FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO



Deep Neural Networks for Face-based Emotion Recognition

Pedro Duarte da Cunha Nunes Lopes

Mestrado em Bioengenharia

Supervisor: Ana Filipa Sequeira

Co-Supervisor: Jaime dos Santos Cardoso

Co-Supervisor: João Ribeiro Pinto

July 29, 2022

© Pedro Duarte Lopes, 2022

Deep Neural Networks for Face-based Emotion Recognition

Pedro Duarte da Cunha Nunes Lopes

Mestrado em Bioengenharia

Aprovado em provas públicas pelo Júri:

Presidente: Ana Maria Rodrigues de Sousa Faria de Mendonça

Arguente: João Monteiro

Vogal: Ana Filipa Sequeira

July 29, 2022

Abstract

Emotions are a very complex concept; they are natural, uncontrollable, and most of the time we reveal what we are feeling without even noticing. This uncanny feature, intrinsic to emotions, is what makes them so important. It is through them that humans can understand each other and communicate without needing to use words. From all the ways emotions are shown, facial expressions are not only the ones which give out more information but also are the most consistent. Moreover, facial expressions transcend the individual subject, making them easier to recognize and associate with a specific emotional state.

In the last century, facial expression have been the focus of multiple studies leading to the development of facial expression recognition methods (FER). These methods have reached surprisingly accurate performance regarding the detection of a small set of emotions - Ekman's basic emotions. The early approaches comprised image processing handcrafted features paired up with machine learning classifiers, but in recent years, with the growth of computer vision, deep learning models have become more widespread and are now the state of the art.

In the present work, we aimed to comprehensively study different approaches in facial expression recognition, firstly, focusing on the use of deep learning methods, and secondly, trying to prove if minimalistic representations based on facial landmarks are a reliable method to recognize human emotions. The proposed work was evaluated using three databases: CK+, JAFFE and SFEW.

As a starting point, a preliminary work was developed using traditional approaches based on handcrafted features: Local Binary Pattern (LBP) and Histogram Oriented Gradient (HOG). The classifiers used for these first methods were Support Vector Machines (SVM), regarded as one of the best traditional machine learning classifiers, and a Multi-Layer Perceptron (MLP), to see how a basic Neural network classifier would perform with handcrafted features.

The following step was an extensive sample modification. Using 468 landmark coordinates three different types of samples were developed. Region of Interest (ROI) Mask was the first approach and it can be described as the application of a binary mask to an original sample, creating a new image where only facial areas of interest are highlighted. Only Landmark (OL) was the second method and originated new samples built by drawing small circles in each one of the 468 key-points positions. At last, Emotion Silhouette (ES) samples correspond to new images made by drawing some facial attributes like eyes and mouth onto black background, creating a different facial expression representation.

The next stage comprised the development of three different Convolutional Neural Networks (CNN), two of them trained from scratch and one built using transfer learning techniques. These models were evaluated with original samples and with the developed landmark-inspired samples. The results have shown that transfer learning models are the best approach when dealing with smaller databases like JAFFE. Moreover, both the developed CNNs were able to reach state-of-the-art performances in both CK+ and SFEW databases. Regarding the landmarks alterations: ROI Mask showed that the eyes, nose and mouth are the most prominent regions to correctly

classify an emotion, but are not enough to reach as good of performances as using the whole face; Although a human face can be accurately represented only its landmarks, OL approach showed that that is not enough to capture all the necessary details to reach state-of-the-art results; Emotion Silhouette (ES) method demonstrated that it is possible to create a minimalist representation of human emotion and still reach good results.

Finally, an attention-based model was built using a Visual Transformer (ViT). Once again, based on the premise of certain areas having more prevalence than others in the final classification, this self-attention model was developed and trained using only the original samples. This model shows great potential when dealing with smaller datasets and using much quicker training easily reached state-of-the-art performances even in more challenging databases.

Thus, the main conclusions of this dissertation were: handcrafted feature extraction is still very reliable reaching high accuracy values when faced with smaller and more imbalanced databases; the tested CNNs proved to be powerful tools when trained with a multitude of different samples; landmarks modifications shown that, however, emotions are dynamic and depend more on the connection between all the elements, than on the elements themselves, we were able to produce simultaneously accurate and minimalist representations of those emotions. Future work should continue to dwell on the potential of the landmark-based approach and focus on developing representations capable of comprising all the necessary traits which define a human emotion. Also, the development of attention based models even more complex can be the future of facial expression recognition models.

Resumo

As emoções são um conceito muito complexo; são naturais, incontroláveis e a maior parte das vezes, revelamos o que estamos a sentir sem nos apercebermos. Esta característica única, intrínseca às emoções, é o que as torna tão importantes. É através das emoções que os humanos conseguem compreender-se e comunicar sem o uso de palavras. De todas as formas que as emoções são demonstradas, as expressões faciais são aquelas que não só nos dão mais informação mas também as mais consistentes. Além disso, expressões faciais ultrapassam o indivíduo, tornando-se mais fácil de as reconhecer e associa-las a um estado emocional específico.

Durante o ultimo século, as expressões faciais foram o foco de múltiplos estudos, levando ao desenvolvimento de mutiplos métodos de reconhecimento de expressão faciais. Estes métodos alcançaram um desempenho surpreendentemente preciso no que diz respeito à deteção de um pequeno conjunto de emoções - as emoções básicas de Ekman. As primeiras abordagens recorrem ao uso de processamento de imagem em conjunto com classificadores de aprendizagem automática. Contudo, nos últimos anos, aliado ao crescimento da área da visão computacional, modelos de *Deep Learning* tornaram-se cada vez mais famosos e são agora o estado da arte.

No presente trabalho, o nosso objetivo é estudar de forma abrangente diferentes abordagens no que toca ao reconhecimento de expressões faciais. Em primeiro lugar, focando na utilização de métodos de *Deep Learning* e, em segundo lugar, tentando provar se representações minimalistas baseadas em marcas faciais são um método fiável de reconhecimento de emoções humanas. O trabalho proposto foi avaliado com base em três bases de dados: CK+, JAFFE e SFEW.

Como ponto de partida, foi desenvolvido um trabalho preliminar utilizando abordagens tradicionais baseadas em características manualmente extraidas: LBP e HOG. Os classificadores utilizados para estes primeiros métodos foram *Support Vector Machines* (SVM), considerados como um dos melhores classificadores dentro de *Machine Learning* e um *Multi-Layer Perceptron* (MLP), para verificar como um classificador básico de redes neuronais se comportaria face ao uso de caracteristicas manualmente extraídas.

O passo seguinte foi uma extensa modificação das amostras. Utilizando 468 coordenadas de referência, foram desenvolvidas três espécies diferentes de amostras. *ROI Mask* foi a primeira abordagem e pode ser descrita como a aplicação de uma máscara binária às amostras originais, criando uma nova imagem onde apenas algumas regiões faciais são reveladas. *Only Landmark (OL)* foi o segundo método e originou novas amostras construídas a partir de pequenos círculos desenhados em cada uma dos 468 dos pontos fiduciais. Por fim, o método *Emotion Silhouette (ES)* corresponde à criação de novas imagens feitas ao desenhar alguns atributos faciais, tais como olhos e boca, num fundo negro, criando uma representação minimalista de uma expressão facial.

A fase seguinte inclui o desenvolvimento de três *Convolutional Neural Networks* (CNN) diferentes, duas delas treinadas desde o início e uma construída usando técnicas de *transfer learning*. Estes modelos foram avaliados com amostras originais e com as amostras inspiradas em marcos faciais.

Os resultados mostraram que os modelos baseados em *transfer learning* são a melhor abordagem quando se trata de bases de dados mais pequenas como a JAFFE. Além disso, ambas as CNNs desenvolvidas foram capazes de alcançar desempenhos ao nivel do estado-da-art, tanto nas bases de dados CK+ como na SFEW. Relativamente às alterações dos marcos faciais: a *ROI Mask* mostrou que os olhos, nariz e boca são as regiões mais proeminentes para corretamente classificar uma emoção, mas não são suficientes para alcançar resultados tão bons como aqueles obitdos utilizando o rosto todo. Embora um rosto humano possa ser representado com precisão apenas utilizando marcos faciais, a abordagem *OL* mostrou que isso não é suficiente para capturar todos os detalhes necessários para alcançar bons resultados. O método *Emotion Silhouette (ES)* demonstrou que é possível criar uma representação minimalista da emoção humana e ainda alcançar bons resultados.

Por fim, um modelo baseado em atenção foi construído usando um Visual Transformer (ViT). Mais uma vez, com base na premissa de que certas áreas tem maior prevalência do que outras para classificação final, este modelo de auto-atenção foi desenvolvido e treinado usando apenas as amostras originais. Este modelo demonstra um enorme potencial ao lidar com conjuntos de dados mais pequenos e, usando treinos muito mais rápidos, facilmente alcançou desempenhos de última geração mesmo em bases de dados mais desafiantes.

