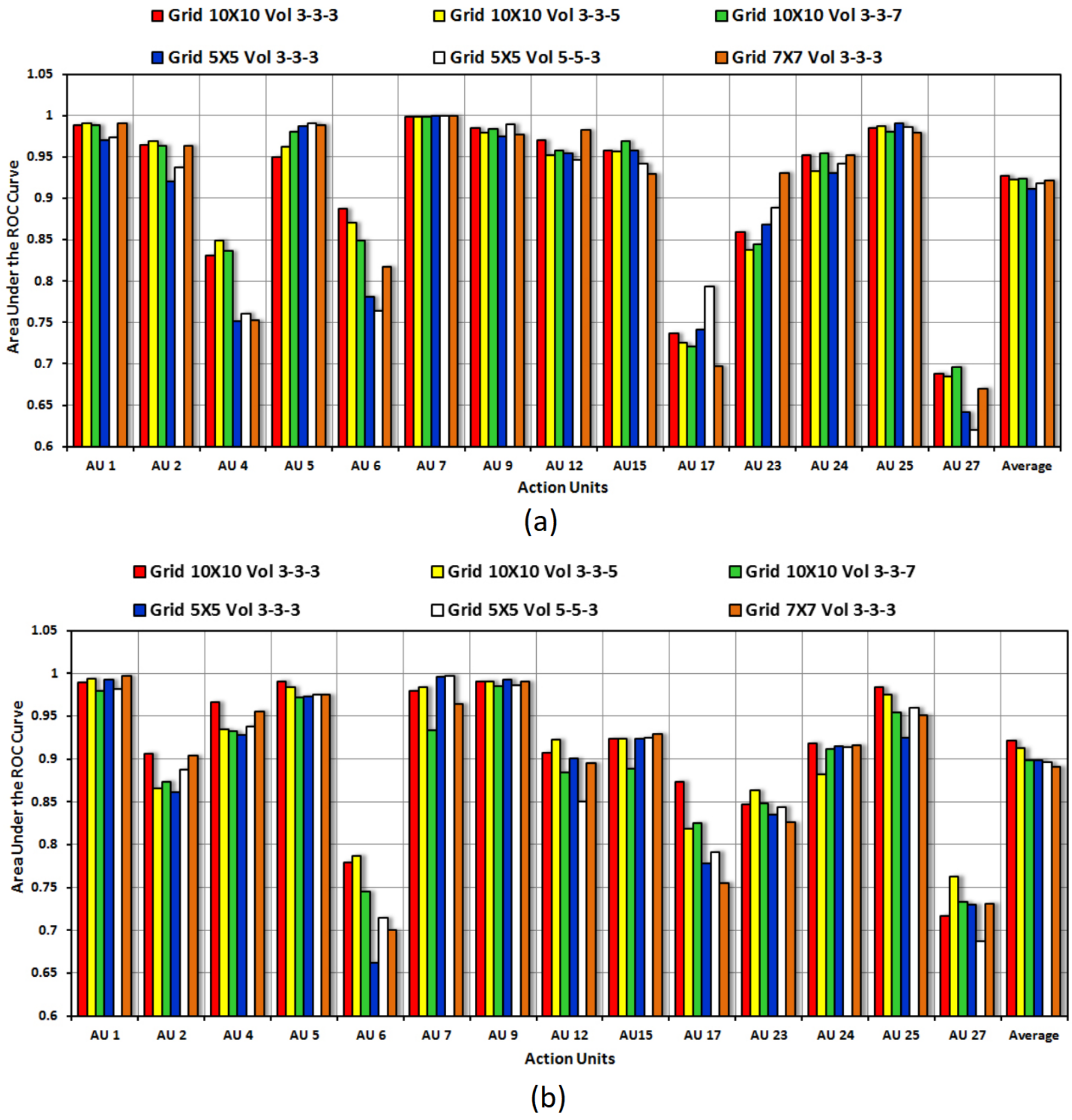

T
X
Y
XY
XT
YT
Feature from three orthogonal planes
....
Feature extracted from the whole sequence

Fig. 3.10 LPQTOP descriptor: Block features extracted from all the three planes are histogram concatenated to create a feature vector which represents the whole sequence.

temporal consistency [616]. More precisely, it aims to minimize the temporal variance of the approximated first order time derivative of the input signal which seeks uncorrelated projections [615] [613]. However, according to [380] state that "Despite its interesting theoretical aspects, the practical applicability of purely unsupervised learning is not clear". Our knowledge shows that until today there is limited interesting work focusing on revealing the dynamics of AUs using nonlinear SFA in an unsupervised way regarding its ability to discover the temporal phases of AUs and their constituent temporal segments (onset, apex, offset) [613]. To do so, we applied the method presented by [584], and this can be accomplished by using an expansion function to extend the input signal data nonlinearly, reducing the dimensionality and track by linear SFA. Fig. 3.13 represents the flow of the algorithms used to detect the activation of AUs using linear and nonlinear Slow Feature Analysis methods. These processes were started by using the features extracted from LPQTOP descriptor on the enhanced CK dataset and then by entering these features to the Principal Component Analysis method to reduce the dimensionality of a matrix of a feature



(a)

(b)

Fig. 3.11 AU activation detection using LPQTOP descriptor with two classifiers: (a) the top is SVM and (b) the bottom is RF based on different parameters.

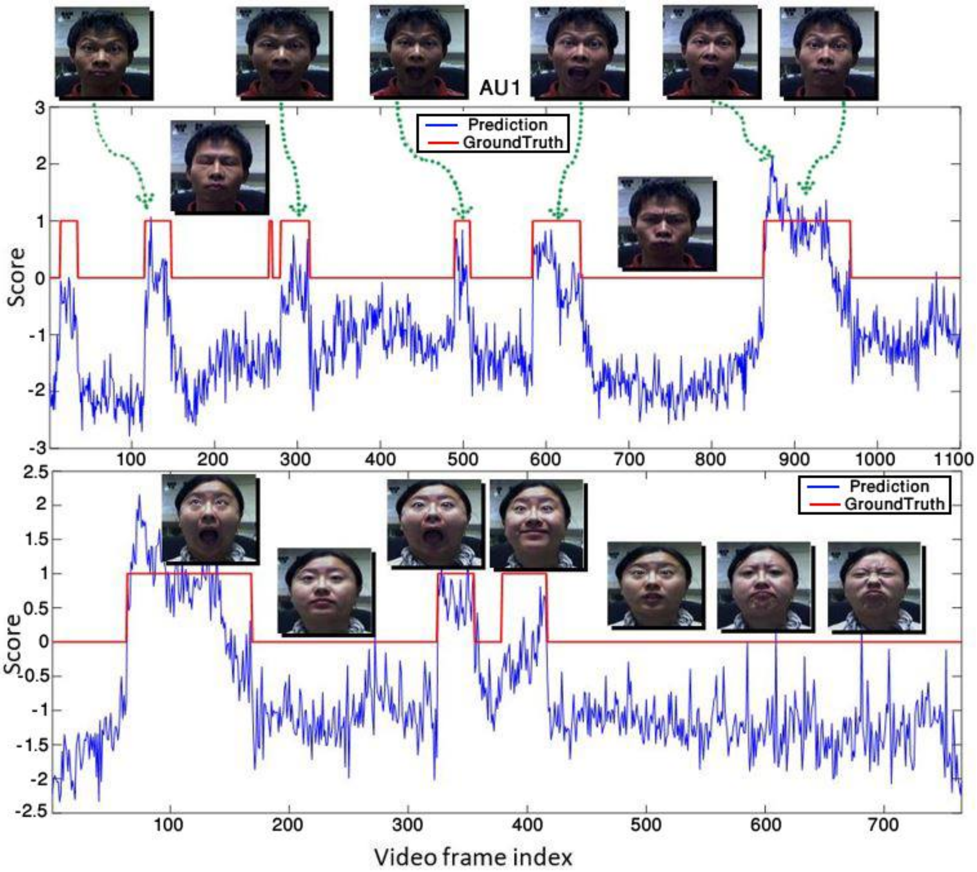


Fig. 3.12 An example of using temporal information. The figure represents a continuous scoring prediction, detection of AU1 on the first half part of the sequence (subject 1), and on the second half part of the sequence (subject 2) in which the feature vector from the enhanced CK dataset was used for training and for testing a feature vector which included a sequence of two videos with two subjects. Each one consists of 900 frames from the ISL Facial Expression dataset using the LPQTOP dynamic descriptor.

vector, then after reducing the features they were entered again to the SVM classifier and the obtained score had been the input to the linear SFA1 method and the resulted features had been entered to the nonlinear SFA2 to be fed finally to the SVM classifier to get the predicted labels of the occurrence activation of the AUs.

3.2.5 LSTM

The Long Short Term Memory is a special type of recurrent neural network modules, proposed by Hochreiter & Schmidhuber [255] to solve the problem of vanishing/exploding gradients (as exploding gradients lead to the weights to be oscillating [209]) encountered by traditional recurrent neural networks, and by adding the gates and a clear definition of

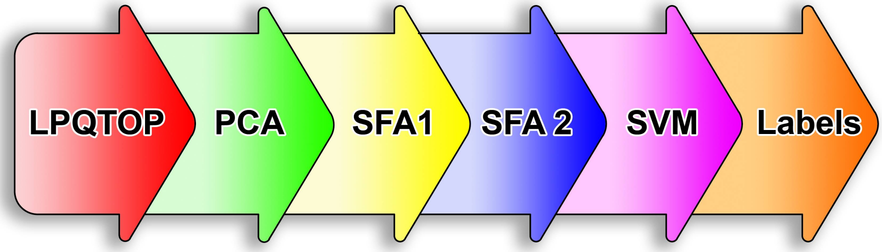


Fig. 3.13 The flow processes of the steps used to predict the occurrence activation of AUs using linear and nonlinear Slow Feature Analysis approaches.

a memory cell, depending on the summation of the memory status [520]. It is embedded to learn long-short dependencies with lower computation cost [255], since LSTM networks are well-designed and suited for classifying and predicting time series data. Notably, LSTM has proven to memorize information for a long time and store context temporal actions, including the previous feature's time step and current states with a time lag [552]. Relative gap-varying length and insensitivity to incompressible noisy sequential data is another benefit of LSTM on RNNs, in contrast with other classifiers such as Hidden Markov Models and sequence learning approaches for numerous applications. Wei et al. [340], assert that having the former state of a facial action expression can absolutely improve the detection of AUs. Recently, LSTMs were used for learning complex sequence processing problems with clear contexts, for example audio analysis [578], speech recognition [197] [472], conversational systems [654] [387], writing [373], machine translation [577], composing of primitive music [148], image caption generation [508], video captioning [607], learning of sequence to sequence [523], forex forecasting [628] [60], video action recognition [340] [370], and signature verification [158] [552].

It likewise possesses two advantages: LSTM is fine-tuned end to end with other models and it supports both fixed and arbitrary length inputs or outputs. A common LSTM architecture is a chain-like figure of a repeated module design of four interacting parts: cell state, forget gate, input gate, and output gate [293] [25] [313]. The LSTM cell and the connection between them called the cell state, the LSTM (long and short term memories can be described by the memory parameters C_t and C_{t-1}) which would take in each time step, the cell state vector, where the information can pass through them without change as shown in Fig. 3.14, which shows a horizontal line located at the top of the main structure of each LSTM unit. The cell remembers the information over arbitrary time intervals through using the point-wise multiplications and the sigmoid function σ. The LSTM can regulate, delete, add and control the information in the cell state through three individual neural layered networks called gates:

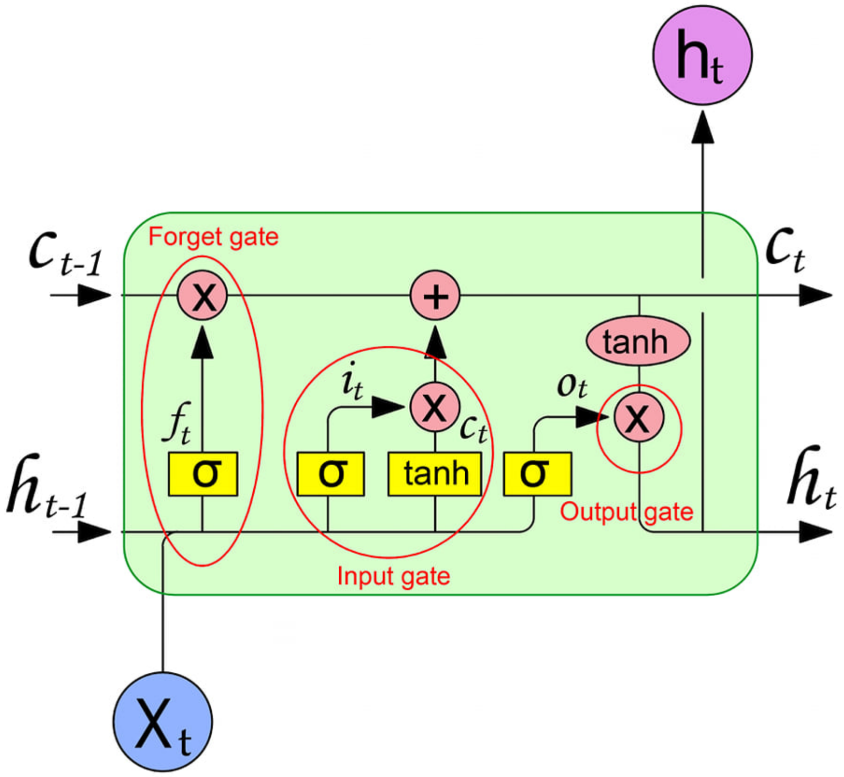


Fig. 3.14 LSTM units cell, where X_t is the new input of the current time step, h_{t-1} is the output from the previous LSTM unit, c_{t-1} is the memory of the previous unit. c_t is the memory of the current unit, h_t is the output of the current network through gates; image adapted from [421].

input gate, output gate and a forget gate of the same structure [293] [550], by generating 1 which means keep or 0 which means remove [209]. The three gates arrange the flow of information into and out of the cell [249]. By combining these three gates, LSTM can model long-term dependencies in a sequence and has been widely employed for video-based expression recognition tasks. The first step in LSTM is to determine what information is going to be saved or erased from the old cell state. This step is controlled by the forget gate [209]. Depending on the previous output of LSTM h_{t-1} and the new input x_t, using the following equation:

$$f_{(t)} = \sigma(W_f \cdot [h_{(t-1)}, x_{(t)}] + b_f) \tag{3.7}$$

W_f and b_f represent the parameters of the forget gate. The next step will be to control what information should be saved or updated for future use in the cell state (C_t), which is controlled by the input gate layer. A sigmoid layer decides which values will be updated.

The three elements C_{t-1}, x_t, h_{t-1} determine the new information in the cell state C_t; these vectors need to pass through the input gate and a tanh layer [340]. Then, a hyperbolic tangent activation function layer will create a vector of new candidate values, $\bar{C}_{(t)}$, that can be added to the state. Then, the updated cell state can be determined from two equations:

$$i_{(t)} = \sigma(W_i \cdot [h_{(t-1)}, x_{(t)}] + b_i) \tag{3.8}$$

$$\bar{C}_{(t)} = \tanh(W_c \cdot [h_{(t-1)}, x_{(t)}] + b_c) \tag{3.9}$$

W_i and b_i are the input gate parameters and W_c, b_c are the candidate gate parameters [340]. The old cell state, $C_{(t-1)}$, is updated into the new cell state $C_{(t)}$ by multiplying the old state by $f_{(t)}$, then $i_{(t)} \cdot \bar{C}_{(t)}$ is added. This given the new candidate values. In the third step the old cell state updates into the new cell state according to the output of the first and second steps.

$$C_{(t)} = f_{(t)} \cdot C_{(t-1)} + i_{(t)} \cdot \bar{C}_{(t)} \tag{3.10}$$

Finally, the output gate is used to enable or prevent an effect according to the new cell state to other neurons depending on whether the information is made visible.

$$O_{(t)} = \sigma(W_o \cdot [h_{(t-1)}, x_{(t)}] + b_o) \tag{3.11}$$

$$h_{(t)} = O_{(t)} \cdot \tanh(C_{(t)}) \tag{3.12}$$

W_o and b_o are the output gate parameters and the output (h_t and C_t) are further used for the next time continuance output generation.

3.2.6 The AlexNet CNN model

Used as a pretrained feature extraction network, this was designed by the Alex Krichevsky [300] and published while he was doing his PhD with Ilya Sutskever and his PhD advisor Geoffrey Hinton. The AlexNet network introduces data augmentation strategy, ReLU activation, local response normalization, and dropout method to prevent over-fitting [519]. It mainly consists of five convolutional layers merged with Rectified Linear Units (ReLU) for the nonlinearity functions to reduce training time as an activation function, with 3 fully connected layers at the top of the layer stack which ended up with 1,000 ways of softmax. Softmax layer is used to predict the label of a video sequence. ReLU is used after each convolutional and fully connected layer. It is interesting to notice that AlexNet was the first to introduce dropout layers suggested by [515] to combat the overfitting risks, and used training time in the fully connected layers, to promote the evolution of huge neural networks.

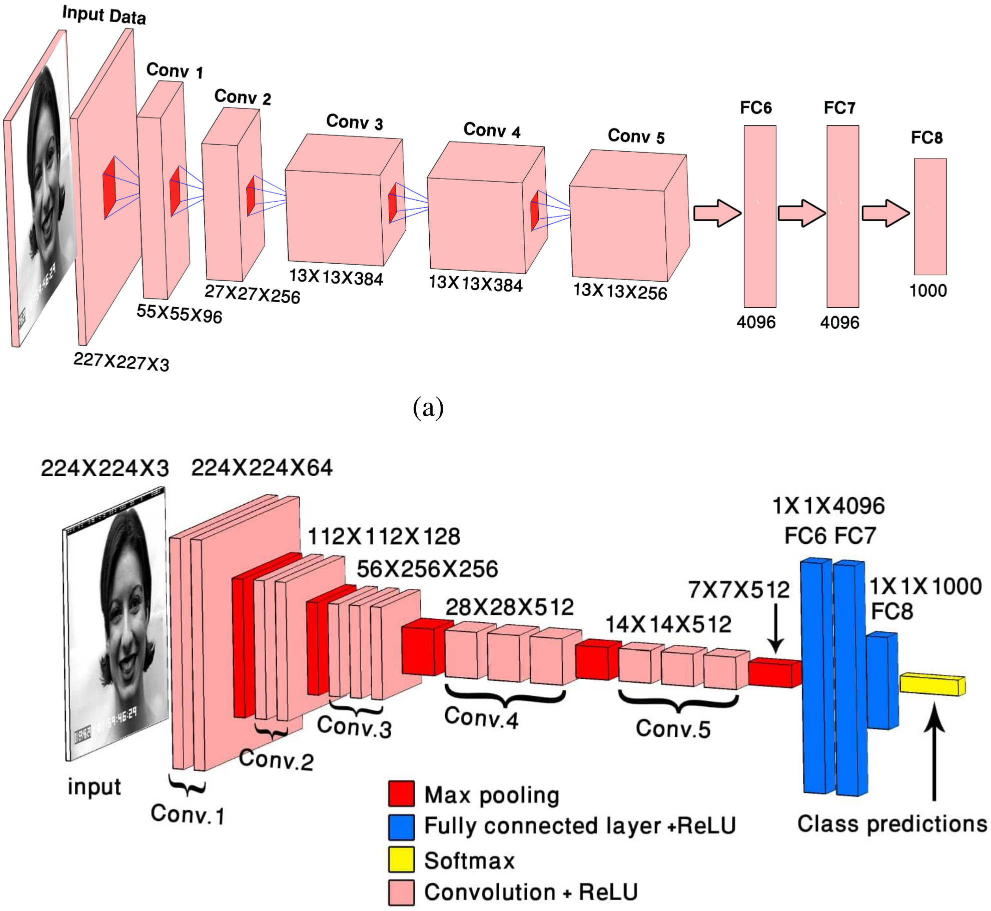
The benefit of Data Augmentation techniques employed during training is to increase more synthetic additional samples to the network by applying in scale rigid image transformations and reflections such as rotation in plane, random perturbation of the pixel values, scaling, and flipping the images [333] [478]. Dropout is implemented before the first and the second fully connected layers. This network was competing solely on ImageNet to classify up to 1,000 various object classes. The input image size to this network should be 227 × 227 × 3. A version of AlexNet model has been pre-trained on the Labelled Faces in the Wild dataset (LFW) and the YouTube Faces dataset for face recognition [342], which was used in this work; therefore, it will be more suitable for facial expression recognition [340] [397] [256] [395].

3.2.7 The VGG16 CNN model

Proposed by the VGG team in the ILSVRC 2014 competition, it differs from AlexNet in that it consists of 16 layers which use rich and complex fixed kernel sized filter banks of 3×3 (11x11 filters in the first layer in AlexNet) for all conventional layers. Using a max pooling of 2x2, the dimensionality of the feature maps is divided by 2 after each max pooling. After the convolutional layers, it is followed by 3 fully connected layers with 1x1 kernel and the output of 512 feature maps [506]. VGG16 is trained on 1.2 million images of size 224 × 224 × 3 belonging to classify 1,000 class categories. The two Fully-Connected layers FC6 and FC7 have been used as a feature extraction layer of depth 4,096 dimensional to learn the deep rich representations of the given targets. A loss layer softmax is added to the end of the network to adjust the back-propagation error and probabilistic predictions [256]. Fig. 3.15 summarises the comparisons between the two Convolutional Neural Networks' proposed architecture chart.

3.2.8 Enhanced Cohn-Kanade database

CK+ is one of the pertinent comprehensive benchmark databases, a baseline for comparing the evaluation performance of cross datasets and generally for evaluating facial expression recognition systems, which is used extensively in the research community. It was introduced by Lucey et al. [366]. It primarily comprises of controlled, frontal and posed on command 593 short emotion sequences from 123 subject participants [334][254]. Only 327 videos from 118 subjects have labels of facial expressions based on the Facial Action Coding System (FACS) [395]. The sequences differ in duration from 10 to 60 frames, starting with a neutral expression phase and ending at the apex phase [205][135], displaying one of the facial emotions: anger, disgust, fear, happiness, sadness, surprise and a non-basic emotion



(a)

(b)

Fig. 3.15 Comparison between: (a) is AlexNet and (b) is VGG16, where the only difference between the two architectures is the use of a small fixed convolution kernel of size 3x3 with a stride of 1, pad of 1 in all layers of VGG16 model and Maxpool filters are of size 2x2 and stride of two. Another important aspect of covenant architecture design as depth, that would range from 11 to 19 layers, compared to eight layers in the AlexNet.

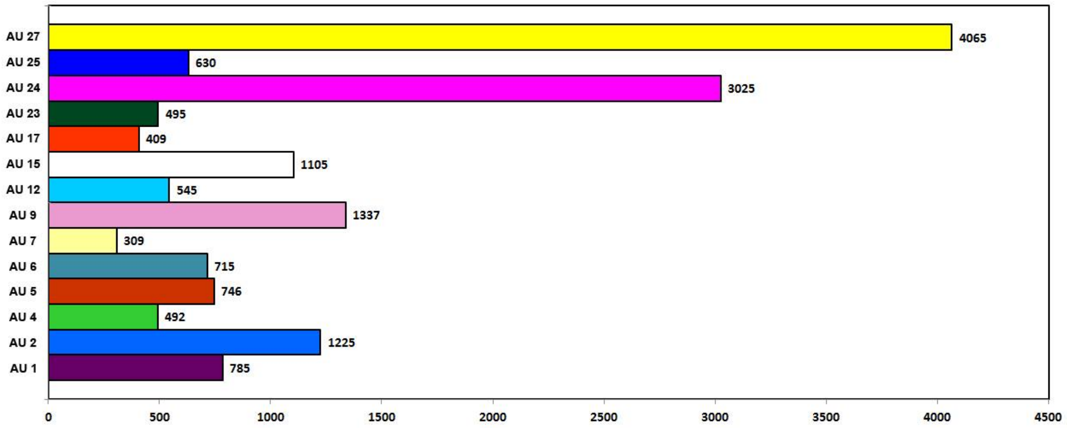


Fig. 3.16 The number of occurrence activations for 14 AUs of the enhanced CK dataset.

(contempt) in addition to 14 AUs. All the sequences within the CK+ dataset were annotated as activation (were FACS coded) [284] for the expressions up to their peak frame [254], and for the presence or absence of AUs by two FACS coders. All the frames were digitized to the size of 640 × 480 pixels, available in both colour and most grayscale images [561], and the coordinates of 68 landmarks were provided to all the images in the CK+ database [640]. Fig. 3.16 represents the number of occurrences for 14 AUs of enhanced CK dataset.

The authors in [380] point out that for more than 10 years, academic researchers have held an all-inclusive range of AU labelling databases but in fact only CK and MMI databases are available. For both of them, the whole sequence is annotated as an active state if the target action unit happens in any frame of the sequence and is classified as a positive of the equivalent video. For instance, AU45 (blink) occurred very quickly in some frames of the video and fundamentally, the entire sequence was labelled as AU45 active, yet the video level annotations (not individual frame level annotations), would not have the same truly frame-by-frame basis for AU annotated ground truth. Also, the information on temporal segment detection annotations is concealed for competition, as mentioned in [555]. For these reasons, for all the experiments in this chapter, the ISL Enhanced Cohn-Kanade AU-coded Facial Expression Database was used, in which the Intelligent System lab at Rensselear Polytechnic Institute produced a new AU manual relabelling which counted by the frame-by-frame annotations, which are mostly used for facial action unit recognition [262].

3.3 Experimental Settings and Evaluation

Three experiments were conducted in this chapter on the available Enhanced CK dataset comparing features extracted by LBP, LPQ, LPQTOP, AlexNet, and VGG16 for each static

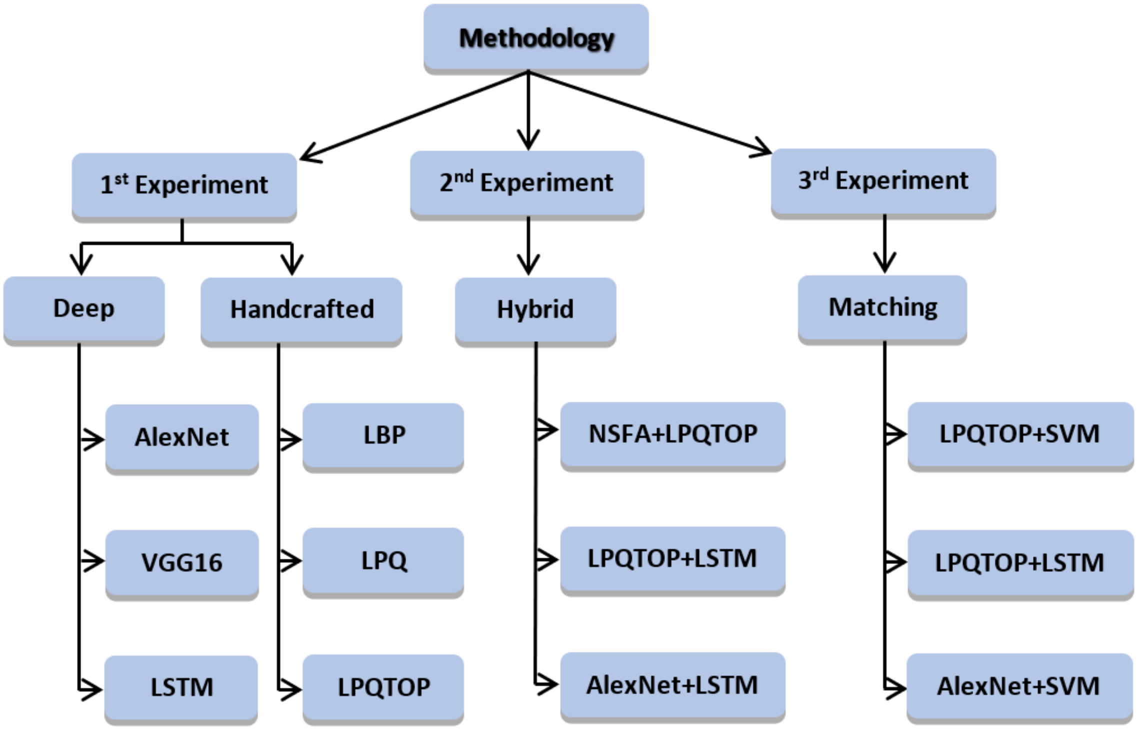


Fig. 3.17 A diagram describes the methodology (three experiments) with all the parts contribute to the whole process.

image of a video for Action Unit activation detection, getting hidden insights of underlying temporal variation detection to be investigated by hybrid nonlinear SFA(NSFA) + LPQTOP, LPQTOP + LSTM, AlexNet + LSTM, from dynamic sequences. Additionally, comparing scoring prediction detection between the features was extracted by LPQTOP + SVM, LPQTOP + LSTM, and AlexNet on the enhanced CK dataset. For the three experiments the system was contrived to extract two types of features from supervised methods, which are extracted by LBP, LPQ, LPQTOP, AlexNet, Vgg16, LSTM and unsupervised methods (linear and nonlinear SFA, PCA) including hand-crafted features represented by LBP, LPQ, LPQTOP and the learned deep visual features extracted by CNN and LSTM on both static and dynamic data. The evaluation was limited to the problem of AU activation detection because there is no similar database with corresponding ground truths tuned to AU target occurrence detection. Fig. 3.17 explains a workflow diagram for summarises the followed methodology which effectively exploited for the FER system with all the related parts that contribute to the whole process.

The evaluation metric, Area Under the ROC (Receiver Operating Characteristic) Curve (AUC), has been extensively used throughout the whole thesis. The ROC curve is a useful tool for predicting the probability of binary classifiers and for visualising its performance.

It is the output of a set of models, which linked to each other through some parameters; a threshold score that one will allocate the positive class label. All the points on the ROC curve corresponding to the values of those parameters (i.e., different test cutoff values). Typically, it utilises the probability estimates (scores or classifier predictions vector) produced from the logistic regression model. The variation in the thresholds would lead to form the continuous curve of connecting points. The number of the chosen thresholds determine the maximum amount of score intervals. In general, the divides of the ROC curve are made where each split interval has the same number of test sample lines. This depends on using all the unique data values or to set how many unique values we need. In this thesis, we used all test scores with all given ground truth labels (true class labels), in addition to the positive class label for each experiment. The ROC curve plots true positive rate (TPR) at the y-axis against false positive rate (FPR) at x-axis using different numbers of classification candidate thresholds between (0.0 -1.0). It plots the TPR (sensitivity) versus the FPR (1 – specificity) for all feasible cut-off values. The AUC has been used as a measure of the model performance, where the highest AUC is equal to 1 represents a perfect classifier. While the diagonal line in the plot indicates that the ROC curve classifies the intended test condition randomly.

3.3.1 First Experiment: Comparing Hand-crafted Features with Deep Learning Representation Approaches

The aim of the first experiment was to predict the presence or absence of AU occurrence at frame level and to test the performance on the supervised proposed model. On this basis, the appearance features were extracted from both static and dynamic information from the same dataset with respect to frame-by-frame base. The experiment was conducted by splitting the dataset into 83% of data for training and 17% of data for testing in which we used 7,000 frames for the training stage and 1,420 frames for testing and the information of test subjects, which was excluded from training while the images of one subject were used in training or testing at the same time. We first located and cropped the face from all the input frame sequences of size 490 $\times$ 640 and utilized an adapted Viola Jones detector. Subsequently, all input frames were resized to be 250 $\times$ 250 pixels (this was also done for experiment two and three). In our experiment, all the images of Cohn-Kanade were in front and this eliminated the problem of head pose non-rigid face registration. Next, to encode texture appearance information for LBP, and similarly for LPQ, and LPQTOP, the images were divided into regions to extract LBP, LPQ, and LPQTOP histograms, respectively. The LBP, LPQ, LPQTOP features extracted from each block are stacked into a single feature histogram. Then, the resulting final histogram is used as a feature vector to represent facial image. For

LBP a region size of 32 × 32 is used. That is, the face image is divided into 10 × 10 blocks. Normalisation was done for the obtained histograms in the range between [-1 : 1], and then we get a feature vector of 256 dimensions. For LPQ a local window of size equal to 7 and 4 × 4 blocks is the optimal choice. For the LPQTOP spatial/temporal descriptor the important parameters are temporal window length (volume size) and spatial block grid size. The average performance is evaluated in a subject independent manner using different parameters. So, the experiment is carried out to find the optimal length and width of the histogram block: ((grid 10x10 Vol 3-3-3), (grid 10x10 vol3-3-5), (grid 10x10 Vol 3-3-7), (grid 5x5 Vol 3-3-3), (grid 5x5 Vol 5-5-3), (grid 7x7 Vol 3-3-3)) as illustrated in Fig. 3.11. Next, the typical linear kernel SVM and RF classifiers are trained separately to detect the occurrence of 14 AUs (AU1, AU2, AU4, AU5, AU6, AU7, AU9, AU12, AU15, AU17, AU23, AU24, AU25, AU27) irrespective of the absences or the presence of other AUs. In general, we can observe from Table 3.1, Table 3.2, and Fig. 3.11 that (a) the Support Vector Machine (top plot) presents better performance than the Random Forest (b) for detecting the activation for all the Action Units. In our case, Area Under the ROC Curve (AUC) is our performance metric on a frame by frame base and is a better ranking-based measure than other metrics, especially in a balanced class binary classification context [505] [92]. In Fig. 3.18, the prominent LBP is clearly superior to LPQ for most action units; similarly, we present the increased relative performance gained by comparing the performance of LBP and LPQ with dynamic features of LPQTOP respectively.

It was reported by [268] and [269] that the LPQTOP dynamic appearance descriptor has been presented as superior for the AU activation detection problem and AUs temporal segments recognition. In addition to that, in [269], it was shown that LPQ achieves higher performance than LBP while [16] concluded that the fixed length window is not appropriate for changing facial actions speed. Our experiment showed that LBP clearly overcomes LPQ and LPQTOP. We also selected two popular pre-trained CNN architecture models: the AlexNet and VGG16 to extract the probability predictions of the cropped faces, in the same way for spatial facial feature representation. Using a pretrained network model can attain very good foremost parameters to expedite the operation of training and testing. It can be observed that the heavy computation burden and the time elapsed of extracting the features using the activations from the fc6 and fc7 fully connected layers as spatial facial learned features are becoming less and reduced significantly. As illustrated in Fig. 3.18, the plots of the AUC for 14 AUs and the five methods (LBP, LPQ, LPQTOP, AlexNet, VGG16), in Table 3.3, show the AUC values for 14 AUs in this experiment as shown in Fig. 3.18; Table 3.4, also represents the accuracy values respectively. Fig. 3.19 explains an example

Table 3.3 AUC values for the first experiment shown in Fig. 3.18.

AU	LBP	LPQ	LPQTOP	AlexNet	VGG16
AU1	0.98793	0.92	0.98841	0.99	**0.99157**
AU2	**0.99297**	0.8277	0.9638	0.99022	0.98671
AU4	0.98925	0.84576	0.82542	**0.99605**	0.98685
AU5	0.98431	0.75515	0.95292	**0.99781**	0.99642
AU6	0.78884	0.7279	0.90291	**0.99605**	0.9911
AU7	**1**	0.92124	**1**	0.99913	0.99909
AU9	0.99283	0.8717	0.98857	**0.99436**	0.99181
AU12	**0.99525**	0.89404	0.97069	0.99478	0.98468
AU15	0.96626	0.88785	0.96493	0.98466	**0.99089**
AU17	0.86467	0.75653	0.73945	0.99206	**0.99286**
AU23	0.95694	0.91858	0.86218	**0.99117**	0.99003
AU24	0.9471	0.70181	0.95272	**0.99508**	0.98716
AU25	0.96899	0.8732	0.9772	**0.99824**	0.9949
AU27	0.76856	0.83406	0.68884	0.97135	**0.97286**
Average	0.943136	0.838251	0.912717	**0.992211**	0.989781

Table 3.4 Accuracy values for the first experiment shown in Fig. 3.18.

AU	LBP	LPQ	LPQTOP	AlexNet	VGG16
AU1	**0.9835**	0.9381	0.9683	0.9553	0.9620
AU2	**0.9611**	0.8917	0.9525	0.9434	0.9220
AU4	0.9647	0.9389	0.9492	**0.9735**	0.9711
AU5	0.9282	0.8972	**0.9775**	0.9727	0.9600
AU6	**0.9666**	0.9514	0.9292	0.9644	0.9612
AU7	**0.9993**	0.9976	0.9943	0.9873	0.9802
AU9	**0.9692**	0.8278	0.9677	0.9232	0.9505
AU12	**0.9753**	0.9654	0.97061	0.9719	0.9663
AU15	0.9036	0.8984	**0.9434**	0.9327	0.9382
AU17	0.9487	0.9128	0.9621	**0.9778**	0.9767
AU23	0.9569	0.9340	0.9653	**0.9782**	0.9715
AU24	0.8972	0.5321	0.8879	**0.9343**	0.9133
AU25	0.9575	0.8950	0.9620	**0.9873**	0.9838
AU27	**0.9957**	**0.9957**	0.6986	0.9240	0.9244
Average	0.957679	0.898293	0.93815	**0.959**	0.9558

Table 3.5 AUC values for the second experiment shown in Fig. 3.21.

AU	NSFA,LPQTOP	LSTM,LPQTOP	AlexNet,LSTM
AU1	0.90607	**0.98616**	0.57259
AU2	0.85748	**0.96482**	0.42819
AU4	0.70504	**0.97462**	0.48538
AU5	0.91509	**0.97646**	0.57634
AU6	0.69804	**0.87946**	0.40306
AU7	0.72461	**0.94159**	0.40704
AU9	0.88865	**0.99087**	0.53262
AU12	0.84715	**0.99183**	0.49982
AU15	0.85302	**0.95114**	0.54175
AU17	0.71731	**0.94882**	0.52354
AU23	0.7245	**0.96361**	0.56597
AU24	0.83395	**0.9655**	0.47486
AU25	0.85649	**0.99641**	0.62957
AU27	0.68757	**0.75688**	0.58425
Average	0.801069	**0.949155**	0.51607

of comparing the accuracy of AU1 versus training data size using LPQTOP descriptor and using two classifiers: SVM and RF.

The best performing method for this task is the AlexNet (in both accuracy and AUC values) which vastly outperforms all others in both training and testing evaluation with an average score of 0.992211, and average accuracy of 0.959 for all the AUs, while the second best score was 0.989781 AUC value achieved by the VGG16 without any need to increase auxiliary GPU units. Our results demonstrate that our models were adept at learning the supervised task; we were therefore able to avoid any risk of overfitting.

Leave One Subject Out cross-validation method is used to analyse every subject separately. Leave-one-subject-out method constantly divides the data as an alternative of doing k-folds, the dataset is divided depending on the number of the available subjects in the dataset. Moreover, a single subject is randomly selected for the testing objectives whilst the other subjects are employed for training the model. The procedure is repeated up until all the subjects were used for the test dataset. In this method, training is performed on the entire dataset while leaving just one subject of the given dataset and subsequently repeats for every subject. Subjects are generally independent except when the model includes uncommon observations such as identical twins and social interactions. Leave one subject out is regarded as the best method for evaluating facial expression recognition system.

This method holds some advantages and disadvantages. The major advantage of utilising this approach is, it requires the usage of all data subjects and consequently, it produces low subject biased estimates because the predictions would be made on new subjects continuously. While the main drawback of this method leads to a higher variation in the testing model because it is testing against one data subject. In case the data point was an outlier, it could be led to a higher variation. One more drawback is that the training algorithm required much computations and execution times since it repeats on all the number of data subjects to do an evaluation. Therefore, leave-one-subject-out cross-validation is not working where there are numerous observations from each individual but small numbers of subjects are available. Likewise, features extracted from datasets are often highly clustered and are not effectively sharing the same distribution across individuals.

3.3.2 Second Experiment: Temporal Modelling by Fusing Multiple Methods

For the second experiment, to provide a better inspection of the performance of the tested methods for modelling the temporal facial behaviours, and to test the hypothesis of dynamic advantages, a new feature integration strategy was employed to preserve the temporal order dependency relations, present in the different frames of the sequences, by feeding the feature vector extracted by LPQTOP and jointly trained them using the LSTM model to classify and yield a prediction per-frame for 14 AUs. This could also show the overall AU activation detection which could capture maximum information from the deep dynamic appearance features construction. The proposed LSTM architecture was trained for 150 epoch iterations on mini-batches of 25 samples. Next, the output scores of CNNs, especially AlexNet and LSTMs, were further aggregated into an averaging fusion network in which both are spatially and temporally deep enough to train CNN and LSTM simultaneously in an end to end framework, accelerating improved future predictions throughout the two networks. Fig. 3.20 depicts an overview of the structure proposed for the hybrid connection system between CNN and LSTM. To this end, the main reason we did not endeavour to establish a relative comparative evaluation baseline of this experiment, with the state-of-the-art deep facial action unit recognition methods, was because there was no existing research paper that could help as the baseline ground truth for AUC results, as the majority of them use an F1 measure for metric evaluation. However, it has constant and discontinued regions making F1 unreliable to use for the gradient based methods.

Between them, the nonlinear Slow Feature analysis method was applied as unsupervised learning on also the LPQTOP feature vector, after reducing the dimensionality of the feature vector using Principle Component Analysis which preserved 85% of explained variability leading to a reduced basis of 1,391 dimensions followed by linear Slow Feature analysis. The first identified latent feature which was obtained corresponded with the most slowly varying one, since the non linear SFA orders the derived latent variables by their temporal slowness. The performance analysis of this model performs well for detecting the temporal information of AUs. Learning a high-level representation from dynamic textures directly by SFA is not practical because of the curse of dimensionality. It was demonstrated that it is possible to use nonlinear SFA for accurately discovering the dynamic of facial action units. Fig. 3.21 and Table 3.5 represent the plots and the values of the Area Under the ROC Curve of 14 AUs for the three methods of this experiment.

3.3.3 Third Experiment: Assessing and Visualising the Maximum Expression Through Continuous Scoring Predictions

To assess the ability for maximum expression of the desired target AUs and the classification quality of the described methods, for the third experiment, three types of validation matching were compared the predicted scores which represented the probability of activation for three methods, and the Area Under the ROC Curve was calculated for all the AUs where AU1 and AU25 were chosen randomly to represent the results of this experiment as shown in Table 3.6. Within every frame in the CK dataset, the AUs were annotated as 0 (not present) ,1 (active) and -1 (not sure). For plotting, in order to make the units standardized for comparison, every frame with -1 ground truth was made equal to 0.5, then three classes had 0, 0.5, and 1 ground truth for the three methods. As can be observed from Fig. 3.22, the time series plot of AU1 (inner eyebrow) and AU25 (lips parted), the detection for each algorithm provides almost different predictions and AU1 and AU25 are unique features that can be compared across all the three algorithms giving them the potential to confidently measure AU1 and AU25 accurately. 317 videos were used for training and 150 videos were used for testing. Therefore, in total 5,891 frames were used during the training phase and 2,529 frames were used for the testing. The representation learned by the proposed methods in Fig. 3.22 was capable of exact prediction of the dynamics of the AU1 and AU25, since it provides more accurate features which in turn matched better with the true label Ground Truth (red line). It seems that the LSTM method is less consistent than the other algorithms. Overall, the performance showed that all the three methods provide better results and are intersected in approximately all the time points that are indicative for detecting and predicting the presence of both AU1 and AU25. To facilitate this analysis further, and to see more accurate matching of the scoring predictions for the three methods, a threshold was applied and drew a bar for

Table 3.6 AUC values for AU1, AU25 and the other AUs for the third experiment.

AUs	LPQTOP + SVM	LPQTOP + LSTM	AlexNet + SVM
AU1	**0.9790**	0.9733	0.9646
AU2	0.9738	0.9748	**0.9912**
AU4	0.83542	0.9746	**0.9961**
AU5	0.9592	0.9765	**0.9979**
AU6	0.9029	0.8795	**0.9961**
AU7	**1**	0.9616	0.9991
AU9	**0.9986**	0.9909	0.9944
AU12	0.9707	0.9918	**0.9950**
AU15	0.9749	0.9511	**0.9947**
AU17	0.7395	0.9588	**0.9921**
AU23	0.8622	0.9636	**0.9912**
AU24	0.9527	0.9755	**0.9951**
AU25	0.9790	0.9579	**0.9985**
AU27	0.6888	0.7569	**0.9714**
Average	0.91548	0.964771	**0.991243**

each method score in Fig. 3.23. Table 3.7 show comparison of the Area Under the ROC Curve values of the chosen proposed methods (D. LBP, E. AlexNet, F. VGG16, and G. LSTM & LPQTOP) with the state-of-the-art approaches (A. SPTS [366], B. Relative AU [284], and C. STM [91]) for AU detection on the extended CK dataset. A comparison of the obtained accuracy was also presented in Table 3.8, with different state-of-the-art techniques on the extended CK dataset including sparse coding, manifold learning, deep and unsupervised learning.

3.4 Conclusion and Future Work

In this chapter, there are three main essential problems which were investigated. First, AU activation detection was done by using a pretrained network model for deep spatial representation that augments reliably the overall unprecedented performance level of recognition rate and accuracy. Significant AU prediction scoring improvements have been gained which increase the demands of using deep learning, in comparison with the traditional nominal hand-crafted and engineered features. Further, temporal modelling was achieved by effective fusing of models which have both temporal and temporal features to retain more long term temporal pattern dependencies. Besides this, we discovered that merging spatial and dynamic

Table 3.7 A comparison of the Area Under the ROC Curve values of the proposed methods (D. LBP, E. AlexNet, F. VGG16, and G. LSTM & LPQTOP) with the state of the art methods (A. SPTS [366], B. Relative AU [284], and C. STM [91]).