Assim, as principais conclusões desta dissertação foram: o uso de metodos tradicionais de *machine learning* é ainda uma forma muito fiável de alcançar valores de elevada precisão mesmo quando confrontados com bases de dados mais pequenas e mais desequilibradas; as CNNs testadas revelaram ser ferramentas poderosas quando treinadas com uma multiplicidade de amostras diferentes; em relação às imagens densevolvidas baseadas em marcos faciais mostraram que, embora as emoções sejam dinâmicas e dependam mais da ligação entre todos os seus elementos, do que dos próprios elementos, é possivel produzir simultaneamente representações rigorosas e minimalistas dessas emoções. Os trabalhos futuros devem continuar a explorar no potencial de abordagens baseadas em marcos faciais e concentrar-se no desenvolvimento de representações capazes de incluir todos os traços necessários que definem uma emoção humana. Além disso, o desenvolvimento de modelos de atenção ainda mais complexos pode ser o futuro no que toca a modelos de reconhecimento de expressões faciais.

Agradecimentos

Em primeiro lugar queria deixar um forte agradecimento à Ana Filipa Sequeira e ao João Pinto por todo o apoio e disponibilidade que me deram ao longo destes últimos meses. Obrigado pela boa disposição em todas as reuniões, obrigado por todas as dicas sempre certeiras e obrigado pela compreensão, mesmo quando não conseguia ir às reuniões. Também uma palavra de apreço ao Professor Jaime que, mesmo menos presente, sempre se mostrou disponivel para me ajudar no quer que fosse preciso.

A seguir, e porque são eles que fazem destes 5 anos os melhores da nossa vida, queria agradecer aos meus amigos que conseguiram a hercúlea tarefa de me aturar durante tanto tempo. Queria agradecer aos que já vinham de "casa" e que magicamente consegui manter. Ao Gui e ao Rui por se manterem as mesmas pessoas e me mostrarem que a distância pouco importa numa amizade. Ao Brasão por mostrar que um bom churrasco resolve tudo. E ao Telha por ter caminhado lado a lado comigo em todas as aventuras ao longo destes 5 anos.

Depois queria deixar um grande abraço ao melhor que a FEUP me deu. Aos Melgas obrigado por terem sido uma verdadeira família que discute e se chateia, mas no fim acaba toda a jantar à mesa. Para fazer jus a tudo o que passamos teria de agradecer individualmente a cada um de vocês, mas como o espaço não permite tenho de deixar tudo para um próximo Celorico.

Aos Chloes, que mesmo aparecendo tão tarde marcaram me de tal forma que biomédica não teria sido o mesmo sem vocês ao meu lado. Teremos sempre Arazede e as nossas discussões de madrugada.

Como não podia deixar de ser, queria deixar uma palavra de apreço para a Maria Mérida Bravo que passou rapidamente de uma escolha louca para uma verdadeira amiga. Para ti um especial obrigado por me teres ensinado que só eu podia decidir o que fazer destes 5 anos.

Também não esquecer um obrigado a todas aqueles que me permitiram aproveitar o Porto mesmo vivendo em Santa Maria da Feira. Ao chão da casa da Patricia, aos tamboretes do Viriato, ao sofá da Nenas, ao quarto do Chico e ao quarto do Tito um especial obrigado, sem vocês teria sido tudo mais difícil.

Acima de tudo, queria agradecer à minha familia. Obrigado aos meus pais pelo apoio e amor incondicional e à minha irmã por estar sempre lá quando preciso.

Sem todos vós estes 5 anos não teriam feito sentido e esta tese de mestrado não teria sido o que foi.

À FEUP, aos FEUPCaffes, ao Arraial e à Queima. Aos jantares em casa da Patricia e prés em Salgueiros. Às noites mal dormidas e aos nós de gravata mal feitos. Às serenatas e às tardes no Kosovo. Aos shots e aos finos na AE. À antiga Sala de Bio e até à nova Sala do NEB. Aos 5 anos que sabem a pouco, mas que vão deixar muita saudade.

Obrigado!

Duarte Lopes

"Now you're looking for the secret...
but you won't find it, because of course you're not really looking.
You don't really want to know.
You want to be fooled."

The Prestige

Contents

1 Introduction **1**
- 1.1 Context . . . 1
- 1.2 Motivation . . . 2
- 1.3 Objectives . . . 3
- 1.4 Main Contributions . . . 3
- 1.5 Document Structure . . . 4

2 Fundamental Concepts **5**
- 2.1 Emotion Recognition . . . 5
  - 2.1.1 Emotion Background Knowledge . . . 5
  - 2.1.2 Facial Expressions . . . 6
- 2.2 Computer Vision and Machine Learning . . . 9
  - 2.2.1 Image Processing Background Knowledge . . . 9
  - 2.2.2 Machine Learning Classifiers . . . 11
  - 2.2.3 Deep Learning Background Knowledge . . . 13

3 Facial Expressions Databases **21**
- 3.1 Introduction . . . 21
- 3.2 Available Databases . . . 22
- 3.3 Summary . . . 26

4 State-Of-The-Art **27**
- 4.1 Face Detection . . . 27
- 4.2 Face Registration . . . 29
- 4.3 Feature Detection/Extraction . . . 30
  - 4.3.1 Traditional Approaches . . . 30
  - 4.3.2 Deep Learning Approaches . . . 32
- 4.4 Emotion Classification . . . 34
- 4.5 Summary . . . 34

5 Methodology **37**
- 5.1 Evaluation Framework . . . 38
- 5.2 Dataset Preparation . . . 39
  - 5.2.1 Face Detector . . . 39
  - 5.2.2 Data Augmentation . . . 40
  - 5.2.3 Landmarks . . . 40
- 5.3 Handcrafted Features . . . 46
  - 5.3.1 Local Binary Pattern . . . 46

5.3.2 Histogram Oriented Gradient 46
5.3.3 Classifiers 47
5.4 Deep Learning - Convolutional Neural Network 48
5.4.1 Convolutional Neural Network with Transfer Learning 49
5.4.2 Convolutional Neural Network 50
5.5 Deep Learning - Attention Mechanism 53
5.6 Summary 56

6 Experimental Work Results **57**
6.1 Datasets descriptions 57
6.2 Experiments Description 58
6.3 Implementation Details 59
6.4 Discussion and Results 60
6.4.1 Experiment A - Handcrafted 60
6.4.2 Experiment B - CNNs 61
6.4.3 Experiment C - Minimalist Approach 65
6.4.4 Experiment D - Attention Model 78

7 Conclusions **83**
7.1 Summary and Conclusions 83
7.2 Future work 85

References **87**

List of Figures

2.1 Duchenne´s FEs stimulation experiment. First image is catalogued as Attention and second as Amazement. [31] . . . 6
2.2 Illustration of the six basic emotions [35]. . . . 7
2.3 Real representation of the six basic emotions, plus contempt. The images are extracted from CK+ Database [65]. . . . 7
2.4 Illustration of individual AUs in FACS [1]. . . . 8
2.5 Illustration of multiple AUs in the same facial expression [1]. . . . 8
2.6 Image sample from CK+ Database. [65] and respective histogram . . . 10
2.7 Illustration of how to calculate LBP value of one pixel given a 3 x 3 neighborhood. Adapted from [75]. . . . 11
2.8 Histogram of LBP of an image sample from CK+ Dataset [65]. Original image Figure 2.6. . . . 11
2.9 HOG image and respective original sample form CK+ Dataset[65]. . . . 12
2.10 Support Vector Machine Hyperplane illustration from [17] . . . 12
2.11 Schematic of a percepton with 3 input values. . . . 13
2.12 Five different activation function graphs. . . . 15
2.13 Schematic of a convolutional operation done by a convolutional layer. . . . 16
2.14 Schematic of a pooling operation done by a Max Pooling layer and an Average Pooling layer [101]. . . . 16
2.15 Schematic of a basic Multi-layer Percepton. . . . 18
2.16 Model overview, schematic imported from Alexey *et al* [30]. . . . 19

3.1 Images from six different datasets: (a) C+[65], (b) MMI[94]; (c) JAFFE[67]; (d) SFEW [29]; (e) FER-2013[40], (f) AffectNet[72] . . . 24

4.1 Block diagram of a typical AFER system. . . . 27
4.2 Architectures of P-Net, R-Net, and O-Net . . . 28
4.3 Face detection performed by Viola&Jones [96] and MTCNN [108] respectively. . 29
4.4 Diagram of the existent ramification for Feature Detection/Extraction stage. . . . 30
4.5 Appearance-based feature extraction examples: (a) LBP histograms, (b) LBQ histograms, (c) HOG, (d) BoW. Adapted from Sariyanidi *et al.*[85]. . . . 31
4.6 Representation of a geometric feature extraction. . . . 32

5.1 Summary of all developed approaches. . . . 37
5.2 Example of a Cross-Validation with 3 folds. . . . 38
5.3 Example of a Nested Cross-Validation with 5 folds in the outer loop and 3 folds in the inner loop. . . . 39
5.4 Face detector computation. Example of a JAFFE Database [67] sample detected and cropped. . . . 39

5.5 Data augmmentation example: (a) original image from SFEW Database [29]; (b) 9 different examples produced from the original sample from SFEW Database [29]. 40
5.6 Landmark Numbering examples: (a) [0-9], (b) [80-89], (c) [220-229], (d) [250-259]. 41
5.7 Three examples from CK+ Database with the chosen landmarks for the Mask. . . 42
5.8 *Region Of Interest Mask* approach: (a) Original Image; (b) Indexes of Landmarks (c) Binary Mask; (d) Final Image. 43
5.9 *Only Landmarks* approach: (a) Original Image; (b) Detection of the Landmarks; (c) Final Image. 43
5.10 Anthropometric landmarks defined by ISO/IEC 19794-5 [49]. 44
5.11 *Emotion Silhouette*: (a) Original Image; (b) Landmarks chosen to delineate the silhouette; (c) Final Image. 45
5.12 Example of a SFEW [29] sample transformed into a LBP Histogram. 46
5.13 Diagram of MLP and SVM classifier. X represents an image, f the corresponding handcrafted features and $\hat{y}$ the predicted class. 47
5.14 Diagram of generic deep learning approach. X represents an image and $\hat{y}$ the predicted class. 48
5.15 Diagram of generic transfer learning approach. X represents an image and $\hat{y}$ the predicted class. 49
5.16 Illustration of the TransVGG16 Architecture. 50
5.17 Illustration of the myCNN Architecture. 51
5.18 Illustration of the myVGG Architecture. 52
5.19 Illustration of Positional Encoding. 53
5.20 Attention Model architecture. 55

6.1 Class distribution for: (a) CK+ [65], (b)JAFFE [67] and (c) SFEW [29]. 58
6.2 Confusion matrix for myVGG. 62
6.3 Evolution of the myVGG without data augmentation accuracy for training and validation phases. 63
6.4 Evolution of the myVGG with data augmentation accuracy for training and validation phases. 64
6.5 Evolution of the TransVGG16 accuracy for training and validation phases. 65
6.6 Evolution of the myCNN accuracy for training and validation phases. 66
6.7 Confusion matrix for myCNN. 66
6.8 ROI Mask samples of CK+ Database for three different emotions: (a) Happiness; (b) Surprise; (c) Disgust. 68
6.9 Confusion matrices of the models trained with ROI Mask samples: (a) TransVGG16; (b) myCNN; (c) myVGG. 69
6.10 ROI Mask samples of CK+ Database of two different emotions: (a) Fear; (b) Contempt. 70
6.11 ROI Mask samples of JAFFE Database for of different emotions: (a) Anger; (b) Surprise; (c) Disgust; (d) Happiness. 70
6.12 ROI Mask less accurate samples of SFEW Database of four different emotions: (a) Fear; (b) Happiness; (c) Anger; (d) Surprise. 71
6.13 ROI Mask more accurate samples of SFEW Database of three different emotions: (a) Sadness; (b) Surprise; (c) Happiness. 71
6.14 Only Landmarks samples of CK+ Database of two different emotions: (a) Surprise; (b) Happiness. 72
6.15 Only Landmarks samples of CK+ Database of three different emotions: (a) Sadness; (b) Anger; (c) Comtempt. 73

6.16 Only Landmarks samples of JAFFE Database of five different emotions: (a) Fear; (b) Anger; (c) Sadness; (d) Surprise; (e) Happiness. 74
6.17 Only Landmarks and ROI Mask samples of SFEW Database of four different emotions: (a) Fear; (b) Surprise; (c) Anger; (d) Happiness. 74
6.18 Confusion matrices of myCNN trained with : (a) ROI Mask Samples; (b) OL samples. 75
6.19 Emotion Silhouette samples of CK+ Database of four different emotions: (a) Surprise; (b) Disgust; (c) Happiness; (d) Anger. 76
6.20 Confusion matrices of myCNN trained with Emotion Silhouette samples. 77
6.21 Emotion Silhouette and Only Landmarks samples of JAFFE Database of four different emotions: (a) Happiness; (b) Surprise; (c) Sadness; (d) Disgust; (e) Fear; (f) Contempt. 78
6.22 Emotion Silhouette samples of SFEW Database of six different emotions: (a) Anger; (b) Fear; (c) Happiness; (d) Surprise. 79
6.23 Confusion matrix for AM-(6x6). 80
6.24 Evolution of the AM-(12x12) accuracy for training and validation phases. 81
6.25 Confusion matrix for AM-(12x12). 81

List of Tables

2.1 AUs coded by FACS systems and respective description. Adapted from [65]. . . . 8
2.2 Characterization of each basic emotion using only AU codes. Adapted from [65]. 9

3.1 Summary of the used AFER Databases 25

4.1 Summary of Traditional Approaches. 35
4.2 Summary of Deep Learning Approaches. 35

5.1 Landmark Indexes for *Region Of Interest Mask*. 42
5.2 Number of key-points used in ISO Documentation and in *Emotion Silhouette* images. 44
5.3 Landmark Indexes used for each Facial Area in *Emotion Silhouette* images. . . . 45
5.4 HOG Descriptor Parameters. 47
5.5 SVM parameters used during optimization step. 48
5.6 MLP parameters used during optimization step. 48
5.7 TransVGG16 architecture details and training parameters. 50
5.8 CNN Baseline architecture details and training parameters. 51
5.9 myVGG architecture details and training parameters. 52
5.10 Attention Mechanism Model architecture details and training parameters. 54
5.11 Attention Model parameters to optimize 54

6.1 Number of samples for each method 59
6.2 Handcrafted (A) - Results accuracy depicted in percentage(%). 60
6.3 CNNs (B) - Results accuracy depicted in percentage(%). 61
6.4 Results from methods with and without data augmentation. 62
6.5 Handcrafted (C) - Results accuracy depicted in percentage(%). 67
6.6 Attention Model (D) - Results accuracy depicted in percentage(%). 78

7.1 Review of the best results presented in Experiment from a to D - Results accuracy depicted in percentage(%) 83

Abbreviations

1-D	Unidimensional
3DCNN	3D Convolutional Neural Network
AC	Affective Computing
AU	Action Unit
AAM	Active Appearance Model
AFER	Automated Facial Expression Recognition
AFEW	Acted Facial Expressions In The Wild
AI	Artificial Intelligence
Artifical Neural Network	ANN
BoW	Bag of Words
CK+	Extended Cohn-Kanade Database
CNN	Convolutional Neural Network
C-LSTM	Convolutional-Long Short-Term Memory
D	Dynamic
EEG	Electroencephalogram
ES	Emotion Silhouette
FACS	Facial Action Coding System
FE	Facial Expression
FER2013	Facial Expression Recognition 2013 Database
FERG-2D	Facial Expression Research Group 2D Database
GAN	Generative Adversarial Network
GPU	Graphics Processing Unit
HCI	Human-Computer Interaction
HOG	Histogram Oriented Gradient
ISO	International Organization for Standardization
IEC	International Electrotechnical Commission
JAFFE	Japanese Female Facial Expression Database
kNN	k-Nearest Neighbor
L	Lab
LBP	Local Binary Pattern
LBQ	Local Phase Quantization
LDP	Local Directional Pattern
LFW	Labeled Faces in the Wild
LeakyReLU	Leaky Rectified Linear Unit
LSTM	Long Short-Term Memory

MA	Multi-Attention
MLP	Multi-Layer Percepton
MSDF	Multi-Scale Dense SIFT
MTCNN	Multi-Task Cascaded Convolutional Neural Network
NLP	Natural Language Processing
NMS	Non Maximum Suppression
NN	Neural Network
OL	Only Landmarks
P	Posed
PCA	Principal Component Analysis
P-Net	Proposal Network
RASnet	Region-Aware Sub-net
ReLU	Rectified Linear Unit
ResNet	Residual Network
RF	Random Forest
RGB	Red Blue Green
ROI Mask	Region Of Interest Mask
R-Net	Refine Network
S	Spontaneous
St	Static
SBFS	Sequential Backwards Feature Selection
SFEW	Static Facial Expressions in the Wild Database
SFFS	Sequential Forward Feature Selection
SIFT	Scale-Invariant Feature Transform
Spyder	Anaconda Scientific Python Development Environment
STTM	Spatiotemporal Texture Map
SVM	Support Vector Machines
T-LSTM	Temporal-Long Short-Term Memory
VGG	Visual Geometry Group
ViT	Vision Transformer

Chapter 1

Introduction

1.1 Context

Emotions can be defined as a subjective experience triggered by someone or some experience, they can be described as a group of involuntary internal (muscle contraction, change of cardiac and respiratory rhythm) or external reactions (facial and vocal expressions, gestures or movements) [24]. Emotions are complex processes, including feelings, body language, cognitive reactions, behaviours or thoughts. Being one of the more primary and natural forms of how we react to the world that surrounds us, emotions always portrayed a fundamental role in the way humans interact with each other.

From an evolutionary point of view, mankind owes its success to its remarkable need of living in groups, as stated by Aristoteles -“A Man is by nature a social animal" - therefore we depend heavily on understanding each other to live our lives [63]. Moreover, and corroborated by historian Yuval Harari in his book ”Sapiens: A Brief History of Humankind" [45], humans only established their dominance over the other species, due to our uncanny ability to share and understand each other. Although, verbal language was the main responsible for this remarkable ability, facial expressions (FEs) come right after, as being one of the most powerful forms of communication.

As it was aforementioned, emotions originate from multiple physical changes that give information about our internal state. Alongside voice, language and posture of the body, facial expressions give out the most clues about the emotional state of someone [20]. Therefore, facial expressions are considered one of the most natural, powerful and universal ways for human beings to express their emotions and intentions [25]. Despite the ability of humans to predict each other’s emotions, there are still limitations to correctly classifying all the facial expressions and mapping them to the emotions those expressions mean. Thus, the development of automated computer vision methodologies becomes an helpful tool to perform emotion recognition tasks more efficiently and automated. Using different features, such as facial expressions [36], speech [44][19], EEG (Electroencephalogram) [80], and even text [98] many different models and systems have been developed. However, facial expressions, due to several reasons - they are visible, they contain many useful features, and it is easier to collect a large dataset of faces - became the most popular

approach. AFER (Automated Facial Expression Recognition) algorithms are currently one of the most active research topics and these methods can be defined as all the systems which can predict what emotion is being felt simply by analysing what is the facial expression shown.