AUs	Area Under the ROC Curve						
	A	B	C	D	E	F	G
AU1	0.94	0.95	0.899	0.98793	0.99	**0.99157**	0.98616
AU2	0.97	0.97	0.875	**0.99297**	0.99022	0.98671	0.96482
AU4	0.86	0.89	0.811	0.98925	**0.99605**	0.98685	0.97462
AU5	0.95	0.97	—	0.98431	**0.99781**	0.99642	0.97646
AU6	0.92	0.94	0.94	0.78884	**0.99605**	0.9911	0.87946
AU7	0.78	0.81	0.916	**1**	0.99913	0.99909	0.94159
AU9	0.98	0.98	—	0.99283	**0.99436**	0.99181	0.99087
AU12	0.91	0.93	0.92	**0.99525**	0.99478	0.98468	0.99183
AU15	0.80	0.83	0.982	0.96626	0.98466	**0.99089**	0.95114
AU17	0.84	0.86	0.96	0.86467	0.99206	**0.99286**	0.94882
AU23	0.91	0.92	—	0.95694	**0.99117**	0.99003	0.96361
AU24	—	—	—	0.9471	**0.99508**	0.98716	0.9655
AU25	0.97	0.97	—	0.96899	**0.99824**	0.9949	0.99641
AU27	**1.00**	**1.00**	—	0.76856	0.97135	0.97286	0.75688
Average	0.899	0.915	0.913	0.943136	**0.992211**	0.989781	0.949155

Table 3.8 Comparison of the accuracy with the state-of-the-art approaches.

Methods	Accuracy%
Baseline [366]	0.833
3DCNN-DAP [353]	0.88
3D Shape [259]	0.868
ExpNet [77]	0.612
ITBN [580]	0.863
DTAN + DTGN [276]	0.95
H-CRF [569]	0.88
NMF + ℓ_1 norm [619]	0.924
Spatio-temporal CNN [208]	0.737
DLBP + LL + TFP [138]	0.929
HOG-TOP, geometric features SVM [80]	0.957
Gabor [372]	0.938
MSDF + BoW [504]	**0.959**
Ours (AlexNet)	**0.959**

features fails to achieve a comparable result, which is better at being issued independently. Next, for the evaluation of the classification quality, a successful comparison of validation matching was achieved for continuous scoring predictions for AUs activation detection which proved to be efficacious. Eventually, the configuration of the different algorithm's parameters was chosen as the best choice after several tries. We found that these parameters (for the three experiments) were able to fit specifically the fundamental aspects of facial expression recognition, which permits a high classification performance. For future work research, modelling of multiple correlated AU activation detection helps with the detection of one AU which will determine the other AUs. This representation seems to construct a fully facial event display for the automatic recognition of a various affective state. In addition, future research could focus on using Generative Adversarial Networks (GANs) to generate data that are sufficiently labelled with the segment-level annotations for learning temporal characteristics for the detection of subtle AU event. Generative Adversarial Networks will be used in the next chapter to reduce the need of training data labels as Deep learning approaches require a lot of training data. The novel contribution of this chapter is published in the IET Journal of Engineering.

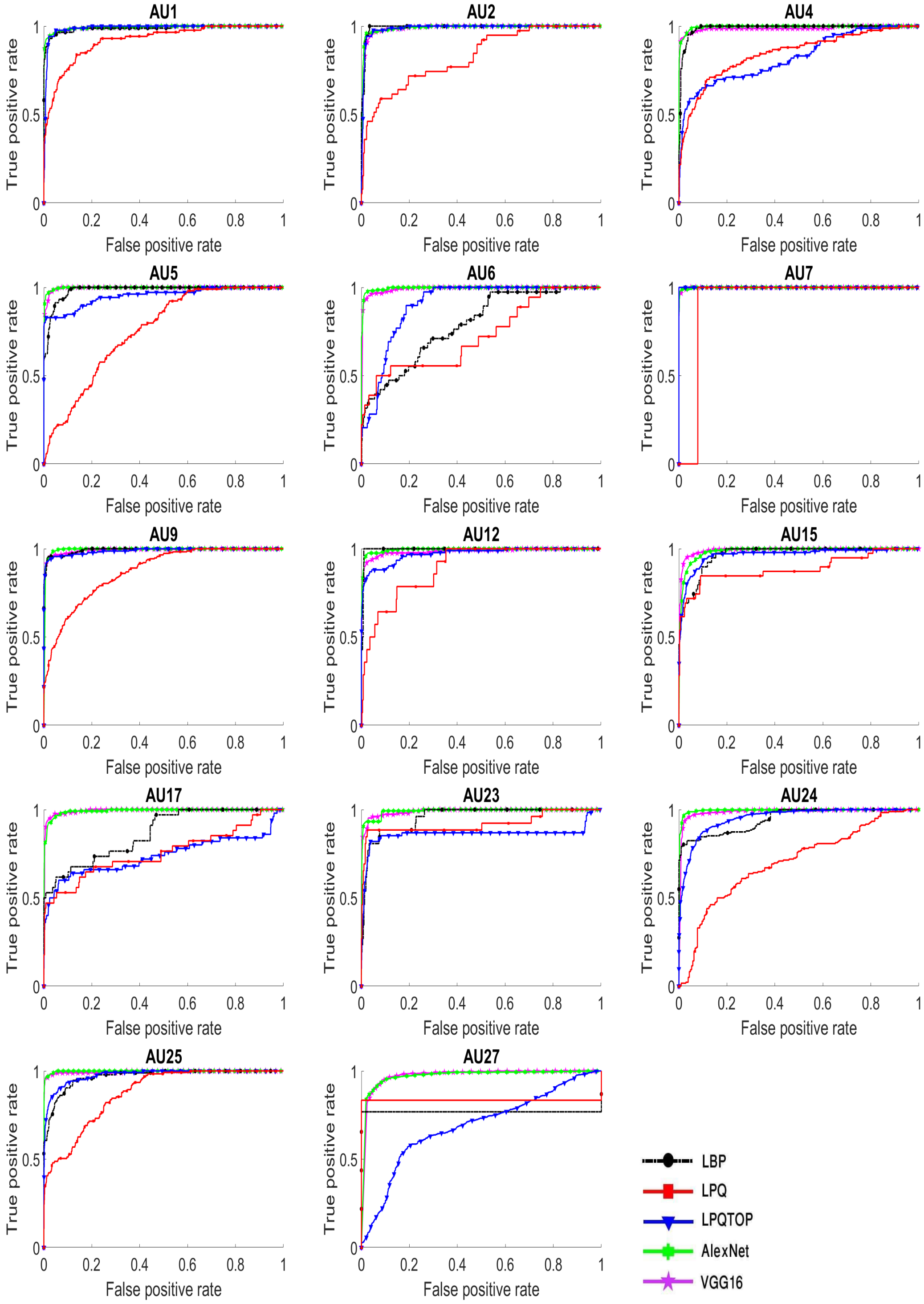


Fig. 3.18 Receiver Operating Curves (ROC) for fourteen Action Units (AU) and five dissimilar methods. Each ROC depicts five methods, black: LBP with SVM, red: LPQ with SVM, blue: LPQTOP with SVM, green: AlexNet with SVM, purple: VGG16.

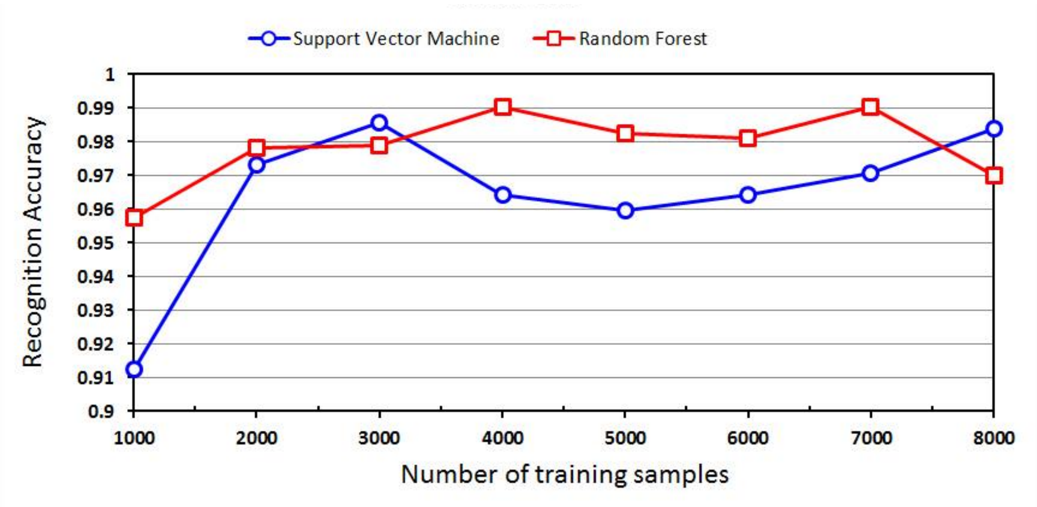


Fig. 3.19 Average recognition accuracy of AU1 versus training data size using two classifiers.

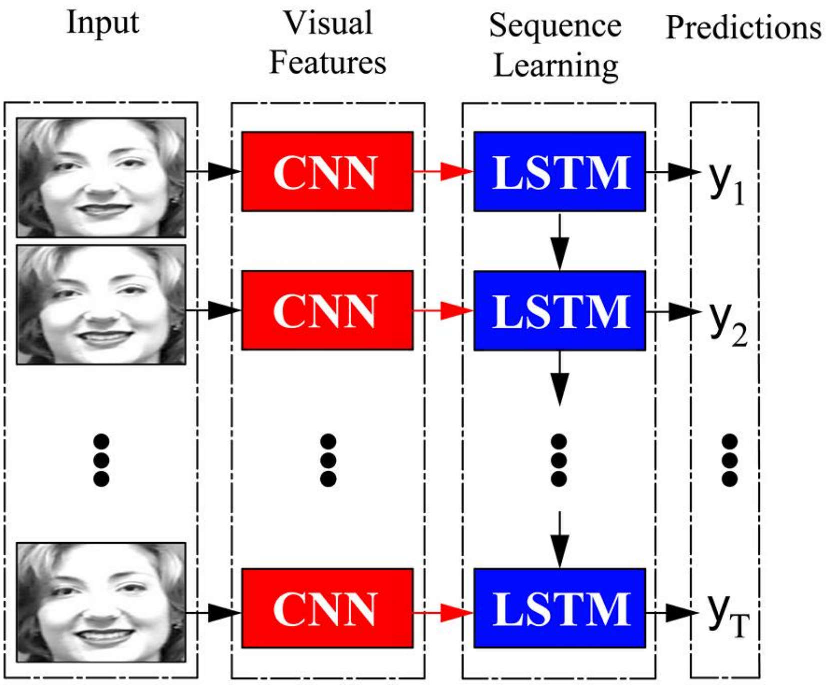


Fig. 3.20 The architecture of the hybrid framework of CNN and LSTM, image adapted from [141].

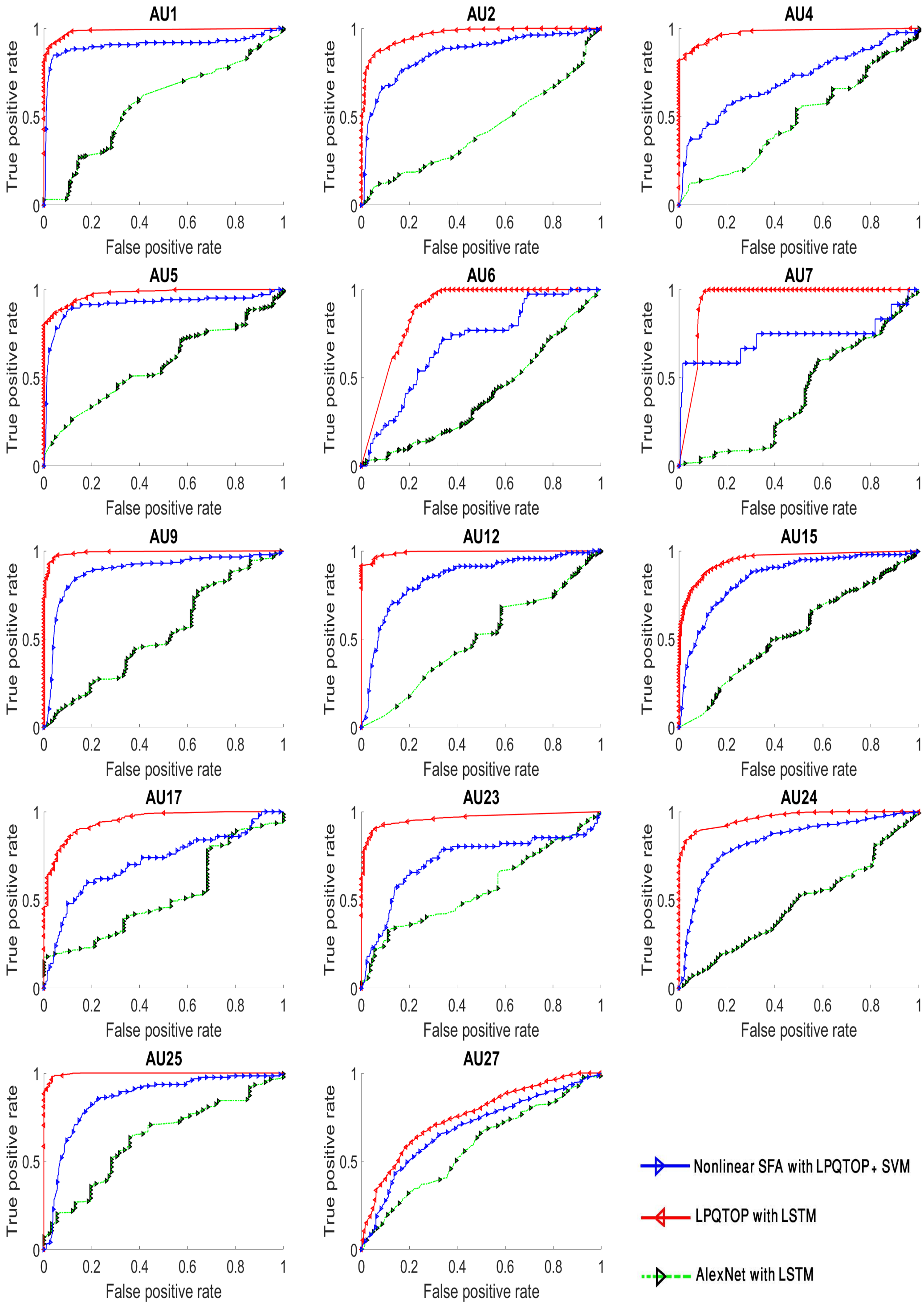


Fig. 3.21 Receiver Operating Curves (ROC) for fourteen Action Units (AU) and three dissimilar methods. Each ROC depicts three methods, red: LPQTOP with LSTM , blue: Nonlinear SFA with LPQTOP, green: AlexNet with LSTM.

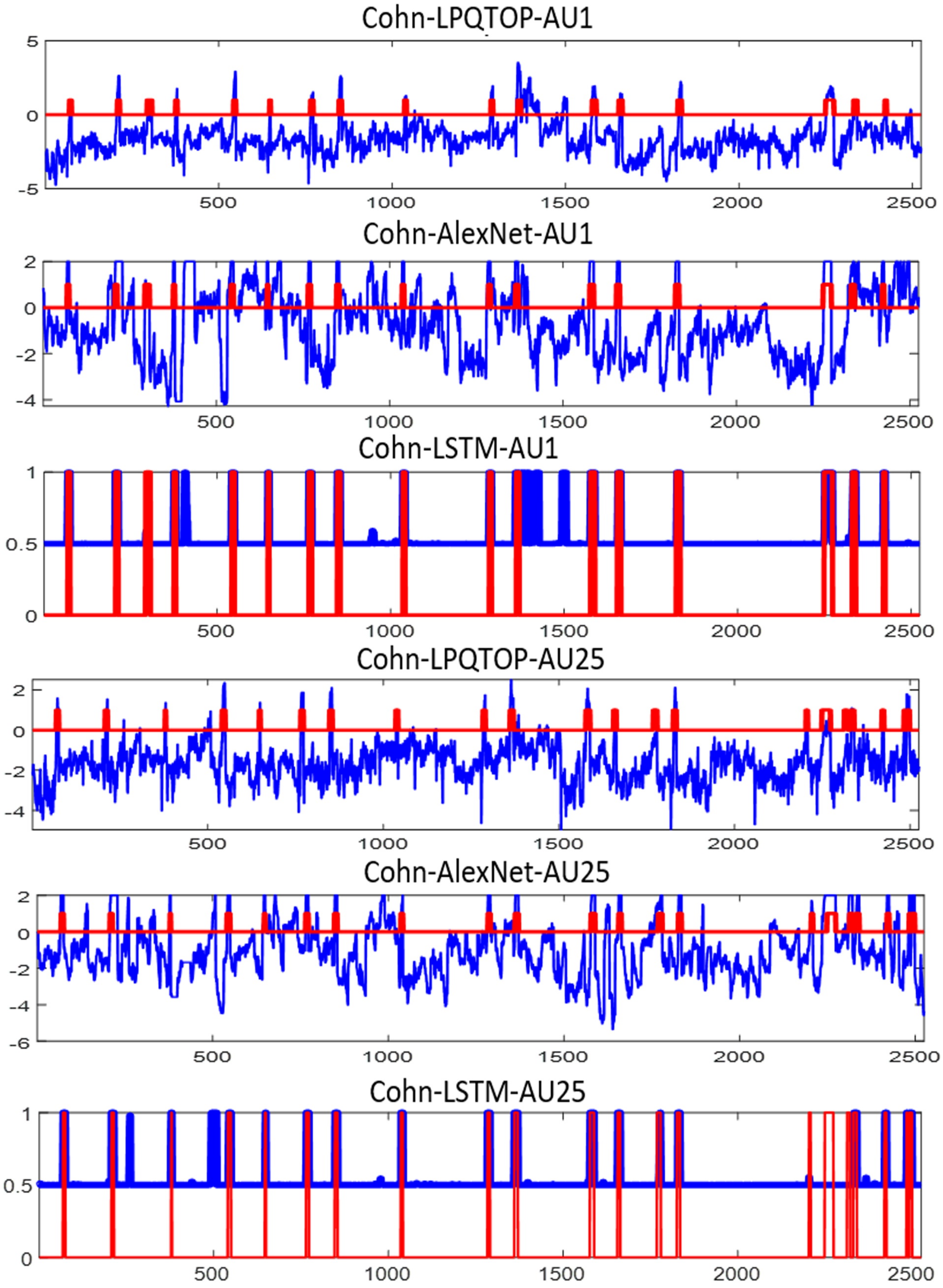


Fig. 3.22 Continuous scoring predictions between the three methods for AU1 and AU25.

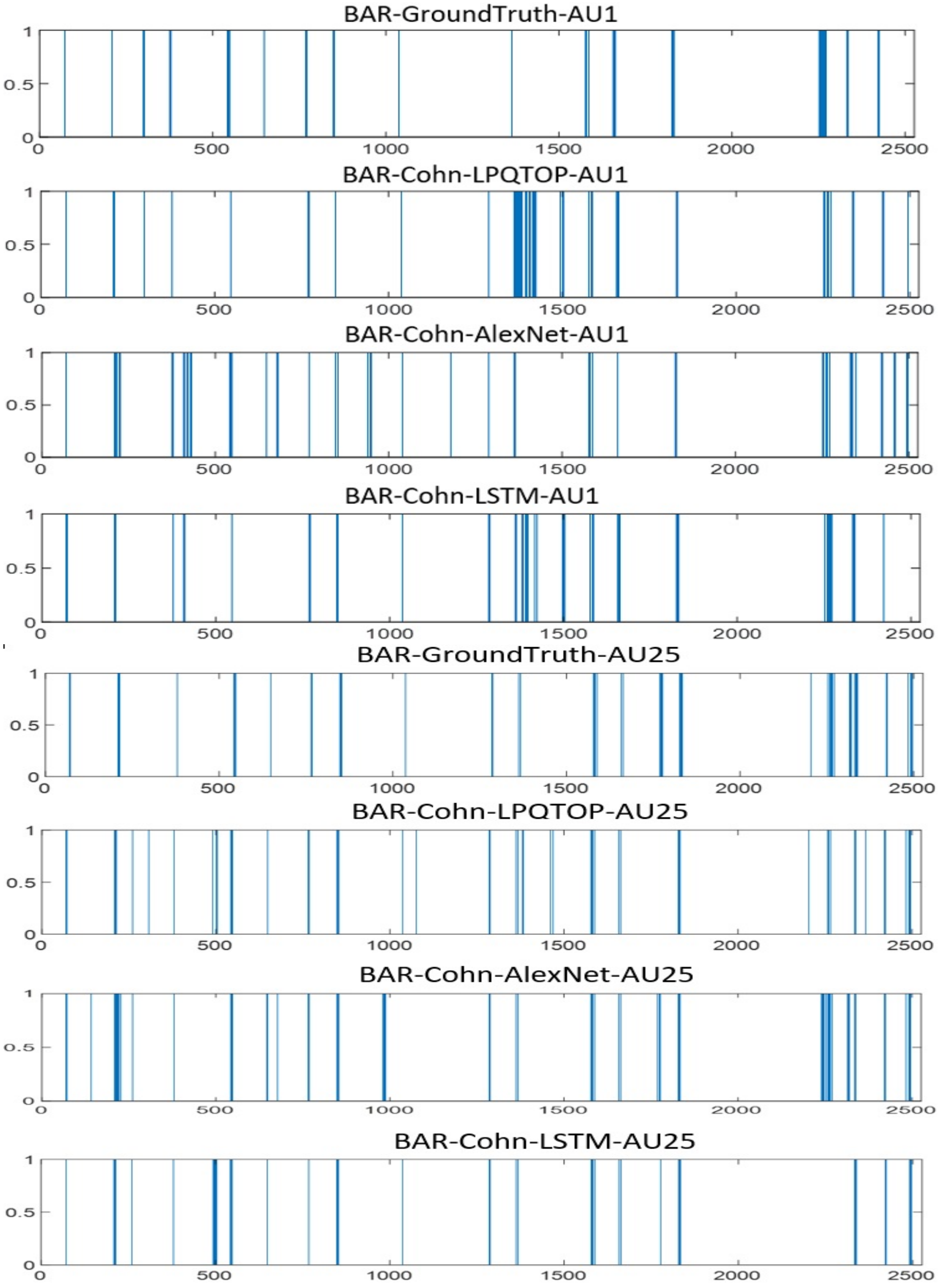


Fig. 3.23 Bars of continuous scoring predictions detection using a threshold for best matching between the three methods.

Chapter 4

Generative Adversarial Networks for Facial Expression Recognition in the Wild

4.1 Introduction

The use of Generative Adversarial Networks (GANs) to improve performance in downstream tasks is a relevant field for facial expression recognition such as transfer of facial expression and synthesize faces from landmark images [332]. Furthermore, the ability to transfer knowledge to associated tasks remains a challenging problem. Deep Neural Networks require an extensive amount of training data. The consequence is that there is no enough data to optimise all parameters, and there are many parameters that are prerequisites that need to be optimised, yet the quantity of labelled data is rarely enough to constrain numerous parameters. The size of the dataset is an important key factor for modelling and training deep Neural Networks. A large dataset prevents overfitting and imparts generalisation by learning better model parameters, and it captures more effectively the complex relationships and patterns inherent in data distribution. The Deep Neural Networks is directly based on the training data size; Fig. 4.1 illustrates how the quantity of used data has a big impact on the performance of deep NNs. Furthermore, the most crucial key factors in training deep NNs are parameter initialization, loss function, data size, learning rates adaptation, and the hyperparameter optimization algorithm [499]. However, accessing or having a dataset that

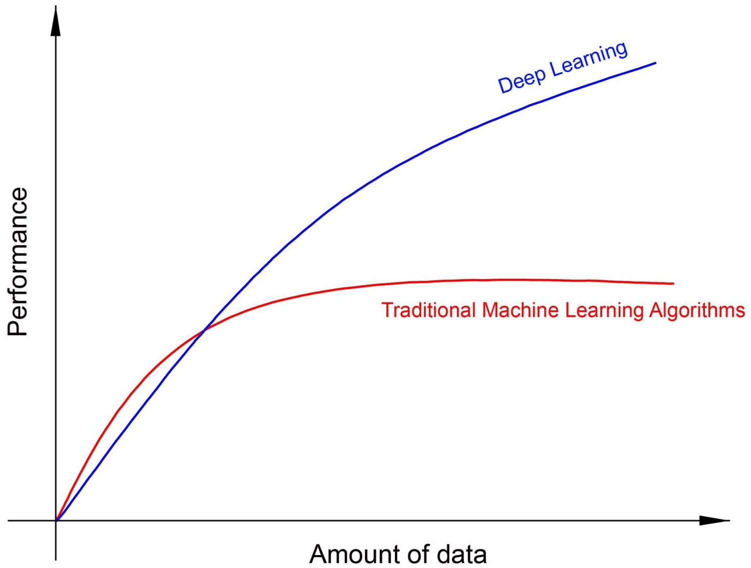


Fig. 4.1 An improvement in the performance of deep learning approaches is enhanced with the given increased amount of data in comparison with the previous traditional machine learning methods, image adapted from [17].

has sufficiently labelled coverage of several situations and conditions is a comparatively large challenge [182].

Facial expression datasets with Action Unit and emotion labels are scarce, limited in size, and imbalanced [473] due to the scarcity of the diversity of certain emotions and AUs. A core problem when learning CNN models is that they are often prone to overfitting and do not generalise well to unseen subjects [217]. Moreover, data labelling of facial expression is difficult, laborious, expensive, time consuming and prone to errors even with expert knowledge [215]. For particular domains, it needs to be performed by qualified experts, such as in remote sensing field the satellite images along with their ground truth are usually not provided in publicly.

It has been shown that deep learning can generalise well once the training data encompasses a large amount of nonlinear facial variability factors such as individual difference, subject identity or facial morphology, different backgrounds, illuminations, occlusions, head pose, which are all fairly common in the unconstrained environment [203]. In this regard, much of the research efforts with deep learning approaches have been focusing on techniques such as balanced batches, ReLU activation functions, training on multiple datasets, dropout regularization, GPU acceleration [157], and data augmentation techniques (rotating, scaling, translation, reflection, over-sampling, horizontal flipping, and photometric transformations), which aim to augment the training data to approximate the true distribution of the problem domain and help to present diversity of data [215]. All the above-mentioned methods assist in

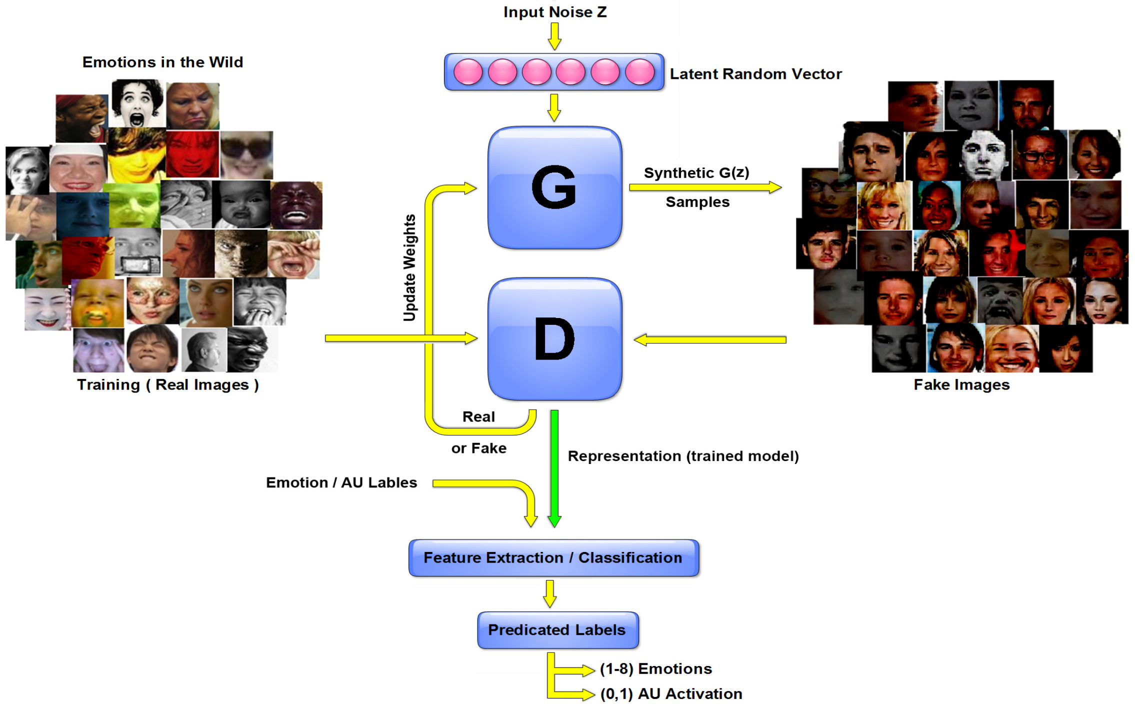


Fig. 4.2 Shows a Deep Convolutional GAN (DCGAN) diagram which combines the main concepts of CNNs for supervised learning as well as the standard ideas of Generative Adversarial Networks (GANs) models for unsupervised learning. The proposed model implements two of deep neural network models (G & D), where G is the Generator, and D is the Discriminator. The Generator is trained to generate fake images that are similar enough to the real ones to fool the Discriminator. The discriminator works as a CNN-based classification network and its output class probabilities. Both models are trained jointly in a competitive min-max process. The process arrives equilibrium once the Discriminator is no longer discriminate the real from fakes images. We added the last part of this architecture which represents the steps of feature extraction and classification of facial expression.

improving the quality performance of deep NNs, increasing the dataset quantity; nevertheless, they are still lacking nonlinear parametric variations among training datasets which may not be addressed by traditional augmentation methods [203]. Another option is to use large unlabelled dataset and unsupervised learning methods. Although there is an increasing amount of available data from the internet, most are unannotated and, therefore, one way to exploit the available unlabelled data and give an incentive to use unsupervised learning is to learn better representations to use these on supervised tasks. Indeed, until recently, supervised learning with CNN has been widely adopted in computer vision applications, while unsupervised learning with CNN has garnered less attention. In addition, generally, the classifier performance of emotion / AU relies mainly on training data [35]. DCGAN is one of the most common approaches that are tailored for this task, more particularly for facial images [451]. This method can be used for many computer vision problems: amendment of facial attributes, exploration in reinforcement learning [436][193], translate synthetic images into realistic photos, image-to-image synthesis, face aging and modelling, image transformation, colour restoration, anime characters generation, texture synthesis [450], datasets augmentation, image denoising, in-painting, super-resolution imaging, structured prediction, advertisements of shop commands [623], sentiment analysis[216], image translation, face generation editing, in door scene modelling, human possess editing [444], image-to-image translation, natural language processing, image colorization and pose adjustment [655].

The question is still uncertain whether using Generative Adversarial Networks such as DCGAN can provide an improvement in the categorisation of emotions in the wild. Can we use the Discriminator features for AUs / emotion recognition? Can we find consistent generalisation across datasets? Fig. 4.2 summarises an overview of the proposed work. The proposed model implements two deep Neural Network models (G & D), where G is the Generator and D is the Discriminator. The Generator is trained to generate fake images which are similar enough to the real ones to fool the Discriminator. Both models are trained in a competitive min-max process at the same rate. In the training, the Generator initially takes a random noise vector and maps into the image domain. The generated images seem like random noise, but as the process of training phase progresses, they will increasingly resemble the real images. The images to the right of the figure show a series of images generated by the Generator, while the Discriminator role is to identify the fake images from the real ones. The scores of the real images are compared with a vector of ones and in the same way for the fake images with a vector of zeros. The loss is computed for both models, and gradient descent have been used to update the two networks. Then, the features are extracted from the Discriminator's convolutional layer 12. Typically, we use the top layer of the network before the output (layer 12). Where these features are linearly separable and the top layer is a logistic regression. The SVM classifier is added at the top of these features to predict and classify the occurrence of AUs and emotion classes. Finally, the trained model will be deployed for the supervised task for the classification of facial expression with the available emotion and AUs labels.

The primary aim of this chapter is to investigate the idea of using deep generative adversarial networks for the extraction of high-level feature representations and the classification of Action Units and the eight emotion classes in the wild, by using the Discriminator network as a feature extractor based on static images and video frames. To examine the ability of DCGAN to generate analogous images from a different perspective (in front, multi-view, and in the wild), indiscernible from their versions in unsupervised manner adaptation. Also, by

Table 4.1 DCGAN Layers input/output details as shown in Fig. 4.3.

Generator		Discriminator	
layers	**output size**	**layers**	**output size**
input - shape	batch size, 100, 1, 1	input - shape	batch size, 3, 64, 64
transposed conv layer	batch size, 1024, 4, 4	conv layer	batch size, 64, 32, 32
transposed conv layer	batch size, 512, 8, 8	conv layer	batch size, 128, 16, 16
transposed conv layer	batch size, 256, 16,16	conv layer	batch size, 256, 8, 8
transposed conv layer	batch size, 128, 32, 32	conv layer	batch size, 512, 4, 4
transposed conv layer	batch size, 3, 64, 64	dense layer	batch size, 1

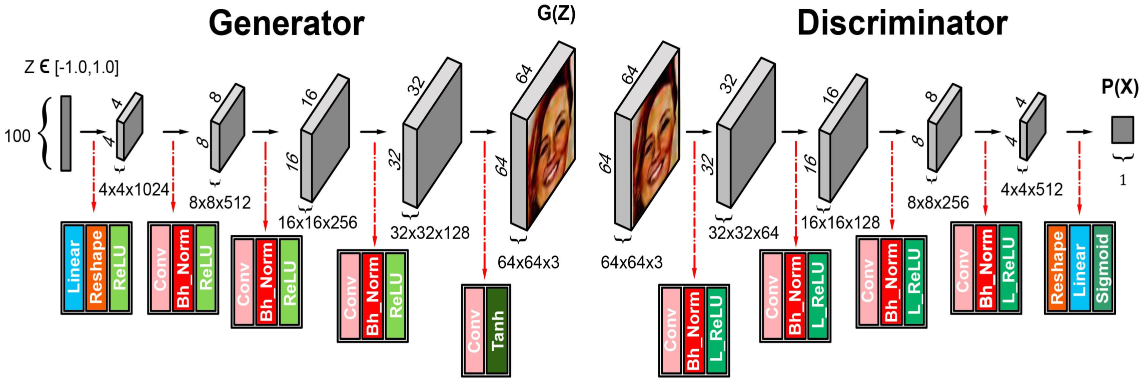


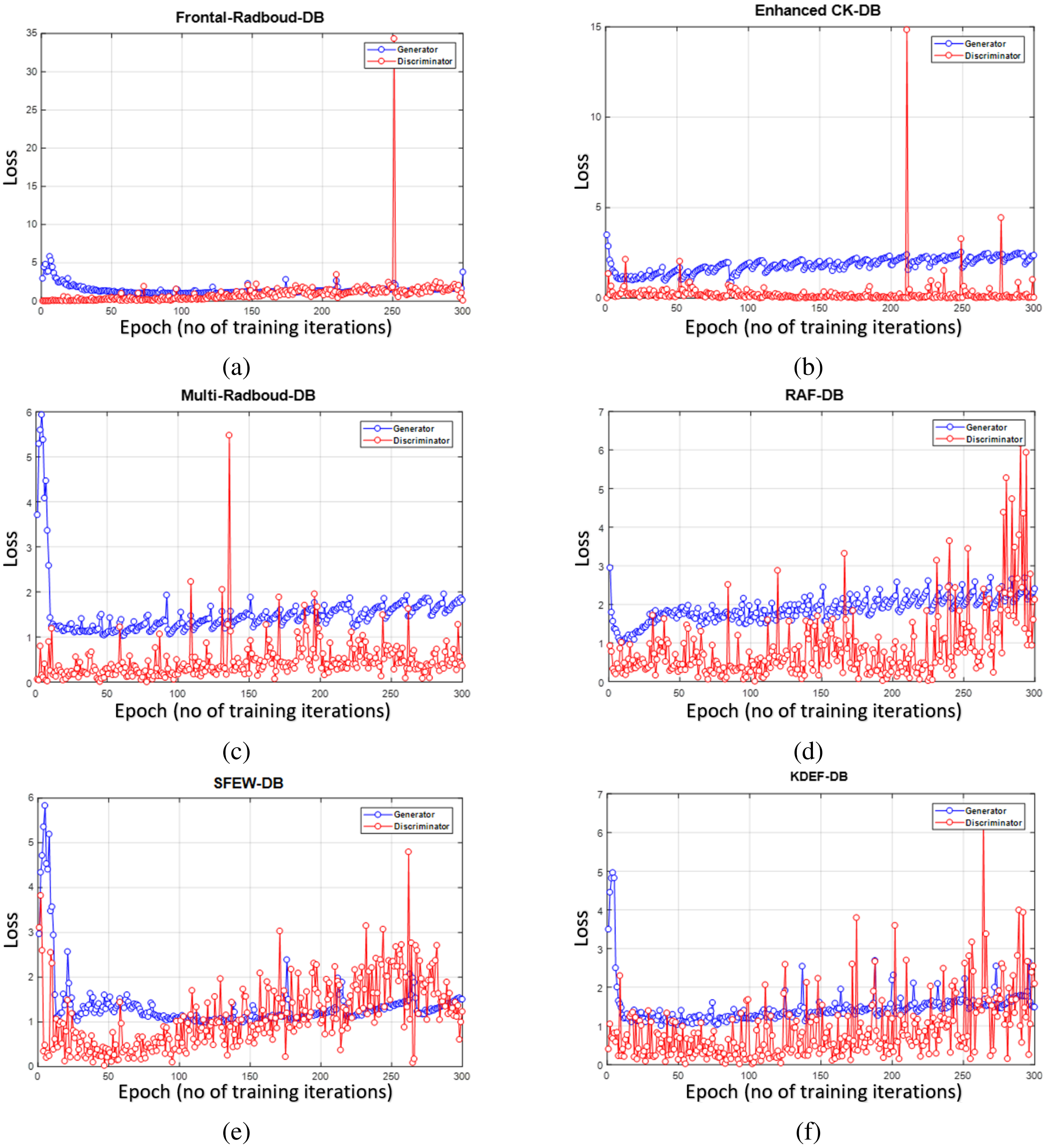
Fig. 4.3 A sketch of DCGAN architecture, image adapted from [83].

depending on the possibility of pre-trained models and transfer learning context, these dissimilar contributions share the same aim of approaching the learned representations. Moreover, training the model on limited specific facial AUs images was also included in this study.

This chapter is structured as follows: Section 4.2 provide details about the DCGAN network architecture design requirements. Then, a description of the frequently used facial expression benchmark datasets is available in section 4.3. Section 4.4 describes the proposed model and the settings of the experiment. In Section 4.5, the experiment results from performances on the selected datasets are also reported, and the extensive results are then analysed. Section 4.6 draws conclusions, summarizes the contributions made in this chapter, and offers recommendations.

4.2 DCGAN Network Architecture Requirements

DCGAN implements a pair of Convolutional NN models to train G the Generator (probability distribution) and D the Discriminator (classification problem); both are participating in a minimax game, competitively, in which they are adversarial. A Generator is trained to model



(a)

(b)

(c)

(d)

(e)

(f)

Fig. 4.4 The logistic loss of G and D during training DCGAN on the datasets used in this chapter: (a) frontal Radboud, (b) Enhanced CK, (c) multi Radboud, (d) RAF, (e) SFEW, (f) KDEF dataset.

the genuine data - the individual class probability distributions. The G network receives an input of 4D tensor of 100×1 random noise vector from latent space, defined as Z, by mapping Z into data space of $Z \sim G(Z)$ and transforms it by progressively increasing its spatial dimensions while decreasing its feature volume depth. The first layer denoted as "project and reshape" in the G is used to extend the random noise, after which it is convolved over it, yielding a $64 \times 64 \times 3$ output image, since the Discriminator can not distinguish the images which came from the Generator or the dataset. A Discriminator is trained to learn to distinguish whether a given sample is an authentic image (assigning high scores) or a generated image (assigning low scores), $y = P(x) \in [0,1]$, where, x is real training images [334]. Both of them (Generator & Discriminator) comprise of four convolutional layers. These replace max-pooling layers and advocate using stride convolutions (Discriminator) and fractional-strided convolutions (Generator) to up-sample spatial feature maps across layers to achieve finer resolutions [446], which allows the two networks to adjust and reduce the spatial dimensionality (down and up-sampling). Batch normalization is employed to all layers in both G and D and is exempt from the last output layer of the Generator and the last input layer of the Discriminator so that the model can learn the correct mean and the scale of the data distribution, in order to keep the gradients more stable during the training process and to circumvent the possibility of a model mode collapse or oscillation [171]. Finally, it is followed by ReLU activation function for all layers in the generator except for the last output layer, where Tanh activation function is suggested. Tanh was used to produce images whose pixel values are scaled to $[-1,1]$. Thus, all image examples in the datasets were pre-processed to have their pixel values $[-1,1]$ [446]. Leaky ReLUs activation function is applied in all the layers of Discriminator to further speed up the training [331]. Finally, the output of the last layer should be flattened using the sigmoid loss activation function network and can be interpreted as the probability of the predictions. The architecture for the generative model of DCGAN is shown in Fig. 4.3, and Table 4.1. Most deep GANs build from a symmetric of two discriminator and generator architectures. The Generator and Discriminator can be both constructed from different layers configuration, such as fully connected, convolutional, and recurrent layers. The Generator is constantly trained to deceive the Discriminator by outputting samples which are very close to real ones. On the other hand, the Discriminator competes to outsmart the opponent and tries to become a better deduct by correctly classifying the images. During training, the Generator and the Discriminator have to contend against each other and both are improved and arrive at an optimum point which is when neither can overpower the other. Largely, this normally signifies the loss which is permanently converged, and the model has learned enough to improve no further. It is reported by Mescheder et al. [389] that the instance noise and label smoothing with

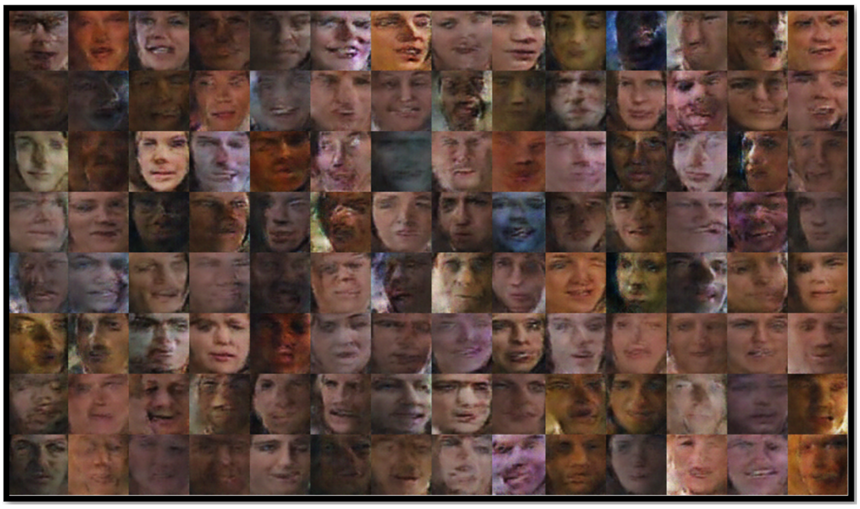


Fig. 4.5 An example of mode collapse of the generated images.

zero-centered gradient penalties and consensus optimization all lead to local convergence [211]. The analysis of the loss of the Generator and the Discriminator respectively over every batch, which resulted during the training of the DCGAN, is shown in Fig. 4.4, for the used datasets. It can be observed that the loss is continuous, oscillating from place to place, while showing high and low bounds, and this implies that the model persistently updates to improve itself. However, in Fig. 4.4, (d), (e), (f), the loss of the D is persistent in fluctuating heavily.

In this context, the adversarial training process is repeated until the Nash equilibrium [406] is reached between the Generator and the Discriminator to achieve good images. The traditional DCGAN model is trained with an aggregation of log loss on the Discriminator output and $\ell_{\mathbf{1}}$ loss between the Generator output and target image. The Discriminator is only trained with log loss. To interpret the loss when training DCGAN, the Discriminator and Generator would adjust their weights with the value function in equation 4.1. The objective requires the Generator to produce data that can match the statistics of the real data. In this case, the Discriminator is only used to match whichever statistics are identical. The G and D sub-network's minimax objective function can be optimized during training by adjusting the loss function:

$$\min_G \max_D V(D,G) = Ex \sim pdata(x)[\log(D(x))] + Ez \sim pz(z)[\log(1-D(G(z)))] \tag{4.1}$$

It is simply a standard cross-entropy function between the Discriminator's output and the actual labels. The first part of this equation is represented by the entropy E passed by the distribution of the real data $pdata(x)$ through the discriminator $D(x)$ and the Discriminator would want to maximise the log probability of predicting one, indicating that the data is real. The second part is represented by the entropy passed by the distribution of the random noise input data $p(z)$ through the Generator $G(z)$ that produces a fake data sample which is further passed to the Discriminator for assessment to identify the counterfeit. In this regard, the discriminator tries to maximise it to 0 (i.e. the log probability that the generated data is fake is equal to 0). On the other hand, the Generator's task is to achieve the opposite by trying to minimise the value function (the log probability of the Discriminator being correct), so that the difference between the real and the fake data is minimal [70].

Two commonly quantitative metrics have been suggested for evaluating the performance of GANs which explained as follows:

The Fréchet Inception Distance (FID) is a metric that was presented by [227] for enhancement of the Inception Score (IS) prevailing metric [474]. FID and IS have been considerably used as the best-adapted benchmarks for evaluating GAN models of non-conditional image creation, though there is no comprehensive study for which metric is the quantitative. Both metrics require to include a specific Inception V3 classification layer as a persistent multivariate Gaussian to evaluate the quality of the images generated by GANs. Particularly, the global spatial coding layer (the last pooling layer before the output classification begins) in a generative model produces a vector of 2048 elements. The activation vector is computed for a set of real and generated images. These activations are estimated by calculating the statistics mean and covariance for both the generated and the real data. The Fréchet distance among these two Gaussians, is known as Wasserstein-2 distance amongst multivariate Gaussians, is fitted to an embedded group of generated images into a feature domain provided by a specific layer of Inception Net. This distance is then used to quantify the quality of generated samples. The distribution of vectors for each of the real and generated image sets are flattened into parameters representing the corresponding distribution (mean and covariance). These parameters are then used to calculate the Fréchet distance between the two vectors. Lower FID score implies lesser distances between real and synthetic distributions, and that associates well with high quality generated images. Finally, this score has been shown it is robustness to noise level and more consistency with human evaluation.