AFER systems aim to quickly and universally classify emotions using the facial features of the user. Although emotion recognition is a task that human beings perform daily, unconsciously and effortlessly, it is not yet easily performed by computers [36]. Besides, not every emotion has been catalogued and understood. Only by 1970, Paul Ekman, a pioneer in the study of emotions and their relation to facial expressions, defined the six basic emotions of human communication [32]. Fear, anger, disgust, happiness, sadness and surprise are considered the six basic raw emotions and are associated with very distinctive features on our faces. Contempt, which was subsequently added, is also regarded as the seventh core emotion [68]. These emotions are not the only ones, but other more complex emotions like guilt, jealousy, pride or shame are even more challenging to classify by just using the face as a source of information. Moreover, these complex emotions can highly depend on micro-expressions which are much more difficult to detect using AFER systems.

The need for automated entities which can predict an emotional state by only examining the face is not purely academic. There is a market which would benefit from a more accurate, generic and impartial mechanism to classify how the different subjects are feeling. The computer science field dedicated to the study of emotions is known as Affective Computing (AC). It is a growing field which comprises subjects like engineering, psychology, cognitive science and even sociology. Also, Affective Computing incorporates numerous applications which include, (1) automated driver assistance [52] - warning tired drivers of the extreme levels of tiredness; (2) healthcare — through wellness monitoring applications identifying causes of stress, anxiety and even of depression [39]; (3) adaptive learning - through a teaching application able to adjust the content delivery rate and many interactions according to the user enthusiasm and frustration; (4) recommendation systems - assisting and asserting personalised content according to the user preferences; and even (5) crowd evaluation systems - evaluating the reaction of a crowd to a movie or music - this evaluation could be really important to infer about what did the subjects felt during a specific experience [13].

1.2 Motivation

In the last few years, Emotion Recognition and other human-computer interaction (HCI) systems attracted a lot of attention because of the potential of human-robot interaction, data-driven animation and interactive games. Therefore, it is natural that a new market for those types of software products is emerging. In fact, according to an article from last year [81], even in a post-pandemic scenario, the Emotion Recognition market is predicted to grow to $37.1 billion by 2026, up from an estimated $19.5 billion in 2020.

As it is proven, the ”world" of AFER systems is already well established, though different challenges and opportunities emerge every day. Ranging from the lack of quality provided by

the state-of-the-art databases to unknown applications of new deep learning techniques, AFER systems are presented as being in constant development.

On the one hand, the most classic software-based approaches include varied machine learning techniques which fail to adapt to real-time situations, on the other hand, neural networks and other deep learning approaches increase robustness but at a high computational cost. For example, by focusing on specific regions of the face during the learning phase, the models could be forced to be more accurate, less overfitted and less computationally expensive.

Overall, there is room, not only, for the development of better and more robust models, but, most importantly, for trying to understand how facial expressions can be deconstructed and see how different facial areas can impact the final result. Understanding how models perceive a human face and seeing which specific areas contribute to better results is the key to developing better models.

1.3 Objectives

The main goal of this dissertation is to perform a comprehensive benchmarking of traditional machine and deep learning domains and develop a variety of methods for facial expression recognition. For this purpose, we aim to investigate and discuss the current literature regarding facial-based emotion recognition models. Moreover, our objective is to recreate already consolidated approaches and come up with new ways of tackling the same problems, focusing mainly on data modifications and forcing the models to certain regions of the face.

Globally, the present work focuses on analysing facial expressions as a set of multiple landmarks and tries to find out if by only using those face regions we can create a good performing classification model. At last, this dissertation will serve as a step towards understudying the impact of using minimalist facial expressions representations as input data, ultimately we want to evaluate if reducing a face to its landmarks can lead to a reliable and accurate facial portrayal

1.4 Main Contributions

Regarding the implements methodologies, this dissertation main contributions are:

- Extensive and thorough study which comprises all the main state-of-the-art approaches, ranging from traditional machine learning to deep learning model.
- Application of attention model based on the vision transformer to a facial expression recognition problem.
- Innovative methodology with the development of three new approaches regarding minimalist representation of human facial expression. Showing empirically that some facial regions are more prominent than other to the final classification, human face can be accurately described only by its facial landmarks and most important high performance models can be developed using minimalist representations of human emotion.

-

1.5 Document Structure

This document contains several chapters which comprise all the information required to achieve the aforementioned goals.

Chapter 2 introduces the background knowledge about the history behind emotions and emotion recognition and image processing and artificial intelligence methods based on machine and deep learning, giving a particular spotlight to the techniques which were used to develop our models.

The following chapters review the existent state-of-the-art approaches on AFER models. In Chapter 3 a review of the most relevant databases available is done, describing and comparing them. Followed by, Chapter 4 in which it is described and further discussed some of the best approaches in facial emotion recognition, through tradictional machine and deep learning techniques.

Chapter 5 comprises the entire methodology of implemented work, followed by all the obtained results and respective discussion in Chapter 6.

Chapter 7 presents the final remarks about this dissertation, as well as an overall discussion and analysis of the results. Final remarks and proposals for future works are also included in this chapter.

Chapter 2

Fundamental Concepts

This chapter will comprise all the theoretical basis needed to fully comprehend this dissertation. A complete background about the history of emotions and facial expression is presented, as well as a brief description of all the required knowledge about computer vision and its respective techniques.

2.1 Emotion Recognition

2.1.1 Emotion Background Knowledge

Understanding the emotional state of someone has always intrigued mankind and has been a complex topic over the years. The first studies date back to ancient Greeks, where Cicero first described emotions as a set of four basic states: metus (fear), aegritudo (pain), libido (lust), and laetitia (pleasure) [41]. Moving forward, during the 19th century, Charles Darwin theorized that emotions, and more specifically facial expressions (FEs), must have had a vital instrument to assure mankind survival and evolution [25]. Understating how one another is feeling is one of the keys to live in society, but also, emotions, can trigger physical actions which act as a response to the environmental challenges. For example, constricting the nostrils in disgust could have served as a way to reduce inhalation of noxious or harmful substances. Or even widening the eyes in surprise in order to increase our visual field [20].

Emotions are defined by the Oxorfd dictionary as ”strong feeling deriving from one’s circumstances, mood, or relationships with others” [3]. However, António Damásio, famous neurologist and pioneer in the field of social cognition and decision-making, argues that emotions and feelings are very different. Feelings are described as being the mental experience which joins the different homeostatic process our body goes through. It is all about the representation of a process of life inside of you, they are happening all the time in one way or another. On the other hand, emotions are so much raw and represent a reaction to the outside world - ”I even think even the word betrays that the root of the word emotion is about motion directed to the outside. (...) It is all about actions." [14]. Saying this, it is possible to conclude that no other human trait gives as many features and information about our emotional state as our faces.

Human faces generally reflect what António Damásio advocates, a reaction towards the outside environments which comprises what we are undergoing. Due to this, facial expressions (FEs) are naturally the most used features. Dr. Albert Mehrabian [69] formulated the 7-38-55 rule, which states only 7% of all communication is done through verbal communication, whereas the nonverbal component of our daily communication, such as the tonality of our voice and body language, make up 38% and 55% respectively. This rule highlights how important our facial expressions are, how much information can be given by our facial traits and why computer vision applications are so much focused on Facial Emotion Recognition systems (FER).

2.1.2 Facial Expressions

As it was previous highlighted, Facial Expressions (FEs) are a subgroup of the big theme which is Emotion Recognition. The interest in facial expressions is traced back to early years of renaissance with even Leonardo da Vinci taking interest in understanding the connection between facial expressions and someone´s personality [16]. However, only by the 17th century more scientific studies start to appear, with John Buwler publishing books like *Pathomyotomia* and or *Dissection of the significant Muscles of the Affections of the Mind*. These publications were the first consistent work in the muscular mechanism of FEs [42]. Moving forward to the 19th century, as it was said previously, Charles Darwin started theorizing about the importance of facial expressions in an evolutionary context [25]. Around the same historical period, Duchenne de Boulogne started performing multiple experiments, electrically stimulating human faces to see what kind of facial expression could be manufactured (Figure 2.1). The French neurologist thought that the human face could be understood as a map of several features which codified multiple mental states of mind [31].



Figure 2.1: Duchenne´s FEs stimulation experiment. First image is catalogued as Attention and second as Amazement. [31]

However, only by 1970, arguably, the most important work within the world of emotions is written. Doctor Paul Ekman introduces to the world his *magnum opus*, *Universal and cultural differences in facial expression of emotion* [32]. In its work, Ekman theorized that, first, there are six basic human emotions: happiness, sadness, anger, fear, surprise, disgust and later contempt was also added as a basic expression (Figure 2.3 and Figure 2.2); and second, and more important, they are innate and shared by everyone. This includes the fact that these emotions are ubiquitous and are accompanied across cultures by universal facial expressions. This is shown by comparing the two presented figures, both the real human faces and the cartoon representation can, by using the same facial attributes, communicate the same emotions.