Inception Score was proposed by [474], which utilises a pre-trained network (the classifier Inception Net [525]) on the generated images. This net trains on the ImageNet [118] to focus on the useful properties of generated samples such as sample diversity with visual quality, and afterward through using information related with the predicted class distribution. Inception Score has been used to measure the mean of Kullback-Leibler divergence (KLD) among the conditional class distribution of images $p(y|x)$ and $p(y)$ the marginal class distribution gained from all the generated images (large entropy of $p(y)$ for high diversity) [18]. However, the Inception Score has a disadvantage is that the statistics are not used to compare between the real and synthetic samples [227]; it is not a suitable distance because of insensitivity to the former distribution of classes. Therefore, this metric is not valuable for our work since it does not compare the generated images to a real validation group, where their performance from the Inception model and our datasets does not resemble the dataset used to train the Inception network. According to [367] the disadvantage of both metrics is, the incapability to identify overfitting.

In practice, Generative adversarial networks (GANs) are still suffering from some limitations that hamper the GANs development such as training instability of the model, non-convergence to a fixed point, undamped oscillations and fluctuate considerably, unpredictable mode-collapsing during training, gradient vanishing, challenging and hard to evaluate GANs performance.

Mode collapse happens when the generator collapses to model the distribution of the training data sufficiently, whereas attempting to increase the complexity of the generator network it does not essentially enhance the image quality. When occasionally a probability distribution of the data is a very complex multimodal (i.e. the generator is loose certain modes of the data was trained on). Mode collapse occurs when the discriminator can differentiate perfectly between real and fake samples before the generator estimates the data distribution. The generator network produces limited types of generated samples or small similar images. As a result, mode collapse would make a poor generalisation to new data. Several methods could be used to obviate the problem of mode collapse. One way is utilising multiple generators, train numerous models with different modes, training by using diverse data samples, and accurate choice of the architectures. Fig. 4.5, represents an example of the mode collapse. As shown in this figure that the model is not able to draw a perfect image, though it draws an image that can be recognised by human eyes. Recently, one of the used qualitative means for evaluating GANs for the quality of the generated samples are human judgment by looking at the produced images. In this work, we demonstrated that the DCGAN model can significantly improve training stability and overpower mode collapse without either increasing the model complexity or degrading the image quality. That is noticeable from the resolution of the generated facial expression images in Fig. 4.10 using different datasets. Furthermore, it was also depending on the size, quality, and quantity of the training dataset samples that we use for training the DCGAN model. As an example of the RAF, KDEF, and SFEW dataset used in this work which they are very challenging datasets themselves. The used method has not suffered from such problems and can produce detectable images.

4.2.1 DCGAN Technique Code Execution Prerequisites

Training on the GPU computing capabilities necessitate Parallel Computing Toolbox and a CUDA implementation of enabled NVIDIA GPU. We performed our framework using Deep Learning Toolbox 2019b MATLAB implementations of Deep Convolutional Generative Adversarial Networks (DCGANs), the execution of the DCGAN model was utilised the MatConvNet library. MatConvNet (CNNs using MATLAB) is a simple, and efficient MATLAB toolbox implementation of the Convolutional Neural Networks (CNNs) models for the applications of computer vision, it can run, learn and implement most state-of-the-art CNNs algorithms. TensorFlow represents a second-generation of flexible arrangement for both the google company and the deployment of numerous machine learning applications, it can be used to create neural networks. These prerequisites for training the DCGAN can be explained in detail as the following:

4.2.1.1 Compute Unified Device Architecture (CUDA)

CUDA was introduced in 2006, as an optional parallel programming model dependency, and was invented by NVIDIA. It has a computing capability with toolkit and drivers [69], suitable standard features for Nvidia GPU. This straightforward heterogeneous implementation accelerated the performance by harnessing the efficiency of the Nvidia GPUs to speed up some parts of the code. CUDA-enabled intensive Computing GPUs systems are widely deployed in many applications such as desktops, data mining, notebooks, workstations, and supercomputers.

4.2.1.2 TensorFlow

TensorFlow [2] is a library of strong numerical computation applicable in a variety of scientific research. It represents an open source interface for implementing a high performance of machine learning algorithms and deep Neural Networks. The original version was developed by a Google Brain researcher, and an engineer's team in the AI organization at Google. It is a flexible architecture that provides computation cores to deal with one or more CPUs, GPUs, and TPU platforms for various devices such as the server, desktop, edge devices, and mobile. The Tensor Flow computation process is represented by a graph and data flows, in which mathematical operations are described in nodes, and tensors are represented as the multidimensional arrays of data flow between these nodes. DCGAN was implemented using the library of Tensor Flow.

4.2.1.3 MatConvNet

This is a novel open source toolbox for integrated implementation of CNNs with a MATLAB environment for several computer vision applications, which combines efficiency and simplicity. VLFeat is one example of an open source library. It was developed by the same team who developed the VLFeat library. MatConvNet is designed to ease the formulation of a new architecture of CNN by virtue the fact that it can run most of the available, pretrained state-of-the-art CNN models (VGG-16, VGG19, VGG Face, VGGText, Fast-R) [563] and by using the building blocks of CNNs as MATLAB functions; mainly, supporting the computations on GPUs, for instance convolutions, linear and non-linear filter bank operations, normalization and features pooling. Concurrently, it provides essential efficient computation on CPU with GPU and CuDNN for training complex models using large datasets, for instance ImageNet ILSVRC.

4.3 Datasets

4.3.1 Radboud Faces Database (RaFD)

RaFD is a laboratory-controlled collection of multi-view and posed facial expression images. A preeminent high quality faces dataset which includes males, females and children from Caucasian ethnicity and males of Moroccan Dutch heritage [586]. It has a total of 4,824 images collected from 67 subjects for eight emotions: angry, disgusted, fearful, happy, sad, surprised, neutral and including contemptuous [205] [334], in which there are 1,608 frontal images from the whole number. Varied head poses were shot from left to right [586]. Each image in the dataset is annotated and captured by asking the participant to do three different gaze directions (front, left and right) [86], using five camera angles at the same time. For the frontal images used in this experiment, for convenience, the images were cropped, then rescaled to 256 × 256, to center the faces, and were changed all the images to greyscale, and then they were resized to 64 x 64, which is the input of the network. In Fig. 4.6, it can be determined that there are equal examples of each emotion class in the Radboud-DB [309].

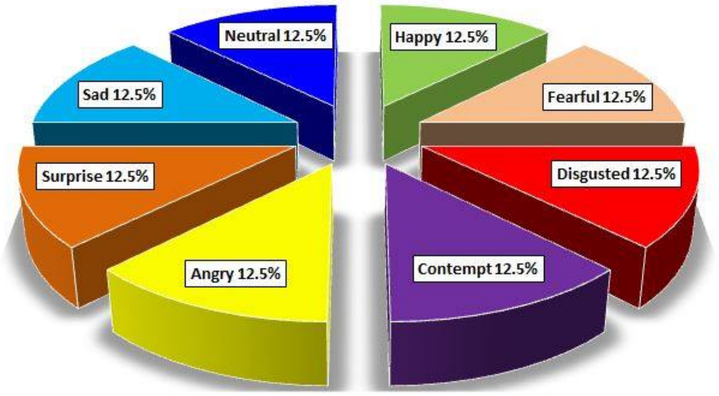


Fig. 4.6 Emotion classes' distribution in Radboud dataset are equal.

4.3.2 Large-scale CelebFaces Attributes (CelebA)

CelebA [359] is currently the largest-scale face recognition dataset with 202,599 celebrity faces, 10,177 identities, and 200k colour images with coarse alignment [597]. It mainly contains frontal faces with various viewpoints and expressions and is particularly biased towards white ethnic groups. It presents very controlled illumination settings and good photo resolution [104]. Each image is annotated with 40 binary labelled attributes which are indicative of gender, facial and hair colour, and five landmark locations [360]. In this work, the pre-trained model of the CelebA dataset was used.

4.3.3 Real-world Affective Faces Database (RAF-DB)

RAF-DB [338] is a large-scale real-world affective faces dataset, comprising of 30,000 face images of great diversity, collected using different search engines and Flickr [157]. It was annotated with two subsets: 12 classes of compound emotions and 7 basic emotion categories of facial expressions annotated by forty independent human coders [388], with the final annotations determined through crowd sourcing methods [625]. The images in this dataset were varied in terms of personal identities such as subject's age range, ethnicity, race, gender attributes, facial hair, glasses, head poses, lighting conditions, and occlusions per image. In this dataset, images have been labelled with seven basic emotions (happiness, anger, sadness, disgust, surprise, fear, and neutral) [119]. It features 15,331 images, and contains 12,271 training samples and 3,068 images in the test set. The distribution of the data is very disparate [354].

4.3.4 Karolinska Directed Emotional Faces (KDEF)

A large dataset was created by Lundqvist et al. [368], at the Karolinska Institute. It includes a set of 4,900 images in facial expressions from 70 subjects (35 female and 35 male), displaying seven emotions (angry, fearful, disgusted, happy, sad, surprised, and neutral) [573], using faces of subjects of age between 20-30 years. During the session, intrusive elements causing any disruption were excluded from their faces such as earrings, facial hair, moustaches, jewelry, makeup, beards, or glasses [19]. Each expression was photographed from five different angles and was recorded two times. Image resolution was 562 × 762 pixels [293]. The participants, all amateur actors, were instructed by being given a description of the expressions that were to be acted out, to try to express the appropriate emotion. They were also asked to rehearse these expressions for an hour before being photographed, in order to create the expression clearly and decisively [114].

4.3.5 Static Facial Expressions in the Wild (SFEW)

SFEW [126] is a very challenging benchmark dataset for conventional facial expression approaches because it has complicated scenes in the videos and the spontaneous facial expressions are more difficult to recognize [134]. It was developed by gathering static subset frames from AFEW dataset video clips [519]. It uses a seven classes-expression semi-automatic labelling process [111]. The database covers natural, unconstrained, versatile, facial expressions [643], which are close to a real wild setting environment and illumination status, varied in head poses [134], with quite a large age range, occlusions, varied focus, and different resolutions of the face. In total, SFEW contains 700 images that have been labelled for seven basic expression categories, including anger, disgust, fear, happiness, sadness, surprise and the neutral class and this was labelled by two independent labellers [235]. The SFEW database is mainly composed of 1,766 images, divided into three sets (958 images for training, 436 images for validation, and 372 images for testing respectively) [388]. Fig. 4.7, shows image samples from the SFEW dataset.

4.4 Methodology

The concept of Generative Adversarial networks (GANs) was originally introduced by Ian Goodfellow et al. [194]. DCGAN was proposed with the aim of improving the training/learning stability of GANs and their performance, and has proven to be a stable, powerful framework for generating synthetic images with reasonable visual fidelity. The main finding of this chapter is explained in detail as the following: the ability of the DCGAN model was

Fig. 4.7 Some image samples from the SFEW dataset, image source from [126].

adapted for the supervised tasks by deep facial features which were extracted and grounded on this model. After training the model, it was observed whether the generated images of AUs and emotions have the same visual fidelity quality of the original images. In terms of assessing their generalisation ability, the trained models were validated on more datasets: RPI ISL Enhanced Cohn-Kanade AU-coded Facial Expression Database (enhanced Cohn-Kanade) [261], Large-scale CelebFaces Attributes (CelebA) [360], Radboud Faces Database (RaFD) [310], Real-world Affective Faces Database (RAF-DB) [339], Karolinska Directed Emotional Faces (KDEF) [368], and Static Facial Expressions in the Wild (SFEW) [126], using the transfer learning approach. Frontal images were cropped using the Viola-Jones method proposed by [566], and the multi-view cropped faces were obtained from the state of the art multi-task CNN utilised approach of MTCNN [635], which was adopted for parsing the face landmarks and bounding boxes. The images were downscaled to an initial resolution of $64 \times 64 \times 3$ pixels before input to the network began. The model was trained for 300 epochs. The features from the Discriminator's convolutional penultimate layer 12 were extracted; this layer gives 512 feature spatial grid maps of size 64×64. Then, the singleton dimensions of size 1 were reshaped and removed from the shape of a tensor (4-dimensional tensor). The nonlinear SVM was used for emotion recognition and the linear SVM was used for AUs activation detection, alongside the emotion/AU labels. SVM was straightforwardly applied at the top of these features to predict and recognize the occurrence of 14 AUs and eight emotion classes. The same steps could be used to extract the features from the Generator, but this can be a task for future research.

This setup was made by using the compatible integration of the Tensor Flow and CUDA toolkit, to empower the parallel calculation and allow better computation execution times and performance. The experiments were carried out on the workstation using the Ubuntu Linux system and all the processes of training and testing were accelerated by the NVIDIA GeForce GTX 980 Ti GPUs. The model was then trained with the following hyper-parameters: mini-batch stochastic gradient descent (SGD) with a batch size of 128, suggested learning rate fixed to 0.0002 for the optimizers, momentum coefficient term β_1 hyper-parameter, chosen to be 0.5 for making the training more stable, and the Adam optimizer (Adaptive Moment Estimation) was adapted as the best choice of an optimization algorithm to decrease the loss functions in the generative models literature [290]. All the weights started from zero-centred Normal distribution and standard deviation of 0.02. Batch normalization was used to normalize the input of each unit having a zero mean and variance [449].

We depended on using the DCGAN architecture with the available adjusted hyper-parameter values as described in their design. They are recommending these hyper-parameters for the training of the model. We identified that using the suggested hyper-parameter values worked well and have been provided a stable training, improving the convergence, and getting convincing results from the quality of the generated images. Other affecting fundamentals which should be considered, such as training time, model complexity, and capacity of the number of parameters to learn. Moreover, cross-validation was done to find the optimal hyperparameters and to evaluate the performance of the model with maximum accuracy.

4.5 Experiments

A study has been carried out to benchmark emotion and AU recognition. The performance of different approaches was compared and analysed.

4.5.1 Experiments on Emotion Recognition

The aim of this experiment is to do emotion recognition and to generate facial expression images. This was briefly divided into eight experiments, to show that the performance gained whether dependent on the specific dataset or was provided from different datasets. Training a new DCGAN and using the ability of a generative model to generate variability in the generated images from different perspectives (frontal, multi-view, and in the wild). The generalisation of cross dataset evaluation was provided from the different dataset based on the fundamental aspects of transfer learning and various pre-training models which was explained as the following:

1) fine-tuning a pre-trained model of the enhanced CK (source dataset) by performing feature extraction and classification for another small collection of images with emotion labels on a frontal Radboud dataset (target dataset). Fine-tuning was used to regularize the networks with the pre-training models which were recently extensively applied and which became the permeated trend technique in deep FER, in order to conquer the problem of

Table 4.2 AUC values for the eight experiments according to emotion recognition shown in Fig. 4.9.

Emotions	Exp1	Exp2	Exp3	Exp4	Exp5	Exp6	Exp7	Exp8
Angry	0.98211	0.999	0.95906	0.95111	**0.99947**	0.93684	0.88826	0.75856
Contemptuous	0.93637	0.99352	0.89695	0.85778	**0.99855**	—	—	—
Disgusted	**1**	0.99992	0.99374	**1**	**1**	0.7904	0.93266	0.82289
Fearful	**0.9998**	0.99765	0.97735	0.97222	0.99809	0.92497	0.723	0.83531
Happy	**1**	**1**	0.99837	0.99294	**1**	0.83292	0.97383	0.71913
Neutral	0.95703	0.99383	0.87509	0.81176	**0.99858**	0.91132	0.90459	0.73661
Sad	0.96665	0.99779	0.93957	0.93176	**0.99944**	0.84292	0.78451	0.68309
Surprised	**1**	0.99934	0.9913	**1**	0.99974	0.85127	0.94278	0.72025
Average	0.980245	0.997631	0.953929	0.93209	**0.999234**	0.870091	0.878519	0.753691

overfitting and lack of training data, and by using secondary data from large or small and general or precise, for instance ImageNet [334].

It was deemed that using a pre-trained model of the enhanced CK to test the frontal Radboud dataset is considered the better adapting initialized weight. As noted above, training a complex model like CNN on a small dataset is challenging [409]. And so, a matrix was produced of size 1,608 × 8,192 dimensions (where 1,608 represents the number of images of the frontal Radboud dataset and 8,192 represents the number of feature vector) for eight emotions.

A multiclass SVM with a Gaussian kernel was used for the classification of all the experiments related to emotion recognition, and the parameters were optimized using the bayesopt optimizer [381] with ten-fold cross validation. Eight ROC curves and a confusion matrix for eight emotions were gained, representing the performance of the classifier. Interestingly, this experiment achieved the best state-of-the-art performance with an accuracy of up to 98.57%. The number of images was not that extensive in this experiment, and the results were really promising. It was expected to leverage the effect of transferring the previous extra hidden feature's learning knowledge. This also may be attributed to the discriminative learned model of CK. One more minor reason for that is the resolution of the images, and every image including the explained reflected exact emotion. See Exp1 in Fig. 4.8, Fig. 4.9 and Table 4.2.

2) in Exp2, a model pre-trained on the CelebA dataset was used to test the frontal Radboud dataset, to examine whether the model is working best when it is trained on a large amount of data. As can be seen, the performance is clearly enhanced, and the result improved which showed that the fine-tuning was the same as in Exp1 but with a different dataset for pre-training accomplishes better results.

3) in Exp3, the performance of DCGAN was examined between frontal and multi-view images of emotions; a pre-trained model of the enhanced CK dataset was used to test

the multi-view images of the Radboud. A matrix of feature vector of size 2,680 × 8,192 was produced. As we can observe from Fig. 4.8, Fig. 4.9 and Table 4.2, the recognition performance is less in comparison with the performance of the frontal Radboud images.

4) Exp4, another experiment was conducted but here the trained model of the multi-view Radboud itself was used to test the multi-view Radboud images. It was found that the performance also significantly decreased and fell again.

5) Exp5 was conducted by train and tested the frontal Radboud images using DCGAN. This experiment achieved promising results (average AUC for all the emotions = 0.999, and accuracy = 97.64%), even with fewer images, for the same reasons mentioned above about the frontal Radboud dataset.

Finally, in 6), 7), 8) DCGAN was trained in the last three experiments (Exp6, Exp7, and Exp8) on the three difficult datasets in the wild (RAF, KDEF, and SFEW), where facial expressions are close to the real wild environment. We can observe that the performance decreased significantly due to the apparent distortion of faces, low resolution imaging in the wild, and insufficient training data, specifically the SFEW image dataset, which limited the capacity to attain accurate results. Other factors make an impact such as different background noise, clutter, colour, robust head pose diversities, non-relevant variations and illumination changes which are difficult to determine and might largely influence the analysis of the DCGAN results. Moreover, the classification of emotions in the wild remains a complex problem and often impairs the performance. Although DCGAN is not designed for facial feature extraction and classification, the results here are promising.

Fig. 4.10 characterized all the generated images for all the datases used in this work. In comparison to the original images, the generated images of the RAF and the SFEW datasets are distorted, have irregular colours, cloudy and not intelligible. Also, the multi-view generated images of the Radboud dataset were slightly distorted because of the introduced noise.

A Receiver Operating Characteristic curve was used frequently in this work to view the optimal performance of the designated classifier for the current tested models under different threshold settings (i.e. with the decision of threshold varies). It displays the true positive rates (hit-rate) on the y-axis (sensitivity), against the false positive rates (false alarm) on the x-axis (1-specificity). It can be observed that a perfect classification is with no misclassified values, and is located on the right angle to the top left of the plot. A poor outcome that represents random number is a line at 45 degrees in the plot. The Area Under Curve number is a measure of the overall efficiency of the classifier performance. Larger AUC values indicate better classifier performance. The ROC Curves for each emotion on all eight experiments are shown in Fig. 4.9 and the AUC values are reported in Table 4.2.

The confusion matrices for eight expressions in all experiments are shown in Fig. 4.8. The correct classified unit for each expression is highlighted in dark blue, while the misclassified units were highlighted in paler blue. The experiments performed very well in recognizing most of the emotions: *surprise*,*fear*, *disgust*, *happiness*, *sadness*, *anger*, *contempt* and *neutral* with a true classification of 94.6% in Exp1 and Exp5. Also, sadness and disgust in Exp1, Exp2 had a correct classification of 100%. Anger and fear showed a relatively low recognition rate in experiments 1, 2, 3, and 4. Moreover, happiness and sadness expressions showed the lowest recognition rate of 38.5% and 38.8% in Exp8 respectively. Table 4.3 represents the obtained accuracy for each dataset compared with the state of the art performance for emotion recognition.

Table 4.3 Comparison of the obtained accuracy for each dataset for emotion recognition with the state-of-the-art methods.

Dataset	Approach	Accuracy	Dataset	Approach	Accuracy
Radboud	(Ali et al., 2017 [11])	85.00%	SFEW	(Zhang et al., 2018 [631])	26.58%
	(Yaddaden et al., 2018 [596])	97.66%		(Dhall et al., 2015 [128])	35.93%
	(Jiang & Jia, 2016 [266])	94.52%		(Levi & Hassner, 2015 [322])	41.92%
	(Wu & Lin , 2018 [586])	96.27%		(Yao et al., 2015 [608])	44.04%
	(Mavani et al., 2017 [385])	95.71%		(Ng et al., 2015 [409])	48.50%
	(Sun et al., 2017 [521])	96.93%		(Yu & Zhang, 2015 [611])	52.29%
	(C.Szegedy et al., 2015 [524])	95.45%		(Mollahosseini et al., 2016 [397])	39.80%
	(Zavarez et al., 2017 [621])	85.97%		(Zhang et al., 2018[643])	**55.27**&
	(Li et al., 2019 [326])	96.11 %		(Mao et al., 2016 [377])	44.72%
	(WANG et al., 2019 [575])	80.69%		(Eleftheriadis et al., 2016 [155])	24.70 %
	ours	**98.57**%		ours	44.52%
KDEF	(Shin et al., 2016 [498])	59.15%	RAF	(Li et al., 2017 [339])	**82.7**%
	(Zavarez et al., 2017 [621])	72.55%		(Li et al., 2019 [337])	74.2%
	(Samara et al., 2017 [476])	**81.84**%		(Fan et al., 2018 [157])	76.73%
	(Yaddaden et al., 2016 [596])	79.69%		(Lin et al., 2018 [350])	75.73%
	(Ali et al., 2017 [11])	78.00%		(Ghosh et al., 2018 [183])	77.48%
	ours	60.44%		ours	61.87%

4.5.2 Experiments on Action Units

4.5.2.1 Action Units on the Enhanced Cohn-Kanade Dataset

This experiment attemptes to answer the question whether features learnt by the layer of a DCGAN and the Discriminator convey information characterising Action Units. For that purpose, the enhanced CK dataset was used which featured extensive AU labelling. We noticed that the generated images in this experiment had a good quality, and high resolution, from which the Generator has learned how to make plausible facial action images. It depicts the personality and identity representation of a face. Fig. 4.10, (a) and (b) indicates the original and the generated images of the enhanced CK dataset. Lastly, the matrix of 4D was flattened and then concatenated, yielding dimensions of 8,422 × 8,192 (in which 8,422 represents the number of the images of the enhanced CK dataset and 8,129 represents the number of feature vectors). We then trained and tested on the enhanced CK using the linear SVM by the LibSVM [76] to recognize the occurrence of 14 AUs (AU1, AU2, AU4, AU5, AU6, AU7, AU9, AU12, AU15, AU17, AU23, AU24, AU25, AU27); the results of Exp.1 are shown in Fig. 4.13, Table 4.4, and Table 4.5; 14 Areas Under the ROC Curve (AUC) values were gained for fourteen Action Units; all the AUC values for all the experiments related to AUs are available in Table 4.4, and a plot of the AUC curves are in Fig. 4.13. Table 4.5 represents all the pretrained models used with the related datasets for cross dataset performance regarding AUs.

Table 4.4 AUC values for all the experiments regarding AUs shown in Fig. 4.13.

AUs	Exp.1	Exp.2	Exp.3	Exp.4	Exp.5	Exp.6	Exp.7	Exp.8
AU1	**0.99813**	0.99388	0.9607	0.90873	0.89586	0.99611	0.8900	0.89578
AU2	0.91755	**0.9994**	0.9483	0.74424	0.63279	0.99813	0.70531	0.67732
AU4	0.78805	**0.99251**	0.88655	0.51451	0.6557	0.98963	0.51881	0.66256
AU5	0.98189	**0.99566**	0.99079	0.67556	0.80839	0.98008	0.62401	0.76701
AU6	0.89533	**1**	0.9076	0.5744	0.51796	0.99959	0.56813	0.54234
AU7	**1**	0.9838	0.99969	0.75741	0.64927	0.97894	0.64154	0.60085
AU9	0.98975	**1**	0.9972	0.66787	0.54624	**1**	0.6828	0.56524
AU12	**0.99965**	0.98976	0.9801	0.5146	0.6329	0.96703	0.54904	0.61531
AU15	0.93152	**0.99186**	0.9629	0.50984	0.6590	0.99019	0.51769	0.64736
AU17	0.78532	**0.98575**	0.8734	0.7084	0.8065	0.98365	0.62314	0.74785
AU23	0.88199	**0.98034**	0.9269	0.8653	0.8277	0.97894	0.77498	0.8135
AU24	0.94762	**0.9842**	0.9018	0.8491	0.6884	0.97894	0.82199	0.72053
AU25	0.99762	**0.99912**	0.99162	0.7994	0.9291	0.99872	0.7552	0.89429
AU27	0.67244	**0.99959**	0.6692	0.6682	0.5406	0.99939	0.67844	0.51883
Average	0.913347	**0.992562**	0.928339	0.696969	0.699315	0.988649	0.667934	0.690626

4.5.2.2 Radboud Emotions Relabelled to AUs

The aim of this experiment is to evaluate whether features trained on a large (potentially unlabelled) dataset can be transferred for supervised training to a different one. This experiment was designed to confirm the results obtained on the CK dataset on a different dataset, namely by Radboud, since Radboud is only annotated for the eight basic emotions, and not for AU. The dataset was re-annotated according to the rules in [553] [558] [380]. Though there are several works present in the area of AU detection, robust methods mapping AUs to emotions are still mostly unexplored. The way to map emotions of the frontal Radboud dataset to AUs is summarised in the Table 4.6, and Fig. 4.11 illustrates an example of images annotated according to the FACS.

Table 4.5 Summaries all the pretrained models obtained from the DCGAN network with the related training and testing datasets regarding AUs.

Experiments	Pretrained models	Training	Testing
Exp.1	enhanced CK	enhanced CK	enhanced CK
Exp.2	enhanced CK	Radboud	Radboud
Exp.3	CelebA	enhanced CK	enhanced CK
Exp.4	CelebA	enhanced CK	Radboud
Exp.5	CelebA	Radboud	enhanced CK
Exp.6	CelebA	Radboud	Radboud
Exp.7	enhanced CK	enhanced CK	Radboud
Exp.8	enhanced CK	Radboud	enhanced CK

Table 4.6 A mapping between emotions and AUs based on rules according to the FACs [145].

Emotions	**AUs**
Happy	{AU6, AU12, AU25}
Sad	{AU1, AU4, AU17, AU15}
Fearful	{AU1, AU2, AU5, AU15, AU25}
Surprised	{AU1, AU2, AU5, AU25, AU27}
Angry	{AU4, AU5, AU7, AU17, AU23, AU24}
Disgusted	{AU9, AU15, AU17, AU25}
Contemptuous	{AU12}

In this regard, a pre-trained model of the enhanced CK dataset was used to extract the features of the frontal Radboud dataset; then, a classification was done using linear SVM. The results look very interesting, although there are action units missed in the annotations which are very important: AU10, AU11, AU14, AU20, AU22, and AU26. For example, the accumulation of AU12, AU22, AU25, AU26 is often considered a sign of happiness [231], and there are no rules for contemptuous emotion (AU12 and AU14 from one side of the face) in the Radboud dataset to map it to AUs as it is not considered among Paul Ekman's six basic emotions. It represents a mixture of disgust and anger emotions. Also, there is no action unit to do lip corner tightening raised on only one side of a face. Exp.2 in Table 4.4, Table 4.5, and Fig. 4.13, show the improvement in the results for all the AUs even with the imbalance and lowest occurrence activations in the dataset. Next, the output feature vector was extracted from each layer of DCGAN on the enhanced Cohn-Kanade dataset (X1 is not within because it represents the input data, and the output responses were gained from each convolutional layer in the Discriminator which is exactly X3, X6, X9, X12) to extract the feature vector (output of layer x3 of dimensions 8,422 × 65,536, output of layer x6 of dimensions 8,422 × 32,768, output of layer x9 of dimensions 8,422 × 16,384 and output of layer x12 of dimensions 8,422 × 8,192) and a classification was done for each output separately using also an SVM classifier (divided the features to train and test), and next, the output features vector was concatenated from each layer to create one feature vector of dimensions 8,422 × 122,880, but the results were the same as the output of layer 12, and there are no improvements at any one of them separately or even when all were merged.

4.5.2.3 Transfer Learning on AUs

The last experiment was conducted for cross dataset evaluation performance research. Cross-dataset means using the images from a dataset for training and for testing using the images from another dataset [335]. Transfer learning is basically conveyed by pre-trained models to cope with the impact of the restricted number of the target dataset and to remove the bias made by an uneven training size. A pre-trained model is a model that was trained on a large benchmark dataset to solve a different problem but a similar and related task to the one that is to be solved. Therefore, due to the computational cost of training those models, it is common practice to import and use models from the published literature. This work explored a relatively simple transfer learning problem because it involved transferring knowledge across identified discrete output classes.

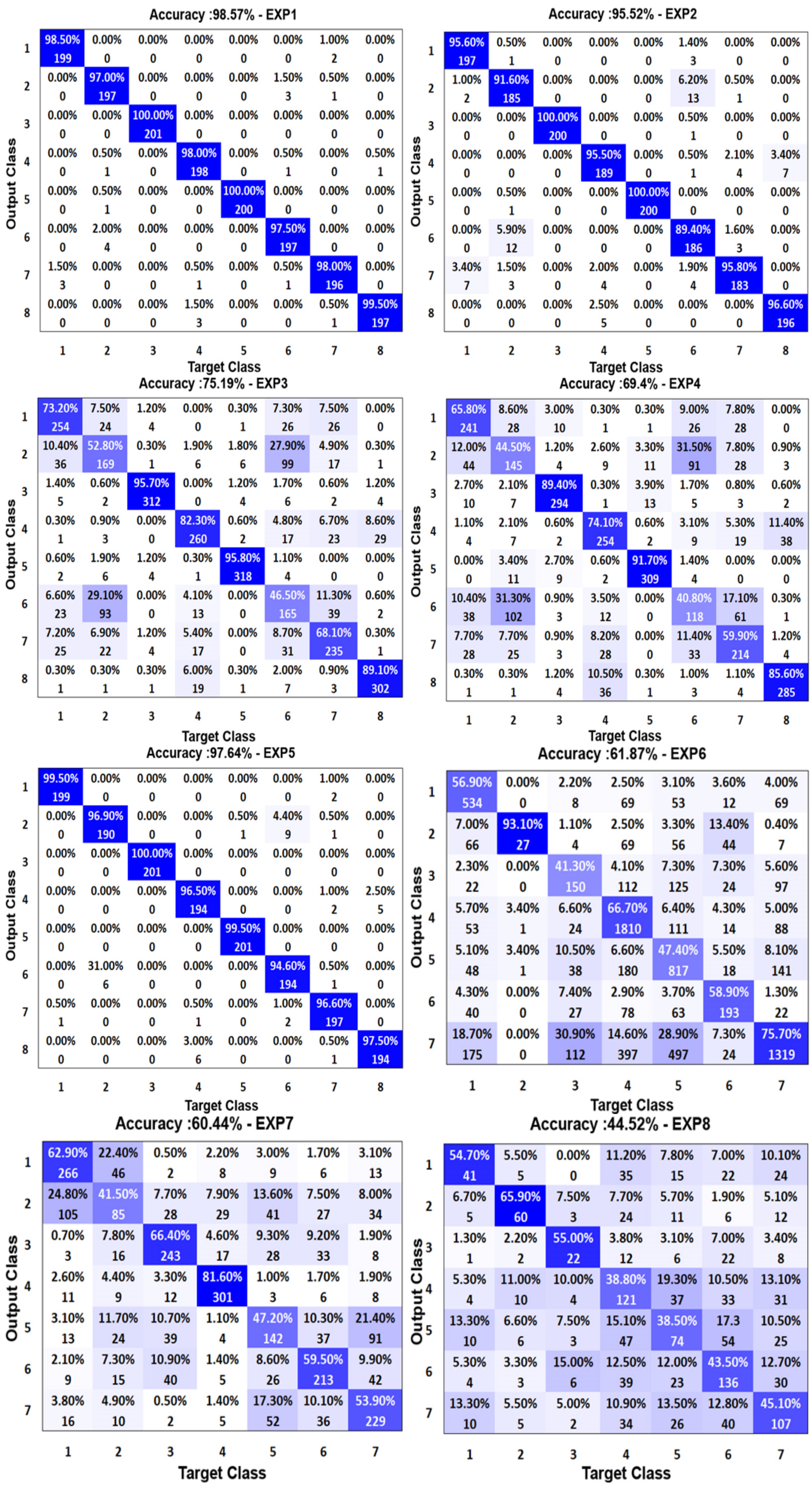
A pre-trained model of the CelebA dataset (pre-training refers to the features in the DCGAN network) was used to train and test the CK dataset. Then, a pre-trained model of the CelebA dataset was used to train and test the CK and the Radboud dataset reciprocally. For more details, refer to Exp.3, Exp.4, Exp.5, Exp.6, Exp.7, and Exp.8 in Table 4.4, Table 4.5, and Fig. 4.13. It was obvious that the model certainly accomplished impressive results when trained and evaluated on the same dataset, such as Exp.2 and Exp.6, in Table 4.4. The cross-dataset performance was quite good, for example Exp.1, Exp.2, Exp.3, and Exp.6; as such AU9 (nose wrinkle) is often considered a distinguishing feature of disgust which occurs frequently, therefore the AUC = 0.990, AUC = 1, AUC = 0.997, AUC = 1 for these experiments respectively; similarly for AU25 (lip parts), the AUC = 0.998, AUC = 0.999, AUC = 0.992, AUC = 0.999, and AU7 (lid lightener). In the the cross-dataset performance of the CNN model, however, training and testing on two different datasets dropped drastically because one of the datasets is quite different and fails to deal with new tasks and further operating settings that have not yet been seen during the training process and development. Moreover, a well-performing Convolutional Neural Network (CNN) model trained on one dataset (source dataset) usually performs poorly on another dataset (target dataset). This is also because the feature distribution of the same emotions / AUs varies in different datasets. Moreover, when the DCGAN is trained on too few images, it is inadequate to fine-tune a model toward the target dataset, whereas when the images are many, the model will do well in recognizing those images.

Notably, the results are encouraging for transferring *some* AUs. As we can observe from Table 4.4, AU1 (inner brow raiser), AU23 (lip tightener), AU24 (lip pressor), and AU25 are transferred and generalised well for all experiments. While for the AU2 (outer brow raiser), and the AU17 (chin raiser) the performance is similar for all the values of AUCs in Exp.4, Exp.5, Exp.6 and Exp.8. The worst transfer appeared for AU4 (brow lowerer), with AUC = 0.515 in Exp.4 and AUC = 0.519 in Exp.7; AU4 is a common feature of confusion which happens in some occasions in our life, as well as AU6 (cheek raiser), AUC = 0.518 in Exp.5, AU12 (lip corner puller), AUC = 0.515 in Exp.4, AU15 (lip corner depressor), AUC = 0.510 in Exp.4, AUC = 0.518 in Exp.7, and AU27 (mouth stretch), AUC = 0.519 in Exp.8. The best generalisation performance of the model was obtained from Exp.2 and Exp.6 with an achieved average best prediction of 0.993 for all the AUs, whilst the second best prediction was 0.989.

The additional experiment for training DCGAN was conducted on a limited number of positive samples having a specific AU (from the CK dataset) which is AU7 (lid tightener). Unfortunately, it was found that the performance became lower in the AUC = 0.58, for testing and for training the AUC = 0.842, since it could be hard detecting AU7 and it is not easy to distinguish it between the other AUs. Fig. 4.12 shows some image samples of AU7 from the enhanced CK dataset.

4.6 Conclusions

DCGAN has emerged as an effective pre-training approach for emotion recognition. The experiments have been achieved on six standard datasets to demonstrate the advantage of this proposed work. This study concluded that training unsupervised DCGAN on a large-scale dataset produces powerful discriminative representation features for predicting and detecting AUs / emotions from frontal face images, which is better than representing the multi-view of facial face images. Results comparative studies with state-of-the-art methods show that this model outperformed, particularly, on the Radboud dataset with an overall accuracy of 98.57%. Future work could involve training a conditional DCGAN to disentangle the subject's expression from identity.



Accuracy :98.57% - EXP1

Output Class	1	2	3	4	5	6	7	8
1	98.50% 199	0.00% 0	0.00% 0	0.00% 0	0.00% 0	0.00% 0	1.00% 2	0.00% 0
2	0.00% 0	97.00% 197	0.00% 0	0.00% 0	0.00% 0	1.50% 3	0.50% 1	0.00% 0
3	0.00% 0	0.00% 0	100.00% 201	0.00% 0	0.00% 0	0.00% 0	0.00% 0	0.00% 0
4	0.00% 0	0.50% 1	0.00% 0	98.00% 198	0.00% 0	0.50% 1	0.00% 0	0.50% 1
5	0.00% 0	0.50% 1	0.00% 0	0.00% 0	100.00% 200	0.00% 0	0.00% 0	0.00% 0
6	0.00% 0	2.00% 4	0.00% 0	0.00% 0	0.00% 0	97.50% 197	0.00% 0	0.00% 0
7	1.50% 3	0.00% 0	0.00% 0	0.50% 1	0.00% 0	0.50% 1	98.00% 196	0.00% 0
8	0.00% 0	0.00% 0	0.00% 0	1.50% 3	0.00% 0	0.00% 0	0.50% 1	99.50% 197

Target Class

Accuracy :95.52% - EXP2

Output Class	1	2	3	4	5	6	7	8
1	95.60% 197	0.50% 1	0.00% 0	0.00% 0	0.00% 0	1.40% 3	0.00% 0	0.00% 0
2	1.00% 2	91.60% 185	0.00% 0	0.00% 0	0.00% 0	6.20% 13	0.50% 1	0.00% 0
3	0.00% 0	0.00% 0	100.00% 200	0.00% 0	0.00% 0	0.50% 1	0.00% 0	0.00% 0
4	0.00% 0	0.00% 0	0.00% 0	95.50% 189	0.00% 0	0.50% 1	2.10% 4	3.40% 7
5	0.00% 0	0.50% 1	0.00% 0	0.00% 0	100.00% 200	0.00% 0	0.00% 0	0.00% 0
6	0.00% 0	5.90% 12	0.00% 0	0.00% 0	0.00% 0	89.40% 186	1.60% 3	0.00% 0
7	3.40% 7	1.50% 3	0.00% 0	2.00% 4	0.00% 0	1.90% 4	95.80% 183	0.00% 0
8	0.00% 0	0.00% 0	0.00% 0	2.50% 5	0.00% 0	0.00% 0	0.00% 0	96.60% 196

Target Class

Accuracy :75.19% - EXP3

Output Class	1	2	3	4	5	6	7	8
1	73.20% 254	7.50% 24	1.20% 4	0.00% 0	0.30% 1	7.30% 26	7.50% 26	0.00% 0
2	10.40% 36	52.80% 169	0.30% 1	1.90% 6	1.80% 6	27.90% 99	4.90% 17	0.30% 1
3	1.40% 5	0.60% 2	95.70% 312	0.00% 0	1.20% 4	1.70% 6	0.60% 2	1.20% 4
4	0.30% 1	0.90% 3	0.00% 0	82.30% 260	0.60% 2	4.80% 17	6.70% 23	8.60% 29
5	0.60% 2	1.90% 6	1.20% 4	0.30% 1	95.80% 318	1.10% 4	0.00% 0	0.00% 0
6	6.60% 23	29.10% 93	0.00% 0	4.10% 13	0.00% 0	46.50% 165	11.30% 39	0.60% 2
7	7.20% 25	6.90% 22	1.20% 4	5.40% 17	0.00% 0	8.70% 31	68.10% 235	0.30% 1
8	0.30% 1	0.30% 1	0.30% 1	6.00% 19	0.30% 1	2.00% 7	0.90% 3	89.10% 302

Target Class

Accuracy :69.4% - EXP4

Output Class	1	2	3	4	5	6	7	8
1	65.80% 241	8.60% 28	3.00% 10	0.30% 1	0.30% 1	9.00% 26	7.80% 28	0.00% 0
2	12.00% 44	44.50% 145	1.20% 4	2.60% 9	3.30% 11	31.50% 91	7.80% 28	0.90% 3
3	2.70% 10	2.10% 7	89.40% 294	0.30% 1	3.90% 13	1.70% 5	0.80% 3	0.60% 2
4	1.10% 4	2.10% 7	0.60% 2	74.10% 254	0.60% 2	3.10% 9	5.30% 19	11.40% 38
5	0.00% 0	3.40% 11	2.70% 9	0.60% 2	91.70% 309	1.40% 4	0.00% 0	0.00% 0
6	10.40% 38	31.30% 102	0.90% 3	3.50% 12	0.00% 0	40.80% 118	17.10% 61	0.30% 1
7	7.70% 28	7.70% 25	0.90% 3	8.20% 28	0.00% 0	11.40% 33	59.90% 214	1.20% 4
8	0.30% 1	0.30% 1	1.20% 4	10.50% 36	0.30% 1	1.00% 3	1.10% 4	85.60% 285

Target Class

Accuracy :97.64% - EXP5

Output Class	1	2	3	4	5	6	7	8
1	99.50% 199	0.00% 0	0.00% 0	0.00% 0	0.00% 0	0.00% 0	1.00% 2	0.00% 0
2	0.00% 0	96.90% 190	0.00% 0	0.00% 0	0.50% 1	4.40% 9	0.50% 1	0.00% 0
3	0.00% 0	0.00% 0	100.00% 201	0.00% 0	0.00% 0	0.00% 0	0.00% 0	0.00% 0
4	0.00% 0	0.00% 0	0.00% 0	96.50% 194	0.00% 0	0.00% 0	1.00% 2	2.50% 5
5	0.00% 0	0.00% 0	0.00% 0	0.00% 0	99.50% 201	0.00% 0	0.00% 0	0.00% 0
6	0.00% 0	31.00% 6	0.00% 0	0.00% 0	0.00% 0	94.60% 194	0.50% 1	0.00% 0
7	0.50% 1	0.00% 0	0.00% 0	0.50% 1	0.00% 0	1.00% 2	96.60% 197	0.00% 0
8	0.00% 0	0.00% 0	0.00% 0	3.00% 6	0.00% 0	0.00% 0	0.50% 1	97.50% 194

Target Class

Accuracy :61.87% - EXP6

Output Class	1	2	3	4	5	6	7
1	56.90% 534	0.00% 0	2.20% 8	2.50% 69	3.10% 53	3.60% 12	4.00% 69
2	7.00% 66	93.10% 27	1.10% 4	2.50% 69	3.30% 56	13.40% 44	0.40% 7
3	2.30% 22	0.00% 0	41.30% 150	4.10% 112	7.30% 125	7.30% 24	5.60% 97
4	5.70% 53	3.40% 1	6.60% 24	66.70% 1810	6.40% 111	4.30% 14	5.00% 88
5	5.10% 48	3.40% 1	10.50% 38	6.60% 180	47.40% 817	5.50% 18	8.10% 141
6	4.30% 40	0.00% 0	7.40% 27	2.90% 78	3.70% 63	58.90% 193	1.30% 22
7	18.70% 175	0.00% 0	30.90% 112	14.60% 397	28.90% 497	7.30% 24	75.70% 1319

Target Class

Accuracy :60.44% - EXP7

Output Class	1	2	3	4	5	6	7
1	62.90% 266	22.40% 46	0.50% 2	2.20% 8	3.00% 9	1.70% 6	3.10% 13
2	24.80% 105	41.50% 85	7.70% 28	7.90% 29	13.60% 41	7.50% 27	8.00% 34
3	0.70% 3	7.80% 16	66.40% 243	4.60% 17	9.30% 28	9.20% 33	1.90% 8
4	2.60% 11	4.40% 9	3.30% 12	81.60% 301	1.00% 3	1.70% 6	1.90% 8
5	3.10% 13	11.70% 24	10.70% 39	1.10% 4	47.20% 142	10.30% 37	21.40% 91
6	2.10% 9	7.30% 15	10.90% 40	1.40% 5	8.60% 26	59.50% 213	9.90% 42
7	3.80% 16	4.90% 10	0.50% 2	1.40% 5	17.30% 52	10.10% 36	53.90% 229

Target Class

Accuracy :44.52% - EXP8

Output Class	1	2	3	4	5	6	7
1	54.70% 41	5.50% 5	0.00% 0	11.20% 35	7.80% 15	7.00% 22	10.10% 24
2	6.70% 5	65.90% 60	7.50% 3	7.70% 24	5.70% 11	1.90% 6	5.10% 12
3	1.30% 1	2.20% 2	55.00% 22	3.80% 12	3.10% 6	7.00% 22	3.40% 8
4	5.30% 4	11.00% 10	10.00% 4	38.80% 121	19.30% 37	10.50% 33	13.10% 31
5	13.30% 10	6.60% 6	7.50% 3	15.10% 47	38.50% 74	17.3 54	10.50% 25
6	5.30% 4	3.30% 3	15.00% 6	12.50% 39	12.00% 23	43.50% 136	12.70% 30
7	13.30% 10	5.50% 5	5.00% 2	10.90% 34	13.50% 26	12.80% 40	45.10% 107

Target Class

Fig. 4.8 The average recognition rate of a Confusion matrix is obtained from SVM classifier for the eight experiments.

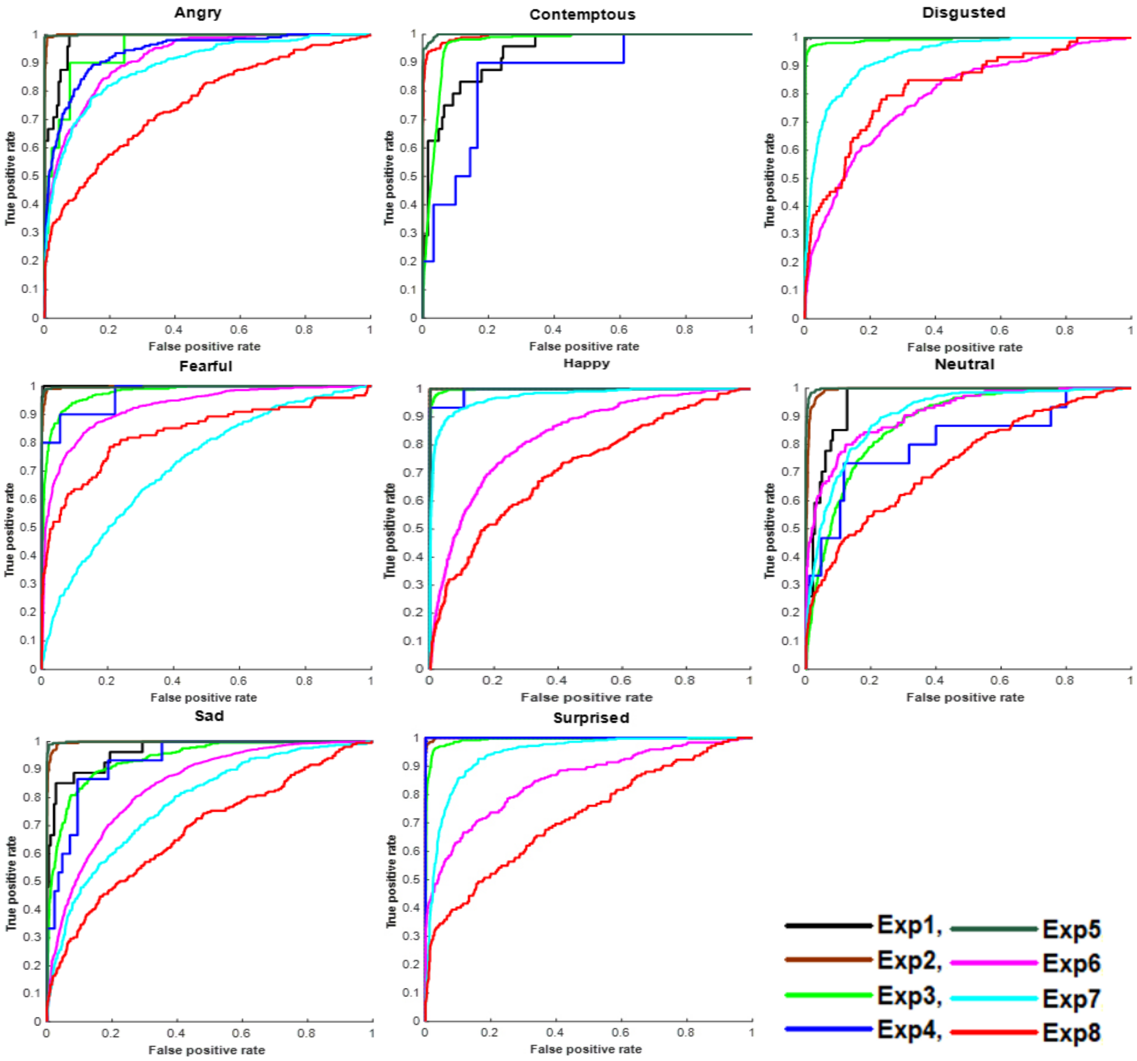


Fig. 4.9 Receiver Operating Curves (ROC) for eight emotions and eight experiments; each figure depicts ROC curve of eight dissimilar experiments.

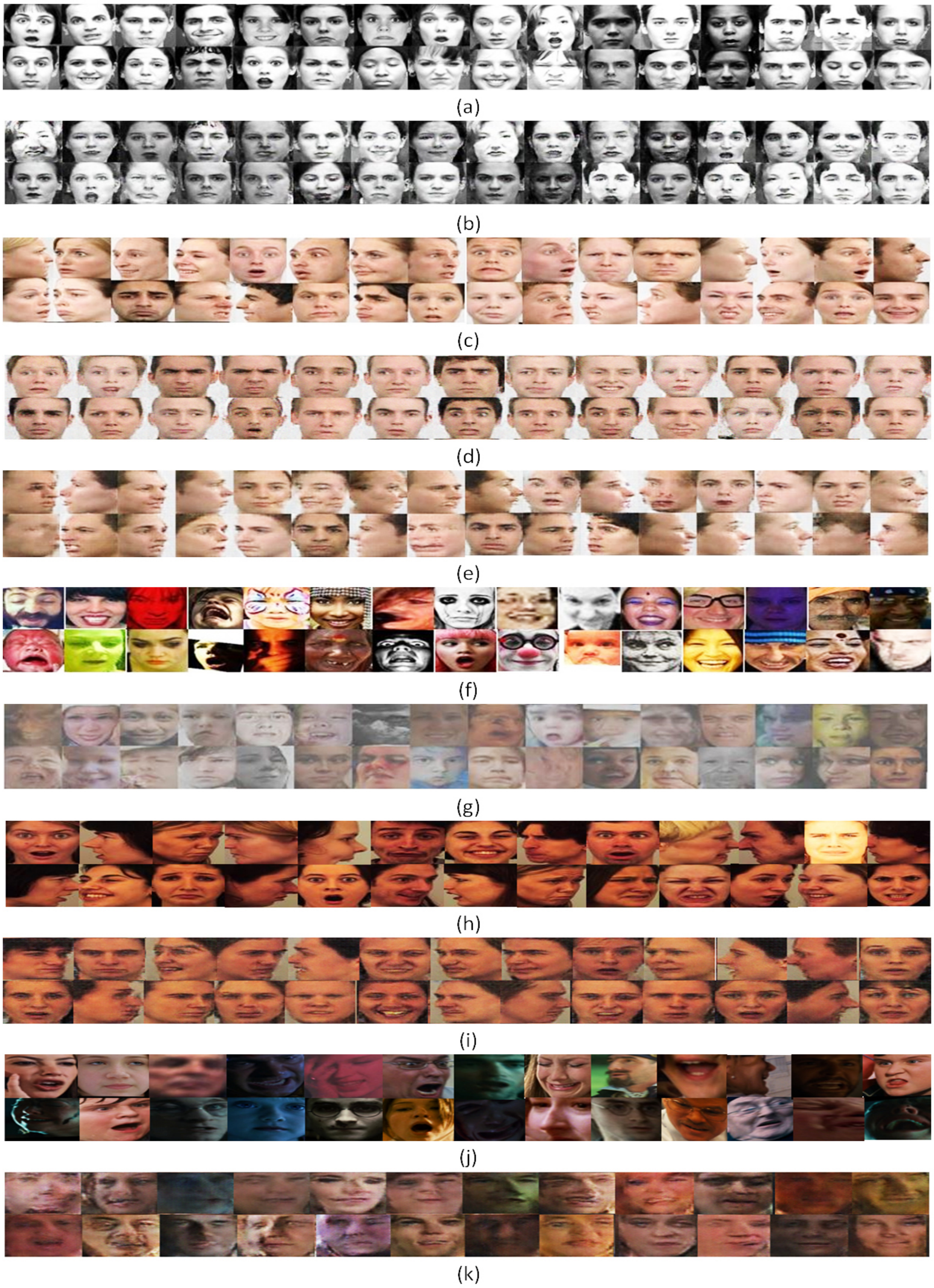


Fig. 4.10 Comparison of the selected samples of the generated facial expression images using DCGAN on different datasets: (a) & (b) enhanced Cohn-Kanade original and generated images, (c) original frontal and multi-view Radboud images, (d) frontal Radboud generated images, (e) multi-view Radboud generated images, (f) & (g) RAF original and generated images, (h) & (i) KDEF original and generated images, (j) & (k) SFEW original and generated images.

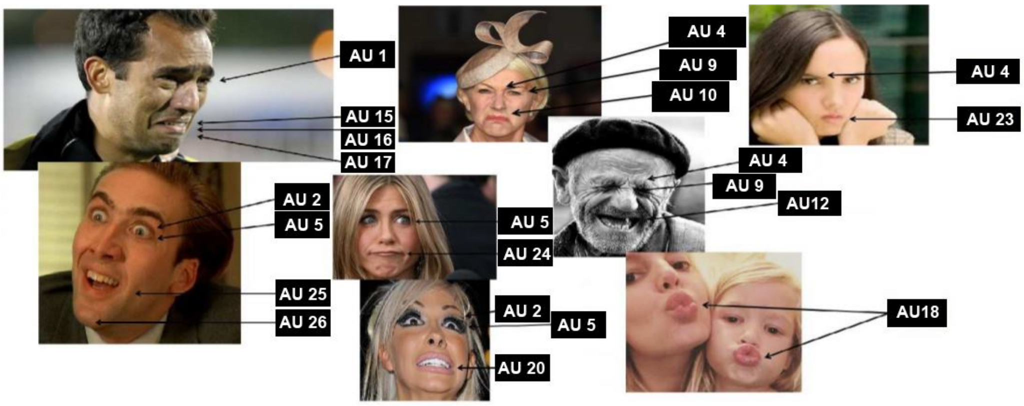


Fig. 4.11 Images labelled by FACs, image adapted from [618].

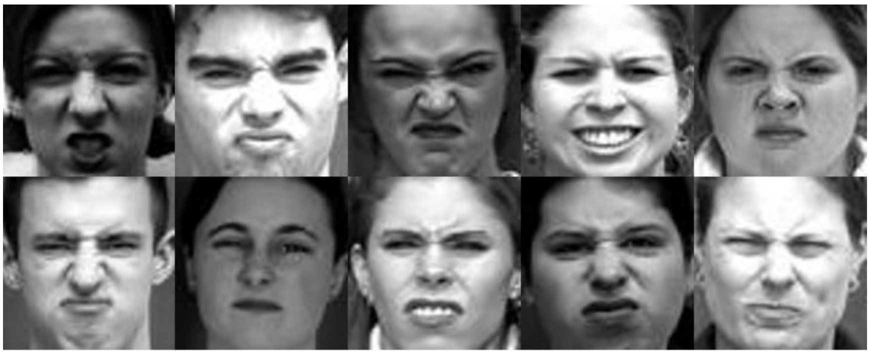


Fig. 4.12 Images from the enhanced CK dataset represent AU7.

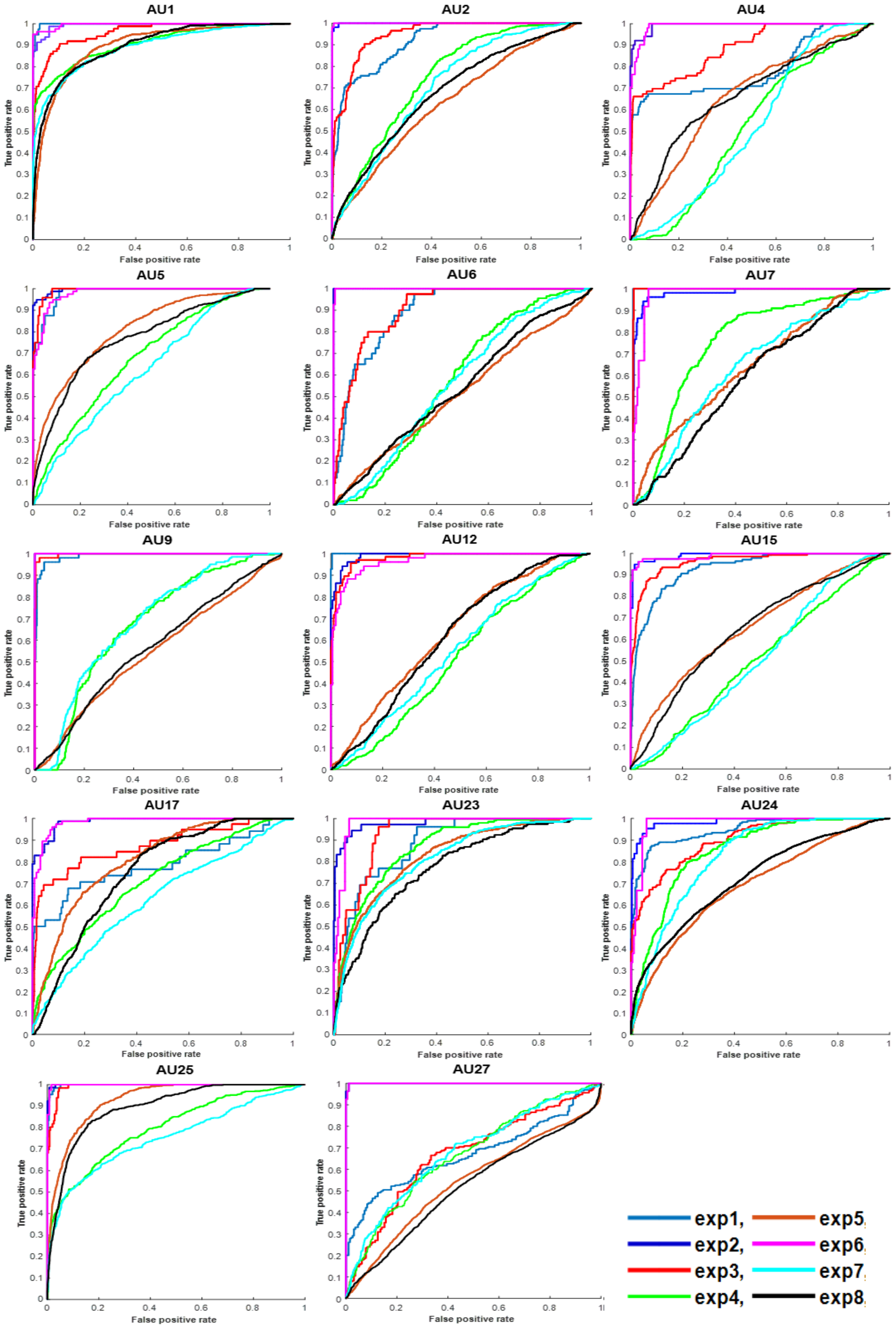


Fig. 4.13 Receiver Operating Curves (ROC) for fourteen Action Units (AU) and eight experiments, each figure depicts ROC curve of eight dissimilar experiments.

Chapter 5

Disentangling Identity from Expression for Facial Emotion Recognition

5.1 Introduction

The approach put forward by this chapter is a novel technique aimed at estimating emotions from "in-the-wild" images using one RGB image to construct a 3D morphable face. The proposed network is trained using a large-scale dataset comprised of facial videos (FaceVid). These videos are rich and abundant in their scope of facial expression dynamics, identities, appearance, and 3D pose changes. Before the network being trained, the images were annotated automatically via the assistance of a 3D face reconstruction method based on particular facial expression parameters represented at an intermediate level. Where image annotations have factors such as pseudo ground truth camera poses (scale, rotation angles, and translation parameters), facial landmarks, and 3D face fitting results. In order to take into account such factors an updated dataset was introduced. The main goal of this database is to take a 3D face model defined by the parameters of identity and facial expression, and later use this database to train the network to focus on regressing the expression. This dataset trains a CNN that is given an RGB image, which it then outputs in the form of the expression parameters. The classification network subsequently takes these parameters and uses them in the prediction of facial expression. After the network has been trained and fed with all the relevant data, a multi-class SVM is used to take a face image and regress an expression vector from it and this is divided into seven emotion classes.

The trained models are surpassed the state-of-the-art approaches when estimating facial expression parameters. It also performs at a comparatively similar degree to state-of-the-art

FER from images, and they can subsequently be applied to other frameworks aiming to address tasks such as recovering the 3D facial images geometry, facial reenactment, and image-to-image translation, to name just a few. In order to make the model more robust when faced with inter-subject variability, as well as capable of relaxing constraints based on appearance changes, it is necessary to put together an extensive dataset that comprises of in-the-wild facial videos that we called 'FaceVid'. This FaceVid is then be annotated frame by frame using a pseudo-ground truth 3D facial reconstruction that is made of separable model-based identity and expression vectors. Following this, the model is trained on the annotated FaceVid dataset as then it can better regress the correct facial expression based on one image. It is still a point of contention, however, as to whether or not simple human emotions can be identified using one RGB image employing geometric 3D-based expression features, and whether this can consequently enhance the classification issue of 7-classes (neutral + six basic emotions) in the wild.

We build on advancements made more recently in 3D face reconstruction from monocular videos as well as Convolutional Neural Networks (CNNs) architectures, which are useful in the field of CVs. The chosen technique is based on separating the identity of the subject from the facial expression using 3D Morphable Models (3DMM) [51]. Per-image a vector is regressed using a Deep Convolutional Neural Network (DCNN), and this vector is representative of the given subject's expression. In terms of being a feature vector, this expression vector works perfectly as a broad range of invariances are successfully reached by it. An emotional classifier that makes use of this vector is trained to identify expressions accurately. The ways in which this chapter contributes to the field are as follows:

- Large-scale datasets of facial videos (6,000 overall) are both gathered and annotated. These are known as FaceVid. Through the assistance of our designated model-based formulation, the following annotation was added to each video per-frame: 1) 68 facial landmarks, 2) a 3D facial shape, which is made up of identity and expression parts, 3) a 3D pose of the head, in addition to the scale of the postulated scaled orthographic camera model.

- The chosen Deep Convolutional Neural Network (DCNN) is robust in terms of using a single input image and subsequently regressing it into the specific expression parameters of a 3D Morphable Model of the facial shape. The network that we have selected is not impacted by changes in view angle, occlusions, and illumination. It is also able to separate independently the expression from the identity of the subject.

- A back-end classification stage is also used with the network for the estimated expression vectors for the Facial Emotion Recognition (FER) task. This means that the framework is subsequently integrated and able to more robustly use single images to identify facial expressions.

There have been many attempts to solve the seven-class problem of static facial expression recognition in the wild, which was first identified by Ekman and Friesen [150]. Many experiments have sought to address this problem. Existing methods for facial expression recognition fall into one of two groups: traditional hand-crafted techniques, such as appearance, geometric, dynamic, and fusion, and deep hierarchical learning representation models. FER has generally taken the hand-crafted methods approach with a reliance on features [266] [127] [650] [651]. However, in terms of practical applications, this method can be limited [292] [316]. More recently, deep learning approaches are more widely used in various vision tasks including image classification, face recognition, and emotion recognition. For more extensive surveys on automatic FER, the reader can look to [627] [483] [426] [160], in addition to [634] [333] as these focus on deep learning related methods.

In contrast to other approaches, the estimated intermediate 3D-based representation of "clear" facial expressions is not unchanged by other factors influencing the input image's formation, such as variations in shape, appearance, strong illumination, relative 3D pose, occlusions, alongside further challenges of in-the-wild conditions. Unlike standard approaches to Deep Learning in CV, the aim is not to solve the problem (recognising facial expression from only one single RGB images) from the perspective of "end-to-end". This is especially so because a large selection of images that are manually edited (labelled) on the basis of facial expression classes would be necessary for this approach. This would therefore be exhaustive and vulnerable to human annotation errors. Rather, we create a very large-scale video dataset, the images from which we will subsequently annotate per frame based on the existing level of expression parameters of a 3DMM. This particular annotation process is something that could successfully automate using a method of reconstructing faces in 3D from existing videos. This technique produces results that are highly accurate as it exploits the diverse and dynamic range of information contained in facial videos. Based on this dataset we are able to identify the process behind regressing expression parameters from single RGB images; the issue of FER is massively made easier because the features given to the emotion classifier are the expression's parameters. There are only 28 dimensions in these parameters and so they are both low-dimensional and able to provide a broad range of potential invariance properties. When being trained, the expression classifier can thus provide optimal results with far less training data that is annotated manually on the basis of expression classes. The resulting FER system from this method is more robust and able to manage images that are significantly more challenging.

The chosen technique is based on using 3D Morphable models (3DMM) to separate the identity of the subject from its expression. Most existing techniques in the discipline identify emotions using features that relate to the appearance of input images. "In the wild" datasets feature wide differences in pose, camera properties, and illumination, leading to large inter-subject and intra-subject variability. Furthermore, a detailed image formation process is used to create these images, which includes scene-related physical interactions of light rays. It can be problematic to depend on RGB values when trying to derive the necessary features. This is because robustness issues often emerge, such as pose, camera viewpoint, sensitivity to illumination, and occlusions. In particular, this is a problem when encountering emotion recognition, which is in-the-wild. On the other hand, it is possible to use a model-based method in order to redeem the geometry of the input face, as long as this model is able to disentangle both the reconstructed identity and expression. Doing so will reduce those parts of the emotion recognition task that are most constraining.

The remainder of this chapter is structured in the following way: Section 5.2 presents a more detailed methodology that deals with 3D face reconstruction, ground truth annotations, dataset collection, pre-processing, and training the model. The experimental findings on the four selected publicly available benchmark datasets are detailed in section 5.3. Section 5.4 ends the chapter with proposed future endeavours.

5.2 Methodology

Figure 5.1 provides an overall summary of the chosen framework and is divided into three sub-figures (a), (b), and (c) to further improve the clarity of the proposed work. Based on the advancements made in the 3D facial reconstruction from images as well as the comprehensive information accompanying videos of facial performances that was acquired, the internet was used to gather a large-scale dataset of facial videos (section 5.2.2). From this dataset, we took the per-frame 3D geometry using 3D Morphable Models (3DMMs) [51] of identity and expression with the assist of the process (section 5.2.1.1). A deep Convolutional Neural Network (CNN) was then trained by the dataset that had already undergone annotation, and this process was supervised in order to regress the expression coefficients vector (e') from one input image I_f (section 5.2.3). The final stage in this process was to add a classifier to the network's output as this would guess the approximate emotion of the given facial expression, and this would then be trained and tested on standard benchmarks for FER (section 5.2.4).

The entire emotion recognition framework was visualised in the last part (c) of Fig. 5.1 and was detailed in section 5.2.4 of this chapter. The details of the previous sub-figures (a and b) are described in section 5.2. In general, this model includes three phases. The

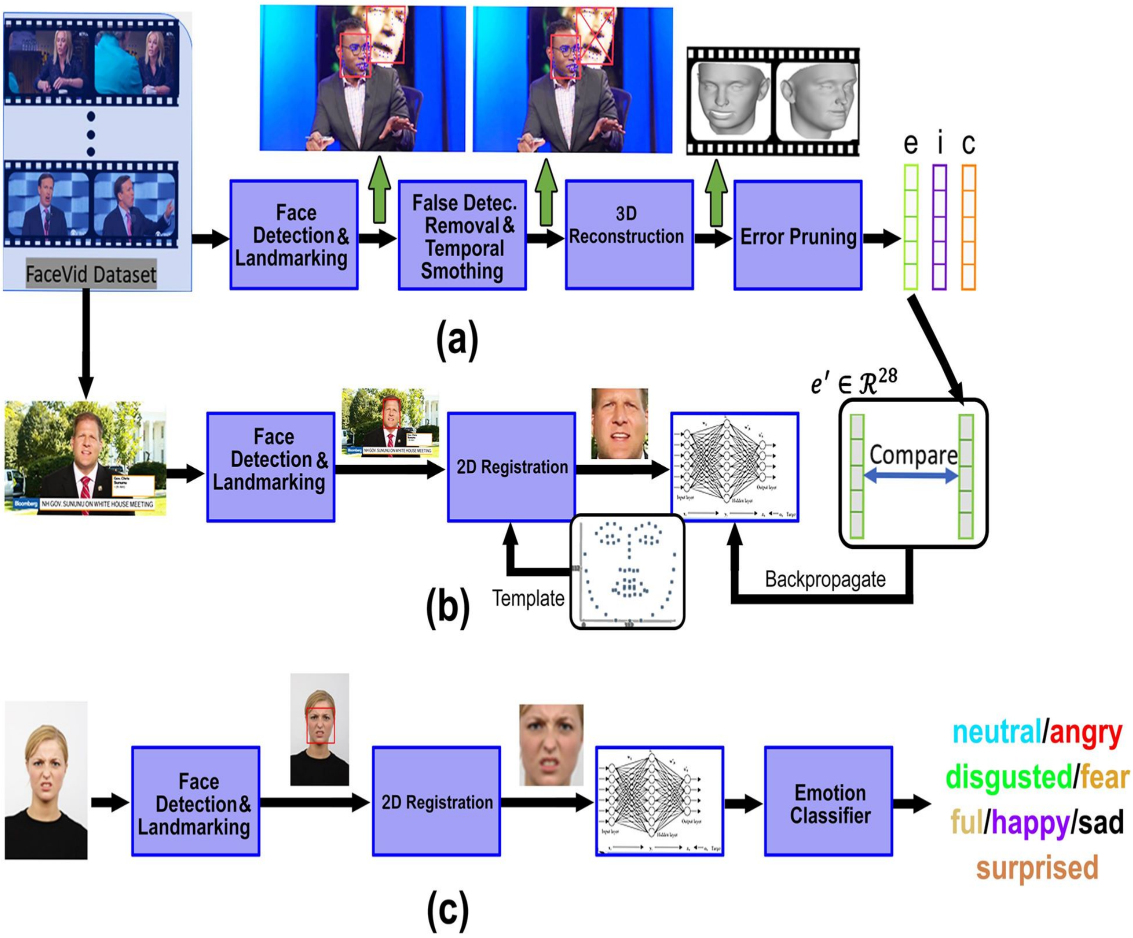


Fig. 5.1 The proposed methodology of Facial Expression Recognition (FER) system from images. Top row (a): the process of the FaceVid annotation (sec. 5.2.1.1 and 5.2.2). Center (b): the training stage of the network (sec. 5.2.3). Bottom row (c): the last framework step of FER (sec. 5.2.4).

first stage starts with pre-processing steps, which consists of face detection, landmarking, registration to a mean face template step, and then the process of 3D reconstruction. The most important step in this phase is the 2D Procrustes registration in the image space. Registration was significantly decreasing the variability of the processed images and can be made the estimations being robust and reliable [397]. More particularly, we utilise the Procrustes alignment to calculate the resemblance transformation parameters that align the elicited 68 landmarks from every employed facial image and have been utilised to register them to a 2D mean face template. These 68 landmarks have come from the projections of a 3D landmarks on the mean face of the used LSFM model. As a result, the input image is afterward warped depending on these gauged resemblance transformation parameters, and this step was a 2D alignment (termed as 3D-aware 2D landmarks which were extracted by [122] in the image space) and it would not be a 3D registration phase. The locating of the 68 landmarks using this approach is identical to the projections of it is corresponding 3D points at the 3D face. This was served by way of a facial normalisation stage of the input image before entered the second stage of deep facial expression 3D CNN.

The second stage (b) of this figure was used to train the facial expression network model in section 5.2.3 by extracting a feature vector of facial expression presented by the input image. Then, a comparison was achieved between the facial expression vector (e) that comes from the first stage (Face Vid pseudo annotations facial expression parameters created in section 5.2.1.1) and facial expression vector produced from the network (e') in order to update the network using ℓ_2 norm error.

Emotion recognition was accomplished at the last classification stage (c) of the whole framework, which classifies the facial expressions that have been already predicted by the trained network model (second stage) on different benchmark datasets (trained and tested) for emotion recognition in section 5.2.4 using an SVM classifier as a final process.

5.2.1 3D Morphable Models (3DMMs)

In this section, we give more details about the suggested method of 3D face reconstruction from videos that are adopted to generate pseudo-ground truth in the training set.

5.2.1.1 3D Face Reconstruction From Videos

The 3D face reconstruction approach has been used in the current work as a recovery of original 3D face geometry. We do not have actual 3D ground truth data, so we use the 3D face reconstruction results of our method as pseudo data. In order to make successful pseudo-ground truth, it is vital that the implementation of 3D face reconstruction is done on a larger-scale video dataset, which will guarantee more accuracy and efficiency. Based on this, the choice was made for the adopted high-quality 3DMM model to be fitted on a series of facial landmarks on each video in the dataset. Given that the aim of this is to create pseudo-ground truth on a broad selection of videos, the needs of online performance are in no way potential limitations. We use the method of [121] (except in regards to the initialization stage, which will be discussed further shortly), which is a batch method that considers the existing information from all video frames alongside making use of the extensive information normally inside the facial videos. The given technique centres on energy minimization in order to align with the joint identity and expression 3DMM model on facial landmarks that was taken from the entirety of the frames of the input video at the same time. After

[121], we show how the existing state-of-the-art facial landmarking is able to attain results in landmark localization that are extremely accurate; it is possible to combine both the landmarks' information and high-quality 3D face models and attain 3D face reconstruction results that are both reliable and robust. The assumption made is that Scaled Orthographic Projection (SOP) is done by the chosen camera; in addition, the parameters for the identity i are rigid yet unknown across the entirety of the frames. Therefore, e_f, the parameters for expression, alongside the parameters for the camera (scale and 3D pose) change per-frame. In summary, then, a cost function comprising of three terms is made smaller. These terms are: 1) a sum of squared 2D landmark reprojection errors across all frames, 2) a shape priors term that imposes a quadratic prior over the identity and per-frame expression parameters, and 3) a temporal smoothness term that enforces temporal smoothness of the expression parameters by using a quadratic penalty of the second temporal derivatives of the expression vector. Alongside this, to manage outliers—outliers in this sense being frames that have occlusions which are strong and subsequently result in gross errors in the landmarks - it is necessary to use box constraints when dealing with parameters related to identity and per-frame expression. Given the assumption that a value has been calculated for the camera parameters during an initialization stage, the consequence of minimising the cost function is a large-scale least squares problem that has box constraints. The Newton method of [99] is used to solve this.

5.2.1.2 Combined Identity and Expression 3D Face Modelling

Following the recent used approaches [617],[121],[296],[172], the 3D face geometry was subsequently based on 3DMMs as well as the additional combined identity and expression variation. More specifically, the initial use of 3D Morphable Models (3DMM) was put forward by the novel work of Blanz and Vetter [51] on the basis of being a linear parametric model that had a compatible framework for the 3D modelling of human faces from surfaces and 2D images. $\mathbf{x} = [x_1, y_1, z_1, ..., x_N, y_N, z_N]^T \in \mathbb{R}^{3N}$ is the vectorized configuration of a 3D facial shape made up of N 3D vertices. It is assumed here that it is possible to create a representation of all facial shapes $\mathbf{x}$ via the given model of shape change:

$$\mathbf{x}(i,e) = \bar{\mathbf{x}} + U_{id}i + U_{exp}e \tag{5.1}$$

In that $\bar{\mathbf{x}} \in \mathbb{R}^{3N}$ is the mean shape vector, specified by $\bar{\mathbf{x}} = \bar{\mathbf{x}}_{id} + \bar{\mathbf{x}}_{exp}$, $\bar{\mathbf{x}}_{id}$ and $\bar{\mathbf{x}}_{\mathbf{exp}}$ are also the mean of identity and expression shape vectors respectively. $\mathbf{U}_{id} \in \mathbb{R}^{3N \times n_i}$ represents the orthonormal basis of $n_i = 157$ principal components ($n_i \ll 3N$) , $\mathbf{U}_{exp} \in \mathbb{R}^{3N \times n_e}$ represents the orthonormal basis of the $n_e = 28$ principal components ($n_e \ll 3N$), and $\mathbf{i} \in \mathbb{R}^{n_i}$, and

$\mathbf{e} \in \mathbb{R}^{n_e}$ are together the identity and expression parameters. Using standard PCA (Principal Component Analysis) as a statistical tool to reduce the dimensions, which represent a core in constructing a 3DMM. PCA is applied to the vectorised training images of facial meshes. It was used to synthesis a parametrisable representation through the dependence on uncorrelated groups of latent coefficients. A fitting of the PCA model often recognised as a fitting of a multivariate Gaussian distribution. This gives the ability to prevent the construction of non-realistic faces because this model can be utilised as a statistical prior. The 3D facial shape $\mathbf{x}$ inside the chosen model (5.1) is a function of both identity and expression coefficients $(\mathbf{x}(\mathbf{i},\mathbf{e}))$. Furthermore, the variations in expression were efficaciously illustrated as offsets from the presented identity shape.

The aspect of the model that is the identity $\{\bar{\mathbf{x}}_{id}, U_{id}\}$ initially stems from the Large Scale Morphable Model (LSFM) [57] [55] that was made via a combination of roughly 10,000 scans of different people. This was the biggest 3DMM ever made and it contained information from a variety of demographics. Alongside this, the aspect of the model that focused on expression $\{\bar{\mathbf{x}}_{exp}, U_{exp}\}$ came from the work of Zafeiriou et. al. [617]. It was created through implementing the blendshapes model of Facewarehouse [67] as well as taking on Nonrigid ICP [85] to register the blendshapes model with the LSFM model.

The successful creation of pseudo-ground truth is contingent on performing 3D face reconstruction on a large video dataset. It is necessary for this dataset to be not only effective but also precise. The decision was made to fit the chosen 3DMM model on a set of facial landmarks for each video of the dataset. Given that this adopted approach [121] is aimed at creating pseudo-ground truth on a wide assemblage of videos, online performance is in no way a limitation. This is a batch method that considers the existing information from all video frames alongside making use of the rich extensive dynamic information normally found inside the facial videos. The chosen approach is based on minimising energy to set both the composed identity and expression of the 3DMM model on facial landmarks from all image frames of those used on video at the same time.

5.2.1.3 Initialization Stage of Estimating Camera Parameters

Rigid Structure from Motion (SfM) was used to approximate the parameters of the camera during the initialization stage of the 3D video reconstruction that was put forward in [121]. Facial videos that have a significant degree of head rotation particularly benefit from this approach, especially as it allows for the necessary amount of variation in the relative 3D pose that SfM generally requires. Nonetheless, when faced with videos that have relatively little head rotation, such as a video where a person is stationary, staring straight at the camera and speaking, the results from SfM provide an unstable approximation of the camera's parameters.

This is because of the level of variances and inexactness that comes from watching the scene from primarily one perspective. Overcoming this constraint and utilising a broader range of facial videos requires implementing a significantly different method at this particular juncture. The chosen approach must use prior 3D face models and successfully address the problem in order to achieve accurate and powerful approximations. More specifically, like in [121], the assumption made by the initialization stage is that throughout the entirety of the video the chosen shape recovered remains solid. Although this assumption is somewhat over-simplified, it is equally sufficient for providing an accurate estimation of camera parameters. This is especially as distortions in the human visage can easily be modelled as restrained shifts from a shape that is otherwise rigid. As opposed to [121], the aim is also not to fully estimate the complete degrees of freedom of the 3D facial shape (i.e. each coordinate of all the points of the 3D shape are differing independent parameters); rather, the goal is to largely reduce the permitted degrees of freedom through assigning the limitation of a 3D face model being used to synthesise it (5.1). From this the camera estimations are much stronger.

In light of this we hope to approximate the identity **i** alongside expression **e** parameters of the rigid facial shape, in addition to the frame by frame camera parameters expressed as the SOP camera projection matrix $\Pi_f \in \mathbb{R}^{2\times3}$ for every frame f (Π_f corresponds to the first 2 rows of the rotation matrix multiplied with the scale parameter). The application of the chosen estimation is done primarily based on the minimisation of the reprojection error term of the overall cost function illustrated in Sec. 5.2.1.1, this being the sole term dependent on the camera parameters. This minimisation is represented as follows:

$$\text{minimise } E_{land}(\Pi_1,..,\Pi_{n_f};\mathbf{i},\mathbf{e}) = \sum_{f=1}^{F}\sum_{j=1}^{L}\left\|\Pi_f\left(\bar{\mathbf{x}}^{(l_j)} + \mathbf{U}_{id}^{(l_j)}\,\mathbf{i} + \mathbf{U}_{exp}^{(l_j)}\,\mathbf{e}\right) - \ell_{j,f}\right\|^2 \tag{5.2}$$

where $\ell_{j,f} \in \mathbb{R}^2$ is the 2D location of the j-th facial landmark ($j = 1,..,L$) in the f-th frame of the input video ($f = 1,..,F$). Additionally, $\bar{\mathbf{x}}^{(l_j)} \in \mathbb{R}^3$, $\mathbf{U}_{id}^{(l_j)} \in \mathbb{R}^{3\times n_i}$ and $\mathbf{U}_{exp}^{(l_j)} \in \mathbb{R}^{3\times n_e}$ are the 3 rows of $\bar{\mathbf{x}}$, $\mathbf{U}_{id}$ and $\mathbf{U}_{exp}$ respectively that correspond to the x, y and z coordinates of the vertex of the dense facial shape with index l_j, which is associated with the j-th landmark. Also, like the central 3D face reconstruction, **box constraints** are applied on the shape parameters of **i** and **e** respectively.

The above-mentioned problem can be resolved via taking on a different approach to minimisation that involves shape parameters $\{\mathbf{i},\mathbf{e}\}$ and camera parameters $\{\Pi_1,..,\Pi_{n_f}\}$. To begin the alternation we set the **i** and **e** to zero vectors; this aligns with the mean shape $\bar{\mathbf{x}}$. Subsequently, there is an alternation from step one to step two and so forth, as defined below:

1. Shape parameters $\{\mathbf{i},\mathbf{e}\}$ remain rigid while the camera parameters are changed $\{\Pi_1,..,\Pi_{n_f}\}$ through reducing E_{land} (5.2) with respect to $\{\Pi_1,..,\Pi_{n_f}\}$. For each frame, the minimisation is decoupled before being estimated via the extended POS method of Bas et al. [39].

2. Camera parameters $\{\Pi_1,..,\Pi_{n_f}\}$ remain rigid this time and shape parameters are modified $\{\mathbf{i},\mathbf{e}\}$ through reducing E_{land} (5.2) with respect to $\{\mathbf{i},\mathbf{e}\}$ via the given box constraints. Once more, this optimisation is a least squares optimisation that has box constraints. These are resolved through a reflective Newton method. The given dataset was previously implemented in the process of training a deep Convolutional Neural Network that identified expression parameters for a 3D Morphable Model alongside joining this with a back-end emotion classifier [99].

Empirical evidence demonstrates how the amount of the above-mentioned iterations only needs to be minimal (e.g. 5), as following the iterations, convergence is reached with negligible updates on the approximated parameters. Given that this is the last process required by the initialisation step, the series of approximate camera parameters (regarding areas like scale, rotation angles and translation parameters) are smoothed by cubic smoothing splines for a short time. It is key to note at the initialisation stage that the estimates made for the rigid shape parameters $\{\mathbf{i},\mathbf{e}\}$ are important only at an auxiliary level for easing the estimation of the camera parameters that happens to be the primary aim at this stage for the final output.

5.2.2 Ground Truth Creation from a Large-scale Videos Dataset

The purpose of this section is to illustrate the method behind using a large-scale dataset for building pseudo-ground truth; this is subsequently implemented for training the network, which is a strong CNN that can take an RGB image and regress the 3DMM expression from it. Initially a collection of 6,000 RGB videos was taken that has 12 million frames overall as well as 1,700 distinct identities. This video compilation was made via collecting roughly 1,500 videos from the $2MF^2$ dataset [89] and attaching an additional 4,500 more public YouTube videos[1] that were downloaded and cropped. In each video a minimum of one person acts in front of a single monocular camera. None of the shots were taken by multiple cameras simultaneously and the faces in the footage have variable resolutions that occurred in different conditions and setups, including conferences, vlogs, TV interviews,

[1] Only the videos have been chosen and facial landmarks from $2MF^2$ were not used as we chose to apply the 3D Aware 2D landmark configuration (3DA-2D) [121]. This, unlike the traditional 2D landmark configuration, provides an estimation of 3D landmarks on the image plane for real projected positions. An example of this is how it does not allocate landmarks of obstructing facial regions on the apparent face edge.

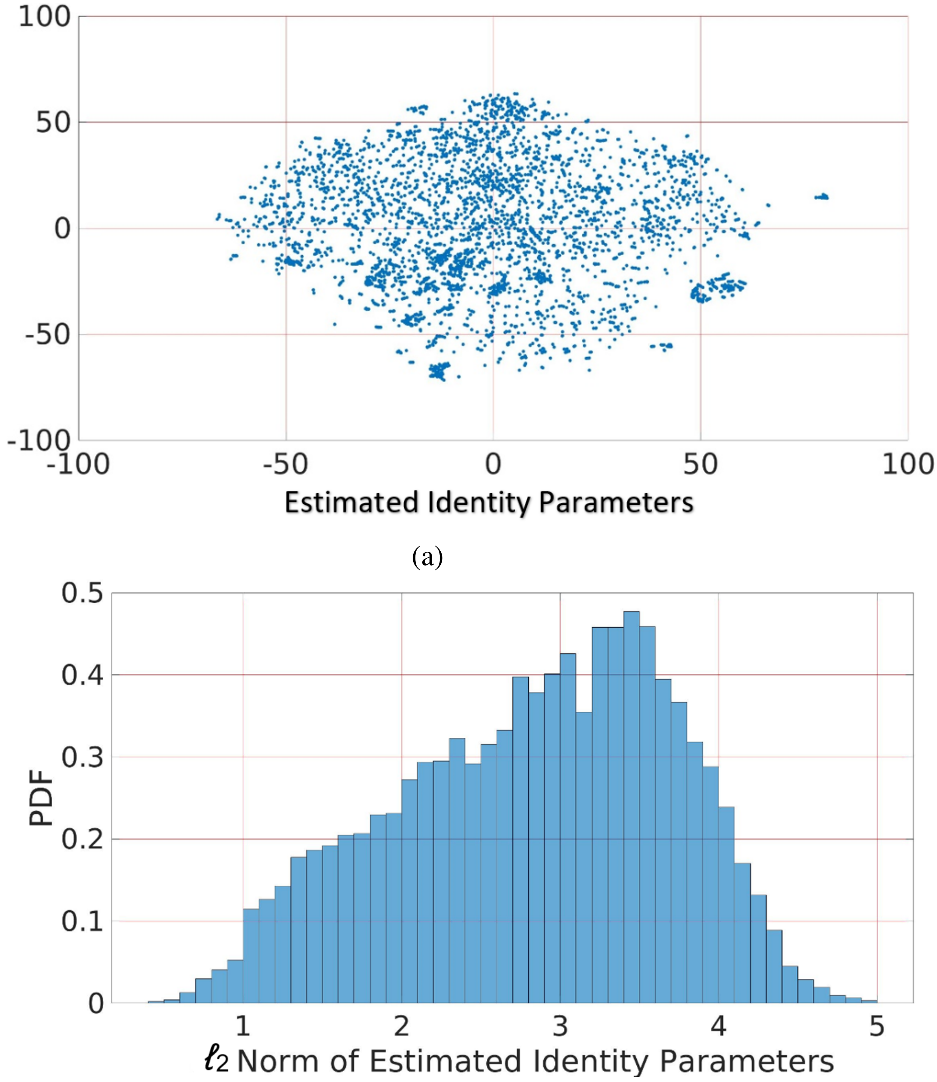
advertisements, and movies, etc. The approach of [122] was used for each frame in each video inside the collection, and was used to both identify faces and extract 68 landmarks from the faces. This was based on the MULTI-PIE markup scheme [201]. Following this, these steps were taken:

False detections removal: The procedure for doing this involved firstly taking the processed video and tracking every detected face in the initial frame from it. If the face's bounding box (BB) remains in a margin that is sensible it is retained. This had to be selected experimentally following the condition that it is half the width of the BB, in comparison to the prior frame's position. Faces that were difficult to track for K frames in a row (the decision was made for this number to be 5) were erased prior to arriving at the amount of tracked frames F that were wished for. This stage was beneficial insofar as it took away false detections that emerged from either a failure in the face detector or the detections that were out-of-context such as a facial photo in video's background, and faces that are getting in or getting out of camera viewing angle, etc.

Temporal smoothing: cubic splines were used short-term to smooth the extracted landmarks. The purpose of this was reducing the effects caused by possible jitters inside extracted landmarks from one frame to the other, alongside filling potential gaps (frames that lost track) that may have continued over a period under K frames.

3D facial reconstruction from videos: The procedure undertaken for each video is explained in Sec. 5.2.1.1 and estimated the parameters of facial shape (i, e_f for $f = 1,...,F$). The ultimate output of pseudo-ground truth generation is a sequence of expression vectors $\{e_f\}$. Nonetheless, identity vector i is still used for error pruning during the next stage.

Error pruning: As the amount of videos is significant, cases will necessarily exist in which reconstruction has been unsuccessful. This inevitable byproduct exists because the chosen landmark localisation, although strong, is not necessarily adequately accurate in situations where facial videos are very challenging. This is compensated for by the proposed approach insofar as there are two steps for pruning problematic videos. These are: **a) Automatic pruning:** This relies on the notion that under the chosen 3D face modelling (5.1), it is expected that the approximated identity vector i per video comes from a multi-variate normal distribution. The entirety of the videos are pruned with a value of $||i||^2$ above a threshold. **b) Manual pruning:** Other potentially problematic videos that "survived" the automatic pruning may exist also. Therefore, reconstructions of the videos that remain are



(a)

(b)

Fig. 5.2 Shows two figures in which (a) represents the t-SNE description that visualises the estimated 3DMM identity parameters (i) of the compiled videos of the dataset. (b) illustrates the ℓ_2 norm of an estimated probability density function of normalized (i), see equation (5.1) in this chapter.

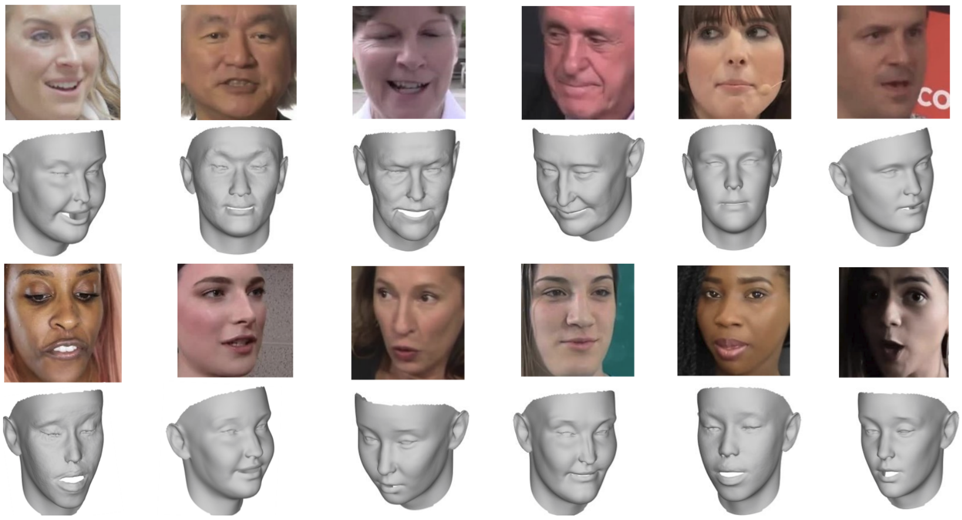


Fig. 5.3 An example of sample frames from the compiled dataset (FaceVid). Top row: display frames from various videos, whereas the bottom row represents the created 3D reconstructions.

examined, especially those where 3D face reconstruction has clearly been unsuccessful. The training set that has been compiled comprises of videos from the collection that survived this video pruning procedure. Within this, there are 5,000 videos, 1,500 different identities and roughly 9M frames.