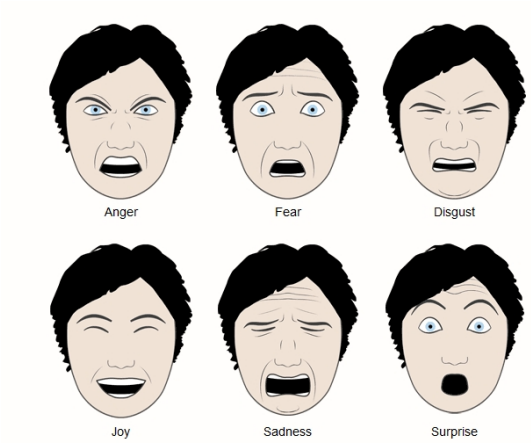


Figure 2.2: Illustration of the six basic emotions [35].

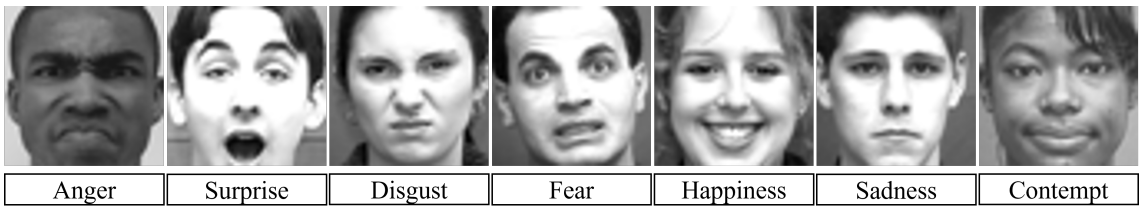


Figure 2.3: Real representation of the six basic emotions, plus contempt. The images are extracted from CK+ Database [65].

The following years, Paul Ekman continued an intense research focusing on facial expression and by 1978, partner with Wallace V. Friesen, published *Facial action coding system* [33]. This work is of the utmost importance because it introduces to the world the Facial Action Coding System (FACS). FACS is a system used in various scientific settings for research. It is a comprehensive, anatomically based system for describing all visually discernible facial movement. It breaks down facial expressions into individual components of muscle movement, called Action Units (AUs) [5] (Table 2.1), which are used as codes for specific areas of a human face. Each AU is defined by the contraction of a certain facial muscle, it can be individual (Figure 2.4) or a combination (Figure 2.5), and produces a unique image feature [93]. The creation of a system capable of dividing the human face like a puzzle was groundbreaking for the emotion recognition

world. Using AUs it was possible to associate very specific facial areas and combination of muscle contractions to a certain facial expression. Table 2.2 highlights the association between the known AUs and the seven basic emotions.

Table 2.1: AUs coded by FACS systems and respective description. Adapted from [65].

AU	Description	AU	Description	AU	Description
1	*Inner Brow Raiser*	13	*Cheek Puller*	25	*Lips Part*
2	*Outer Brow Raiser*	14	*Dimpler*	26	*Jaw Drop*
4	*Brow Lowerer*	15	*Lip Corner Depressor*	27	*Mouth Strech*
5	*Upper Lip Raiser*	16	*Lower Lip Depressor*	28	*Lip Suck*
6	*Cheek Raiser*	17	*Chin Raiser*	29	*Jaw Thrust*
7	*Lid Tightener*	18	*Lip Puckerer*	31	*Jaw Clencher*
9	*Nose Wrinkler*	20	*Lip Stretcher*	34	*Cheek Puff*
10	*Upper Lip Raiser*	21	*Neck Tightener*	38	*Nostril Dilator*
11	*Nasolabial Deepener*	23	*Lip Tightener*	39	*Nostril Compressor*
12	*Lip Corner Puller*	24	*Lip Pressor*	43	*Eyes Closed*

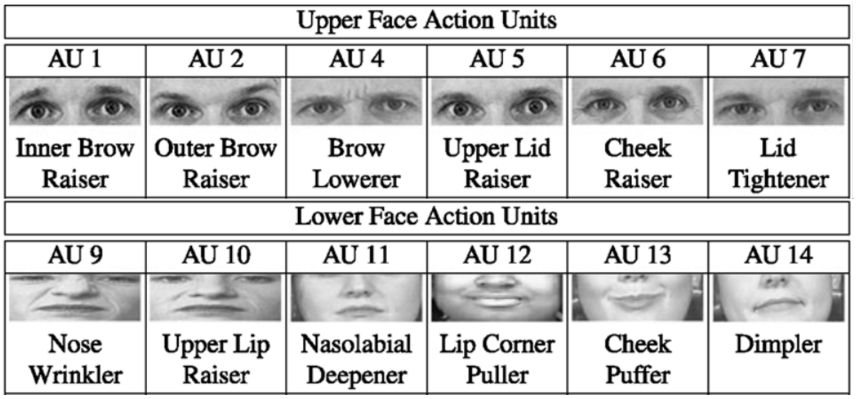


Figure 2.4: Illustration of individual AUs in FACS [1].

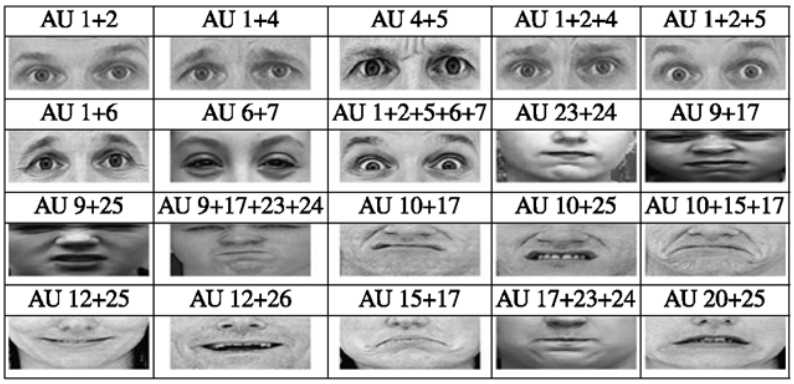


Figure 2.5: Illustration of multiple AUs in the same facial expression [1].

Finally, it is believed that the Ekamn´s seven basic emotions are very limited in their ability to describe more complex emotions like guilt, pleasure or even relief. However, for automatic

Table 2.2: Characterization of each basic emotion using only AU codes. Adapted from [65].

Emotion	Criteria
Anger	AU23 and AU24 must be present in the AU combination
Surprise	Either AU1+2 or 5 must be present
Disgust	Either AU9 or A10 must be present
Fear	AU combination of AU1+2+4 must be present
Happiness	AU12 must be present
Sadness	Either AU1+4+15 or AU11 must me present. Exception is AU6+15
Contempt	AU14 must be present (either unilateral or bilateral)

recognition systems like AFER these basic emotions and respective facial representations are the most used.

2.2 Computer Vision and Machine Learning

2.2.1 Image Processing Background Knowledge

AFER models are framed as a solution to a computer vision problem, and as it happens in most of classification problems based on images, specially the more conventional approaches, the samples cannot be fed as an image directly to the classification models. Thus, a pre-processing is required, which can transform our samples into usable features. Those processes are commonly known by Image Processing and aim to transform an image into a set of features which embody the necessary properties.

2.2.1.1 Image Histogram

One of the most common ways of representing an image is by an histogram. The histogram of an image is a statistical distribution of its pixels. The number associated to each ”bin" is the total number of pixels from a specific intensity value, hence each ”bin" corresponds to the digital value of the pixels. In a image that value can range from 0 to 255 or if the image is normalized from 0 to 1. In Figure 2.6 it is possible to see a image and its respective graphical representation. Histograms are a powerful tool to define an image using the information about the number of pixels that share the same intensity.

2.2.1.2 Local Binary Patterns

The Local Binary Patterns (LBP) [77] is a useful tool to detect local spatial patterns and contrasts in the image like edges, spots and even flat areas. The LBP operator looks at points surrounding a central point and tests whether the surrounding points are greater or less than the central point. If the surrounding point is higher it is turned to 1, if it is lower it is turned to 0. After turning the neighbors points into binary values we compute the LBP value of the central pixel [11]. Finally, we apply the same approach to every pixel of the image. The LBP descriptor is defined by the Equation 2.1 and illustrated in Figure 2.7.

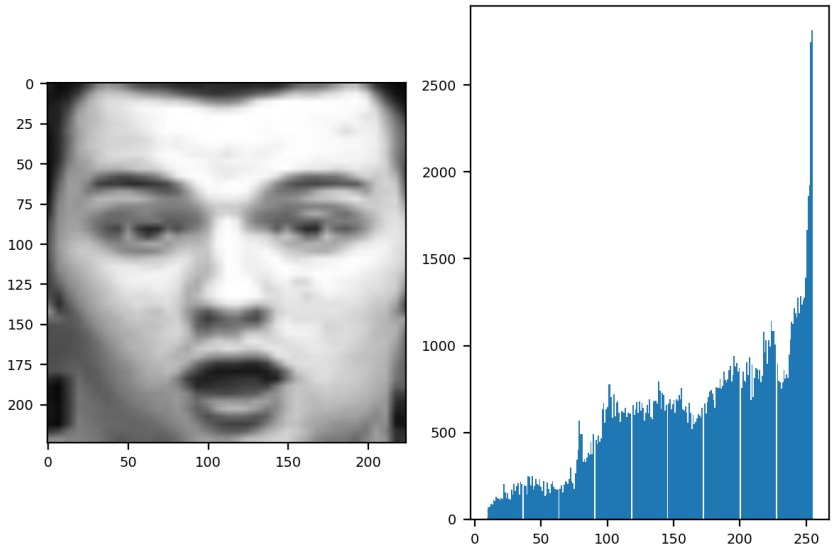


Figure 2.6: Image sample from CK+ Database. [65] and respective histogram

$$LBP = \sum_{p=0}^{P-1} f(p-c) \times 2^p \tag{2.1}$$

$$f(x) = \begin{cases} 1 & \text{if } x \geq 0 \\ 0, & \text{otherwise.} \end{cases} \tag{2.2}$$

In Equation 2.1, P is the number of neighbor points to consider and c our central pixel. Equation 2.2 is a binary function that outputs one when the value of the pixel is higher than the value of the central pixel and 0 otherwise. At last, the binary value is multiplied by the 2^p and replaces the numeric value of the original pixel.