The automatic pruning step detailed above relies on the idea that while the chosen 3D face modelling is used, the identity vector i that has been approximated for each video is presumably taken from a multi-variate normal distribution with zero mean and variance given by the identity matrix. Therefore, $||i||^2$ produces a squared Mahalanobis distance among the recent identity shape and the average (mean) identity (pursuing a chi-square distribution). From this it can be ascertained as to how plausible the estimated identity vector is under the assumed face model. These are therefore labelled as outliers and the videos that align with the estimated value of $||i||^2$ above a threshold θ are pruned automatically. θ was selected so that $||i||^2$ is assumed to be lower than θ with extremely high probability, beneath the estimated multi-variate normal distribution of i. The visualisation in Fig 5.2 shows a t-SNE [371] depiction of the estimated 3DMM identity parameters (i) of the gathered dataset videos as well as the estimated probability density function (PDF) of their Mahalanobis distance. In Fig 5.3 there is the opportunity to see a selection of randomly chosen frames from the dataset, in addition to their created 3D pseudo GT.

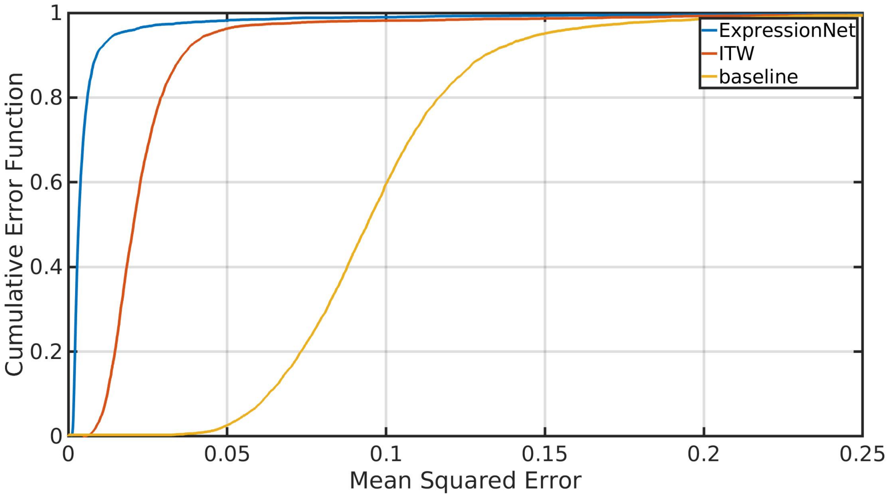


Fig. 5.4 Cumulative Error function of the mean squared error (MSE) between the facial expression coefficients estimation on the test split of FaceVid dataset by the trained Expression Network, ITW [56], and baseline [244][243] methods, and the ground truth. The average MSE values are 0.007, 0.026, 0.098 for the trained Expression Network, ITW, and baseline respectively.

5.2.3 The Trained Expression Network Model

The dataset of training videos that has been put together is filled with facial expressions that are diverse and viewable from both different angles and various illumination. This is a condition present throughout the video and overall there are 1,500 identities. Through this, training the Convolutional Neural Network (CNN) is made a lot easier, $N : I \to e$, and the purpose of this is to take an RGB image, I, and from it regress the 3DMM facial expression coefficients [in equation (5.1) it is referred to as e]. Throughout the training phase, the network's $N(I)$ ability to map from the image space to the facial expression space regardless of the identity of the subject demonstrated in image I is learned. It is possible to achieve this through facial 3DMM, which is representative of the reconstructed face as a combination of both the identity and the expression parts on top of the mean face of the model ($\bar{\mathrm{x}}$), equation (5.1). Vectors e from the dataset are extracted because of the fitting method detailed in section 5.2.1.1, which is then used as pseudo annotations to train N in a way that is supervised. The ResNet 50 [223] network structure was chosen for this stage and subsequently trained following the replacement of the output softmax layer via a linear regression layer with $n_e = 28$ neurons. Prior to the beginning of training, dataset frames were scaled to the template of size 224 × 224 with the 68 points landmark [470], which were projected from the 3D face mean ($\bar{\mathrm{x}}$) to the image space. Overall, the trained expression

Table 5.1 A detailed description of the emotion datasets used in this chapter. B stands for basic, N for neutral and C for compound emotions.

Dataset	# images	# subjects	Emotions	Elicitation	Resolution
RadFD[309]	8040	67	7 B+ 1 N	posed	681×1024
KDEF[368]	4900	70	6 B+ 1 N	posed	562×762
RAF-DB[338]	29672	N/A	6 B+ 1 N, 12 C	posed& spon.	web images
CFEE[145]	5060	230	6 B+ 1 N, 15 C	posed	1000×750

network model is a mapping: $R^{224\times224} \to R^{28}$. In the end, 70% of the dataset was used to train and what remained was subsequently halved for both testing and validation. During the training, the ℓ_2 norm error between the network output and the pseudo annotations created in section 5.2.1.1 was minimized by the network.

5.2.4 Emotion Recognition

To resolve the 7-categories classification problem (six common basic emotions with neutral), the Error Correcting Output Codes (ECOC) technique [131] was used for the classification of the facial expression vector $e \in R^{28}$, which resulted from the network N. Various binary learners are joined together via the ECOC strategy to find a solution for the multi-class classification issue. The chosen binary learner was Support-Vector Machines (SVM) [58]. The 10-fold cross validation as well as the one-versus-all [413] encoding scheme were applied to the classifier for both training and testing purposes. The improvement of SVM hyper-parameters was done through the Bayesian optimization approach [507]. 68 landmarks taken from the images from the entire compilation of used emotion datasets in table 5.1 were employed to register them to the mean face template, in the same way as section 5.2.3.

5.3 Experimental Results

The following section details qualitative and quantitative experiments that both assess and compare the pipeline on estimation of facial expression parameters in addition to the ability to discern emotions from facial images. The ResNet [223] CNN structure that had 50 layers was employed in this approach and was implemented using TensorFlow [1] and trained with the Nvidia Tesla V100 GPU.

A further aim of this section is to detail further visualisations and results based on the experimental findings. It can be seen from Fig. 5.4, that the Cumulative Distribution Function (CDF) of the Mean Squared Error (MSE) is a consequence from the comparing of facial expression parameters using the trained CNN, ITW [56], and baseline [244] [243] on the test split of FaceVid dataset as well as the ground truth. A small portion of these findings that were obtained via the three methods are illustrated in Fig. 5.5. When combined with the mean face of the model the sole distinction is that of the making of the facial expression vector e. When the 3D reconstructions are observed in detail, it is clear that the chosen approach can result in greater accuracy in visual results, especially when capturing the eyes opening/closure and eyebrows/mouth movements. This differs from the ITW and baseline methods that are often unsuccessful in this and also often produce results that are stiff.

In figures 5.6 and 5.7 it is possible to see facial expressions that have been predicted by the trained model on RadFD [309], KDEF [368], RAF-DB [338], CFEE [145]. When observed in more specific detail, it is clear that the chosen method successfully captures the facial expressions regardless of factors such as identity. This can especially be noted as being evident along each column. Through this, the emotion classifier is more successful in separating the tested images according to their given emotion. This is especially possible because of the facial expression features that the trained network was able to find. Nonetheless, sometimes emotions are particularly difficult to identify, even to the human eye, as their manifestation varies from subject to subject.

Additional analysis regarding how well the framework performs was done through the following steps. First, the emotion classifier was trained and applied to each facial expression feature that emerged from the trained network through employing all the four benchmarks (RadFD [309], KDEF [368], RAF-DB [338], CFEE [145]). Second, the whole framework underwent testing on new and unfamiliar video that was taken from YouTube following it being parsed into frames that were then put onto the system one at a time. The results show how well the framework performs when automatically recognising emotions on the human face.

5.3.1 Facial Expression Recognition

Assessing the FER approach was done via four publicly available datasets for emotion recognition. These were: Radboud [309], KDEF [368], RAF-DB [338], CFEE [145]. The chosen datasets all contained basic emotion [151] annotations (angry, happy, fearful, surprised, sad, and disgusted), in addition to an expression that was neutral. The findings for each dataset are detailed below:

1) The **Radboud dataset** [309] had 67 subjects each captured from 5 different angles simultaneously. Two experiments were done for evaluating how well it performed as a

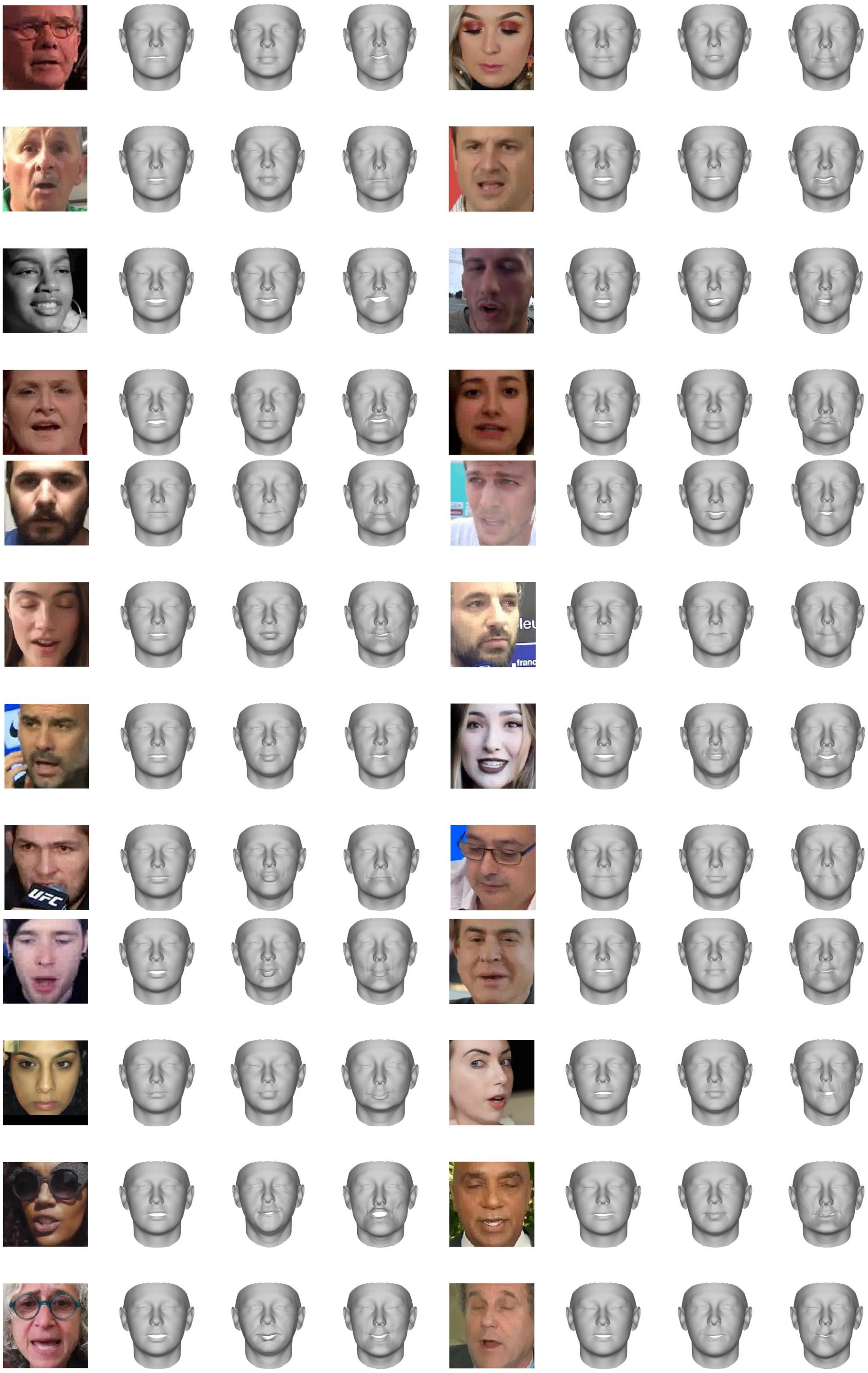


Fig. 5.5 The estimated facial expressions have been obtained using the trained network, ITW [56], and baseline [244] [243] on the selected image samples from the test partition of FaceVid.

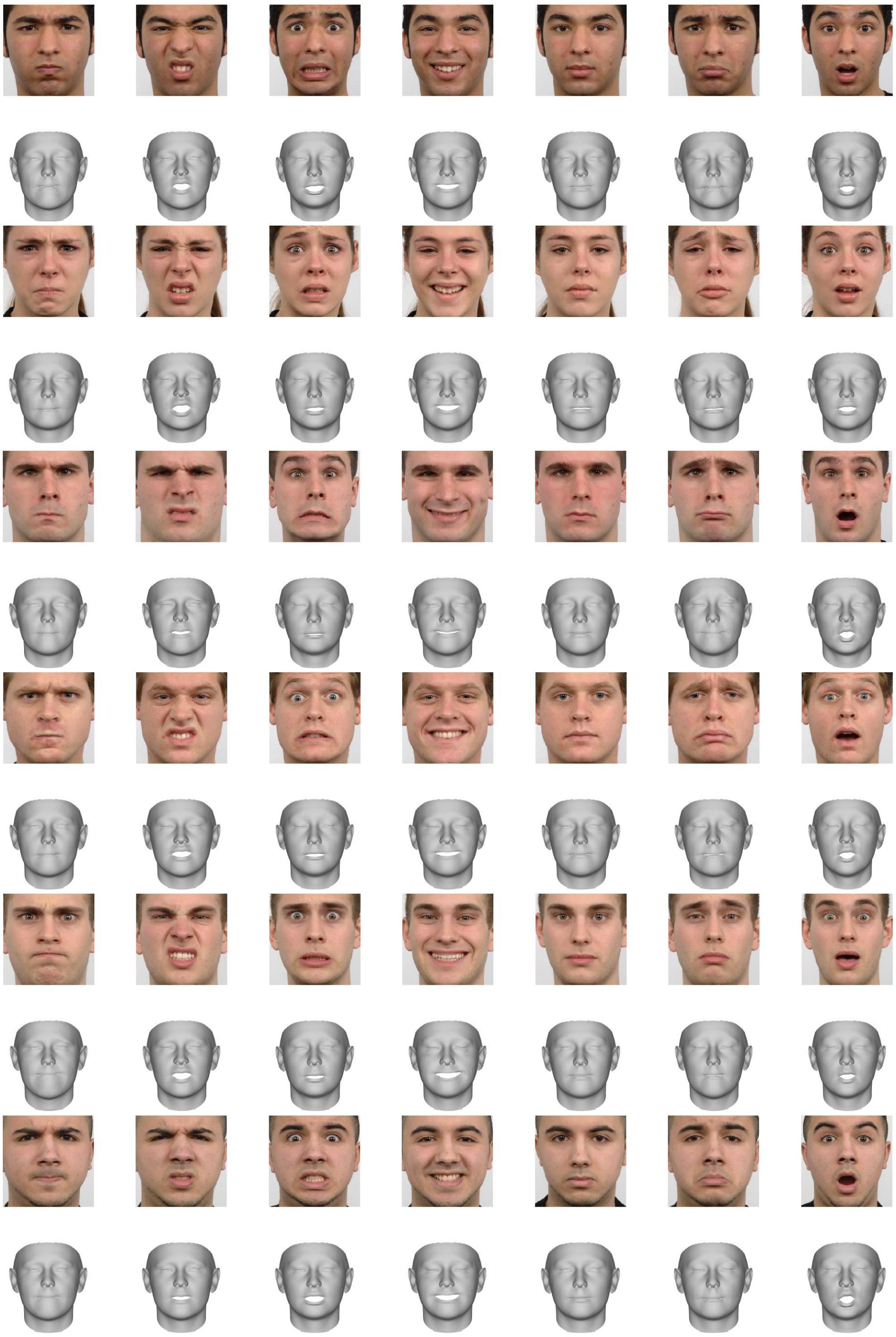


Fig. 5.6 The estimated facial expression obtained from the trained model using some randomly chosen images from the RadFD [309] dataset, where all the expressions are visualised at the top of mean face $\bar{\mathbf{x}}$ of the 3DMM. The visualisation order of emotions starting from the left is anger, disgust, fear, happiness, neutral, sad, and ends with surprised.

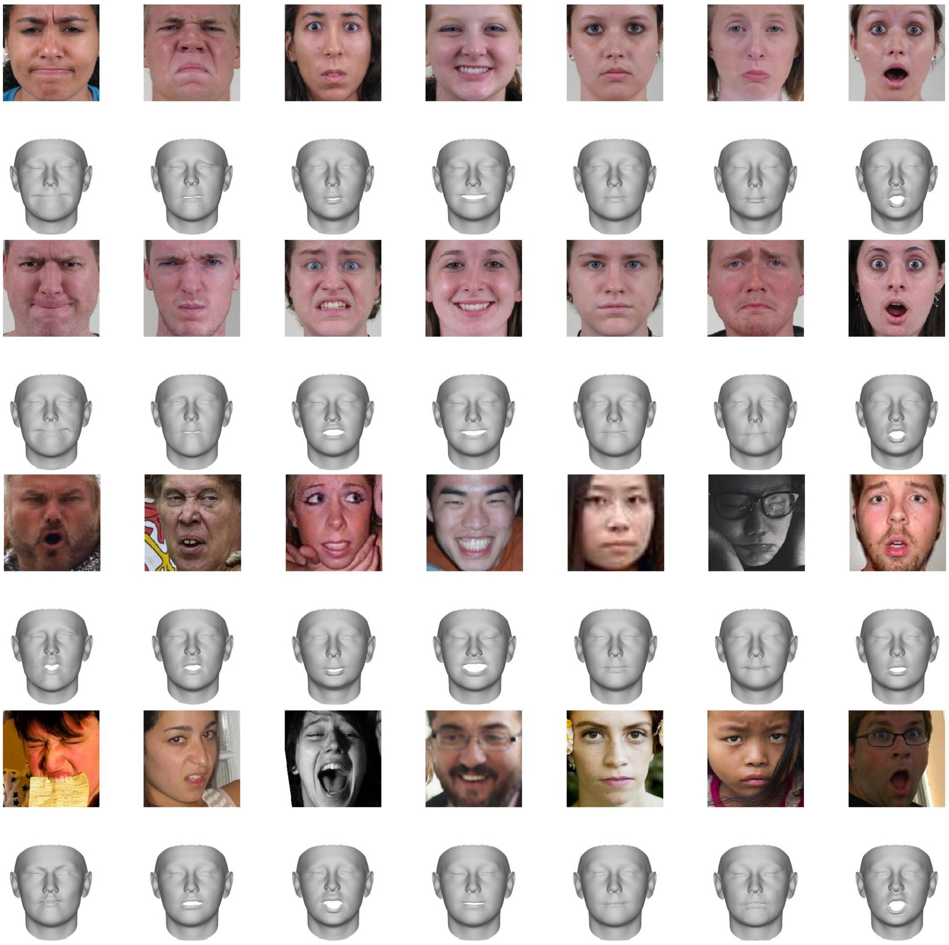


Fig. 5.7 The estimated facial expression produced using the trained model on the selected image samples from the CFEE [309], and the RAF-DB[338] datasets.

network when trying to recognise one emotion from the same person from dissimilar view angles. During the initial experiment, the frontal image of each subject was that of a specific emotion (in total $67 \times 7 = 469$ images), whereas during the next experiment, one of the semi-profile faces (imaged from 45/135 degrees) of each subject was randomly selected and employed for 10-fold cross validation. Fig. 5.8 details the confusion matrices and accuracies that resulted from each experiment. Across all subjects, the average MSE of expression parameters that resulted from semi-profile and frontal images was 0.008. Both the generated accuracies in Fig. 5.8 and the small MSE are illustrative of how the used expression network is able to create view-angle autonomous expressions estimations. It is clear from table 5.2 that the chosen method results in a competitive degree of accuracy on the RadFD [309] dataset, reaching the first-best performance among the state of the art.

The **KDEF dataset** [368] structure does not differ much from that of RadDF [309] insofar as for both the pictures, each of the 70 subjects were taken from five different angles simultaneously (0°, 45°, 90°, 135°, and 180°). It was requested from each subject that they do the emotion twice, and then one was randomly chosen. Of the images, those selected were frontal images (in total 7×70 = 490) in table 5.2, which details the reported findings. A high level of accuracy was achieved in comparison to alternative state-of-the-art approaches (92.24%), and this demonstrated how powerful the model is in generating separable facial expressions based on the basic emotion label. Fig 5.9 shows: a) the t-SNE [371] visualisation of predicted facial expression vectors from the KDEF dataset images, b) a confusion matrix produced from the emotion classifier. The confusion matrix in addition to the projection of expression vectors show how the level of separability between happy and disgusted emotions is high, especially in comparison to the other labels (100% and 97.1% respectively); alongside this, it appears that particular emotions such as sadness and neutral can be grouped together (78.6% and 84.3% respectively).

We visualised the obtained feature vector of the KDEF dataset using t-SNE, a valuable tool to calculate the reduced features dimensions and to visualise the high dimensional data.

Fig. 5.9 (a), visualises the distribution of the predicted facial expression feature vector extracted by our work via standard t-SNE [371]. The colours in the legend match to every label of the dataset, where each colour specifies single facial expression class (Angry, Disgusted, Fearful, Happy, Neutral, Sad, and Surprised). The data positions were automatically gathered and identified by the proposed method which is well-preserved, and the features learned were separated corresponding to each label. To explain further, different sample tests of the facial expressions tend to get together into a cluster. In general, the quantity of the overlap between these categories should provide the expanse of distances and similarities between them. There are however several examples of different facial expressions that are enclosed to each other, which implies that the test samples do not differentiate between an expression from another. For example, the facial emotions sad, neutral, and angry have extreme overlap which means that they could be perplexed simply, and those expressions have less overlap with the other emotions; surprised, happy, and disgusted. It seems visually that the emotions happy, disgusted, fearful, and surprised have very distinct clusters from the others. In addition to the scatter of each class displays the variation of this category. As an instance of both disgusted and surprised classes which were mapped to another region, suggesting that more modes contained by each category. It is noticeable from the visualisations of the result that it has not been useful and report better separation for all the clustered of facial expressions with t-SNE.

Fig. 5.10 shows, comparing with Fig. 5.9 (a), the visualisation of t-SNE plots for the output representations which are extracted by different methods using different datasets. As shown in Fig. 5.10 (a) that merges multi-scale facial expression representations using a Pose-adaptive Hierarchical attention Network (PhaNet) and Scale Attention Learning (SAL) for the detection of facial expressions with poses. It can learn more facial expression discriminative and pose-invariant representations for multi-view facial expression recognition. A trainable end-to-end PhaNet finds the most pertinent regions of facial expression through an attention mechanism in a hierarchical scale, and so chooses the highly informative scales to learn the expression-discriminative and pose invariant representations. It can successfully preserve more local identity information structure and are clearly separate well distinguished facial expression clusters [358].

Whereas Fig. 5.10 (b) and (c) show relative t-SNE which was used to visualise the learned features obtained from the RAF-DB uses both the proposed SoftMax loss and the Separate loss, respectively. Separate loss increases intra-class resemblance while decreasing the similarity between different classes. It includes of two term conditions. In terms of similarity, Separate loss is employed to minimise the inter-class and maximise the intra-class. These two terms are corresponding since no need to add extra hyper parameter to balance them. Moreover, the features learned by using this strategy are described by intra-category compactness and inter- category separation for the efficacy of this loss. The proposed Separate loss was utilised to train the CNN to learn spatial discriminative features for both basic and compound facial expression recognition in the wild. As can be seen from Fig. 5.10 (c), the learned features are more discriminative and the features within the similar class are closer. While the features learned only using the SoftMax loss are not discriminating enough for highly accurate facial expression recognition in the wild, especially in the compound facial expression recognition [343].

Fig. 5.10 (d) illustrates a comparative's t-SNE plot-based visualisation of the data distribution, which is formed by the latent representations for the seven clusters identical to the six emotions and the neutral state. A conditional generative model has been used to synthesis a system of emotional text-audio-visual speech that learns the underlying implanting space of emotions. This method was applied as an unsupervised approach to create acoustic, duration, and visual features aspects of speech with no use of emotion labels. The Conditional Variational Auto-Encoder (CVAE) approach was used to combine emotions jointly. The CVAE generative nature architecture permitted to generate better detected subtle differences of the six basic emotions and to combine different emotions simultaneously. This model can generate additional emotions that are not available in the dataset with nuances of the given emotion. The data samples are clustered differently depending on the type of data [107].

Fig. 5.10 (e) shows the t-SNE described by Convolutional Neural Network (CNN) features were extracted from the entire channel images of the CMU Multi-PIE dataset with different illumination variations [257]. Facial expression samples were represented by the colours in the figure. A colour channel-wise sequencing and Recurrent Neural Network was proposed to learn facial features consecutively alongside colour channels. The extracted spatial features from the sequential colour channels of the CNN were entered to the Long Short-Term Memory (LSTM). This technique preserves the discriminating of facial expression and colour using the sequence of spatial features [257].

The **CFEE dataset** [145] comprises of 230 subjects that are divided into two groups: basic and compound of labelled images. The images that were assigned basic emotions (in total of 1836) were passed through the network before being trained and tested by the emotion classifier. The average accuracy per class can be compared on this dataset with the state-of-the-art by Du et. al. [145], see table 5.2.

The **RAF benchmark** [338] happens to be the most demanding. Initially the compilation of this dataset was done via the internet and it was without conditions that were lab-controlled. Experienced annotated were used by the creators of [338] to accurately divide the dataset into the two categories of emotion: basic and compound. Overall there were 13,395 basic emotion images employed, and these were used to predict facial expressions via the network. Splits between train/test given by the authors of [338] were utilised when training/testing the emotion classifier. The average accuracy attained by the proposed experiment per class can be compared to the highest accuracy otherwise attained on this dataset (79.45% vs 81.83%).

Like as with other benchmarks, the emotion classifier that was trained was able to identify the happy, fearful, surprised and disgusted emotions at a higher level of accuracy than the remaining emotions (angry, sad, and neutral). This can be attributed to two key aspects; firstly, how intense the given AUs are particularly when deconstructing the emotions based on the Emotional Facial Action Coding System (EFACS) [166], secondly, the capability of the used expression basis (U_{exp}) in attaining the pertinent AUs. Motions that are related to mouth-jaw-and cheeks are generally accurately captured by the trained network (AU6, AU12, AU14, AU15, AU16, AU20, and AU26 [166]), for example lip corner puller/depressor, lower lip depressor, lip stretcher, jaw drop, etc. These are key in identifying emotions such as happy, surprised, disgusted and fearful. In contrast, more nuanced variance like subtle details around the eyes, such as inner brow raiser, brow lowerer, upper lid raiser, lid tightener, etc., are important when classifying and distinguishing between sadness and anger, and these posed a greater level of difficulty for the network. It can discern the reason for this through the way that AU6, AU12, AU14, AU15, AU16, AU20, and 26 have a more accurate representation inside the original Face Warehouse model, and these were used for annotating the compiled video dataset in section 5.2.2. In addition, automatically extracted landmarks can lead to a degradation in performance due to inaccuracies in landmark positioning /

Table 5.2 Represents a comparison which was done between the generated accuracies of the proposed method and the state-of-the-art approaches of the utilised benchmarks.

Dataset	Approach	Acc.(%)	Dataset	Approach	Ave. Acc.(%)
RadFD	Ali et al. [11]	85.00	RAF-DB	Li& Deng [337]	74.20
	Zavarez et al. [622]	85.97		Lin et al. [351]	75.73
	Jiang& Jia [266]	94.52		Fan et al. [157]	76.73
	Mavani et al. [385]	95.71		Gosh et al. [184]	77.48
	Wu& Lin [586]	95.78		Shen et al. [497]	78.60
	Sun et al. [522]	96.93		**Proposed**	**79.45**
	Proposed	**97.65±1.00**		Vielzeuf et al. [564]	80.00
	Yaddaden et al. [596]	97.57±1.33		Deng et al. [120]	81.83
KDEF	Zavarez et al. [622]	72.55	CFEE	**Proposed**	**96.43±1.1**
	Ali et al. [11]	78.00		Du et al. [145]	96.84±9.73
	Ruiz-Garcia et al. [469]	86.73			
	Yaddaden et al. [596]	90.62±1.60			
	Proposed	**92.24±0.70**			

extraction especially if they are unable to accurately annotate the 68 targeted locations on the face.

5.3.2 Experiment on Stress Detection

In this work, we examine the capability of our suggested framework utilising solely facial videos for detecting stress conditions. Stress can be identified by biosignals, which is commonly conceived as a complex emotional state [185]. But the recording of biosignals is not always been appropriate for daily monitoring. Therefore, researchers pursuit stress recognition when depending on only facial cues which completely constitutes a challenging mission. Literature related to the combination of biosignals with deep learning techniques [247][189] or visual cues [186][188] are limited. The performance of our approach versus other state-of-the-art techniques in stress identification is evaluated in this study. Towards achieving that, we use the dataset (SRD'15) which has been employed in [186]. This dataset consist of 24 subjects (their age between 47.3±9.3 years) have been performed eleven experimental tasks with 288 videos as total, which is either neutral or stressful. Four phases were divided from the whole experiments: social exposure, emotional recall, stressful images/mental task, and stressful videos. Each video frame was annotated as 'stressful' or as 'non-stressful', in according to the task under examination. Following, our approach was utilised to recognise facial expressions from each frame. To assure that the frames for the same subject are not included at the same time in both training and testing sets, where 5-fold cross-validation was used. After repeating the experiments ten times, the average accuracy of each phase for stress recognition is reported in table 5.3, where the stage (social exposure) comprises a task with speech was not undertaken since it have an effect on 'per se' head motility in comparison to the neutral un-speech task as clarified in [187]. For comparisons, we have applied the same protocol of [186] which utilise head motility, also we

Table 5.3 Comparison of accuracy for stress detection using the dataset in [186]

Phase	**Head motility [186]**	**DWNet1D [189]**	**Ours**
Emotional Recall	82.99%	83.50%	**86.70**%
Stressful images	85.42	**92.60**%	88.42%
Stressful videos	85.83%	85.90%	**88.83**%
Average	84.75%	87.33%	**87.98**%

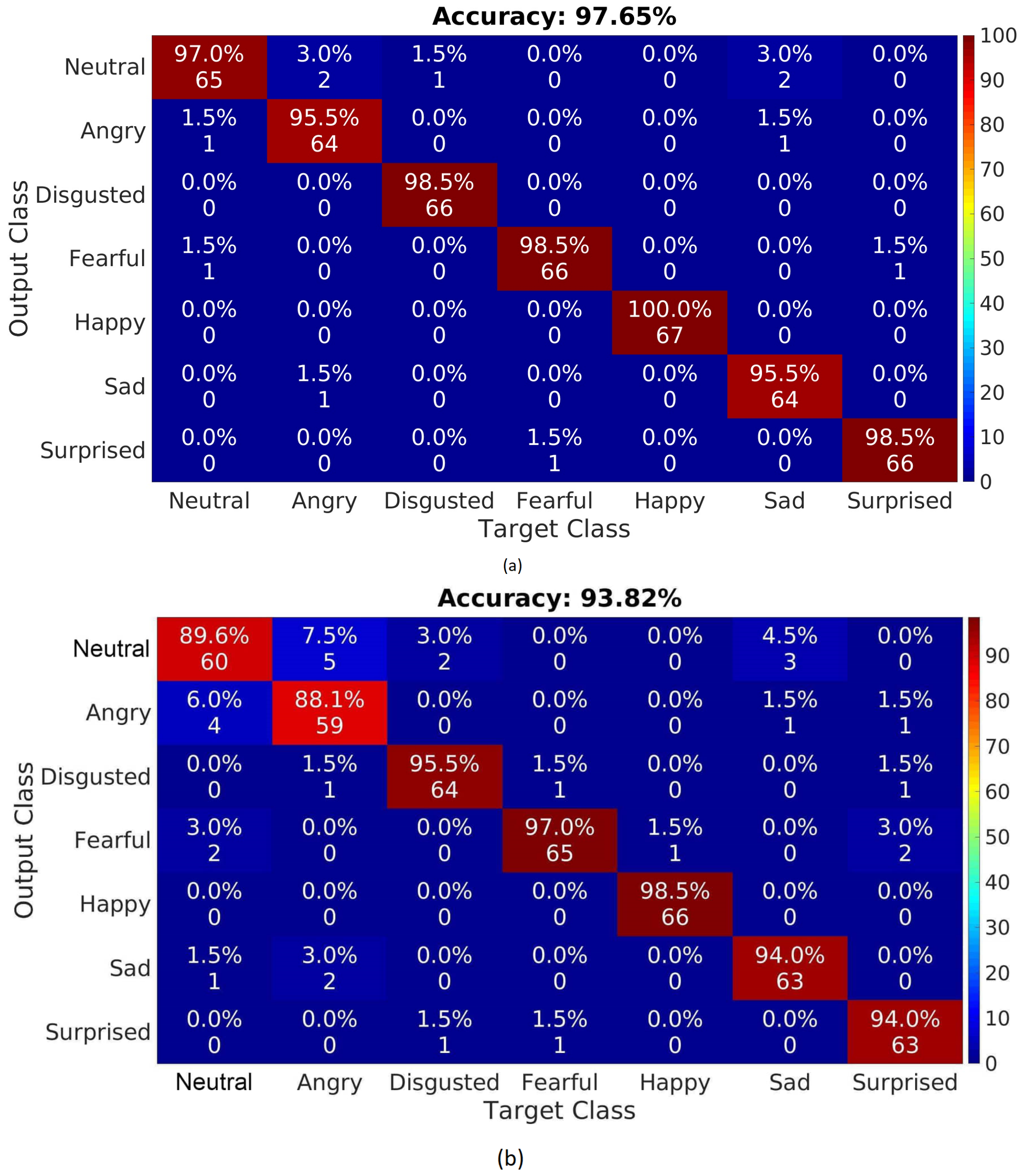
have followed the method of [189] that has been used the heart activity signals (IBI), where their results are also shown in table 5.3. It can be noticed from table 5.3 that the proposed approach accomplishes high accuracy and outperforms the other examined methods. For stress analysis, this represents an indicate of a promising result, where our approach utilises solely un-invasive and frame based appearance features.

5.3.3 Estimation of Expression Parameters

Although the destined outcome of the chosen pipeline is the emotion class, a further study is put forward by the work based on both a qualitative and quantitive evaluation of the overall accuracy of identifying parameters related to facial expressions. This was an intermediate pipeline stage removed from the network's output. We tested its performance on the test split of the FaceVid dataset and these were quantitative as well as qualitative. It can be observed from Fig. 5.11 that the qualitative results which emerged from the test split and a comparison can be made with the original Ground Truth (GT) reconstructions. The predicted GT expressions on top of the mean face ($\bar{x}$), in addition to the GT identity parameters, are demonstrated here. It has been identified that the estimations given via the approach visually have some similarity to the GT. The network is also compared to: 1) a baseline method that follows a linear shape model fitting put forward in [244] [243], and 2) the state-of-the-art 3D reconstruction approach from in-the-wild images (ITW) [56]. Both of these approaches are given the same facial expression model (FaceWarehouse [67]) and the average Mean Squared Error is computed (MSE) between the estimations and the ground truth for the test split. The chosen approach attains a much lower (best) MSE of 0.007, whereas the ITW [56] and the baseline [244] [243] obtain 0.026 and 0.098, respectively.

5.4 Conclusion

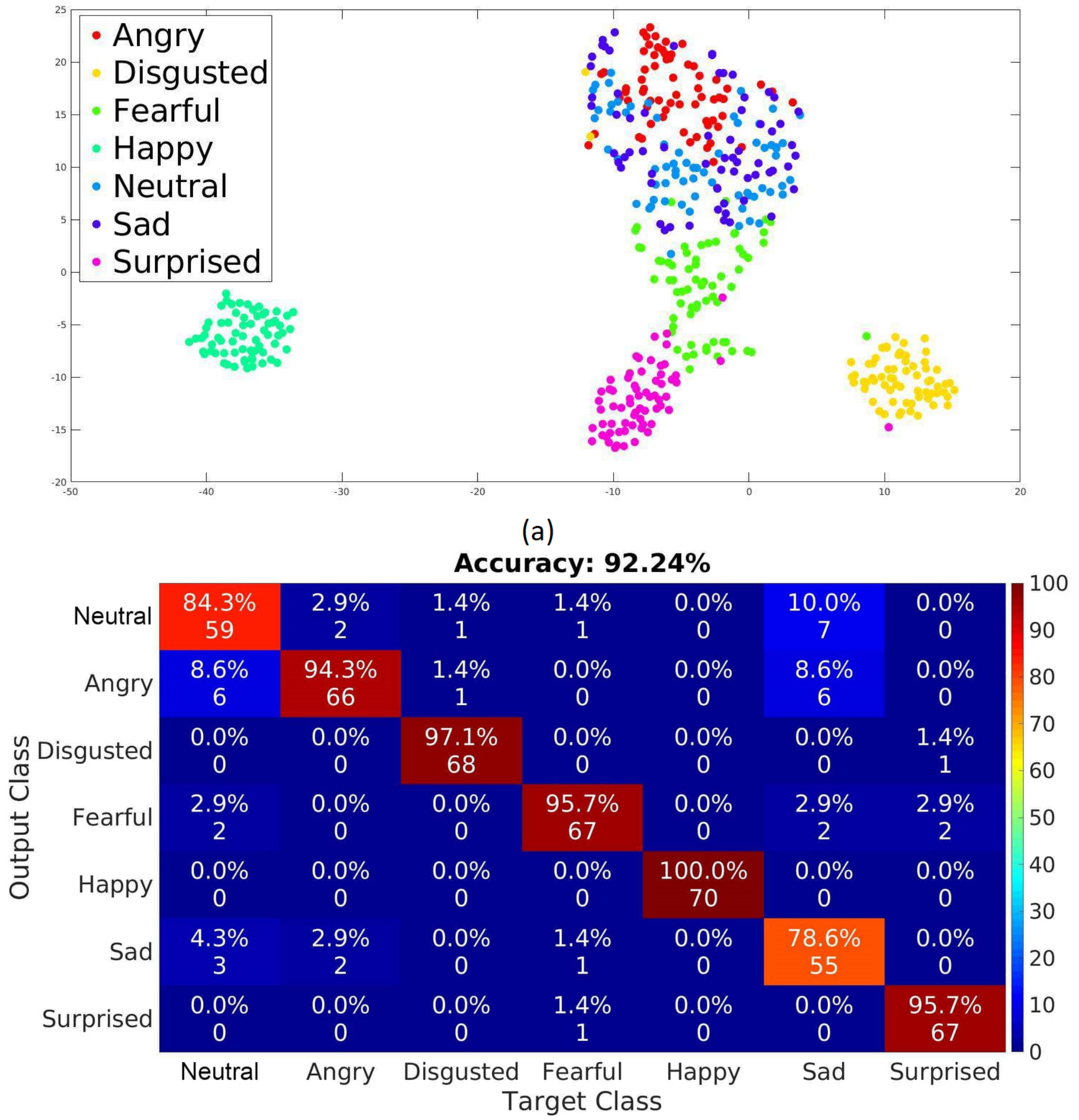
This chapter has detailed the chosen framework for the purpose of automatically recognising human emotions from monocular images. A well-trained deep CNN is used and implemented by the proposed framework. It is able to predict, even in more difficult situations, the facial expression from one image. The trained model was both explained and also compared with state-of-the-art methods for 3D reconstruction from in-the-wild images. The proposed model performs at a more superior level insofar as it is able to regress the facial expression coefficients, especially during training on the similar facial expression model as the contestants. The trained network's ability to recognise emotions in combination with an "m-SVM" classifier on four widely used benchmarks (it was taken together under either controlled settings or in-the-wild conditions) in the field was also tested. The emotion recognition findings which were attained show that the proposed framework, in comparison to state of the art approaches on the same datasets, performs competitively. The highest accuracy on the RadFD [309] dataset was reported as (97.65%) and the KDEF [368] dataset was reported as (92.24%) and also the CFEE [145], and the second best was on the RAF-DB [338] with a 2.38% gap in comparison with the first. The novel contribution of this chapter is published in the FG 2020 by the 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2020), as a collaborative work.



(a)

(b)

Fig. 5.8 Confusion matrices were generated by the emotion classifier using 10-fold cross validation over RadFD [309], (up matrix) showing solely on frontal images, (down matrix) semi-profile images for each subject.



(a)

Accuracy: 92.24%

	Neutral	Angry	Disgusted	Fearful	Happy	Sad	Surprised
Neutral	84.3% 59	2.9% 2	1.4% 1	1.4% 1	0.0% 0	10.0% 7	0.0% 0
Angry	8.6% 6	94.3% 66	1.4% 1	0.0% 0	0.0% 0	8.6% 6	0.0% 0
Disgusted	0.0% 0	0.0% 0	97.1% 68	0.0% 0	0.0% 0	0.0% 0	1.4% 1
Fearful	2.9% 2	0.0% 0	0.0% 0	95.7% 67	0.0% 0	2.9% 2	2.9% 2
Happy	0.0% 0	0.0% 0	0.0% 0	0.0% 0	100.0% 70	0.0% 0	0.0% 0
Sad	4.3% 3	2.9% 2	0.0% 0	1.4% 1	0.0% 0	78.6% 55	0.0% 0
Surprised	0.0% 0	0.0% 0	0.0% 0	1.4% 1	0.0% 0	0.0% 0	95.7% 67

(b)

Fig. 5.9 Top: t-SNE displays the KDEF dataset estimated facial expression parameters using our proposed method. Bottom: a confusion matrix of the KDFE dataset.

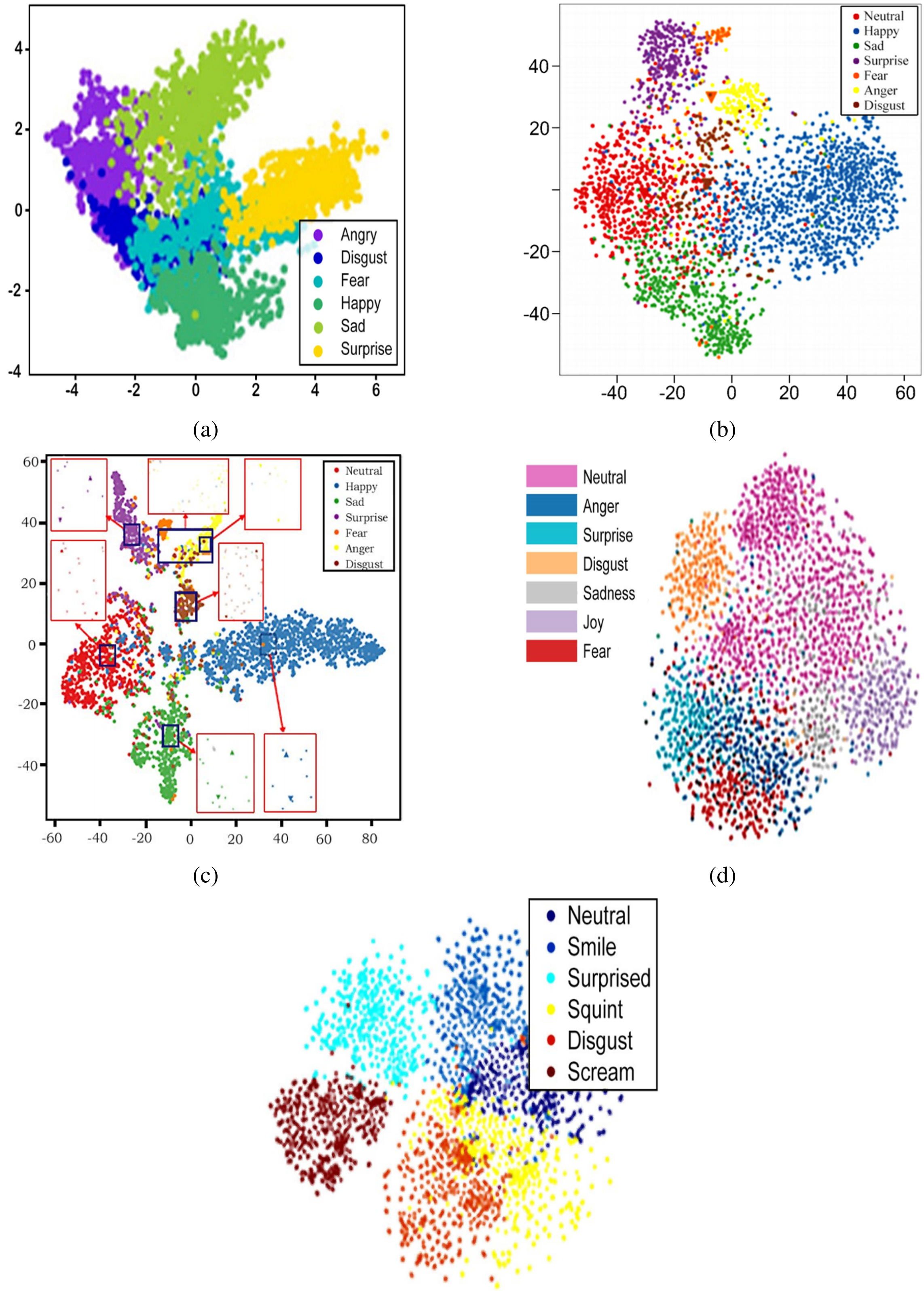


Fig. 5.10 A set of t-SNE plots were resulted from different methods for facial expression recognition.

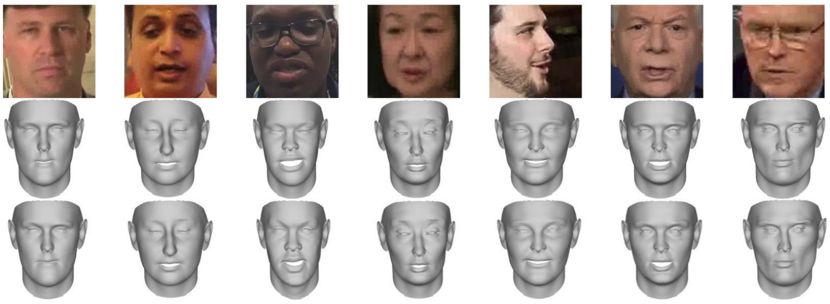


Fig. 5.11 Some of the quantitative results were obtained from the selected test sample images of the dataset (first row) and compared with the original Ground Truth (GT) expression reconstructions (bottom row). The estimated (second row) and the GT original expressions are both visualized using the 3D GT identity.

Chapter 6

Conclusions and Future Work

6.1 Summery and Discussion of Thesis Achievements

Various studies in the Affective Computing discipline have focused on human emotions analysis. The rising advancement of both the Computer Vision and Machine Learning fields, especially in light of their focus on studying human faces, have garnered a great deal of attention given that this focus is vital in the development of a range of significant applications. The popularity of human emotion recognition research has risen in importance and many studies centre around this field. The majority of these studies seek to use a captured image in order to automate the recognition of a person's emotions. The successful implementation of an approach that resolves this issue will benefit a wide range of applications - including, but not limited to, human-computer intelligent interaction, affective computing, security, health care, computer games, stress analysis, facial expression transfer, and animations. Facial expression analysis is an interesting problem that humans can solve quite easily. The ultimate goal of face emotion recognition is to enable computers to mimic the human visual system. To date, none of the three intriguing, different but interrelated aspects of the problems have been solved yet. These are: face detection, facial expression information extraction, and facial expression classification. As presented earlier, not all the facial expressions are classified under the six cardinal emotion expressions (*happiness*, *sadness*, *surprise*, *fear*, *anger*, and *disgust*), that posit a distinct number of well-defined affective states which are represented by identified patterns of facial action. FACs may provide a solution to this dilemma, as it classifies facial actions prior to any interpretation attempts [160]. The cutting-edge advances in various aspects of facial expression recognition and the associated areas (such as the potential difficulties) were introduced, then followed by the characteristics of an ideal proposed system.

The problem of AUs recognition was mainly the focus at the beginning of this thesis. The automatic detection of AUs can provide explicit benefits because it considers both static frame-based and dynamic facial features methods. Three experiments of AU occurrence activation detection were introduced by extracting features using spatial and temporal approaches, using hand-crafted and deep learning representation from each static image of a video. We showed that deep learning produces a more discriminative representation for facial expression recognition. The first experiment of AU activation detection confirms the superior ability of a pre-trained AlexNet that augments the overall unprecedented level of recognition accuracy, and shows significant improvements in AU prediction, in contrast to the traditional hand-crafted approaches. Overall, the performance indicates that CNN is very promising and outperformed all other methods. The second experiment was conducted for temporal modelling by testifying that hybrid systems have both temporal and temporal features are achieved to gain more long term temporal pattern information and to get effective fusing of temporal features. In addition, we discovered the importance of stacking time windowed LSTM on top of CNN for learning temporal information by combining the spatial and temporal schemes simultaneously. However, we showed that merging spatial and dynamic features fails to achieve a comparable result. Moreover, we also hypothesised that using the unsupervised Slow Feature Analysis method can leach invariant information from dynamic textures. Additionally, performing a successful comparison of continuous scoring predictions of AUs activation detection was achieved which was shown to be efficacious. A competitive substantial performance evaluation was carried out on the ISL Enhanced Cohn-Kanade AU-coded Facial Expression dataset.

Facial expressions video frames are inherently temporal and provide better real wild conditions than still images. Approaches using dynamic features are more efficient in recognising facial expressions from frontal poses. That is because the variation among expressions are frequently conveyed using dynamic transitions between different stages of the facial expression in a better way than a static state described by a single key frame. Nevertheless, these methods are not able to handle the effect of out-of-plane rotations with a robust manner sufficiently well and can only be achieved by increasing the differences between data and by using deep learning methods. Deep learning has also made this process easier given its broad range of big data and, more specifically, facial data online. The ease of availability of visuals (images/videos) and datasets of unlabelled human faces has enabled researchers to identify benchmarks that they have subsequently evaluated their proposed approaches against. As a result of these combined factors, the industry has moved toward modelling the above-mentioned problems more extensively. Consequently, more competitive techniques have emerged, which are capable of identifying facial emotions based on web images that are both controlled or 'in-the-wild'. Nonetheless, it was found that deep learning requires massive amounts of training data since many parameters must be tuned by the learning algorithms. The performance of deep learning approaches can be inhibited with the limited amounts of training data because of the problem of overfitting. We found that Generative Adversarial Networks (GANs) endow a novel way to provide additional expressive facial images from training data by generating synthetic samples by having the manifestation of real images, that are not restricted by the need to provide well-labelled data when training them. GANs models are a powerful unsupervised learning technique. In particular, the suggested DCGAN model boosts the topic of unsupervised learning, which mainly desired to join the new concept of GANs with the recent achievement of CNNs. Obtaining large amounts of Ground Truth information with accurate facial expression labels is a labour-intensive and time-consuming process if the annotations are made by highly human, trained coders. Thus, a solution that is increasing in popularity is to reduce the need of labelled data using Generative Adversarial Networks by synthetically creating such training data. A recent research trend, which is to use unsupervised learning through Adversarial Networks was investigated. The ability of generative modelling of Deep Convolutional GAN (DCGAN) was peculiarly exploited for facial features extraction and classification. Particularly, the generation of facial expression images under different arbitrary poses (in front, multi-view, and in the wild), with the recognition of emotion categories and Action Units. The comprehensive experimentation of the proposed cross-dataset experiments has demonstrated the convincing generality which performed better. Also, we showed that the features learnt by the DCGAN process are poorly suited to encoding facial expressions when observed under multiple views, or when trained from a limited number of examples. However, in this issue, an action based on parametric models such as 3DMMs and its formation has to play a big role in solving and generating multi-view facial images, which are the most promising.

This research has been concentrated on the purpose of disentangling identity from the expression for automatic facial expression recognition. A novel technique was implemented for emotion recognition from a single monocular image. The approach raised by this work is aimed at estimating facial expressions from in-the-wild images using RGB images to construct a 3D morphable face. A new Large-scale dataset collection of facial image videos (which was taken as 6,000 individual, distinguishable, facial identities) known as FaceVid, were both gathered and annotated. The compiled dataset was rich in variations and distribution of facial dynamics, appearance, identities, expressions, and 3D poses. Through the help of the accurate proposed model-based formulation which was utilized during the training, the following annotation was added to each video per-frame: 68 facial landmarks, a 3D facial shape, which is made up of both identity and expression parts, and a 3D pose of the head, in addition to the scale of the postulated scaled orthographic camera model. A well-trained deep CNN (ResNet) is used and implemented by the proposed framework. The selected DCNN design is robust regarding its ability to use a single input image and afterwards to predict and regress the specific expression parameters of a 3DMM of the facial shape. It was not affected by challenging viewing conditions such as changes in camera view angle, occlusions, variations in position of facial expression, dim lights and illumination. With the substantial development of deep learning architectures, typically, the trend has been that they are likely to become deeper; for example, ResNet 50, which was awarded the champion of ILSVRC 2015, is about 20 times deeper than AlexNet and 8 times deeper than VGGNet. In which case, when incrementing the depth, the network can approximate the target expression by increasing the non-linearity and ultimately gaining better feature representations. Moreover, it was also able to segregate the expression autonomously from the identity of the subject. The estimated facial expression vectors were also used to link the network with a back-end classification stage for the facial expression recognition task. This signifies that the model was subsequently structured to be able to robustly use single images to identify facial expressions. The trained network's ability to recognize emotions in combination with an m-SVM classifier on four widely used benchmarks (it was taken together under either controlled settings or in-the-wild conditions) in the field was also tested and reported state-of-the-art performance. Experimental results show that the proposed method significantly reduces reconstruction errors. It is noteworthy to state that the best performance has been created by synthesizing the detailed 3D facial shape model from 2D dynamic images. By modifying the 3DMMs to explicitly model other intra-personal variations in addition to pose, illumination and expression with the combination of the 3DMM with the subsequent deep learning methods have delivered the best performances. With the wide availability of parallel processing hardwares such as the use of omnipresent GPUs, the robustness and reliability of the facial expression recognition system has been largely improved and was a compromise, due to the fast 3D technical scanning being limited at the present time.

6.2 Future work

The accuracy of 3DMMs in facial reconstruction of an individual has not been without its limitations, such as producing overstated and impossible to perceive deformations or extremely generic images if the weight is not adapted accordingly; this is because the expressions could be either less or more subtle. Furthermore, the construction of the facial morphology process could be affected by the landmark detection errors, and generally the fitting of face shapes is a substantial problem and requires a non-linear optimization solution. Moreover, it may be increasingly sensitive to uncontrolled viewing settings. Reconstructing a face using a single 2D image is a one-sided problem because of the likelihood of losing the information during the camera projection, where the estimation of both intrinsic and extrinsic camera coefficients (light, 3D shape, and texture parameter) are involved. Confined acquisition conditions exist such as low-resolution imaging, large facial expression variations, the size and type of the datasets and how it impacts model performance. In addition to the linear bases, 3DMMs are linear models representing each unseen face as a linear combination of the training faces. This makes it difficult for the 3D reconstructed faces to differ from the training ones. Robust occlusion, the precision of the feature representation approaches, capturing fine-scale and high-frequency geometrical details (e.g wrinkles and teeth) can all be challenging. This can be improved by dividing the face into small, distinct regions, defining a 3DMM for each region, and combining them in the result. Popular conventional methods of 3DDM often based on monocular cues, these cues represent a limited number of feature pixels by using their position information and largely neglecting the surface textural information (the texture is less detailed or more repetitious) of the input image. Another limitation is that although the model was trained using a large training set acquired in the wild conditions, nevertheless, it could not cover the huge, sheer variety of the human population's features; the size and variability (gender, age, ethnicity etc.) of the training set of faces which had been used to construct the 3DMM, which then acted as a bottleneck. This seems to imply the need to make it as large and varied as possible to capture real-world faces accurately. It is important to make it comprehensive and able to showcase all the general combinations. Consequently, the quantity of facial appearances and their variability has played an instrumental role in the learning of significant representations and improving the fitting performance. Dense 3D reconstruction instead of relying on only sparse (68) landmarks can be done with the help of optical flow which is able to track dense points on the face in the image space. Facial features components can be characterized as holistic in both shape and texture parameters, these features must be strengthened by fusing of other representations such as capitalising on both geometry and texture while applying the 3D reconstruction, or adding sparse learning into the 3DMM fitting to achieve more robust and accurate results. It would be useful to examine more robust and precise fusion strategies. Moreover, the regressed parameters could be efficiently hybridized to generate a more complex expression model. In addition, new ways should be thoroughly investigated for further improving the accuracy and fitting efficiency. For future work, adopting some improvements should be considered such an extension of the proposed technique by using different methods of 3DMM reconstruction and by incorporating the texture component of the images. For example, a high-quality texture model will learn to improve the quality of 3D reconstruction of facial images from a single input image, which is beneficial to disentangle the expression from identity. Simultaneously, it completely explores the implicit multi-view 3D facial information images, and the robustness of the performances would be also enhanced. Furthermore, using non-linear methods for 3DMM face reconstruction such as deep Neural Networks learning approaches. Lastly, a conditional DCGAN will train also to disentangle the subject's expression from identity.

Building a new collection of dynamic 3D facial expression datasets completes the objectives of this thesis: higher resolution videos containing general spontaneous facial expressions and involving micro-expressions with more complex affective states rather than fixed, acted still images. A large bespoke range of various ages, genders or ethnic groups and thorough abundant demographic information for each subject. A possible future work would be improving the face model by including the larger diversity of face datasets.

DCGAN network model will be used for the visual representation of the saliency maps for facial expression image saliency detection to better estimate the ground-truth, in addition, to be used as a feature learning model. The feature learning model of DCGAN can be extended also for the detection of compound emotions by enhancing the representation of facial expressions which would be a complement to the categorical model.

Currently, the performance gap between machines (technology and academia) and humans is still wide and the recognition accuracy within acceptable execution time (real-time) is not yet up to the required standard. There is a large gap between the available academic solutions and commercial systems [601]. In order to enhance the spontaneous facial expression recognition rate, an integrated facial expression recognition system with developing methods of features selection and extraction should work in different environments without manual intervention during the initialisation and deployment of the system. Additional improvements to the classification step would lead to a robust automatic labelling framework. This is a very important step because without establishing an efficient facial expression recognition system this research cannot progress any further. A strong and coordinated effort and further investigation between the computer vision, computer graphics, signal processing, real-time face tracking involving cognitive representation and to take into account study the social and cultural context will enable in the building of a high-quality digital facial communication system through reconstructing and analysing human faces depending on visual input, all of which are needed to attain this objective.

References

[1] Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., and Zheng, X. (2015). TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org.

[2] Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I. J., Harp, A., Irving, G., Isard, M., Jia, Y., Józefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D. G., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P. A., Vanhoucke, V., Vasudevan, V., Viégas, F. B., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., and Zheng, X. (2016). Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *CoRR*, abs/1603.04467.

[3] Abate, A. F., Nappi, M., Riccio, D., and Sabatino, G. (2007). 2d and 3d face recognition: A survey. *Pattern recognition letters*, 28(14):1885–1906.

[4] Abdat, F., Maaoui, C., and Pruski, A. (2008). Real time facial feature points tracking with pyramidal lucas-kanade algorithm. In *RO-MAN 2008-The 17th IEEE International Symposium on Robot and Human Interactive Communication, Munich, doi: 10.1109/RO-MAN.2008.4600645*, pages 71–76. IEEE.

[5] Agarwal, S. and Mukherjee, D. P. (2017). Facial expression recognition through adaptive learning of local motion descriptor. *Multimedia Tools and Applications*, 76(1):1073–1099.

[6] Agrawal, A. K. and Singh, Y. N. (2015). Evaluation of face recognition methods in unconstrained environments. *Procedia Computer Science*, 48:644–651.

[7] Al-gawwam, S. and Benaissa, M. (2018). Robust eye blink detection based on eye landmarks and savitzky–golay filtering. *Information*, 9(4):93.

[8] Alabort-i Medina, J. and Zafeiriou, S. (2014). Bayesian active appearance models. In *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, doi: 10.1109/CVPR.2014.439*, pages 3438–3445.

[9] Aldrian, O. and Smith, W. A. (2012). Inverse rendering of faces with a 3d morphable model. *IEEE transactions on pattern analysis and machine intelligence*, 35(5):1080–1093.

[10] Alemy, R., Shiri, M. E., Didehvar, F., and Hajimohammadi, Z. (2012). New facial feature localization algorithm using adaptive active shape model. *International Journal of Pattern Recognition and Artificial Intelligence*, 26(01):1256003.

[11] Ali, A. M., Zhuang, H., and Ibrahim, A. K. (2017). An approach for facial expression classification. *IJBM*, 9(2):96–112.

[12] Ali, G., Iqbal, M. A., and Choi, T.-S. (2016). Boosted nne collections for multicultural facial expression recognition. *Pattern Recognition*, 55:14–27.

[13] Ali, S. and Shah, M. (2008). Human action recognition in videos using kinematic features and multiple instance learning. *IEEE transactions on pattern analysis and machine intelligence*, 32(2):288–303.

[14] Allen, B., Curless, B., Curless, B., and Popović, Z. (2003). The space of human body shapes: reconstruction and parameterization from range scans. In *ACM transactions on graphics (TOG)*, volume 22, pages 587–594. ACM.

[15] Almaev, T., Martinez, B., and Valstar, M. (2015). Learning to transfer: transferring latent task structures and its application to person-specific facial action unit detection. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, doi: 10.1109/ICCV.2015.430*, pages 3774–3782.

[16] Almaev, T. R. and Valstar, M. F. (2013). Local gabor binary patterns from three orthogonal planes for automatic facial expression recognition. In *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction, Geneva, doi: 10.1109/ACII.2013.65*, pages 356–361. IEEE.

[17] Alom, M. Z., Taha, T. M., Yakopcic, C., Westberg, S., Sidike, P., Nasrin, M. S., Hasan, M., Van Essen, B. C., Awwal, A. A., and Asari, V. K. (2019). A state-of-the-art survey on deep learning theory and architectures. *Electronics*, 8(3):292.

[18] Alqahtani, H., Kavakli-Thorne, M., Kumar, G., and SBSSTC, F. (2019). An analysis of evaluation metrics of gans. In *International Conference on Information Technology and Applications (ICITA)*.

[19] Alshamsi, H., Kepuska, V., and Meng, H. (2017). Real time automated facial expression recognition app development on smart phones. In *2017 8th IEEE Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON), Vancouver, BC, doi: 10.1109/IEMCON.2017.8117150*, pages 384–392. IEEE.

[20] Alsubari, A., Satange, D., and Ramteke, R. (2017). Facial expression recognition using wavelet transform and local binary pattern. In *2017 2nd International Conference for Convergence in Technology (I2CT), Mumbai, doi: 10.1109/I2CT.2017.8226147*, pages 338–342. IEEE.

[21] Ambadar, Z., Schooler, J. W., and Cohn, J. F. (2005). Deciphering the enigmatic face: The importance of facial dynamics in interpreting subtle facial expressions. *Psychological science*, 16(5):403–410.

[22] Amberg, B., Knothe, R., and Vetter, T. (2008). Expression invariant 3d face recognition with a morphable model. In *2008 8th IEEE International Conference on Automatic Face & Gesture Recognition, Amsterdam, doi: 10.1109/AFGR.2008.4813376*, pages 1–6. IEEE.

[23] Amberg, B., Romdhani, S., and Vetter, T. (2007). Optimal step nonrigid icp algorithms for surface registration. In *2007 IEEE Conference on Computer Vision and Pattern Recognition, Minneapolis, MN, doi: 10.1109/CVPR.2007.383165*, pages 1–8. IEEE.

[24] Amores, J. (2013). Multiple instance classification: Review, taxonomy and comparative study. *Artificial intelligence*, 201:81–105.

[25] An, F. and Liu, Z. (2020). Facial expression recognition algorithm based on parameter adaptive initialization of cnn and lstm. *The Visual Computer*, 36(3):483–498.

[26] Antipov, G., Baccouche, M., and Dugelay, J.-L. (2017). Face aging with conditional generative adversarial networks. In *2017 IEEE International Conference on Image Processing (ICIP), Beijing, doi: 10.1109/ICIP.2017.8296650*, pages 2089–2093. IEEE.

[27] Arandjelovic, O., Shakhnarovich, G., Fisher, J., Cipolla, R., and Darrell, T. (2005). Face recognition with image sets using manifold density divergence. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 581–588. IEEE.

[28] Arjovsky, M., Chintala, S., and Bottou, L. (2017). Wasserstein generative adversarial networks. In Precup, D. and Teh, Y. W., editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 214–223, International Convention Centre, Sydney, Australia. PMLR.

[29] Asiedu, L., Mettle, F. O., and Nortey, E. N. (2014). Recognition of facial expressions using principal component analysis and singular value decomposition. *International Journal of Statistics and Systems*, 9(2):157–172.

[30] Babenko, B. (2008). Multiple instance learning: algorithms and applications. *View Article PubMed/NCBI Google Scholar*, pages 1–19.

[31] Babenko, B., Yang, M.-H., and Belongie, S. (2009). Visual tracking with online multiple instance learning. In *2009 IEEE Conference on computer vision and Pattern Recognition, Miami, doi: 10.1109/CVPR.2009.5206737*, pages 983–990. IEEE.

[32] Bagautdinov, T., Wu, C., Saragih, J., Fua, P., and Sheikh, Y. (2018). Modeling facial geometry using compositional vaes. In *Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, doi: 10.1109/CVPR.2018.00408*, pages 3877–3886.

[33] Bagherzadeh, J. and Asil, H. (2019). A review of various semi-supervised learning models with a deep learning and memory approach. *Iran Journal of Computer Science*, 2(2):65–80.

[34] Bakshi, U. and Singhal, R. (2014). A survey on face detection methods and feature extraction techniques of face recognition. *International Journal of Emerging Trends & Technology in Computer Science (IJETTCS)*, 3(3):233–237.

[35] Baltrušaitis, T., Mahmoud, M., and Robinson, P. (2015). Cross-dataset learning and person-specific normalisation for automatic action unit detection. In *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, volume 6, pages 1–6. IEEE.

[36] Bargal, S. A., Barsoum, E., Ferrer, C. C., and Zhang, C. (2016). Emotion recognition in the wild from videos using images. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, pages 433–436. ACM.

[37] Barr, J. R. (2015). *Gallery-free methods for detecting and recognizing people and groups of interest" in the wild"*. Citeseer.

[38] Bartlett, M. S., Littlewort, G., Frank, M., Lainscsek, C., Fasel, I., and Movellan, J. (2005). Recognizing facial expression: machine learning and application to spontaneous behavior. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 2, pages 568–573. IEEE.

[39] Bas, A., Smith, W. A. P., Bolkart, T., and Wuhrer, S. (2017). Fitting a 3d morphable model to edges: A comparison between hard and soft correspondences. In Chen, C.-S., Lu, J., and Ma, K.-K., editors, *Computer Vision – ACCV 2016 Workshops*, pages 377–391, Cham. Springer International Publishing.

[40] Beinema, T. (2014). The detection of facial expressions for action coordination.

[41] Bejaoui, H., Ghazouani, H., and Barhoumi, W. (2017). Fully automated facial expression recognition using 3d morphable model and mesh-local binary pattern. In *International Conference on Advanced Concepts for Intelligent Vision Systems*, pages 39–50. Springer.

[42] Belahcene, M., Laid, M., Chouchane, A., Ouamane, A., and Bourennane, S. (2016). Local descriptors and tensor local preserving projection in face recognition. In *2016 6th European Workshop on Visual Information Processing (EUVIP)*, pages 1–6. IEEE.

[43] Berretti, S., Amor, B. B., Daoudi, M., and Del Bimbo, A. (2011). 3d facial expression recognition using sift descriptors of automatically detected keypoints. *The Visual Computer*, 27(11):1021.

[44] Berretti, S., Del Bimbo, A., Pala, P., Amor, B. B., and Daoudi, M. (2010). A set of selected sift features for 3d facial expression recognition. In *2010 20th International Conference on Pattern Recognition, Istanbul, Volume: 1, doi: 10.1109/ICPR.2010.1002*, pages 4125–4128. IEEE.

[45] Bettadapura, V. (2012). Face expression recognition and analysis: The state of the art. *CoRR*, abs/1203.6722.

[46] Beveridge, J. R., Givens, G. H., Phillips, P. J., Draper, B. A., and Lui, Y. M. (2008). Focus on quality, predicting frvt 2006 performance. In *2008 8th IEEE International Conference on Automatic Face & Gesture Recognition, Amsterdam, NL, doi: 10.1109/AFGR.2008.4813375*, pages 1–8. IEEE.

[47] Bhati, D. and Gupta, V. (2015). Survey-a comparative analysis of face recognition technique. *Int. J. Eng. Res. General Sci.*, 3(2):597–609.

[48] Bigdeli, A., Sim, C., Biglari-Abhari, M., and Lovell, B. C. (2007). Face detection on embedded systems. In Lee, Y.-H., Kim, H.-N., Kim, J., Park, Y., Yang, L. T., and Kim, S. W., editors, *Embedded Software and Systems*, pages 295–308, Berlin, Heidelberg. Springer Berlin Heidelberg.

[49] Blanz, V., Basso, C., Poggio, T., and Vetter, T. (2003). Reanimating faces in images and video. In *Computer graphics forum*, volume 22, pages 641–650. Wiley Online Library.

[50] Blanz, V. and Vetter, T. (2003). Face recognition based on fitting a 3d morphable model. *IEEE Transactions on pattern analysis and machine intelligence*, 25(9):1063–1074.

[51] Blanz, V., Vetter, T., et al. (1999). A morphable model for the synthesis of 3d faces. In *Siggraph*, volume 99, pages 187–194.

[52] Boccignone, G., Bodini, M., Cuculo, V., and Grossi, G. (2018). Predictive sampling of facial expression dynamics driven by a latent action space. In *2018 14th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS), Las Palmas de Gran Canaria, Spain, doi: 10.1109/SITIS.2018.00031*, pages 143–150. IEEE.

[53] Bodini, M. (2019). A review of facial landmark extraction in 2d images and videos using deep learning. *Big Data and Cognitive Computing*, 3(1):14.

[54] Bolkart, T. and Wuhrer, S. (2015). A groupwise multilinear correspondence optimization for 3d faces. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Santiag, doi: 10.1109/ICCV.2015.411*, pages 3604–3612.

[55] Booth, J., Roussos, A., Ponniah, A., Dunaway, D., and Zafeiriou, S. (2018a). Large scale 3d morphable models. *International Journal of Computer Vision*, 126(2-4):233–254.

[56] Booth, J., Roussos, A., Ververas, E., Antonakos, E., Ploumpis, S., Panagakis, Y., and Zafeiriou, S. (2018b). 3d reconstruction of "in-the-wild" faces in images and videos. *IEEE transactions on pattern analysis and machine intelligence*, 40(11):2638–2652.

[57] Booth, J., Roussos, A., Zafeiriou, S., Ponniah, A., and Dunaway, D. (2016). A 3d morphable model learnt from 10,000 faces. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5543–5552.

[58] Boser, B. E., Guyon, I. M., and Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual ACM workshop on Computational learning theory*, pages 144–152. ACM Press.

[59] Bouaziz, S., Wang, Y., and Pauly, M. (2013). Online modeling for realtime facial animation. *ACM Transactions on Graphics (ToG)*, 32(4):40.

[60] Bouktif, S., Fiaz, A., Ouni, A., and Serhani, M. (2018). Optimal deep learning lstm model for electric load forecasting using feature selection and genetic algorithm: Comparison with machine learning approaches. *Energies*, 11(7):1636.

[61] Bousmalis, K., Silberman, N., Dohan, D., Erhan, D., and Krishnan, D. (2017). Unsupervised pixel-level domain adaptation with generative adversarial networks. *In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, doi: 10.1109/CVPR.2017.18*, pages 95–104.

[62] Bowyer, K. W., Chang, K., and Flynn, P. (2006). A survey of approaches and challenges in 3d and multi-modal 3d+ 2d face recognition. *Computer vision and image understanding*, 101(1):1–15.

[63] Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

[64] Bulat, A. and Tzimiropoulos, G. (2017). How far are we from solving the 2d 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks). *2017 IEEE International Conference on Computer Vision (ICCV), Venice, doi: 10.1109/ICCV.2017.116*, pages 1021–1030.

[65] Bull, P. (2013). *Communication under the microscope: The theory and practice of microanalysis*. Routledge.

[66] Cao, C., Hou, Q., and Zhou, K. (2014a). Displaced dynamic expression regression for real-time facial tracking and animation. *ACM Transactions on graphics (TOG)*, 33(4):43.

[67] Cao, C., Weng, Y., Zhou, S., Tong, Y., and Zhou, K. (2014b). Facewarehouse: A 3d facial expression database for visual computing. *IEEE Transactions on Visualization and Computer Graphics*, 20(3):413–425.

[68] Cao, J., Hu, Y., Zhang, H., He, R., and Sun, Z. (2018). Learning a high fidelity pose invariant model for high-resolution face frontalization. *CoRR*, abs/1806.08472.

[69] Caramihale, T., Popescu, D., and Ichim, L. (2018a). Emotion classification using a tensorflow generative adversarial network implementation. *Symmetry*, 10(9):414.

[70] Caramihale, T., Popescu, D., and Ichim, L. (2018b). Emotion classification using a tensorflow generative adversarial network implementation. *Symmetry*, 10(9).

[71] Carbonneau, M.-A., Cheplygina, V., Granger, E., and Gagnon, G. (2018). Multiple instance learning: A survey of problem characteristics and applications. *Pattern Recognition*, 77:329–353.

[72] Carcagnì, P., Del Coco, M., Leo, M., and Distante, C. (2015). Facial expression recognition and histograms of oriented gradients: a comprehensive study. *SpringerPlus*, 4(1):645.

[73] Celiktutan, O., Ulukaya, S., and Sankur, B. (2013). A comparative study of face landmarking techniques. *EURASIP Journal on Image and Video Processing*, 2013(1):13.

[74] Chai, X., Wang, Q., Zhao, Y., and Li, Y. (2016). Robust facial landmark detection based on initializing multiple poses. *International Journal of Advanced Robotic Systems*, 13(5):1729881416662793.

[75] Chang, C.-C. and Lin, C.-J. (2011a). Libsvm: A library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)*, 2(3):27.

[76] Chang, C.-C. and Lin, C.-J. (2011b). LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3):27:1–27:27. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[77] Chang, F.-J., Tran, A. T., Hassner, T., Masi, I., Nevatia, R., and Medioni, G. (2018). Expnet: Landmark-free, deep, 3d facial expressions. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 122–129. IEEE.

[78] Chang, Y., Hu, C., Feris, R., and Turk, M. (2006). Manifold based analysis of facial expression. *Image and Vision Computing*, 24(6):605–614.

[79] Chen, D., Chen, Q., Wu, J., Yu, X., and Jia, T. (2019a). Face swapping: Realistic image synthesis based on facial landmarks alignment. *Mathematical Problems in Engineering*, 2019.

[80] Chen, J., Chen, Z., Chi, Z., and Fu, H. (2016a). Facial expression recognition in video with multiple feature fusion. *IEEE Transactions on Affective Computing*, 9(1):38–50.

[81] Chen, J., Luo, Z., Takiguchi, T., and Ariki, Y. (2016b). Multithreading cascade of surf for facial expression recognition. *EURASIP Journal on Image and Video Processing*, 2016(1):37.

[82] Chen, X., Duan, Y., Houthooft, R., Schulman, J., Sutskever, I., and Abbeel, P. (2016c). Infogan: Interpretable representation learning by information maximizing generative adversarial nets. *CoRR*, abs/1606.03657.

[83] Chen, Z., Jin, M., Deng, Y., Wang, J.-S., Huang, H., Deng, X., and Huang, C.-M. (2019b). Improvement of a deep learning algorithm for total electron content maps: Image completion. *Journal of Geophysical Research: Space Physics*, 124(1):790–800.

[84] Cheng, F., Yu, J., and Xiong, H. (2010). Facial expression recognition in jaffe dataset based on gaussian process classification. *IEEE Transactions on Neural Networks*, 21(10):1685–1690.

[85] Cheng, S., Marras, I., Zafeiriou, S., and Pantic, M. (2017). Statistical non-rigid icp algorithm and its application to 3d face alignment. *Image and Vision Computing*, 58:3–12.

[86] Choi, Y., Choi, M., Kim, M., Ha, J.-W., Kim, S., and Choo, J. (2018). Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, doi: 10.1109/CVPR.2018.00916*, pages 8789–8797.

[87] Christou, N. and Kanojiya, N. (2019). Human facial expression recognition with convolution neural networks. In *Third International Congress on Information and Communication Technology*, pages 539–545. Springer.

[88] Chrysos, G. G., Antonakos, E., Snape, P., Asthana, A., and Zafeiriou, S. (2018a). A comprehensive performance evaluation of deformable face tracking “in-the-wild”. *International Journal of Computer Vision*, 126(2-4):198–232.

[89] Chrysos, G. G., Favaro, P., and Zafeiriou, S. (2018b). Motion deblurring of faces. *CoRR*, abs/1803.03330.

[90] Chu, W.-S. (2017). *Automatic Analysis of Facial Actions: Learning from Transductive, Supervised and Unsupervised Frameworks*. PhD thesis.

[91] Chu, W.-S., De la Torre, F., and Cohn, J. F. (2016). Selective transfer machine for personalized facial expression analysis. *IEEE transactions on pattern analysis and machine intelligence*, 39(3):529–545.

[92] Chu, W.-S., De la Torre, F., and Cohn, J. F. (2017). Learning spatial and temporal cues for multi-label facial action unit detection. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017), Washington, doi: 10.1109/FG.2017.13*, pages 25–32. IEEE.

[93] Chuan, T., Xinrui, H., Zhicheng, W., Yu, Z., Mingyu, X., and Xin, W. (2019). Head pose estimation via multi-task cascade cnn. In *Proceedings of the 2019 3rd High Performance Computing and Cluster Technologies Conference*, pages 123–127. ACM.

[94] Claes, P., Liberton, D. K., Daniels, K., Rosana, K. M., Quillen, E. E., Pearson, L. N., McEvoy, B., Bauchet, M., Zaidi, A. A., Yao, W., et al. (2014). Modeling 3d facial shape from dna. *PLoS genetics*, 10(3):e1004224.

[95] Cohen, I., Garg, A., Huang, T. S., et al. (2000). Emotion recognition from facial expressions using multilevel hmm. In *Neural information processing systems*, volume 2. Citeseer.

[96] Cohen, I., Sebe, N., Garg, A., Chen, L. S., and Huang, T. S. (2003a). Facial expression recognition from video sequences: temporal and static modeling. *Computer Vision and image understanding*, 91(1-2):160–187.

[97] Cohen, I., Sebe, N., Gozman, F., Cirelo, M. C., and Huang, T. S. (2003b). Learning bayesian network classifiers for facial expression recognition both labeled and unlabeled data. In *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings.*, volume 1, pages I–I. IEEE.

[98] Cohn, J. F. and De la Torre, F. (2014). Automated face analysis for affective computing. In *The Oxford handbook of affective computing. In R. A. Calvo, S. K. D'Mello, J. Gratch, A. Kappas (Eds.), Oxford library of psychology, Oxford University Press.*, pages 131–150.

[99] Coleman, T. F. and Li, Y. (1996). A reflective newton method for minimizing a quadratic function subject to bounds on some of the variables. *SIAM Journal on Optimization*, 6(4):1040–1058.

[100] Connie, T., Al-Shabi, M., Cheah, W. P., and Goh, M. (2017). Facial expression recognition using a hybrid cnn–sift aggregator. In *International Workshop on Multi-disciplinary Trends in Artificial Intelligence, Volume 10607, ISBN : 978-3-319-69455-9*, pages 139–149. Springer.

[101] Cootes, T. F., Edwards, G. J., and Taylor, C. J. (1998). Active appearance models. In *European conference on computer vision*, pages 484–498. Springer.

[102] Corneanu, C. A., Simón, M. O., Cohn, J. F., and Guerrero, S. E. (2016). Survey on rgb, 3d, thermal, and multimodal approaches for facial expression recognition: History, trends, and affect-related applications. *IEEE transactions on pattern analysis and machine intelligence*, 38(8):1548–1568.

[103] Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20(3):273–297.

[104] Curtó, J. D., Zarza, I. C., la Torre, F. D., King, I., and Lyu, M. R. (2017). High-resolution deep convolutional generative adversarial networks. *CoRR*, abs/1711.06491.

[105] Dahmane, M. and Meunier, J. (2011a). Continuous emotion recognition using gabor energy filters. In *international conference on Affective computing and intelligent interaction, ACII 2011. Lecture Notes in Computer Science, vol 6975*, pages 351–358. Springer.

[106] Dahmane, M. and Meunier, J. (2011b). Emotion recognition using dynamic grid-based hog features. In *Face and Gesture 2011, Santa Barbara, CA, doi: 10.1109/FG.2011.5771368*, pages 884–888. IEEE.

[107] Dahmani, S., Colotte, V., Girard, V., and Ouni, S. (2019). Conditional variational auto-encoder for text-driven expressive audiovisual speech synthesis. *Proc. Interspeech 2019*, pages 2598–2602.

[108] Dai, H., Pears, N. E., Smith, W. A. P., and Duncan, C. (2017). A 3d morphable model of craniofacial shape and texture variation. In *ICCV*, pages 3104–3112.

[109] Dalal, N. and Triggs, B. (2005). Histograms of oriented gradients for human detection. In *international Conference on computer vision & Pattern Recognition (CVPR'05)*, volume 1, pages 886–893. IEEE Computer Society.

[110] Dapogny, A., Bailly, K., and Dubuisson, S. (2015). Pairwise conditional random forests for facial expression recognition. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, doi: 10.1109/ICCV.2015.431*, pages 3783–3791.

[111] Dapogny, A., Bailly, K., and Dubuisson, S. (2018). Confidence-weighted local expression predictions for occlusion handling in expression recognition and action unit detection. *International Journal of Computer Vision*, 126(2-4):255–271.

[112] Davison, A. K. (2016). *Micro-facial movement detection using spatio-temporal features*. PhD thesis, Manchester Metropolitan University.

[113] Davison, A. K., Lansley, C., Costen, N., Tan, K., and Yap, M. H. (2016). Samm: A spontaneous micro-facial movement dataset. *IEEE transactions on affective computing*, 9(1):116–129.

[114] Dawel, A., Wright, L., Irons, J., Dumbleton, R., Palermo, R., O'Kearney, R., and McKone, E. (2017). Perceived emotion genuineness: normative ratings for popular facial expression stimuli and the development of perceived-as-genuine and perceived-as-fake sets. *Behavior research methods*, 49(4):1539–1562.

[115] De la Torre, F., Campoy, J., Ambadar, Z., and Cohn, J. F. (2007). Temporal segmentation of facial behavior. In *2007 IEEE 11th International Conference on Computer Vision, Rio de Janeiro, doi: 10.1109/ICCV.2007.4408961*, pages 1–8. IEEE.

[116] Del Coco, M., Carcagnì, P., Leo, M., and Distante, C. (2015). A minimax framework for gender classification based on small-sized datasets. In *International Conference on Advanced Concepts for Intelligent Vision Systems. ACIVS 2015. Lecture Notes in Computer Science, vol 9386*, pages 415–427. Springer, Cham.

[117] Den Uyl, M. and Van Kuilenburg, H. (2005). The facereader: Online facial expression recognition. In *Proceedings of measuring behavior*, volume 30, pages 589–590. Citeseer.

[118] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.

[119] Deng, J., Pang, G., Zhang, Z., Pang, Z., Yang, H., and Yang, G. (2019). cgan based facial expression recognition for human-robot interaction. *IEEE Access*, 7:9848–9859.

[120] Deng, J., Pang, G., Zhang, Z., Pang, Z., Yang, H., and Yang, G. (2019). cgan based facial expression recognition for human-robot interaction. *IEEE Access*, 7:9848–9859.

[121] Deng, J., Roussos, A., Chrysos, G., Ververas, E., Kotsia, I., Shen, J., and Zafeiriou, S. (2019). The menpo benchmark for multi-pose 2d and 3d facial landmark localisation and tracking. *International Journal of Computer Vision*, 127(6-7):599–624.

[122] Deng, J., Zhou, Y., Cheng, S., and Zaferiou, S. (2018). Cascade multi-view hourglass model for robust 3d face alignment. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), Xi'an, doi: 10.1109/FG.2018.00064*, pages 399–403. IEEE.

[123] Denton, E. L., Chintala, S., szlam, a., and Fergus, R. (2015). Deep generative image models using a laplacian pyramid of adversarial networks. In Cortes, C., Lawrence, N. D., Lee, D. D., Sugiyama, M., and Garnett, R., editors, *Advances in Neural Information Processing Systems 28*, pages 1486–1494. Curran Associates, Inc.

[124] Dhall, A., Asthana, A., Goecke, R., and Gedeon, T. (2011). Emotion recognition using phog and lpq features. In *Face and Gesture 2011, Santa Barbara, CA, doi: 10.1109/FG.2011.5771366*, pages 878–883. IEEE.

[125] Dhall, A., Goecke, R., Joshi, J., Wagner, M., and Gedeon, T. (2013). Emotion recognition in the wild challenge (emotiw) challenge and workshop summary. In *ICMI 2013 - Proceedings of the 2013 ACM International Conference on Multimodal Interaction*, pages 371–372. ACM.

[126] Dhall, A., Goecke, R., Lucey, S., and Gedeon, T. (2011). Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark. In *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, pages 2106–2112.

[127] Dhall, A., Kaur, A., Goecke, R., and Gedeon, T. (2018). Emotiw 2018: audio-video, student engagement and group-level affect prediction. In Mower Provost, E., Soleymani, M., and Worsley, M., editors, *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, pages 653–656, United States of America. Association for Computing Machinery (ACM). International Conference on Multimodal Interfaces 2018, ICMI 2018 ; Conference date: 16-10-2019 Through 20-10-2019.

[128] Dhall, A., Ramana Murthy, O., Goecke, R., Joshi, J., and Gedeon, T. (2015). Video and image based emotion recognition challenges in the wild: Emotiw 2015. In Bohus, D., Horaud, R., and Meng, H., editors, *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, pages 423–426, United States of America. Association for Computing Machinery (ACM). International Conference on Multimodal Interfaces 2015, ICMI 2015 ; Conference date: 09-11-2015 Through 13-11-2015.

[129] Dhavalikar, A. S. and Kulkarni, R. K. (2014). Facial expression recognition using euclidean distance method. *Journal of Telematics and Informatics*, 2(1):1–6.

[130] Dhote, S. Y., Gotmare, A., and Nimbarte, M. (2015). Differentiating identical twins by using conditional face recognition algorithms. *International Journal of Science and Research*, 4(3).

[131] Dietterich, T. G. and Bakiri, G. (1994). Solving multiclass learning problems via error-correcting output codes. *Journal of artificial intelligence research*, 2:263–286.

[132] Dietterich, T. G., Lathrop, R. H., and Lozano-Pérez, T. (1997). Solving the multiple instance problem with axis-parallel rectangles. *Artificial intelligence*, 89(1-2):31–71.

[133] Ding, H., Sricharan, K., and Chellappa, R. (2018). Exprgan: Facial expression editing with controllable expression intensity. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

[134] Ding, H., Zhou, S., and Chellappa, R. (2017a). Facenet2expnet: Regularizing a deep face recognition net for expression recognition. In *2017 12th IEEE International Conference on Automatic Face Gesture Recognition (FG 2017)*, pages 118–126, Los Alamitos, CA, USA. IEEE Computer Society.

[135] Ding, X., Chu, W.-S., De la Torre, F., Cohn, J. F., and Wang, Q. (2013). Facial action unit event detection by cascade of tasks. In *Proceedings of the IEEE 2013 international conference on computer vision, Sydney, NSW, doi: 10.1109/ICCV.2013.298*, pages 2400–2407.

[136] Ding, X., Chu, W.-S., De la Torre, F., Cohn, J. F., and Wang, Q. (2016). Cascade of tasks for facial expression analysis. *Image and Vision Computing*, 51:36–48.

[137] Ding, X. and Fang, C. (2004). Discussions on some problems in face recognition. In *Chinese Conference on Biometric Recognition*, pages 47–56. Springer.

[138] Ding, Y., Zhao, Q., Li, B., and Yuan, X. (2017b). Facial expression recognition from image sequence based on lbp and taylor expansion. *IEEE Access*, 5:19409–19419.

[139] Dobs, K., Bülthoff, I., and Schultz, J. (2018). Use and usefulness of dynamic face stimuli for face perception studies–a review of behavioral findings and methodology. *Frontiers in psychology*, 9:1355.

[140] Dollár, P., Appel, R., Belongie, S., and Perona, P. (2014). Fast feature pyramids for object detection. *IEEE transactions on pattern analysis and machine intelligence*, 36(8):1532–1545.

[141] Donahue, J., Hendricks, L. A., Rohrbach, M., Venugopalan, S., Guadarrama, S., Saenko, K., and Darrell, T. (2017). Long-term recurrent convolutional networks for visual recognition and description. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(4):677–691.

[142] Doran, G. and Ray, S. (2016). Multiple-instance learning from distributions. *The Journal of Machine Learning Research*, 17(1):4384–4433.

[143] Dou, P., Shah, S. K., and Kakadiaris, I. A. (2017). End-to-end 3d face reconstruction with deep neural networks. *CoRR*, abs/1704.05020.

[144] Du, S. and Martinez, A. M. (2015). Compound facial expressions of emotion: from basic research to clinical applications. *Dialogues in clinical neuroscience*, 17(4):443.

[145] Du, S., Tao, Y., and Martinez, A. M. (2014). Compound facial expressions of emotion. *Proceedings of the National Academy of Sciences*, 111(15):E1454–E1462.

[146] Duan, F., Huang, D., Tian, Y., Lu, K., Wu, Z., and Zhou, M. (2015). 3d face reconstruction from skull by regression modeling in shape parameter spaces. *Neurocomputing*, 151:674–682.

[147] Duong, C. N., Luu, K., Quach, K. G., and Bui, T. D. (2019). Deep appearance models: A deep boltzmann machine approach for face modeling. *International Journal of Computer Vision*, 127(5):437–455.

[148] Eck, D. and Schmidhuber, J. (2002). A first look at music composition using lstm recurrent neural networks. *Istituto Dalle Molle Di Studi Sull Intelligenza Artificiale*, 103:48.

[149] EDWARDS, V. V. (2017). The definitive guide to reading microexpressions (facial expressions).

[150] Ekman, P. (1976). Pictures of facial affect. *Consulting Psychologists Press*.

[151] Ekman, P. (1993). Facial expression and emotion. *American psychologist*, 48(4):384.

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

[153] Ekmen, B. and Ekenel, H. K. (2019). From 2d to 3d real-time expression transfer for facial animation. *Multimedia Tools and Applications*, 78(9):12519–12535.

[154] El Maghraby, A., Abdalla, M., Enany, O., and El Nahas, M. (2014). Detect and analyze face parts information using viola-jones and geometric approaches. *International Journal of Computer Applications*, 101(3):23–28.

[155] Eleftheriadis, S., Rudovic, O., and Pantic, M. (2014). Discriminative shared gaussian processes for multiview and view-invariant facial expression recognition. *IEEE transactions on image processing*, 24(1):189–204.

[156] Etemad, K. and Chellappa, R. (1997). Discriminant analysis for recognition of human face images. *Josa a*, 14(8):1724–1733.

[157] Fan, Y., Lam, J. C., and Li, V. O. (2018). Multi-region ensemble convolutional neural network for facial expression recognition. In *International Conference on Artificial Neural Networks*, pages 84–94. Springer.