At last, all the values of the LBP descriptors can be rearranged into a histogram. The frequency of each LBP is computed and a histogram is created. This histogram can be used as a good texture descriptor since it comprises the information of the LBP of each image pixel (Figure 2.8).

2.2.1.3 Histogram of Oriented Gradient

The use of this approach goes back to the article written by Dalal *et al.* [22], where this descriptor was used to detect human faces. Histogram Oriented Gradient (HOG) is a feature descriptor similar to to scale-invariant feature transform (SIFT) [64]. However, this method is based on evaluating well-normalized local histograms. An image is divided into equally sized cells and a local 1-D histogram of gradient directions is computed. After obtaining the gradient of each pixel, the gradient matrices (magnitude and angle matrix) are divided into 8x8 cells to form a block. For each block, a 9-point histogram is calculated. Finally, combining all these local histograms results in a histogram that represents the whole image.

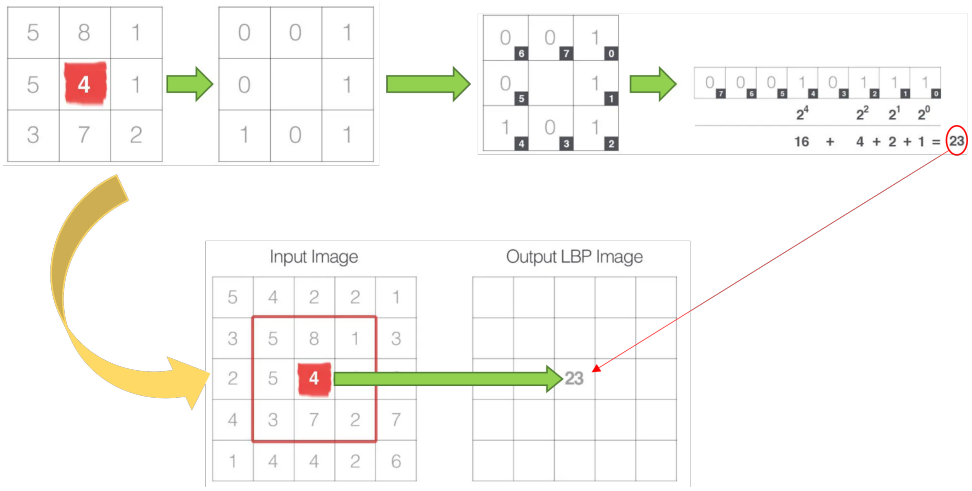


Figure 2.7: Illustration of how to calculate LBP value of one pixel given a 3 x 3 neighborhood. Adapted from [75].

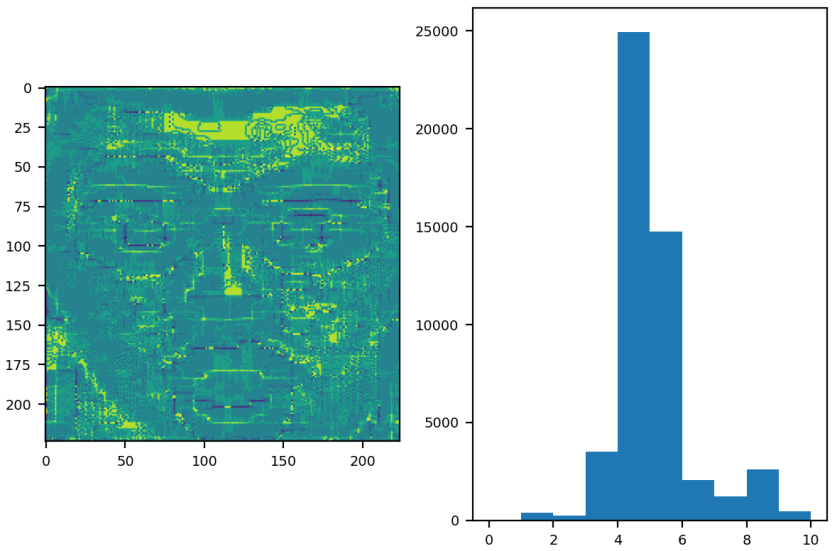


Figure 2.8: Histogram of LBP of an image sample from CK+ Dataset [65]. Original image Figure 2.6.

All in all, HOG decomposes an image into small squared cells, computes an histogram of oriented gradients in each cell, normalizes the result using a block-wise pattern, and return a descriptor for each cell. This descriptor focuses on the structure or the shape of an object and it is better than any edge descriptor as it uses magnitude as well as angle of the gradient to compute the features, Figure 2.9.

2.2.2 Machine Learning Classifiers

In machine learning, a classifier is an algorithm which automatically sorts or categorizes data into one or more "classes". In this section some of the most relevant classifiers are described and their

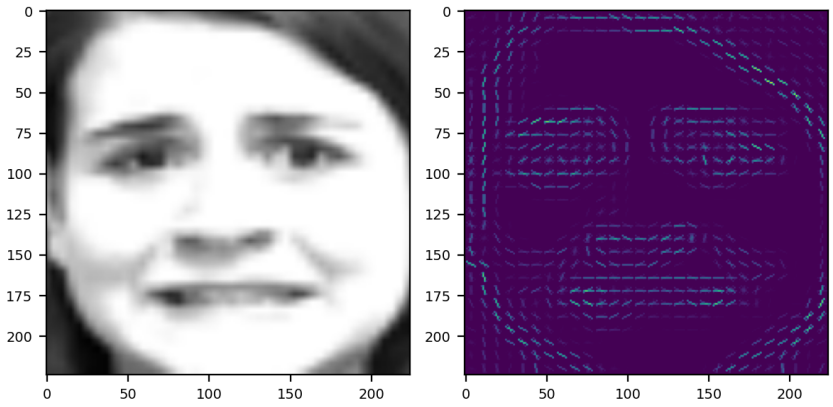


Figure 2.9: HOG image and respective original sample form CK+ Dataset[65].

operation is explained.

2.2.2.1 Support Vector Machines

Support Vector Machines (SVM) [21] are one of the most popular machine leaning classifiers. This type of classifier is a common supervised learning technique that may be used to solve both classification and regression issues. Its main purpose is to find an optimal way to categorize a n-dimensional space (N — the number of features) creating a decision boundary known as hyperplane. Using data points that are closer to the hyperplane, known as support vectors, this classifier builds a margin between points from different classes. The hyperplane is defined assuring that the same class points are close to each other and as far possible to support vectors of each class, maximizing the margin inter-classes(Figure 2.10).

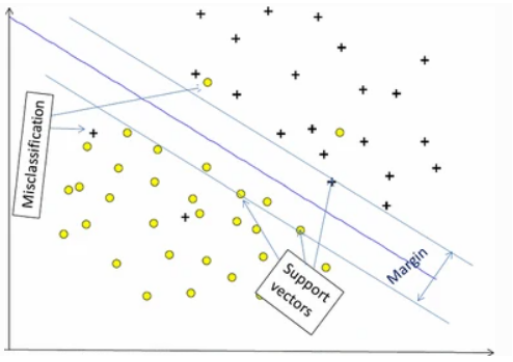


Figure 2.10: Support Vector Machine Hyperplane illustration from [17]

2.2.3 Deep Learning Background Knowledge

Deep Learning is defined as a type of machine learning and artificial intelligence (AI) which tries to imitate the way humans gain certain types of knowledge. While traditional machine learning algorithms are linear, deep learning algorithms are stacked in a hierarchy of increasing complexity and abstraction. Using powerful "weapons" called Neural Network (NN), deep learning mechanism mimic the way our brain works and without any prior knowledge gathers useful information from the data. These models are composed of consecutive processing layers that learn representations of data with multiple levels of abstraction, capturing features that traditional methods could not compute [54]. As such, deep learning has recently become state-of-the-art in various tasks, including image classification, having achieved excellent results, sometimes even surpassing the human ability [46].