[158] Fan, Y., Lu, X., Li, D., and Liu, Y. (2016). Video-based emotion recognition using cnn-rnn and c3d hybrid networks. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, pages 445–450. ACM.

[159] Fang, T., Zhao, X., Ocegueda, O., Shah, S. K., and Kakadiaris, I. A. (2011). 3d facial expression recognition: A perspective on promises and challenges. In *2011 IEEE International Conference on Automatic Face Gesture Recognition (FG 2011)*, pages 603–610, Los Alamitos, CA, USA. IEEE Computer Society.

[160] Fasel, B. and Luettin, J. (2003). Automatic facial expression analysis: a survey. *Pattern recognition*, 36(1):259–275.

[161] Felzenszwalb, P. F., Girshick, R. B., McAllester, D., and Ramanan, D. (2009). Object detection with discriminatively trained part-based models. *IEEE transactions on pattern analysis and machine intelligence*, 32(9):1627–1645.

[162] Feng, Y., Wu, F., Shao, X., Wang, Y., and Zhou, X. (2018). Joint 3d face reconstruction and dense alignment with position map regression network. In *Proceedings of the European Conference on Computer Vision (ECCV), ECCV 2018. Lecture Notes in Computer Science, vol 11218*, pages 534–551. Springer.

[163] Fischer, A. and Igel, C. (2014). Training restricted boltzmann machines: An introduction. *Pattern Recognition*, 47(1):25–39.

[164] Forsyth, D. A. and Ponce, J. (2002). *Computer vision: a modern approach*. Prentice Hall Professional Technical Reference.

[165] Friesen, E. and Ekman, P. (1978). Facial action coding system: a technique for the measurement of facial movement. *Palo Alto*, 3.

[166] Friesen, W. V., Ekman, P., et al. (1983). Emfacs-7: Emotional facial action coding system. *Unpublished manuscript, University of California at San Francisco*, 2(36):1.

[167] Galleguillos, C., Babenko, B., Rabinovich, A., and Belongie, S. (2008). Weakly supervised object localization with stable segmentations. In *European Conference on Computer Vision – ECCV 2008. Lecture Notes in Computer Science, vol 5302*, pages 193–207. Springer.

[168] Garrido, P., Valgaerts, L., Sarmadi, H., Steiner, I., Varanasi, K., Perez, P., and Theobalt, C. (2015). Vdub: Modifying face video of actors for plausible visual alignment to a dubbed audio track. In *Computer graphics forum*, volume 34, pages 193–204. Wiley Online Library.

[169] Garrido, P., Zollhöfer, M., Casas, D., Valgaerts, L., Varanasi, K., Pérez, P., and Theobalt, C. (2016). Reconstruction of personalized 3d face rigs from monocular video. *ACM Transactions on Graphics (TOG)*, 35(3):28.

[170] Gauthier, J. (2014). Conditional generative adversarial nets for convolutional face generation. *Class Project for Stanford CS231N: Convolutional Neural Networks for Visual Recognition, Winter semester*, 2014(5):2.

[171] Ge, H., Xia, Y., Chen, X., Berry, R., and Wu, Y. (2018). Fictitious gan: Training gans with historical models. In Ferrari, V., Hebert, M., Sminchisescu, C., and Weiss, Y., editors, *Computer Vision – ECCV 2018*, pages 122–137, Cham. Springer International Publishing.

[172] Gecer, B., Ploumpis, S., Kotsia, I., and Zafeiriou, S. (2019). Ganfit: Generative adversarial network fitting for high fidelity 3d face reconstruction. *arXiv preprint arXiv:1902.05978*.

[173] Gehrig, T., Al-Halah, Z., Ekenel, H. K., and Stiefelhagen, R. (2015). Action unit intensity estimation using hierarchical partial least squares. In *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, volume 1, pages 1–6. IEEE.

[174] Geng, C. and Jiang, X. (2012). Face alignment based on the multi-scale local features. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1517–1520. IEEE.

[175] Genova, K., Cole, F., Maschinot, A., Sarna, A., Vlasic, D., and Freeman, W. T. (2018). Unsupervised training for 3d morphable model regression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8377–8386.

[176] Georgescu, M.-I., Ionescu, R. T., and Popescu, M. (2018). Local learning with deep and handcrafted features for facial expression recognition. *arXiv preprint arXiv:1804.10892*.

[177] Georgescu, M.-I., Ionescu, R. T., and Popescu, M. (2019). Local learning with deep and handcrafted features for facial expression recognition. *IEEE Access*, 7:64827–64836.

[178] Gerig, T., Morel-Forster, A., Blumer, C., Egger, B., Luthi, M., Schönborn, S., and Vetter, T. (2018). Morphable face models-an open framework. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), Xi'an, doi: 10.1109/FG.2018.00021*, pages 75–82. IEEE.

[179] Gers, F. A., Schmidhuber, J., and Cummins, F. (2000). Learning to forget: Continual prediction with lstm. *Neural Computation*, 12(10):2451–2471.

[180] Geyer, G., Peel, J. S., Streng, M., Voigt, S., Fischer, J., and PREUßE, M. (2014). A remarkable amgan (middle cambrian, stage 5) fauna from the sauk tanga, madygen region, kyrgyzstan. *Bulletin of Geosciences*, 89(2):375–400.

[181] Gharsalli, S., Emile, B., Laurent, H., and Desquesnes, X. (2016). Feature selection for emotion recognition based on random forest. In *VISIGRAPP (4: VISAPP)*, pages 610–617.

[182] Ghosh, A., Bhattacharya, B., and Chowdhury, S. B. R. (2016). Sad-gan: Synthetic autonomous driving using generative adversarial networks. *arXiv preprint arXiv:1611.08788*.

[183] Ghosh, S., Dhall, A., and Sebe, N. (2018). Automatic group affect analysis in images via visual attribute and feature networks. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 1967–1971.

[184] Ghosh, S., Dhall, A., and Sebe, N. (2018). Automatic group affect analysis in images via visual attribute and feature networks. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 1967–1971. IEEE.

[185] Giannakakis, G., Grigoriadis, D., Giannakaki, K., Simantiraki, O., Roniotis, A., and Tsiknakis, M. (2019a). Review on psychological stress detection using biosignals. *IEEE Transactions on Affective Computing*.

[186] Giannakakis, G., Manousos, D., Chaniotakis, V., and Tsiknakis, M. (2018a). Evaluation of head pose features for stress detection and classification. In *2018 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*, pages 406–409. IEEE.

[187] Giannakakis, G., Manousos, D., Simos, P., and Tsiknakis, M. (2018b). Head movements in context of speech during stress induction. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 710–714. IEEE.

[188] Giannakakis, G., Pediaditis, M., Manousos, D., Kazantzaki, E., Chiarugi, F., Simos, P. G., Marias, K., and Tsiknakis, M. (2017). Stress and anxiety detection using facial cues from videos. *Biomedical Signal Processing and Control*, 31:89–101.

[189] Giannakakis, G., Trivizakis, E., Tsiknakis, M., and Marias, K. (2019b). A novel multi-kernel 1d convolutional neural network for stress recognition from ecg. In *2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, pages 1–4. IEEE.

[190] Girard, J. M., Cohn, J. F., Jeni, L. A., Lucey, S., and De la Torre, F. (2015). How much training data for facial action unit detection? In *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, volume 1, pages 1–8. IEEE.

[191] Girshick, R., Donahue, J., Darrell, T., and Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587.

[192] Gonzalez, I., Cartella, F., Enescu, V., and Sahli, H. (2015). Recognition of facial actions and their temporal segments based on duration models. *Multimedia Tools and Applications*, 74(22):10001–10024.

[193] Goodfellow, I. (2016). Nips 2016 tutorial: Generative adversarial networks. *arXiv preprint arXiv:1701.00160*.

[194] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680.

[195] Goodfellow, I. J., Erhan, D., Carrier, P. L., Courville, A., Mirza, M., Hamner, B., Cukierski, W., Tang, Y., Thaler, D., Lee, D.-H., et al. (2013). Challenges in representation learning: A report on three machine learning contests. In *International Conference on Neural Information Processing*, pages 117–124. Springer.

[196] Gopalan, N., Bellamkonda, S., and Chaitanya, V. S. (2018). Facial expression recognition using geometric landmark points and convolutional neural networks. In *2018 International Conference on Inventive Research in Computing Applications (ICIRCA)*, pages 1149–1153. IEEE.

[197] Graves, A., Jaitly, N., and Mohamed, A.-r. (2013). Hybrid speech recognition with deep bidirectional lstm. In *2013 IEEE workshop on automatic speech recognition and understanding*, pages 273–278. IEEE.

[198] Grewe, C. M. and Zachow, S. (2016). Fully automated and highly accurate dense correspondence for facial surfaces. In *European Conference on Computer Vision*, pages 552–568. Springer.

[199] Grgic, M. and Delac, K. (2013). Face recognition homepage. *Zagreb, Croatia (www. face-rec. org/databases)*, 324.

[200] Gross, R. (2005). *Face Databases*, pages 301–327. Springer New York, New York, NY.

[201] Gross, R., Matthews, I., Cohn, J., Kanade, T., and Baker, S. (2010). Multi-pie. *Image and Vision Computing*, 28(5):807–813.

[202] Grother, P. J., Grother, P. J., and Ngan, M. (2014). *Face recognition vendor test (FRVT)*. US Department of Commerce, National Institute of Standards and Technology.

[203] Gu, G., Kim, S. T., Kim, K. H., Baddar, W. J., and Ro, Y. M. (2017). Differential generative adversarial networks: Synthesizing non-linear facial variations with limited number of training data. *CoRR*, abs/1711.10267.

[204] Gu, J., Wang, Z., Kuen, J., Ma, L., Shahroudy, A., Shuai, B., Liu, T., Wang, X., Wang, G., Cai, J., et al. (2018). Recent advances in convolutional neural networks. *Pattern Recognition*, 77:354–377.

[205] Gui, L., Baltrušaitis, T., and Morency, L.-P. (2017). Curriculum learning for facial expression recognition. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, pages 505–511. IEEE.

[206] Guo, Y., Zhang, J., Cai, J., Jiang, B., and Zheng, J. (2018). Cnn-based real-time dense face reconstruction with inverse-rendered photo-realistic face images. *IEEE transactions on pattern analysis and machine intelligence*.

[207] Guo, Y., Zhao, G., and Pietikäinen, M. (2016). Dynamic facial expression recognition with atlas construction and sparse representation. *IEEE Transactions on Image Processing*, 25(5):1977–1992.

[208] Gupta, O., Raviv, D., and Raskar, R. (2018). Illumination invariants in deep video expression recognition. *Pattern Recognition*, 76:25–35.

[209] Haamer, R. E., Kulkarni, K., Imanpour, N., Haque, M. A., Avots, E., Breisch, M., Nasrollahi, K., Escalera, S., Ozcinar, C., Baro, X., et al. (2018). Changes in facial expression as biometric: A database and benchmarks of identification. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 621–628. IEEE.

[210] Hadjiabadi, D. H. (2018a). Generating realistic morphologies of neurons in rodent hippocampus with dcgan. *bioRxiv*, page 363481.

[211] Hadjiabadi, D. H. (2018b). Generating realistic morphologies of neurons in rodent hippocampus with dgcan. *bioRxiv*.

[212] Hamm, J., Kohler, C. G., Gur, R. C., and Verma, R. (2011). Automated facial action coding system for dynamic analysis of facial expressions in neuropsychiatric disorders. *Journal of neuroscience methods*, 200(2):237–256.

[213] Hammal, Z. (2006). *Facial features segmentation, analysis and recognition of facial expressions using the transferable belief model*. PhD thesis, PhD thesis, LIS Laboratory, Grenoble, France.

[214] Hammond, P. and Suttie, M. (2012). Large-scale objective phenotyping of 3d facial morphology. *Human mutation*, 33(5):817–825.

[215] Han, J., Zhang, Z., Cummins, N., and Schuller, B. W. (2018). Adversarial training in affective computing and sentiment analysis: Recent advances and perspectives [review article]. *IEEE Computational Intelligence Magazine*, 14:68–81.

[216] Han, J., Zhang, Z., and Schuller, B. (2019). Adversarial training in affective computing and sentiment analysis: Recent advances and perspectives. *IEEE Computational Intelligence Magazine*, 14(2):68–81.

[217] Han, S., Meng, Z., Khan, A.-S., and Tong, Y. (2016). Incremental boosting convolutional neural network for facial action unit recognition. In *Advances in neural information processing systems*, pages 109–117.

[218] Hariharan, A. and Adam, M. T. P. (2015). Blended emotion detection for decision support. *IEEE Transactions on Human-Machine Systems*, 45(4):510–517.

[219] Hasani, B., Arzani, M. M., Fathy, M., and Raahemifar, K. (2016). Facial expression recognition with discriminatory graphical models. In *2016 2nd International Conference of Signal Processing and Intelligent Systems (ICSPIS)*, pages 1–7. IEEE.

[220] Hasani, B. and Mahoor, M. H. (2017). Spatio-temporal facial expression recognition using convolutional neural networks and conditional random fields. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, pages 790–795. IEEE.

[221] Hastie, T., Tibshirani, R., Friedman, J., and Franklin, J. (2005). The elements of statistical learning: data mining, inference and prediction. *The Mathematical Intelligencer*, 27(2):83–85.

[222] Hatem, H., Beiji, Z., and Majeed, R. (2015). A survey of feature base methods for human face detection. *International Journal of Control and Automation*, 8(5):61–78.

[223] He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

[224] He, L., Jiang, D., Yang, L., Pei, E., Wu, P., and Sahli, H. (2015). Multimodal affective dimension prediction using deep bidirectional long short-term memory recurrent neural networks. In *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge*, pages 73–80. ACM.

[225] Henry, H. (2014). 3-d model links facial features and dna.

[226] Henry, H. (2017). Dan jurafsky lecture 6: Emotion cs 424p/ linguist 287 extracting social meaning and sentiment.

[227] Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. (2017). Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in neural information processing systems*, pages 6626–6637.

[228] Hinton, G. E., Osindero, S., and Teh, Y.-W. (2006). A fast learning algorithm for deep belief nets. *Neural computation*, 18(7):1527–1554.

[229] Hinton, G. E. and Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *science*, 313(5786):504–507.

[230] Hjelmås, E. and Low, B. K. (2001). Face detection: A survey. *Computer vision and image understanding*, 83(3):236–274.

[231] Hoque, M. E., El Kaliouby, R., and Picard, R. W. (2009). When human coders (and machines) disagree on the meaning of facial affect in spontaneous videos. In *International Workshop on Intelligent Virtual Agents*, pages 337–343. Springer.

[232] Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 24(6):417.

[233] Hou, X., Shen, L., Sun, K., and Qiu, G. (2017). Deep feature consistent variational autoencoder. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1133–1141. IEEE.

[234] Hsu, C., Chang, C., and Lin, C. (2003). A practical guide to support vector classification (taipei).

[235] Hu, G., Liu, L., Yuan, Y., Yu, Z., Hua, Y., Zhang, Z., Shen, F., Shao, L., Hospedales, T., Robertson, N., et al. (2018). Deep multi-task learning to recognise subtle facial expressions of mental states. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 103–119.

[236] Hu, P., Cai, D., Wang, S., Yao, A., and Chen, Y. (2017). Learning supervised scoring ensemble for emotion recognition in the wild. In *Proceedings of the 19th ACM international conference on multimodal interaction*, pages 553–560. ACM.

[237] Huang, D., Shan, C., Ardabilian, M., Wang, Y., and Chen, L. (2011). Local binary patterns and its application to facial image analysis: a survey. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 41(6):765–781.

[238] Huang, G. B., Mattar, M., Berg, T., and Learned-Miller, E. (2008). Labeled faces in the wild: A database for studying face recognition in unconstrained environments.

[239] Huang, R., Zhang, S., Li, T., and He, R. (2017). Beyond face rotation: Global and local perception gan for photorealistic and identity preserving frontal view synthesis. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2439–2448.

[240] Huang, X., Wang, S.-J., Zhao, G., and Piteikainen, M. (2015). Facial micro-expression recognition using spatiotemporal local binary pattern with integral projection. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 1–9.

[241] Huang, X. and Zhao, G. (2017). Spontaneous facial micro-expression analysis using spatiotemporal local radon-based binary pattern. In *2017 International Conference on the Frontiers and Advances in Data Science (FADS)*, pages 159–164. IEEE.

[242] Huang, X., Zhao, G., Zheng, W., and Pietikäinen, M. (2012). Towards a dynamic expression recognition system under facial occlusion. *Pattern Recognition Letters*, 33(16):2181–2191.

[243] Huber, P., Hu, G., Tena, R., Mortazavian, P., Koppen, P., Christmas, W. J., Ratsch, M., and Kittler, J. (2016). A multiresolution 3d morphable face model and fitting framework. In *Proceedings of the 11th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*.

[244] Huber, P., Kopp, P., Christmas, W., Rätsch, M., and Kittler, J. (2017). Real-time 3d face fitting and texture fusion on in-the-wild videos. *IEEE Signal Processing Letters*, 24(4):437–441.

[245] Hui, Z. and Chu, W.-S. (2016). An empirical study of dimensional reduction techniques for facial action units detection. *arXiv preprint arXiv:1603.08039*.

[246] Hupont, I. and Chetouani, M. (2019). Region-based facial representation for real-time action units intensity detection across datasets. *Pattern Analysis and Applications*, 22(2):477–489.

[247] Hwang, B., You, J., Vaessen, T., Myin-Germeys, I., Park, C., and Zhang, B.-T. (2018). Deep ecgnet: An optimal deep learning framework for monitoring mental stress using ultra short-term ecg signals. *TELEMEDICINE and e-HEALTH*, 24(10):753–772.

[248] Ichim, A. E., Bouaziz, S., and Pauly, M. (2015). Dynamic 3d avatar creation from hand-held video input. *ACM Transactions on Graphics (ToG)*, 34(4):45.

[249] Ilyas, C. M. A., Haque, M. A., Rehm, M., Nasrollahi, K., and Moeslund, T. B. (2018). Facial expression recognition for traumatic brain injured patients. In *VISIGRAPP (4: VISAPP)*, pages 522–530.

[250] Ionescu, R. T., Popescu, M., and Grozea, C. (2013). Local learning to improve bag of visual words model for facial expression recognition. In *Workshop on challenges in representation learning, ICML.*

[251] Isola, P., Zhu, J.-Y., Zhou, T., and Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134.

[252] Jackson, A. S., Bulat, A., Argyriou, V., and Tzimiropoulos, G. (2017). Large pose 3d face reconstruction from a single image via direct volumetric cnn regression. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1031–1039.

[253] Jain, V. and Learned-Miller, E. (2010). Fddb: A benchmark for face detection in unconstrained settings.

[254] Jaiswal, S., Martinez, B., and Valstar, M. F. (2015). Learning to combine local models for facial action unit detection. In *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, volume 6, pages 1–6. IEEE.

[255] Jaiswal, S. and Valstar, M. (2016). Deep learning the dynamic appearance and shape of facial action units. In *2016 IEEE winter conference on applications of computer vision (WACV)*, pages 1–8. IEEE.

[256] Jan, A., Meng, H., Gaus, Y. F. B. A., and Zhang, F. (2018). Artificial intelligent system for automatic depression level analysis through visual and vocal expressions. *IEEE Transactions on Cognitive and Developmental Systems*, 10(3):668–680.

[257] Jang, J., Kim, D. H., Kim, H.-I., and Ro, Y. M. (2017). Color channel-wise recurrent learning for facial expression recognition. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1233–1237. IEEE.

[258] Jeni, L. A., Cohn, J. F., and Kanade, T. (2015). Dense 3d face alignment from 2d videos in real-time. In *2015 11th IEEE international conference and workshops on automatic face and gesture recognition (FG)*, volume 1, pages 1–8. IEEE.

[259] Jeni, L. A., Lőrincz, A., Nagy, T., Palotai, Z., Sebők, J., Szabó, Z., and Takács, D. (2012). 3d shape estimation in video sequences provides high precision evaluation of facial expressions. *Image and Vision Computing*, 30(10):785–795.

[260] Jensen, O. H. (2008). Implementing the viola-jones face detection algorithm. Master's thesis, Technical University of Denmark, DTU, DK-2800 Kgs. Lyngby, Denmark.

[261] Ji, Q. (2019a). RPI Intelligent Systems Lab (ISL) Image Databases/ ISL Enhanced Cohn-Kanade AU-coded Facial Expression Database.

[262] Ji, Q. (2019b). Rpi intelligent systems lab (isl) image databases.

[263] Jia, J., Xu, Y., Zhang, S., and Xue, X. (2016). The facial expression recognition method of random forest based on improved pca extracting feature. In *2016 IEEE International Conference on Signal Processing, Communications and Computing (ICSPCC)*, pages 1–5. IEEE.

[264] Jia, Y. and Zhang, C. (2008). Instance-level semisupervised multiple instance learning. In *AAAI*, pages 640–645.

[265] Jiang, B. (2014). Spatial and temporal analysis of facial actions.

[266] Jiang, B. and Jia, K. (2016). Robust facial expression recognition algorithm based on local metric learning. *Journal of Electronic Imaging*, 25(1):013022.

[267] Jiang, B., Valstar, M., Martinez, B., and Pantic, M. (2013). A dynamic appearance descriptor approach to facial actions temporal modeling. *IEEE transactions on cybernetics*, 44(2):161–174.

[268] Jiang, B., Valstar, M., Martinez, B., and Pantic, M. (2014). A dynamic appearance descriptor approach to facial actions temporal. *IEEE transactions on cybernetics*, 44(2):161–174.

[269] Jiang, B., Valstar, M. F., and Pantic, M. (2011). Action unit detection using sparse appearance descriptors in space-time video volumes. In *Face and Gesture 2011*, pages 314–321. IEEE.

[270] Jin, X. and Tan, X. (2017). Face alignment in-the-wild: A survey. *Computer Vision and Image Understanding*, 162:1–22.

[271] Jolliffe, I. (2003). Principal component analysis. *Technometrics*, 45(3):276.

[272] Jolliffe, I. T. and Cadima, J. (2016). Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2065):20150202.

[273] Joshi, D. (2018). A brief review of facial expressions recognition system. *ASIAN JOURNAL FOR CONVERGENCE IN TECHNOLOGY (AJCT)-UGC LISTED.*

[274] Jourabloo, A. and Liu, X. (2015). Pose-invariant 3d face alignment. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3694–3702.

[275] Jourabloo, A. and Liu, X. (2016). Large-pose face alignment via cnn-based dense 3d model fitting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4188–4196.

[276] Jung, H., Lee, S., Yim, J., Park, S., and Kim, J. (2015). Joint fine-tuning in deep neural networks for facial expression recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 2983–2991.

[277] Kahou, S. E., Pal, C., Bouthillier, X., Froumenty, P., Gülçehre, Ç., Memisevic, R., Vincent, P., Courville, A., Bengio, Y., Ferrari, R. C., et al. (2013). Combining modality specific deep neural networks for emotion recognition in video. In *Proceedings of the 15th ACM on International conference on multimodal interaction*, pages 543–550. ACM.

[278] Kanade, T., Cohn, J. F., and Tian, Y. (2000). Comprehensive database for facial expression analysis. In *Proceedings Fourth IEEE International Conference on Automatic Face and Gesture Recognition (Cat. No. PR00580)*, pages 46–53. IEEE.

[279] Kaneko, T., Hiramatsu, K., and Kashino, K. (2017). Generative attribute controller with conditional filtered generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6089–6098.

[280] Kankanamge, M. and Madushika, S. (2018). *Facial analysis models for face and facial expression recognition*. PhD thesis, Queensland University of Technology.

[281] Kaya, H., Gürpınar, F., and Salah, A. A. (2017). Video-based emotion recognition in the wild using deep transfer learning and score fusion. *Image and Vision Computing*, 65:66–75.

[282] Kazemi, V. and Sullivan, J. (2014). One millisecond face alignment with an ensemble of regression trees. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1867–1874.

[283] Kervadec, C., Vielzeuf, V., Pateux, S., Lechervy, A., and Jurie, F. (2018). Cake: Compact and accurate k-dimensional representation of emotion. *arXiv preprint arXiv:1807.11215*.

[284] Khademi, M. and Morency, L.-P. (2014). Relative facial action unit detection. In *IEEE Winter Conference on Applications of Computer Vision*, pages 1090–1095. IEEE.

[285] Khan, M. H., McDonagh, J., and Tzimiropoulos, G. (2017). Synergy between face alignment and tracking via discriminative global consensus optimization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 3811–3819. IEEE.

[286] Khan, R. A. (2013). *Detection of emotions from video in non-controlled environment*. PhD thesis.

[287] Kim, B.-K., Dong, S.-Y., Roh, J., Kim, G., and Lee, S.-Y. (2016a). Fusing aligned and non-aligned face information for automatic affect recognition in the wild: a deep learning approach. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 48–57.

[288] Kim, B.-K., Roh, J., Dong, S.-Y., and Lee, S.-Y. (2016b). Hierarchical committee of deep convolutional neural networks for robust facial expression recognition. *Journal on Multimodal User Interfaces*, 10(2):173–189.

[289] King, D. E. (2009). Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, 10(Jul):1755–1758.

[290] Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

[291] Kingma, D. P., Mohamed, S., Rezende, D. J., and Welling, M. (2014). Semi-supervised learning with deep generative models. In *Advances in neural information processing systems*, pages 3581–3589.

[292] Kingma, D. P. and Welling, M. (2014). Stochastic gradient vb and the variational auto-encoder. In *Second International Conference on Learning Representations, ICLR*.

[293] Ko, B. (2018). A brief review of facial emotion recognition based on visual information. *sensors*, 18(2):401.

[294] Koelstra, S., Pantic, M., and Patras, I. (2010). A dynamic texture-based approach to recognition of facial actions and their temporal models. *IEEE transactions on pattern analysis and machine intelligence*, 32(11):1940–1954.

[295] Koujan, M. R., Akram, A., McCool, P., Westerfeld, J., Wilson, D., Dhaliwal, K., McLaughlin, S., and Perperidis, A. (2018). Multi-class classification of pulmonary endomicroscopic images. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 1574–1577. IEEE.

[296] Koujan, M. R. and Roussos, A. (2018). Combining dense nonrigid structure from motion and 3d morphable models for monocular 4d face reconstruction. In *Proceedings of the 15th ACM SIGGRAPH European Conference on Visual Media Production*, CVMP '18, pages 2:1–2:9, New York, NY, USA. ACM.

[297] Kowalski, M., Naruniec, J., and Trzcinski, T. (2017). Deep alignment network: A convolutional neural network for robust face alignment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 88–97.

[298] Krestinskaya, O. and James, A. P. (2017). Facial emotion recognition using min-max similarity classifier. In *2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, pages 752–758. IEEE.

[299] Krig, S. (2016). Ground truth data, content, metrics, and analysis. In *Computer Vision Metrics*, pages 247–271. Springer.

[300] Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105.

[301] Krumhuber, E. G., Skora, L., Küster, D., and Fou, L. (2017). A review of dynamic datasets for facial expression research. *Emotion Review*, 9(3):280–292.

[302] Kulkarni, K., Corneanu, C., Ofodile, I., Escalera, S., Baro, X., Hyniewska, S., Allik, J., and Anbarjafari, G. (2018). Automatic recognition of facial displays of unfelt emotions. *IEEE transactions on affective computing*.

[303] Kumari, J., Rajesh, R., and Pooja, K. (2015). Facial expression recognition: A survey. *Procedia Computer Science*, 58:486–491.

[304] Kurutach, T., Tamar, A., Yang, G., Russell, S. J., and Abbeel, P. (2018). Learning plannable representations with causal infogan. In *Advances in Neural Information Processing Systems*, pages 8733–8744.

[305] Kyrkou, C. (2017). Object detection using local binary patterns.

[306] la Torre De, F., Chu, W.-S., Xiong, X., Vicente, F., Ding, X., and Cohn, J. (2015). Intraface. In *IEEE International Conference on Automatic Face & Gesture Recognition and Workshops*, volume 1. NIH Public Access.

[307] Labati, R. D., Muñoz, E., Piuri, V., Sassi, R., and Scotti, F. (2019). Deep-ecg: Convolutional neural networks for ecg biometric recognition. *Pattern Recognition Letters*, 126:78–85.

[308] Laine, S., Karras, T., Aila, T., Herva, A., Saito, S., Yu, R., Li, H., and Lehtinen, J. (2017). Production-level facial performance capture using deep convolutional neural networks. In *Proceedings of the ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, page 10. ACM.

[309] Langner, O., Dotsch, R., Bijlstra, G., Wigboldus, D. H., Hawk, S. T., and Van Knippenberg, A. (2010a). Presentation and validation of the radboud faces database. *Cognition and emotion*, 24(8):1377–1388.

[310] Langner, O., Dotsch, R., Bijlstra, G., Wigboldus, D. H. J., Hawk, S. T., and van Knippenberg, A. (2010b). Presentation and validation of the radboud faces database. *Cognition and Emotion*, 24(8):1377–1388.

[311] Larochelle, H. and Murray, I. (2011). The neural autoregressive distribution estimator. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 29–37.

[312] Larsen, A. B. L., Sønderby, S. K., Larochelle, H., and Winther, O. (2015). Autoencoding beyond pixels using a learned similarity metric. *arXiv preprint arXiv:1512.09300*.

[313] Le, X.-H., Ho, H. V., Lee, G., and Jung, S. (2019). Application of long short-term memory (lstm) neural network for flood forecasting. *Water*, 11(7):1387.

[314] Learned-Miller, E., Huang, G. B., RoyChowdhury, A., Li, H., and Hua, G. (2016). Labeled faces in the wild: A survey. In *Advances in face detection and facial image analysis*, pages 189–248. Springer.

[315] Lee, K.-C., Ho, J., Yang, M.-H., and Kriegman, D. (2003). Video-based face recognition using probabilistic appearance manifolds. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 1, pages I–313. Citeseer.

[316] Lee, Y. J., Lee, S. J., Park, K. R., Jo, J., and Kim, J. (2012). Single view-based 3d face reconstruction robust to self-occlusion. *EURASIP Journal on Advances in Signal Processing*, 2012(1):176.

[317] Lemaire, P., Ardabilian, M., Chen, L., and Daoudi, M. (2013). Fully automatic 3d facial expression recognition using differential mean curvature maps and histograms of oriented gradients. *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pages 1–7.

[318] Lemerise, E. A. and Dodge, K. A. (2008). The development of anger and hostile interactions. *Handbook of emotions*, 3:730–741.

[319] Leonovych, O., Koujan, M. R., Akram, A., Westerfeld, J., Wilson, D., Dhaliwal, K., McLaughlin, S., and Perperidis, A. (2018). Texture descriptors for classifying sparse, irregularly sampled optical endomicroscopy images. In *Annual Conference on Medical Image Understanding and Analysis*, pages 165–176. Springer.

[320] Leopold, D. A., O'Toole, A. J., Vetter, T., and Blanz, V. (2001). Prototype-referenced shape encoding revealed by high-level aftereffects. *Nature neuroscience*, 4(1):89.

[321] Leung, T., Song, Y., and Zhang, J. (2011). Handling label noise in video classification via multiple instance learning. In *2011 International Conference on Computer Vision*, pages 2056–2063. IEEE.

[322] Levi, G. and Hassner, T. (2015). Emotion recognition in the wild via convolutional neural networks and mapped binary patterns. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, ICMI '15, pages 503–510, New York, NY, USA. ACM.

[323] Levine, M. D. and Yu, Y. C. (2009). State-of-the-art of 3d facial reconstruction methods for face recognition based on a single 2d training image per person. *Pattern Recognition Letters*, 30(10):908–913.

[324] Li, C. (2013). Probability estimation in random forests.

[325] Li, C., Zhou, K., and Lin, S. (2015). Simulating makeup through physics-based manipulation of intrinsic image layers. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4621–4629.

[326] Li, D., Li, Z., Luo, R., Deng, J., and Sun, S. (2019a). Multi-pose facial expression recognition based on generative adversarial network. *IEEE Access*, 7:143980–143989.

[327] Li, H., Chen, L., Huang, D., Wang, Y., and Morvan, J.-M. (2012). 3d facial expression recognition via multiple kernel learning of multi-scale local normal patterns. *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*, pages 2577–2580.

[328] Li, H., Sun, J., Xu, Z., and Chen, L. (2017a). Multimodal 2d+ 3d facial expression recognition with deep fusion convolutional neural network. *IEEE Transactions on Multimedia*, 19(12):2816–2831.

[329] Li, H. and Wen, G. (2019). Sample awareness-based personalized facial expression recognition. *Applied Intelligence*, pages 1–14.

[330] Li, K. and Huang, Q. (2019). Cross-pose face recognition by integrating regression iteration and interactive subspace. *EURASIP Journal on Wireless Communications and Networking*, 2019(1):105.

[331] Li, L. and Vakanski, A. (2018). Generative adversarial networks for generation and classification of physical rehabilitation movement episodes. *International journal of machine learning and computing*, 8(5):428.

[332] Li, M., Zuo, W., and Zhang, D. (2016a). Deep identity-aware transfer of facial attributes. *ArXiv*, abs/1610.05586.

[333] Li, S. and Deng, W. (2018a). Deep facial expression recognition: A survey. *arXiv preprint arXiv:1804.08348*.

[334] Li, S. and Deng, W. (2018b). Deep facial expression recognition: A survey. *CoRR*, abs/1804.08348.

[335] Li, S. and Deng, W. (2018c). Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition. *IEEE Transactions on Image Processing*, 28(1):356–370.

[336] Li, S. and Deng, W. (2019a). A deeper look at facial expression dataset bias. *arXiv preprint arXiv:1904.11150*.

[337] Li, S. and Deng, W. (2019b). Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition. *IEEE Transactions on Image Processing*, 28(1):356–370.

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

[339] Li, S., Deng, W., and Du, J. (2017c). Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2584–2593.

[340] Li, W., Abtahi, F., and Zhu, Z. (2017d). Action unit detection with region adaptation, multi-labeling learning and optimal temporal fusing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1841–1850.

[341] Li, W., Wu, G., Zhang, F., and Du, Q. (2016b). Hyperspectral image classification using deep pixel-pair features. *IEEE Transactions on Geoscience and Remote Sensing*, 55(2):844–853.

[342] Li, X., Chen, S., and Jin, Q. (2017e). Facial action units detection with multi-features and-aus fusion. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, pages 860–865. IEEE.

[343] Li, Y., Lu, Y., Li, J., and Lu, G. (2019b). Separate loss for basic and compound facial expression recognition in the wild. In Lee, W. S. and Suzuki, T., editors, *Proceedings of The Eleventh Asian Conference on Machine Learning*, volume 101 of *Proceedings of Machine Learning Research*, pages 897–911, Nagoya, Japan. PMLR.

[344] Li, Y., Wang, S., Zhao, Y., and Ji, Q. (2013). Simultaneous facial feature tracking and facial expression recognition. *IEEE Transactions on Image Processing*, 22(7):2559–2573.

[345] Li, Z. (2010). *Video-based facial expression analysis*. PhD thesis, Rutgers University-Graduate School-New Brunswick.

[346] Li, Z., Imai, J.-i., and Kaneko, M. (2009). Facial-component-based bag of words and phog descriptor for facial expression recognition. In *2009 IEEE International Conference on Systems, Man and Cybernetics*, pages 1353–1358. IEEE.

[347] Li, Z. and Luo, Y. (2017). Generate identity-preserving faces by generative adversarial networks. *arXiv preprint arXiv:1706.03227*.

[348] Liaw, A., Wiener, M., et al. (2002). Classification and regression by randomforest. *R news*, 2(3):18–22.

[349] Lienhart, R. and Maydt, J. (2002). An extended set of haar-like features for rapid object detection. In *Proceedings. international conference on image processing*, volume 1, pages I–I. IEEE.

[350] Lin, F., Hong, R., Zhou, W., and Li, H. (2018). Facial expression recognition with data augmentation and compact feature learning. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 1957–1961.

[351] Lin, F., Hong, R., Zhou, W., and Li, H. (2018). Facial expression recognition with data augmentation and compact feature learning. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 1957–1961. IEEE.

[352] Liong, S.-T., See, J., Wong, K., and Phan, R. C.-W. (2016). Automatic micro-expression recognition from long video using a single spotted apex. In *Asian Conference on Computer Vision*, pages 345–360. Springer.

[353] Liu, M., Li, S., Shan, S., Wang, R., and Chen, X. (2014a). Deeply learning deformable facial action parts model for dynamic expression analysis. In *Asian conference on computer vision*, pages 143–157. Springer.

[354] Liu, M., Shan, S., Wang, R., and Chen, X. (2014b). Learning expressionlets on spatio-temporal manifold for dynamic facial expression recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1749–1756.

[355] Liu, M., Wang, R., Li, S., Shan, S., Huang, Z., and Chen, X. (2014c). Combining multiple kernel methods on riemannian manifold for emotion recognition in the wild. In *Proceedings of the 16th International Conference on multimodal interaction*, pages 494–501. ACM.

[356] Liu, X., Vijaya Kumar, B., You, J., and Jia, P. (2017a). Adaptive deep metric learning for identity-aware facial expression recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 20–29.

[357] Liu, Y., Jourabloo, A., Ren, W., and Liu, X. (2017b). Dense face alignment. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1619–1628.

[358] Liu, Y., Peng, J., Zeng, J., and Shan, S. (2019). Pose-adaptive hierarchical attention network for facial expression recognition. *arXiv preprint arXiv:1905.10059*.

[359] Liu, Z., Luo, P., Wang, X., and Tang, X. (2015a). Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pages 3730–3738.

[360] Liu, Z., Luo, P., Wang, X., and Tang, X. (2015b). Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*.

[361] Lopes, A. T., de Aguiar, E., De Souza, A. F., and Oliveira-Santos, T. (2017). Facial expression recognition with convolutional neural networks: coping with few data and the training sample order. *Pattern Recognition*, 61:610–628.

[362] Lu, G.-m., He, J.-l., Yan, J.-j., and Li, H. (2016). Convolutional neural network for facial expression recognition. *Journal of Nanjing University of Posts and Telecommunications*, 36(1):16–22.

[363] Lu, Y., Tai, Y.-W., and Tang, C.-K. (2018). Attribute-guided face generation using conditional cyclegan. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 282–297.

[364] Lu, Y., Zhou, J., and Yu, S. (2012). A survey of face detection, extraction and recognition. *Computing and informatics*, 22(2):163–195.

[365] Lu, Z., Li, Z., Cao, J., He, R., and Sun, Z. (2017). Recent progress of face image synthesis. In *2017 4th IAPR Asian Conference on Pattern Recognition (ACPR)*, pages 7–12. IEEE.

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

[367] Lucic, M., Kurach, K., Michalski, M., Gelly, S., and Bousquet, O. (2018). Are gans created equal? a large-scale study. In *Advances in neural information processing systems*, pages 700–709.

[368] Lundqvist, D., Flykt, A., and Öhman, A. (1998). The karolinska directed emotional faces (kdef). *CD ROM from Department of Clinical Neuroscience, Psychology section, Karolinska Institutet*, 91:630.

[369] Lyons, M. J., Budynek, J., and Akamatsu, S. (1999). Automatic classification of single facial images. *IEEE transactions on pattern analysis and machine intelligence*, 21(12):1357–1362.

[370] Ma, S., Sigal, L., and Sclaroff, S. (2016). Learning activity progression in lstms for activity detection and early detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1942–1950.

[371] Maaten, L. v. d. and Hinton, G. (2008). Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605.

[372] Mahoor, M. H., Zhou, M., Veon, K. L., Mavadati, S. M., and Cohn, J. F. (2011). Facial action unit recognition with sparse representation. In *Face and Gesture 2011*, pages 336–342. IEEE.

[373] Makarenkov, V., Rokach, L., and Shapira, B. (2019). Choosing the right word: Using bidirectional lstm tagger for writing support systems. *Engineering Applications of Artificial Intelligence*, 84:1–10.

[374] Mandal, B., Lee, D., and Ouarti, N. (2016). Distinguishing posed and spontaneous smiles by facial dynamics. In *Asian Conference on Computer Vision*, pages 552–566. Springer.

[375] Mandal, B., Lee, D., and Ouarti, N. (2017). Distinguishing posed and spontaneous smiles by facial dynamics. In Chen, C.-S., Lu, J., and Ma, K.-K., editors, *Computer Vision – ACCV 2016 Workshops*, pages 552–566, Cham. Springer International Publishing.

[376] Mandel, M. I. and Ellis, D. P. (2008). Multiple-instance learning for music information retrieval.

[377] Mao, Q., Rao, Q., Yu, Y., and Dong, M. (2016). Hierarchical bayesian theme models for multipose facial expression recognition. *IEEE Transactions on Multimedia*, 19(4):861–873.

[378] Mao, Q., Rao, Q., Yu, Y., and Dong, M. (2017). Hierarchical bayesian theme models for multipose facial expression recognition. *IEEE Transactions on Multimedia*, 19(4):861–873.

[379] Martinez, B. and Valstar, M. F. (2016). Advances, challenges, and opportunities in automatic facial expression recognition. In *Advances in face detection and facial image analysis*, pages 63–100. Springer.

[380] Martinez, B., Valstar, M. F., Jiang, B., and Pantic, M. (2017). Automatic analysis of facial actions: A survey. *IEEE transactions on affective computing*.

[381] Martinez-Cantin, R. (2014). Bayesopt: A bayesian optimization library for nonlinear optimization, experimental design and bandits. *The Journal of Machine Learning Research*, 15(1):3735–3739.

[382] Marusca, L. (2014). What every body is saying. an ex-fbi agent's guide to speed-reading people. *Journal of Media Research*, 7(3):89.

[383] Mathe, E., Mitsou, A., Spyrou, E., and Mylonas, P. (2018). Arm gesture recognition using a convolutional neural network. In *2018 13th International Workshop on Semantic and Social Media Adaptation and Personalization (SMAP)*, pages 37–42. IEEE.

[384] Mavadati, S. M., Mahoor, M. H., Bartlett, K., Trinh, P., and Cohn, J. F. (2013). Disfa: A spontaneous facial action intensity database. *IEEE Transactions on Affective Computing*, 4(2):151–160.

[385] Mavani, V., Raman, S., and Miyapuram, K. P. (2017). Facial expression recognition using visual saliency and deep learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2783–2788.

[386] Mayya, V., Pai, R. M., and Pai, M. M. (2016). Automatic facial expression recognition using dcnn. *Procedia Computer Science*, 93:453–461.

[387] Mei, H., Bansal, M., and Walter, M. R. (2017). Coherent dialogue with attention-based language models. In *Thirty-First AAAI Conference on Artificial Intelligence*.

[388] Meng, Z., Liu, P., Cai, J., Han, S., and Tong, Y. (2017). Identity-aware convolutional neural network for facial expression recognition. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, pages 558–565. IEEE.

[389] Mescheder, L., Nowozin, S., and Geiger, A. (2018). Which training methods for gans do actually converge? In *International Conference on Machine Learning (ICML)*.

[390] Messer, K., Kittler, J., Sadeghi, M., Hamouz, M., Kostin, A., Cardinaux, F., Marcel, S., Bengio, S., Sanderson, C., Poh, N., et al. (2004). Face authentication test on the banca database. In *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, volume 4, pages 523–532. IEEE.

[391] Messer, K., Kittler, J., Sadeghi, M., Marcel, S., Marcel, C., Bengio, S., Cardinaux, F., Sanderson, C., Czyz, J., Vandendorpe, L., et al. (2003). Face verification competition on the xm2vts database. In *International Conference on Audio-and Video-Based Biometric Person Authentication*, pages 964–974. Springer.

[392] Mirza, M. and Osindero, S. (2014). Conditional generative adversarial nets. cite arxiv:1411.1784.

[393] Mishra, S. and Dubey, A. (2015). Face recognition approaches: a survey. *International Journal of Computing and Business Research (IJCBR)*, 6(1).

[394] Mitiche, I., Morison, G., Nesbitt, A., Hughes-Narborough, M., Stewart, B., and Boreham, P. (2018). Imaging time series for the classification of emi discharge sources. *Sensors*, 18(9):3098.

[395] Mohammad Mahoor, B. H. et al. (2017). Facial expression recognition using enhanced deep 3d convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 30–40.

[396] Mohseni, S., Kordy, H. M., and Ahmadi, R. (2013). Facial expression recognition using dct features and neural network based decision tree. In *Proceedings ELMAR-2013*, pages 361–364. IEEE.

[397] Mollahosseini, A., Chan, D., and Mahoor, M. H. (2016). Going deeper in facial expression recognition using deep neural networks. In *2016 IEEE Winter conference on applications of computer vision (WACV)*, pages 1–10. IEEE.

[398] Mollahosseini, A., Hasani, B., and Mahoor, M. H. (2017). Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*.

[399] Mortazavian, P. (2013). *Face recognition in low resolution using a 3D morphable model*. PhD thesis, PhD Thesis, University of Surrey.

[400] Moujahid, A., Abanda, A., and Dornaika, F. (2016). Feature extraction using block-based local binary pattern for face recognition. *Electronic Imaging*, 2016(10):1–6.

[401] Munasinghe, M. (2018). Facial expression recognition using facial landmarks and random forest classifier. In *2018 IEEE/ACIS 17th International Conference on Computer and Information Science (ICIS)*, pages 423–427. IEEE.

[402] Murphy-Chutorian, E. and Trivedi, M. M. (2008). Head pose estimation in computer vision: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 31(4):607–626.

[403] Naeem, M., Qureshi, I., and Azam, F. (2015). Face recognition techniques and approaches: A survey. *Science International*, 27(1).

[404] Nanni, L., Lumini, A., and Brahnam, S. (2012). Survey on lbp based texture descriptors for image classification. *Expert Systems with Applications*, 39(3):3634–3641.

[405] Nanni, L., Paci, M., Brahnam, S., Ghidoni, S., and Menegatti, E. (2013). Local phase quantization texture descriptor for protein classification. In *Proceedings of the International Conference on Bioinformatics & Computational Biology (BIOCOMP)*, page 1. The Steering Committee of The World Congress in Computer Science, Computer ....

[406] Nash, J. F. et al. (1950). Equilibrium points in n-person games. *Proceedings of the national academy of sciences*, 36(1):48–49.

[407] Nezami, O. M., Dras, M., Hamey, L., Richards, D., Wan, S., and Paris, C. (2018a). Automatic recognition of student engagement using deep learning and facial expression.

[408] Nezami, O. M., Hamey, L., Richards, D., and Dras, M. (2018b). Deep learning for domain adaption: Engagement recognition. *arXiv preprint arXiv:1808.02324*.

[409] Ng, H.-W., Nguyen, V. D., Vonikakis, V., and Winkler, S. (2015). Deep learning for emotion recognition on small datasets using transfer learning. In *Proceedings of the 2015 ACM on international conference on multimodal interaction*, pages 443–449. ACM.

[410] Nhan Duong, C., Luu, K., Gia Quach, K., and Bui, T. D. (2015). Beyond principal components: Deep boltzmann machines for face modeling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4786–4794.

[411] Nicolle, J., Bailly, K., and Chetouani, M. (2016). Real-time facial action unit intensity prediction with regularized metric learning. *Image and Vision Computing*, 52:1–14.

[412] Nie, S., Wang, Z., and Ji, Q. (2015). A generative restricted boltzmann machine based method for high-dimensional motion data modeling. *Computer Vision and Image Understanding*, 136:14–22.

[413] Nilsson, N. and Machines, L. (1965). Foundations of trainable pattern classifying systems. *McGraw-Hill, New York OBrien RM (2007) A caution regarding rules of thumb for variance ination factors. Qual Quant*, 41:673.

[414] Nonis, F., Dagnes, N., Marcolin, F., and Vezzetti, E. (2019). 3d approaches and challenges in facial expression recognition algorithms—a literature review. *Applied Sciences*, 9(18):3904.

[415] Odena, A. (2016). Semi-supervised learning with generative adversarial networks. *arXiv preprint arXiv:1606.01583*.

[416] Odena, A., Olah, C., and Shlens, J. (2017). Conditional image synthesis with auxiliary classifier gans. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 2642–2651. JMLR. org.

[417] Oh, Y.-H., See, J., Le Ngo, A. C., Phan, R. C.-W., and Baskaran, V. M. (2018a). A survey of automatic facial micro-expression analysis: Databases, methods and challenges. *Frontiers in psychology*, 9:1128.

[418] Oh, Y.-H., See, J., Le Ngo, A. C., Phan, R. C. W., and Baskaran, V. M. (2018b). A survey of automatic facial micro-expression analysis: Databases, methods, and challenges. *Frontiers in Psychology*, 9:1128.

[419] Ojala, T., Pietikäinen, M., and Mäenpää, T. (2002). Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (7):971–987.

[420] Ojansivu, V. and Heikkilä, J. (2008). Blur insensitive texture classification using local phase quantization. In *International conference on image and signal processing*, pages 236–243. Springer.

[421] Olah, C. (2015). Understanding lstm networks.

[422] Ouyang, W., Zhang, R., and kuen Cham, W. (2010). Fast pattern matching using orthogonal haar transform. *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 3050–3057.

[423] Pan, X., Shi, J., Luo, P., Wang, X., and Tang, X. (2018). Spatial as deep: Spatial cnn for traffic scene understanding. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

[424] Pantic, M. (2009). Machine analysis of facial behaviour: Naturalistic and dynamic behaviour. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1535):3505–3513.

[425] Pantic, M. and Patras, I. (2006). Dynamics of facial expression: recognition of facial actions and their temporal segments from face profile image sequences. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 36(2):433–449.

[426] Pantic, M. and Rothkrantz, L. J. (2000). Automatic analysis of facial expressions: The state of the art. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (12):1424–1445.

[427] Pantic, M. and Rothkrantz, L. J. (2001). Affect-sensitive multi-modal monitoring in ubiquitous computing: Advances and challenges. In *ICEIS (1)*, pages 466–474.

[428] Pantic, M., Valstar, M., Rademaker, R., and Maat, L. (2005). Web-based database for facial expression analysis. In *2005 IEEE international conference on multimedia and Expo*, pages 5–pp. IEEE.

[429] Parkhi, O. M., Vedaldi, A., Zisserman, A., et al. (2015). Deep face recognition. In *bmvc*, volume 1, page 6.

[430] Patel, A. and Smith, W. A. (2009). 3d morphable face models revisited. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1327–1334. IEEE.

[431] Patel, R., Rathod, N., Shah, A., and Sevak, M. (2014). Face recognition using eye distance and pca approaches. *International Journal of Computer Science and Information Technologies*, 5(1).

[432] Paysan, P., Knothe, R., Amberg, B., Romdhani, S., and Vetter, T. (2009). A 3d face model for pose and illumination invariant face recognition. In *2009 Sixth IEEE International Conference on Advanced Video and Signal Based Surveillance*, pages 296–301. Ieee.

[433] Pearson, K. (1901). Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572.

[434] Peng, X., Feris, R. S., Wang, X., and Metaxas, D. N. (2016). A recurrent encoder-decoder network for sequential face alignment. In *European conference on computer vision*, pages 38–56. Springer.

[435] Peng, X., Yu, X., Sohn, K., Metaxas, D. N., and Chandraker, M. (2017). Reconstruction-based disentanglement for pose-invariant face recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 1623–1632.

[436] Pfau, D. and Vinyals, O. (2016). Connecting generative adversarial networks and actor-critic methods. *arXiv preprint arXiv:1610.01945*.

[437] Pfister, T., Li, X., Zhao, G., and Pietikäinen, M. (2011). Recognising spontaneous facial micro-expressions. In *2011 international conference on computer vision*, pages 1449–1456. IEEE.

[438] Phillips, P. J., Grother, P., Micheals, R. J., Blackburn, D. M., Tabassi, E., and Bone, M. (2003). Face recognition vendor test 2002: overview and summary. In *IEEE Interational Workshop on Analysis and Modeling of Faces and Gestures*, page 44.

[439] Phillips, P. J., Moon, H., Rizvi, S. A., and Rauss, P. J. (2000). The feret evaluation methodology for face-recognition algorithms. *IEEE Transactions on pattern analysis and machine intelligence*, 22(10):1090–1104.

[440] Pons, G. and Masip, D. (2018). Multi-task, multi-label and multi-domain learning with residual convolutional networks for emotion recognition. *arXiv preprint arXiv:1802.06664*.

[441] Porter, S., Ten Brinke, L., and Wallace, B. (2012). Secrets and lies: Involuntary leakage in deceptive facial expressions as a function of emotional intensity. *Journal of Nonverbal Behavior*, 36(1):23–37.

[442] Pramerdorfer, C. and Kampel, M. (2016). Facial expression recognition using convolutional neural networks: state of the art. *arXiv preprint arXiv:1612.02903*.

[443] Priya Sharma, Apoorva Arora, S. P. S. G. (2014). Study of different aspects in machine learning. *International Journal of Engineering and Computer Science*, 3(05).

[444] Pumarola, A., Agudo, A., Martinez, A. M., Sanfeliu, A., and Moreno-Noguer, F. (2018). Ganimation: Anatomically-aware facial animation from a single image. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 818–833.

[445] Punitha, A. and Geetha, M. K. (2013). Hmm based real time facial expression recognition. *International Journal of Emerging Technology and Advanced Engineering*, 3(1):180–185.

[446] Qi, G.-J. (2017). Loss-sensitive generative adversarial networks on lipschitz densities. *CoRR*, abs/1701.06264.

[447] Qiao, M., Liu, L., Yu, J., Xu, C., and Tao, D. (2017). Diversified dictionaries for multi-instance learning. *Pattern Recognition*, 64:407–416.

[448] Quattoni, A., Wang, S., Morency, L.-P., Collins, M., and Darrell, T. (2007). Hidden conditional random fields. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (10):1848–1852.

[449] Radford, A., Metz, L., and Chintala, S. (2015a). Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. pages 1–16.

[450] Radford, A., Metz, L., and Chintala, S. (2015b). Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*.

[451] Radford, A., Metz, L., and Chintala, S. (2016). Unsupervised representation learning with deep convolutional generative adversarial networks. *CoRR*, abs/1511.06434.

[452] Ramadhan, W., Novianty, S. A., and Setianingsih, S. C. (2017). Sentiment analysis using multinomial logistic regression. In *2017 International Conference on Control, Electronics, Renewable Energy and Communications (ICCREC)*, pages 46–49. IEEE.

[453] Ramkumar, G. and Logashanmugam, E. (2016). An effectual facial expression recognition using hmm. In *2016 International Conference on Advanced Communication Control and Computing Technologies (ICACCCT)*, pages 12–15. IEEE.

[454] Rao, Q., Qu, X., Mao, Q., and Zhan, Y. (2015). Multi-pose facial expression recognition based on surf boosting. In *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 630–635. IEEE.

[455] Rattani, A., Chen, C., and Ross, A. (2014). Evaluation of texture descriptors for automated gender estimation from fingerprints. In *European Conference on Computer Vision*, pages 764–777. Springer.

[456] Rawlinson, T., Bhalerao, A., and Wang, L. (2010). Principles and methods for face recognition and face modelling. In *Handbook of Research on Computational Forensics, Digital Crime, and Investigation: Methods and Solutions*, pages 53–78. IGI Global.

[457] Ray, S. and Craven, M. (2005). Supervised versus multiple instance learning: An empirical comparison. In *Proceedings of the 22nd international conference on Machine learning*, pages 697–704. ACM.

[458] Ren, S., He, K., Girshick, R., and Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99.

[459] Ren, Y. (2008). Facial expression recognition system. Master's thesis, University of Waterloo.

[460] Richardson, E., Sela, M., and Kimmel, R. (2016). 3d face reconstruction by learning from synthetic data. In *2016 Fourth International Conference on 3D Vision (3DV)*, pages 460–469. IEEE.

[461] Richardson, E., Sela, M., Or-El, R., and Kimmel, R. (2017). Learning detailed face reconstruction from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1259–1268.

[462] Richoz, A.-R., Lao, J., Pascalis, O., and Caldara, R. (2018). Tracking the recognition of static and dynamic facial expressions of emotion across the life span. *Journal of vision*, 18(9):5–5.

[463] Ringeval, F., Schuller, B., Valstar, M., Jaiswal, S., Marchi, E., Lalanne, D., Cowie, R., and Pantic, M. (2015). Av+ ec 2015: The first affect recognition challenge bridging across audio, video, and physiological data. In *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge*, pages 3–8. ACM.

[464] Romdhani, S. (2005). *Face image analysis using a multiple features fitting strategy*. PhD thesis, University_of_Basel.

[465] Romdhani, S., Torr, P. H. S., Schölkopf, B., and Blake, A. (2001). Computationally efficient face detection. *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*, 2:695–700 vol.2.

[466] Romdhani, S. and Vetter, T. (2005). Estimating 3d shape and texture using pixel intensity, edges, specular highlights, texture constraints and a prior. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 2, pages 986–993. IEEE.

[467] Roth, J., Tong, Y., and Liu, X. (2015). Unconstrained 3d face reconstruction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2606–2615.

[468] Rudovic, O., Pavlovic, V., and Pantic, M. (2012). Kernel conditional ordinal random fields for temporal segmentation of facial action units. In *European Conference on Computer Vision*, pages 260–269. Springer.

[469] Ruiz-Garcia, A., Elshaw, M., Altahhan, A., and Palade, V. (2017). Stacked deep convolutional auto-encoders for emotion recognition from facial expressions. In *2017 International Joint Conference on Neural Networks (IJCNN)*, pages 1586–1593. IEEE.

[470] Sagonas, C., Antonakos, E., Tzimiropoulos, G., Zafeiriou, S., and Pantic, M. (2016). 300 faces in-the-wild challenge: Database and results. *Image and vision computing*, 47:3–18.

[471] Saha, A. and Wu, Q. J. (2010). Curvelet entropy for facial expression recognition. In *Pacific-Rim Conference on Multimedia*, pages 617–628. Springer.

[472] Sak, H., Senior, A., and Beaufays, F. (2014). Long short-term memory recurrent neural network architectures for large scale acoustic modeling. In *Fifteenth annual conference of the international speech communication association*.

[473] Salehinejad, H., Valaee, S., Dowdell, T., Colak, E., and Barfett, J. (2018). Generalization of deep neural networks for chest pathology classification in x-rays using generative adversarial networks. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 990–994. IEEE.

[474] Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., and Chen, X. (2016). Improved techniques for training gans. In *Advances in neural information processing systems*, pages 2234–2242.

[475] Samadiani, N., Huang, G., Cai, B., Luo, W., Chi, C.-H., Xiang, Y., and He, J. (2019). A review on automatic facial expression recognition systems assisted by multimodal sensor data. *Sensors*, 19(8):1863.

[476] Samara, A., Galway, L., Bond, R., and Wang, H. (2017). Affective state detection via facial expression analysis within a human–computer interaction context. *Journal of Ambient Intelligence and Humanized Computing*.

[477] Samir, C., Srivastava, A., and Daoudi, M. (2006). Three-dimensional face recognition using shapes of facial curves. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(11):1858–1863.

[478] Sánchez, E., Tzimiropoulos, G., and Valstar, M. F. (2018). Joint action unit localisation and intensity estimation through heatmap regression. *ArXiv*, abs/1805.03487.

[479] Sandbach, G., Zafeiriou, S., Pantic, M., and Yin, L. (2012). Static and dynamic 3d facial expression recognition: A comprehensive survey. *Image and Vision Computing*, 30(10):683–697.

[480] Saqur, R. and Vivona, S. (2019). Capsgan: Using dynamic routing for generative adversarial networks. In *Science and Information Conference*, pages 511–525. Springer.

[481] Saragih, J. and Goecke, R. (2007). A nonlinear discriminative approach to aam fitting. In *2007 IEEE 11th International Conference on Computer Vision*, pages 1–8. IEEE.

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

[483] Sariyanidi, E., Gunes, H., and Cavallaro, A. (2015). Automatic analysis of facial affect: A survey of registration, representation, and recognition. *IEEE transactions on pattern analysis and machine intelligence*, 37(6):1113–1133.

[484] Sariyanidi, E., Gunes, H., and Cavallaro, A. (2017). Learning bases of activity for facial expression recognition. *IEEE Transactions on Image Processing*, 26(4):1965–1978.

[485] Scovanner, P., Ali, S., and Shah, M. (2007). A 3-dimensional sift descriptor and its application to action recognition. In *Proceedings of the 15th ACM international conference on Multimedia*, pages 357–360. ACM.

[486] Seckington, M. (2011). Using dynamic bayesian networks for posed versus spontaneous facial expression recognition.

[487] Sela, M., Richardson, E., and Kimmel, R. (2017). Unrestricted facial geometry reconstruction using image-to-image translation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1576–1585.

[488] Senechal, T., McDuff, D., and Kaliouby, R. (2015). Facial action unit detection using active learning and an efficient non-linear kernel approximation. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 10–18.

[489] Senechal, T., Rapp, V., Salam, H., Seguier, R., Bailly, K., and Prevost, L. (2011). Combining aam coefficients with lgbp histograms in the multi-kernel svm framework to detect facial action units. In *Face and Gesture 2011*, pages 860–865. IEEE.

[490] Shamir, L. (2008). Evaluation of face datasets as tools for assessing the performance of face recognition methods. *International Journal of Computer Vision*, 79(3):225.

[491] Shan, C. and Braspenning, R. (2010). Recognizing facial expressions automatically from video. In *Handbook of ambient intelligence and smart environments*, pages 479–509. Springer.

[492] Shan, C., Gong, S., and McOwan, P. W. (2009). Facial expression recognition based on local binary patterns: A comprehensive study. *Image and vision Computing*, 27(6):803–816.

[493] Sharma, R. and Savakis, A. (2015a). Lean histogram of oriented gradients features for effective eye detection. *Journal of Electronic Imaging*, 24(6):063007.

[494] Sharma, R. and Savakis, A. E. (2015b). Lean histogram of oriented gradients features for effective eye detection. *J. Electronic Imaging*, 24:063007.

[495] Shbib, R. and Zhou, S. (2015). Facial expression analysis using active shape model. *International Journal of Signal Processing, Image Processing and Pattern Recognition*, 8(1):9–22.

[496] Sheerman-Chase, T. (2012). *On the Automatic Recognition of Facial Non-verbal Communication Meaning in Informal, Spontaneous Conversation*. PhD thesis, University of Surrey.

[497] Shen, F., Liu, J., and Wu, P. (2018). Double complete d-lbp with extreme learning machine auto-encoder and cascade forest for facial expression analysis. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 1947–1951. IEEE.

[498] Shin, M., Kim, M., and Kwon, D. (2016). Baseline cnn structure analysis for facial expression recognition. In *2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pages 724–729.

[499] Shrestha, A. and Mahmood, A. (2019). Review of deep learning algorithms and architectures. *IEEE Access*, 7:53040–53065.

[500] Siddiqi, M. H., Ali, R., Khan, A. M., Park, Y.-T., and Lee, S. (2015). Human facial expression recognition using stepwise linear discriminant analysis and hidden conditional random fields. *IEEE Transactions on Image Processing*, 24(4):1386–1398.

[501] Sikka, K., Dhall, A., and Bartlett, M. (2013a). Weakly supervised pain localization using multiple instance learning. In *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pages 1–8. IEEE.

[502] Sikka, K., Dykstra, K., Sathyanarayana, S., Littlewort, G., and Bartlett, M. (2013b). Multiple kernel learning for emotion recognition in the wild. In *Proceedings of the 15th ACM on International conference on multimodal interaction*, pages 517–524. ACM.

[503] Sikka, K., Sharma, G., and Bartlett, M. (2016). Lomo: Latent ordinal model for facial analysis in videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5580–5589.

[504] Sikka, K., Wu, T., Susskind, J., and Bartlett, M. (2012). Exploring bag of words architectures in the facial expression domain. In *European Conference on Computer Vision*, pages 250–259. Springer.

[505] Simon, T., Nguyen, M. H., De La Torre, F., and Cohn, J. F. (2010). Action unit detection with segment-based svms. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 2737–2744. IEEE.

[506] Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

[507] Snoek, J., Larochelle, H., and Adams, R. P. (2012). Practical bayesian optimization of machine learning algorithms. In *Advances in neural information processing systems*, pages 2951–2959.

[508] Soh, M. (2016). Learning cnn-lstm architectures for image caption generation. *Dept. Comput. Sci., Stanford Univ., Stanford, CA, USA, Tech. Rep.*

[509] Sonka, M., Hlavac, V., and Boyle, R. (2014). *Image processing, analysis, and machine vision*. Cengage Learning.

[510] Soukupova, T. and Cech, J. (2016). Eye blink detection using facial landmarks. In *21st Computer Vision Winter Workshop, Rimske Toplice, Slovenia*.

[511] Söylemez, Ö. F., Ergen, B., and Söylemez, N. H. (2017). A 3d facial expression recognition system based on svm classifier using distance based features. In *2017 25th Signal Processing and Communications Applications Conference (SIU)*, pages 1–3. IEEE.

[512] Springenberg, J. T. (2015). Unsupervised and semi-supervised learning with categorical generative adversarial networks. *CoRR*, abs/1511.06390.

[513] Spurr, A., Aksan, E., and Hilliges, O. (2017). Guiding infogan with semi-supervision. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 119–134. Springer.

[514] Srinivasan, R., Golomb, J. D., and Martinez, A. M. (2016). A neural basis of facial action recognition in humans. *Journal of Neuroscience*, 36(16):4434–4442.

[515] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958.

[516] Staal, F. C., Ponniah, A. J., Angullia, F., Ruff, C., Koudstaal, M. J., and Dunaway, D. (2015). Describing crouzon and pfeiffer syndrome based on principal component analysis. *Journal of Cranio-Maxillofacial Surgery*, 43(4):528–536.

[517] Sumathi, C., Santhanam, T., and Mahadevi, M. (2012). Automatic facial expression analysis a survey. *International Journal of Computer Science and Engineering Survey*, 3(6):47.

[518] Sun, B., Cao, S., He, J., Yu, L., and Li, L. (2017a). Automatic temporal segment detection via bilateral long short-term memory recurrent neural networks. *Journal of Electronic Imaging*, 26(2):020501.

[519] Sun, B., Li, L., Zhou, G., and He, J. (2016a). Facial expression recognition in the wild based on multimodal texture features. *Journal of Electronic Imaging*, 25(6):061407.

[520] Sun, B., Wei, Q., Li, L., Xu, Q., He, J., and Yu, L. (2016b). Lstm for dynamic emotion and group emotion recognition in the wild. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, pages 451–457. ACM.

[521] Sun, W., Zhao, H., and Jin, Z. (2017b). An efficient unconstrained facial expression recognition algorithm based on stack binarized auto-encoders and binarized neural networks. *Neurocomputing*, 267:385 – 395.

[522] Sun, W., Zhao, H., and Jin, Z. (2017c). An efficient unconstrained facial expression recognition algorithm based on stack binarized auto-encoders and binarized neural networks. *Neurocomputing*, 267:385–395.

[523] Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

[524] Szegedy, C., , , Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2015). Going deeper with convolutions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9.

[525] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826.

[526] Taherkhani, A., Cosma, G., and McGinnity, T. M. (2018). Deep-fs: A feature selection algorithm for deep boltzmann machines. *Neurocomputing*, 322:22–37.

[527] Tarhini, A. (2014). Face recognition: An introduction.

[528] Tax, D. M., Hendriks, E., Valstar, M. F., and Pantic, M. (2010). The detection of concept frames using clustering multi-instance learning. In *2010 20th International Conference on Pattern Recognition*, pages 2917–2920. IEEE.

[529] Tewari, A., Zollhoefer, M., Bernard, F., Garrido, P., Kim, H., Perez, P., and Theobalt, C. (2018a). High-fidelity monocular face reconstruction based on an unsupervised model-based face autoencoder. *IEEE transactions on pattern analysis and machine intelligence*.

[530] Tewari, A., Zollhöfer, M., Garrido, P., Bernard, F., Kim, H., Pérez, P., and Theobalt, C. (2018b). Self-supervised multi-level face model learning for monocular reconstruction at over 250 hz. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2549–2559.

[531] Tewari, A., Zollhofer, M., Kim, H., Garrido, P., Bernard, F., Perez, P., and Theobalt, C. (2017). Mofa: Model-based deep convolutional face autoencoder for unsupervised monocular reconstruction. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1274–1283.

[532] Thies, J., Zollhofer, M., Stamminger, M., Theobalt, C., and Nießner, M. (2016). Face2face: Real-time face capture and reenactment of rgb videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2387–2395.

[533] Thies, J., Zollhöfer, M., Stamminger, M., Theobalt, C., and Nießner, M. (2018). Facevr: Real-time gaze-aware facial reenactment in virtual reality. *ACM Transactions on Graphics 2018 (TOG)*.

[534] Thomas, L. A., De Bellis, M. D., Graham, R., and LaBar, K. S. (2007). Development of emotional facial recognition in late childhood and adolescence. *Developmental science*, 10(5):547–558.

[535] Tian, Y.-I., Kanade, T., and Cohn, J. F. (2001). Recognizing action units for facial expression analysis. *IEEE Transactions on pattern analysis and machine intelligence*, 23(2):97–115.

[536] Tian, Y.-L., Kanade, T., and Cohn, J. F. (2005a). Facial expression analysis. In *Handbook of face recognition*, pages 247–275. Springer.

[537] Tian, Y.-L., Kanade, T., and Cohn, J. F. (2005b). *Facial Expression Analysis*, pages 247–275. Springer New York, New York, NY.

[538] Tie, Y. and Guan, L. (2013). Automatic landmark point detection and tracking for human facial expressions. *EURASIP Journal on Image and Video Processing*, 2013(1):8.

[539] Tobji, R., Di, W., and Ayoub, N. (2019). Fmnet: Iris segmentation and recognition by using fully and multi-scale cnn for biometric security. *Applied Sciences*, 9(10):2042.

[540] TOME, D. (2015). Convolutional neural network based method for pedestrian detection.

[541] Topi, M., Timo, O., Matti, P., and Maricor, S. (2000). Robust texture classification by subsets of local binary patterns. In *Proceedings 15th International Conference on Pattern Recognition. ICPR-2000*, volume 3, pages 935–938. IEEE.

[542] Tran, L., Liu, F., and Liu, X. (2019). Towards high-fidelity nonlinear 3d face morphable model. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1126–1135.

[543] Tran, L. and Liu, X. (2018). Nonlinear 3d face morphable model. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7346–7355.

[544] Tran, L. and Liu, X. (2019). On learning 3d face morphable model from in-the-wild images. *IEEE transactions on pattern analysis and machine intelligence*.

[545] Tran, L., Yin, X., and Liu, X. (2017). Disentangled representation learning gan for pose-invariant face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1415–1424.

[546] Tu, X., Zhao, J., Jiang, Z., Luo, Y., Xie, M., Zhao, Y., He, L., Ma, Z., and Feng, J. (2019). Joint 3d face reconstruction and dense face alignment from a single image with 2d-assisted self-supervised learning. *arXiv preprint arXiv:1903.09359*.

[547] Tu, Y., Lin, Y., Wang, J., and Kim, J.-U. (2018). Semi-supervised learning with generative adversarial networks on digital signal modulation classification. *Comput. Mater. Continua*, 55(2):243–254.

[548] Tuan Tran, A., Hassner, T., Masi, I., and Medioni, G. (2017). Regressing robust and discriminative 3d morphable models with a very deep neural network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5163–5172.

[549] Tzimiropoulos, G. and Pantic, M. (2013). Optimization problems for fast aam fitting in-the-wild. In *Proceedings of the IEEE international conference on computer vision*, pages 593–600.

[550] Ullah, A., Ahmad, J., Muhammad, K., Sajjad, M., and Baik, S. W. (2018). Action recognition in video sequences using deep bi-directional lstm with cnn features. *IEEE Access*, 6:1155–1166.

[551] Ur-Rahman, N. and Harding, J. A. (2012). Textual data mining for industrial knowledge management and text classification: A business oriented approach. *Expert Systems with Applications*, 39(5):4729–4739.

[552] Vadapalli, H. (2014). Facial action unit recognition from video streams with recurrent neural networks. *International Journal of Computer Applications*, 96(19).

[553] Valstar, M. and Pantic, M. (2006a). Biologically vs. logic inspired encoding of facial actions and emotions in video. volume 2006, pages 325–328.

[554] Valstar, M. and Pantic, M. (2006b). Fully automatic facial action unit detection and temporal analysis. In *2006 Conference on Computer Vision and Pattern Recognition Workshop (CVPRW'06)*, pages 149–149. IEEE.

[555] Valstar, M. F. (2008). Timing is everything: A spatio-temporal approach to the analysis of facial actions.

[556] Valstar, M. F., Almaev, T., Girard, J. M., McKeown, G., Mehu, M., Yin, L., Pantic, M., and Cohn, J. F. (2015). Fera 2015-second facial expression recognition and analysis challenge. In *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, volume 6, pages 1–8. IEEE.

[557] Valstar, M. F., Jiang, B., Mehu, M., Pantic, M., and Scherer, K. (2011). The first facial expression recognition and analysis challenge. In *Face and Gesture 2011*, pages 921–926. IEEE.

[558] Valstar, M. F. and Pantic, M. (2006c). Biologically vs. logic inspired encoding of facial actions and emotions in video. In *2006 IEEE International Conference on Multimedia and Expo*, pages 325–328. IEEE.

[559] Valstar, M. F. and Pantic, M. (2007). Combined support vector machines and hidden markov models for modeling facial action temporal dynamics. In *International workshop on human-computer interaction*, pages 118–127. Springer.

[560] Valstar, M. F., Patras, I., and Pantic, M. (2005). Facial action unit detection using probabilistic actively learned support vector machines on tracked facial point data. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)-Workshops*, pages 76–76. IEEE.

[561] van der Maaten, L. and Hendriks, E. (2012). Action unit classification using active appearance models and conditional random fields. *Cognitive processing*, 13(2):507–518.

[562] Vapnik, V. and Vapnik, V. (1998). Statistical learning theory wiley. *New York*, pages 156–160.

[563] Vedaldi, A. and Lenc, K. (2015). Matconvnet: Convolutional neural networks for matlab. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 689–692. ACM.

[564] Vielzeuf, V., Kervadec, C., Pateux, S., Lechervy, A., and Jurie, F. (2018). An occam's razor view on learning audiovisual emotion recognition with small training sets. In *Proceedings of the 2018 on International Conference on Multimodal Interaction*, pages 589–593. ACM.

[565] Viola, P., Jones, M., et al. (2001). Rapid object detection using a boosted cascade of simple features. *CVPR (1)*, 1(511-518):3.

[566] Viola, P. and Jones, M. J. (2004). Robust real-time face detection. *International journal of computer vision*, 57(2):137–154.

[567] Vo, D. M. and Le, T. H. (2016). Deep generic features and svm for facial expression recognition. In *2016 3rd National Foundation for Science and Technology Development Conference on Information and Computer Science (NICS)*, pages 80–84. IEEE.

[568] Walecki, R., Rudovic, O., Pavlovic, V., and Pantic, M. (2015). Variable-state latent conditional random fields for facial expression recognition and action unit detection. In *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, volume 1, pages 1–8. IEEE.

[569] Walecki, R., Rudovic, O., Pavlovic, V., and Pantic, M. (2017). Variable-state latent conditional random field models for facial expression analysis. *Image and Vision Computing*, 58:25–37.

[570] Wang, D., Fu, R., and Luo, Z. (2017). Classroom attendance auto-management based on deep learning. In *2017 2nd International Conference on Education, Sports, Arts and Management Engineering (ICESAME 2017)*. Atlantis Press.

[571] Wang, J., Li, X., and Yang, J. (2018a). Stacked conditional generative adversarial networks for jointly learning shadow detection and shadow removal. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1788–1797.

[572] Wang, L., Li, R.-F., Wang, K., and Chen, J. (2014). Feature representation for facial expression recognition based on facs and lbp. *International Journal of Automation and Computing*, 11(5):459–468.

[573] Wang, X., Li, W., Mu, G., Huang, D., and Wang, Y. (2018b). Facial expression synthesis by u-net conditional generative adversarial networks. In *Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval*, pages 283–290. ACM.

[574] Wang, X., Wang, X., and Ni, Y. (2018c). Unsupervised domain adaptation for facial expression recognition using generative adversarial networks. *Computational intelligence and neuroscience*, 2018.

[575] Wang, X., Wang, Y., and Li, W. (2019). U-net conditional gans for photo-realistic and identity-preserving facial expression synthesis. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 15(3s):1–23.

[576] Wang, Y., Ai, H., Wu, B., and Huang, C. (2004). Real time facial expression recognition with adaboost. In *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, volume 3, pages 926–929. IEEE.

[577] Wang, Y., Huang, M., Zhu, X., and Zhao, L. (2016a). Attention-based lstm for aspect-level sentiment classification. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 606–615.

[578] Wang, Y., Neves, L., and Metze, F. (2016b). Audio-based multimedia event detection using deep recurrent neural networks. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2742–2746. IEEE.

[579] Wang, Y.-Q. (2014). An analysis of the viola-jones face detection algorithm. *Image Processing On Line*, 4:128–148.

[580] Wang, Z., Wang, S., and Ji, Q. (2013). Capturing complex spatio-temporal relations among facial muscles for facial expression recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3422–3429.

[581] Wei, W., Jia, Q., and Chen, G. (2016). Real-time facial expression recognition for affective computing based on kinect. In *2016 IEEE 11th Conference on Industrial Electronics and Applications (ICIEA)*, pages 161–165. IEEE.

[582] Wen, G., Hou, Z., Li, H., Li, D., Jiang, L., and Xun, E. (2017). Ensemble of deep neural networks with probability-based fusion for facial expression recognition. *Cognitive Computation*, 9(5):597–610.

[583] Weyrauch, B., Heisele, B., Huang, J., and Blanz, V. (2004). Component-based face recognition with 3d morphable models. In *2004 Conference on Computer Vision and Pattern Recognition Workshop*, pages 85–85. IEEE.

[584] Wiskott, L. and Sejnowski, T. J. (2002). Slow feature analysis: Unsupervised learning of invariances. *Neural computation*, 14(4):715–770.

[585] Wolf, L., Hassner, T., and Maoz, I. (2011). *Face recognition in unconstrained videos with matched background similarity*. IEEE.

[586] Wu, B.-F. and Lin, C.-H. (2018). Adaptive feature mapping for customizing deep learning based faci-al expression recognition model. *IEEE access*, 6:12451–12461.

[587] Wu, J., Huang, Z., Thoma, J., Acharya, D., and Van Gool, L. (2018). Wasserstein divergence for gans. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 653–668.

[588] Wu, J., Yu, Y., Huang, C., and Yu, K. (2015). Deep multiple instance learning for image classification and auto-annotation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3460–3469.

[589] Wu, T., Bartlett, M. S., and Movellan, J. R. (2010). Facial expression recognition using gabor motion energy filters. In *2010 IEEE computer society conference on computer vision and pattern recognition-workshops*, pages 42–47. IEEE.

[590] Wu, Y. and Ji, Q. (2016). Constrained joint cascade regression framework for simultaneous facial action unit recognition and facial landmark detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3400–3408.

[591] Wu, Y., Wang, Z., and Ji, Q. (2013). Facial feature tracking under varying facial expressions and face poses based on restricted boltzmann machines. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3452–3459.

[592] Xia, L. (2014). Facial expression recognition based on svm. In *2014 7th International Conference on Intelligent Computation Technology and Automation*, pages 256–259. IEEE.

[593] Xia, W., Yin, S., and Ouyang, P. (2013). A high precision feature based on lbp and gabor theory for face recognition. *Sensors*, 13(4):4499–4513.

[594] Xiao, C., Choi, E., and Sun, J. (2018). Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review. *Journal of the American Medical Informatics Association*, 25(10):1419–1428.

[595] Xiong, X. and De la Torre, F. (2013). Supervised descent method and its applications to face alignment. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 532–539.

[596] Yaddaden, Y., Adda, M., Bouzouane, A., Gaboury, S., and Bouchard, B. (2018). User action and facial expression recognition for error detection system in an ambient assisted environment. *Expert Systems with Applications*, 112:173–189.

[597] Yan, Q. and Wang, W. (2017). Dcgans for image super-resolution, denoising and debluring. *Adv. Neural Inf. Process. Syst.*, 8:487–495.

[598] Yan, S., Zhu, X., Liu, G., and Wu, J. (2017). Sparse multiple instance learning as document classification. *Multimedia tools and applications*, 76(3):4553–4570.

[599] Yan, W.-J., Wu, Q., Liang, J., Chen, Y.-H., and Fu, X. (2013). How fast are the leaked facial expressions: The duration of micro-expressions. *Journal of Nonverbal Behavior*, 37(4):217–230.

[600] Yang, B., Cao, J., Ni, R., and Zhang, Y. (2017a). Facial expression recognition using weighted mixture deep neural network based on double-channel facial images. *IEEE Access*, 6:4630–4640.

[601] Yang, B., Yan, J., Lei, Z., and Li, S. Z. (2015). Fine-grained evaluation on face detection in the wild. In *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, volume 1, pages 1–7. IEEE.

[602] Yang, J. (2005). Review of multi-instance learning and its applications. *Technical report, School of Computer Science Carnegie Mellon University*.

[603] Yang, J. (2008). Mill: A multiple instance learning library.

[604] Yang, J., Franco, J., Hétroy-Wheeler, F., Wuhrer, S., and Bogo, F. (2018a). Kanazawa, a., black, mj, jacobs, dw, malik, j.: End-to-end recovery of human shape and pose. in: Cvpr (2018) loper, m., mahmood, n., romero, j., pons-moll, g., black, m.: Smpl: a skinned multi-person linear model. in: Siggraph (2015) krizhevsky, a., sutskever, i., hinton, ge: Imagenet classification with deep con. In *Computer Vision–ECCV 2018: 15th European Conference, Munich, Germany, September 8–14, 2018, Proceedings*, volume 11211, page 35. Springer.

[605] Yang, J., Kannan, A., Batra, D., and Parikh, D. (2017b). Lr-gan: Layered recursive generative adversarial networks for image generation. *arXiv preprint arXiv:1703.01560*.

[606] Yang, M.-H., Kriegman, D. J., and Ahuja, N. (2002). Detecting faces in images: A survey. *IEEE Transactions on pattern analysis and machine intelligence*, 24(1):34–58.

[607] Yang, Y., Zhou, J., Ai, J., Bin, Y., Hanjalic, A., Shen, H. T., and Ji, Y. (2018b). Video captioning by adversarial lstm. *IEEE Transactions on Image Processing*, 27(11):5600–5611.

[608] Yao, A., Shao, J., Ma, N., and Chen, Y. (2015). Capturing au-aware facial features and their latent relations for emotion recognition in the wild. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, ICMI '15, pages 451–458, New York, NY, USA. ACM.

[609] Yin, X., Yu, X., Sohn, K., Liu, X., and Chandraker, M. (2017). Towards large-pose face frontalization in the wild. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3990–3999.

[610] Yu, Z. and Zhang, C. (2015a). Image based static facial expression recognition with multiple deep network learning. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, pages 435–442. ACM.

[611] Yu, Z. and Zhang, C. (2015b). Image based static facial expression recognition with multiple deep network learning. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, ICMI '15, pages 435–442, New York, NY, USA. ACM.

[612] Zafeiriou, L., Antonakos, E., Zafeiriou, S., and Pantic, M. (2014). Joint unsupervised face alignment and behaviour analysis. In *European Conference on Computer Vision*, pages 167–183. Springer.

[613] Zafeiriou, L., Nicolaou, M. A., Zafeiriou, S., Nikitidis, S., and Pantic, M. (2013). Learning slow features for behaviour analysis. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2840–2847.

[614] Zafeiriou, L., Nicolaou, M. A., Zafeiriou, S., Nikitidis, S., and Pantic, M. (2015a). Probabilistic slow features for behavior analysis. *IEEE transactions on neural networks and learning systems*, 27(5):1034–1048.

[615] Zafeiriou, L., Nicolaou, M. A., Zafeiriou, S., Nikitidis, S., and Pantic, M. (2016a). Probabilistic slow features for behavior analysis. *IEEE transactions on neural networks and learning systems*, 27(5):1034–1048.

[616] Zafeiriou, L., Zafeiriou, S., and Pantic, M. (2017a). Deep analysis of facial behavioral dynamics. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 42–50.

[617] Zafeiriou, S., Chrysos, G. G., Roussos, A., Ververas, E., Deng, J., and Trigeorgis, G. (2017b). The 3d menpo facial landmark tracking challenge. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2503–2511.

[618] Zafeiriou, S., Papaioannou, A., Kotsia, I., Nicolaou, M., and Zhao, G. (2016b). Facial affect“in-the-wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 36–47.

[619] Zafeiriou, S. and Petrou, M. (2010). Sparse representations for facial expressions recognition via l 1 optimization. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops*, pages 32–39. IEEE.

[620] Zafeiriou, S., Zhang, C., and Zhang, Z. (2015b). A survey on face detection in the wild: past, present and future. *Computer Vision and Image Understanding*, 138:1–24.

[621] Zavarez, M. V., Berriel, R. F., and Oliveira-Santos, T. (2017). Cross-database facial expression recognition based on fine-tuned deep convolutional network. In *2017 30th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, pages 405–412.

[622] Zavarez, M. V., Berriel, R. F., and Oliveira-Santos, T. (2017). Cross-database facial expression recognition based on fine-tuned deep convolutional network. In *2017 30th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, pages 405–412. IEEE.

[623] Zeid Baker, M. (2018). Generation of synthetic images with generative adversarial networks.

[624] Zeng, J., Chu, W.-S., De la Torre, F., Cohn, J. F., and Xiong, Z. (2015). Confidence preserving machine for facial action unit detection. In *Proceedings of the IEEE international conference on computer vision*, pages 3622–3630.

[625] Zeng, J., Shan, S., and Chen, X. (2018). Facial expression recognition with inconsistently annotated datasets. In *Proceedings of the European conference on computer vision (ECCV)*, pages 222–237.

[626] Zeng, Z., Pantic, M., Roisman, G. I., and Huang, T. S. (2008). A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE transactions on pattern analysis and machine intelligence*, 31(1):39–58.

[627] Zeng, Z., Pantic, M., Roisman, G. I., and Huang, T. S. (2009). A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE transactions on pattern analysis and machine intelligence*, 31(1):39–58.

[628] Zhang, B. (2018). Foreign exchange rates forecasting with an emd-lstm neural networks model. In *Journal of Physics: Conference Series*, volume 1053, page 012005. IOP Publishing.

[629] Zhang, C., Platt, J. C., and Viola, P. A. (2006). Multiple instance boosting for object detection. In *Advances in neural information processing systems*, pages 1417–1424.

[630] Zhang, F., Zhang, T., Mao, Q., and Xu, C. (2018a). Joint pose and expression modeling for facial expression recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3359–3368.

[631] Zhang, F., Zhang, T., Mao, Q., and Xu, C. (2018b). Joint pose and expression modeling for facial expression recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[632] Zhang, H., Xu, T., Li, H., Zhang, S., Wang, X., Huang, X., and Metaxas, D. N. (2018c). Stackgan++: Realistic image synthesis with stacked generative adversarial networks. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):1947–1962.

[633] Zhang, J.-B., Sun, Y.-X., and Zhan, D.-C. (2017a). Multiple-instance learning for text categorization based on semantic representation. *Big Data & Information Analytics*, 2(1):69–75.

[634] Zhang, K., Huang, Y., Du, Y., and Wang, L. (2017b). Facial expression recognition based on deep evolutional spatial-temporal networks. *IEEE Transactions on Image Processing*, 26(9):4193–4203.

[635] Zhang, K., Zhang, Z., Li, Z., and Qiao, Y. (2016). Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503.

[636] Zhang, L. and Tjondronegoro, D. (2011). Facial expression recognition using facial movement features. *IEEE Transactions on Affective Computing*, 2(4):219–229.

[637] Zhang, M., Fu, Q., Chen, Y.-H., and Fu, X. (2014). Emotional context influences micro-expression recognition. *PloS one*, 9(4):e95018.

[638] Zhang, M., Teck Ma, K., Hwee Lim, J., Zhao, Q., and Feng, J. (2017c). Deep future gaze: Gaze anticipation on egocentric videos using adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4372–4381.

[639] Zhang, Q. and Goldman, S. A. (2002). Em-dd: An improved multiple-instance learning technique. In Dietterich, T. G., Becker, S., and Ghahramani, Z., editors, *Advances in Neural Information Processing Systems 14*, pages 1073–1080. MIT Press.

[640] Zhang, X. (2015). Facial expression analysis via transfer learning.

[641] Zhang, X. and Mahoor, M. H. (2016). Task-dependent multi-task multiple kernel learning for facial action unit detection. *Pattern Recognition*, 51:187–196.

[642] Zhang, X., Qu, S., Huang, J., Fang, B., and Yu, P. (2018d). Stock market prediction via multi-source multiple instance learning. *IEEE Access*, 6:50720–50728.

[643] Zhang, Z., Luo, P., Loy, C. C., and Tang, X. (2018e). From facial expression recognition to interpersonal relation prediction. *International Journal of Computer Vision*, 126(5):550–569.

[644] Zhao, G. and Li, X. (2019). Automatic micro-expression analysis: Open challenges. *Frontiers in psychology*, 10:1833.

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

[646] Zhao, S., Rudzicz, F., Carvalho, L. G., Márquez-Chin, C., and Livingstone, S. (2014). Automatic detection of expressed emotion in parkinson's disease. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4813–4817. IEEE.

[647] Zhao, X. and Zhang, S. (2011). Facial expression recognition based on local binary patterns and kernel discriminant isomap. *Sensors*, 11(10):9573–9588.

[648] Zhao, Z., Fu, G., Liu, S., Elokely, K. M., Doerksen, R. J., Chen, Y., and Wilkins, D. E. (2013). Drug activity prediction using multiple-instance learning via joint instance and feature selection. In *BMC bioinformatics*, volume 14, page S16. BioMed Central.

[649] Zheng, Z., Zheng, L., and Yang, Y. (2017). Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3754–3762.

[650] Zhi, R., Flierl, M., Ruan, Q., and Kleijn, W. B. (2011). Graph-preserving sparse nonnegative matrix factorization with application to facial expression recognition. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 41(1):38–52.

[651] Zhong, L., Liu, Q., Yang, P., Liu, B., Huang, J., and Metaxas, D. N. (2012). Learning active facial patches for expression analysis. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2562–2569. IEEE.

[652] Zhou, H., Sun, J., Yacoob, Y., and Jacobs, D. W. (2018). Label denoising adversarial network (ldan) for inverse lighting of faces. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6238–6247.

[653] Zhou, S.-R., Yin, J.-P., and Zhang, J.-M. (2013). Local binary pattern (lbp) and local phase quantization (lbq) based on gabor filter for face representation. *Neurocomputing*, 116:260–264.

[654] Zhou, X., Dong, D., Wu, H., Zhao, S., Yu, D., Tian, H., Liu, X., and Yan, R. (2016). Multi-view response selection for human-computer conversation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 372–381.

[655] Zhou, Y. and Shi, B. E. (2017). Photorealistic facial expression synthesis by the conditional difference adversarial autoencoder. In *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 370–376. IEEE.

[656] Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A. (2017a). Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232.

[657] Zhu, M., Cao, Z., Xiao, Y., and Xie, X. (2015a). Beyond local phase quantization: Mid-level blurred image representation using fisher vector. In *2015 IEEE International Conference on Image Processing (ICIP)*, pages 1890–1894. IEEE.

[658] Zhu, X., Lei, Z., Liu, X., Shi, H., and Li, S. Z. (2016). Face alignment across large poses: A 3d solution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 146–155.

[659] Zhu, X., Lei, Z., Yan, J., Yi, D., and Li, S. Z. (2015b). High-fidelity pose and expression normalization for face recognition in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 787–796.

[660] Zhu, X., Liu, X., Lei, Z., and Li, S. Z. (2017b). Face alignment in full pose range: A 3d total solution. *IEEE transactions on pattern analysis and machine intelligence*, 41(1):78–92.

[661] Zhu, X. and Ramanan, D. (2012). Face detection, pose estimation, and landmark localization in the wild. In *2012 IEEE conference on computer vision and pattern recognition*, pages 2879–2886. IEEE.

[662] Zollhöfer, M., Thies, J., Garrido, P., Bradley, D., Beeler, T., Pérez, P., Stamminger, M., Nießner, M., and Theobalt, C. (2018). State of the art on monocular 3d face reconstruction, tracking, and applications. In *Computer Graphics Forum*, volume 37, pages 523–550. Wiley Online Library.