As it was aforementioned, deep learning models, especially artificial neural network (ANNs), aim to mimic the way our brain works. Therefore, and as it happens in our brain, these models depend on units and connections between units which act like neurons and synapses. The most basic version of a neural network is the Perceptron [84], composed of a single unit, which is represented in Figure 2.11. Mathematically, the "neuron" of the Percepton acts like an operator that receives a set of inputs and applies to them a linear transformation. The unit contains a set of weights (w), applied to its inputs, and a bias θ , which is a constant value. The perceptron linearly combines its weighted inputs, adds the bias, and applies a nonlinear function $a(X)$, also known as activation function. The resulting output can be seen in Equation 2.3

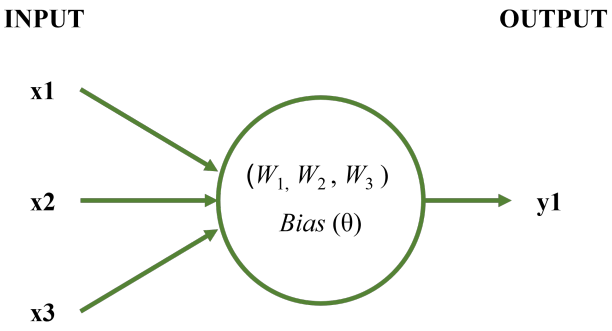


Figure 2.11: Schematic of a percepton with 3 input values.

$$y = a(\sum_{i=0}^{n} w_i \times x_i + \Theta) \tag{2.3}$$

The activation function acts as the final decision before the final value is outputted. This function, decides whether the neuron should or not be activated and represented in Equation 2.3 by $a(X)$. The commonly used activation functions are presented as follows [2] and shown in Figure 2.12:

- **Linear Activation:** The most basic activation function is Linear Activation, which is known as ”identity function" and its output is proportional to the input. The input x will be returned as it is.

- **Hyperbolic Tangent (Tanh):** Tanh function is very similar to the sigmoid function and even has the same S-shape. However, it ouputs a value between -1 and 1. In this activation, the larger the input (more positive), the closer the output value will be to 1, whereas the smaller the input (more negative), the closer the output will be to 1. (Equation 2.4)

$$a(x) = \frac{2}{1+e^{-2x}} - 1 \tag{2.4}$$

- **Sigmoid:** This activation function outputs a number between 0 and 1 and has a S shape. In general, the larger the input (more positive), the closer the output value will be to 1, whereas the smaller the input (more negative), the closer the output will be to 0. This function can be very good when performing a classification task primarily binary ones, since it pushes the predictions to extreme values. (Equation 2.8)

$$a(x) = \frac{1}{1+e^{-x}} \tag{2.5}$$

- **Rectified Linear Unit(ReLU):** ReLU, which stands for Rectified Linear Unit, is one of the most simple and popular activation function. All non-positive inputs ($x \leq 0$) are outputted as 0 and every other input is outputted as x. Only specific neurons are activated, making the network sparse and efficient for computation. All the neurons will only be deactivated if the output 0 or less, and activated if the output is a positive value. (Equation 2.6)

$$a(x) = \begin{cases} x & \text{if } x > 0 \\ 0, & \text{if } x \leq 0. \end{cases} \tag{2.6}$$

- **Leaky Rectified Linear Unit(LeakyReLU):** This function is an improved version of ReLU. It is a linear function with a small positive slope for negative inputs, which is able to solve the dead ReLU problem (all the negative units are deactivated by the ReLU activation function).(Equation 2.7)

$$a(x) = \begin{cases} x & \text{if } x > 0 \\ ax, & \text{if } x \leq 0. \end{cases} \tag{2.7}$$

- **Softmax:** Softmax activation functions is very different from all the aforementioned functions. This activation is described as being a combination of multiple sigmoids and is commonly used for multi-class classification problems. It calculates the relative probabilities.

Similar to the sigmoid/logistic activation function, the SoftMax function returns the probability of each class, it is frequently used in the last layer of neural network designed for multi-class classification. Mathematically, the softmax function is shown below:

$$a(x) = \frac{e^{z_j}}{\sum_{k=1}^{K} e^{z_k}} \tag{2.8}$$

where z is a vector of the inputs to the output layer and j indexes the output units, so j = 1, 2, ..., K.

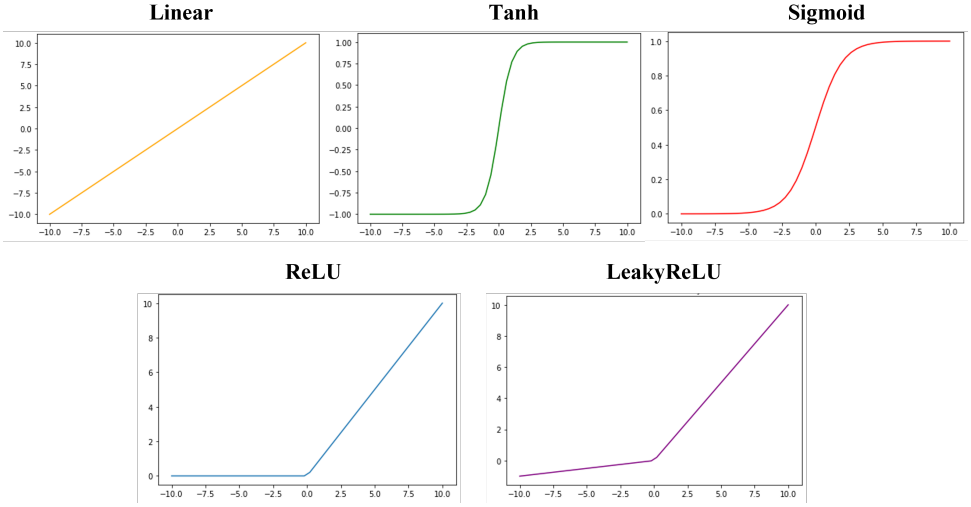


Figure 2.12: Five different activation function graphs.

Having one percepton is like having one neuron, one neuron does not have the power to interpret the world like our brain, but millions of neurons connected together become an organ capable of learning to recognize people, to speak, to write poems and even to have abstract thoughts [83]. Thus, the ultimate objective of deep learning models is to combine and organise multiple perceptons and build a network, whose goal is to receive an input, like an image, discover and fetch the most important feature and output a final result.

2.2.3.1 Convolution Neural Networks

Convolutional Neural Networks (CNNs) are one of the most used type of deep learning model, largely due to their accomplishment of reaching good performances and extracting valuable information from data, mainly images. CNNs are, now, state-of-the-art approaches for challenge like image classification, signal processing and even natural language processing and speech recognition. CNNS like VGG [89], Inception [91][92] and ResNet [47] have been developed and are widely used in multiple image classification problems. These powerful classifiers are composed essentially by convolutional and pooling layers, which are used alternative to detect and extract the most prominent features of the data.

- **Convolutional Layers:** Convolutional layers are the fundamental layers of CNNs. These type of layers are composed by filter, or kernels, and transform the input, usually an image, and output a feature map through a process of convolution between the input and the filter. The filter are grids, commonly smaller than the input image, that act like a sliding window over the input. As it is illustrated in Figure 2.13, a filter with the size of 3x3 slides over the input image (6x6) applying a convolution operation to each pixel and originating a feature map with size 4x4.

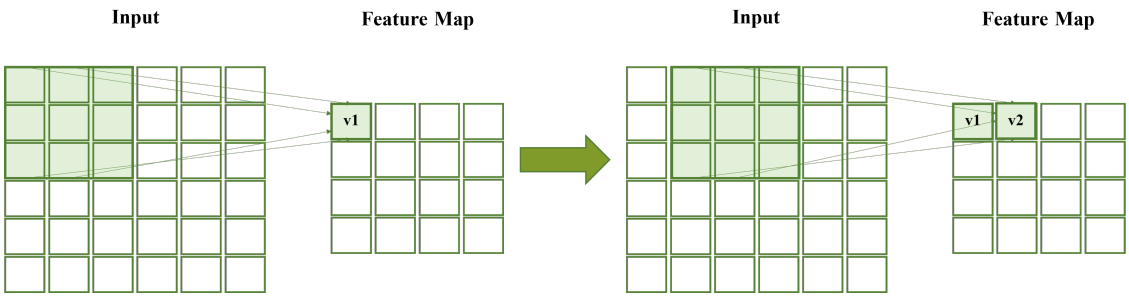


Figure 2.13: Schematic of a convolutional operation done by a convolutional layer.

- **Pooling Layer:** There are two types of pooling layers: Average Pooling or Max Pooling; both types have the same objective, transform a number of values into a single value. Average Pooling receives the number of values embraced in the window of the layer, defined by a specific size (for example 2x2) and outputs the average value of those values. Max Pooling works with the same principle, but instead of using the average, the layer outputs the maximum value. In Figure 2.14 a pooling layer of 2x2 slides the feature map and outputs a new feature map. These layers are used to "down sample" the data, which makes the representations more compact and invariant to modifications.

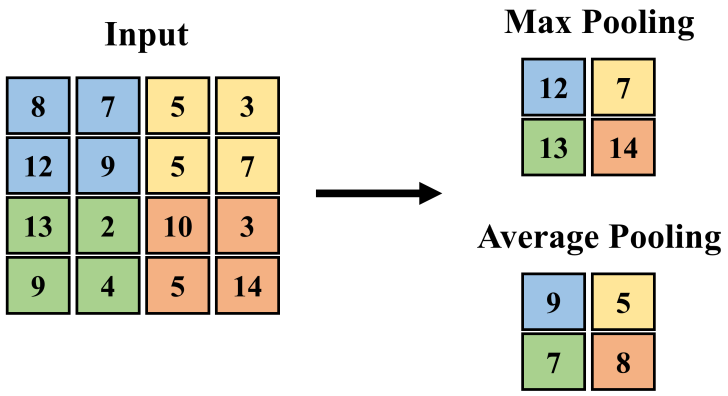


Figure 2.14: Schematic of a pooling operation done by a Max Pooling layer and an Average Pooling layer [101].

- **Fully Connected Layers:** Fully Connected layers are those layers where all the inputs from one layer are connected to every activation unit of the next layer. In most deep learning

models, the last few layers are fully connected layers which compile the data extracted by previous layers to form the final output. These layers are often used after convolutional and pooling layers, taking all the previous layers as input and outputing a final prediction.

- **Dropout:** Dropout layers are widely used to regulate overfitting and underfitting of a CNN model and are defined by a value between 0 and 1. These values define the probability of activation of the neurons. This layers introduce more variability into the model so the model learns to generalize better.

Before being ready for training, CNNs are, like it was aforestated for perceptons, followed by an activation function. Regarding training phase, there are multiple other parameters which can help reduce overfitting:

- **Data Augmentation:** This technique is most commonly used on images. Having more and diverse data can help convolutional neural network to be more robust and general. Thus, using operations like flipping, rotating and cropping we can enlarge the dataset and feed the model more samples to make it more resistant to overfitting [87][79].

- **Batch Normalization:** It is a way of normalizing the output of the previous activation layer and increase the stability of the neural network. This way, the distributions inside the model are maintained similar, which helps it to learn and consequently reduces significantly the time needed to converge. Batch normalization applies a transformation that maintains the mean output close to 0 and the output standard deviation close to 1.

- **L1 and L2 regularization:** these are traditional regularization strategies that consist in adding a penalty term to the objective function and control the model complexity. They are parameters which are added to the loss function and penalise the model´s weights. More specifically L1 uses a penalty which encourages the sum of the absolute values of the parameters to be minimum; L2 can be perceived as adaptive minimization of the square error with a penalisation.

- **Early stopping:** it is a strategy that consists of stopping the training process before the performance of the network drops for the validation set. This way our model´s training stops before it enters a phase where it is already overfitting.

2.2.3.2 Multi-layer Perceptron

A Multi-Layer Perceptron is a basic combination of perceptrons. MLPs are a sub-type of Artificial Neural Network (ANN) which have some of its neurons, or perceptons, hidden. An MLP consists of at least three layers of nodes: an input layer, a hidden layer and an output layer. Figure 2.15 shows an example of a Multi-layer Perceptron with one hidden layer.

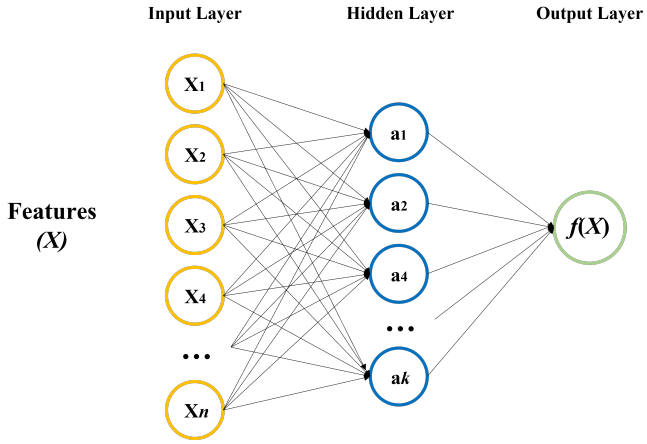


Figure 2.15: Schematic of a basic Multi-layer Percepton.

2.2.3.3 Attention Mechanism

Attention Networks are somewhat recent deep learning approaches. In 2017, Ashish introduced to the world the paper ”Attention Is All You Need" and the concept of Transformer [95]. In his paper, the author proposes a transformer architecture, made by an encoder and decoder, which can substitute a complex Recurrent Neural Netowrk (RNN). This type of network has gained very good results mainly in natural language processing (NLP) problems. The dominant approach is to pre-train on a large text corpus and then fine-tune on a smaller task-specific dataset. All in all, thanks to Vaswani´s Transformers computational efficiency and scalability, it has become possible to train models of unprecedented size, with over 100 billion parameters [27].

Taking into consideration these good results in language processing problems, many researchers tried to implement the same principle to image. The basic assumption of Vaswani´s is that a sentence can be dismantled into words and then apply the transformer to the words. Alexey Dosovitskiy applied the same logic to image classification, dividing each image into small patches and then applying a Vision Transformer derived from the original Vaswani´s Transformer [30].

As it depicted in Figure 2.16, the first step of Vision Transformer (ViT) is to transform an image into fixed-size patches, linearly embed each one of them and add position embeddings. Subsequently, this new data format is fed to a standard Transformer encoder [95]. Finally, in order to perform classification, we use the standard approach of adding an extra learnable “classification token” to the sequence.

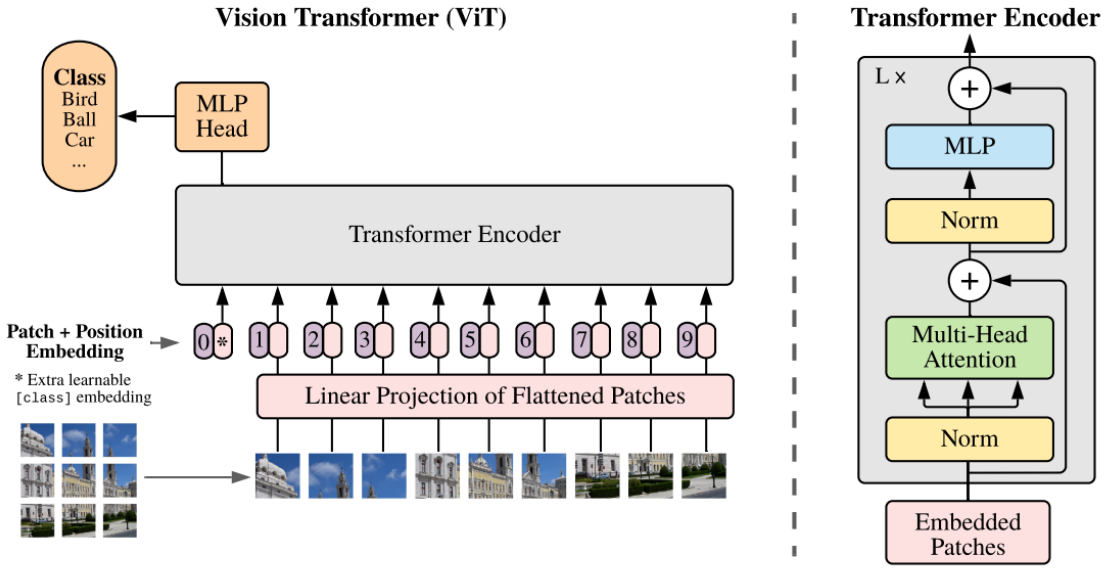


Figure 2.16: Model overview, schematic imported from Alexey *et al* [30].

Chapter 3

Facial Expressions Databases

3.1 Introduction

As it happens in most computer vision problems and especially those based on machine and deep learning algorithms, a model is just as good as the data with which it was trained. Hence, the choice of what database to use is very important. As of today, there are multiple databases for Facial Emotion Recognition purposes, including a range of different populations and environments. In this section, we select the most used databases and highlight their pros and cons.

First, it is necessary to separate the two types of data used in Facial Expression Recognition: static and dynamic. Static databases are compilations of images of people experiencing one of six emotions, this kind of database is the most common and usually produces great results. However, for detecting more complex emotions there is a need for not only an image but also a sequence of images. In those cases, the more suitable data is video sequences, also known as dynamic databases. Then, we have other two subgroups: Posed or Spontaneous. As the name indicates posed databases are full of images where the subjects were asked to make a specific facial expression. Those databases are great to make the model understand the details of each emotion. For example, representing surprise the subject should drop the jaw, raise the eyebrows and widen the eyes. However, a database which only has staged examples will not be useful to train a model that will be tested in completely different conditions. Hence, there is a need for spontaneous databases. This type of database comprises examples from the most random situations. The examples will be much noisier and will have problems regarding the illumination, face positioning and even face angles. But if the model learns from these examples, it will be much more prepared for real-life scenarios. Finally, there are two more categories directly connected to the last ones: Lab or non-Lab. This information just completes the origin of the data. Normally if the data comes from a lab environment, it is associated with posed images and non-lab environment is associated with images in real-life situations and therefore spontaneous poses. Talking more precisely about the different databases that exist for AFER purposes we select the ones that are the most used or have special characteristics which make them unique. Table 3.2 presents a summary of some of the most prominent and used databases in the literature. This table highlights the main aspects of

each database: number of samples, number of subjects, recording conditions, subject position and even the number of expressions labelled. To avoid any ambiguity, for the purpose of the presented table, ”seven basic expressions" comprise Ekman´s six different emotional states, plus the neutral expression. For some of the databases Comtempt is also present.

3.2 Available Databases

CK+ (Extended Cohn-Kanade) Database

One of the first datasets available for the study of emotion recognition, The Extended Cohn-Kanade (CK+) [65] is one of the most extensively used laboratory-controlled databases. This database contains 201 different subjects with posed and spontaneous examples. The Extended Cohn–Kanade database is the most extensively used laboratory-controlled database for evaluating FER systems. CK+ contains sequences that show a shift from a neutral expression to a peak expression. The database is a set of videos where the three frames with peak formation and the first frame of each sequence are saved.

JAFFE (The Japanese Female Facial Expression) Database

JAFFE [67], The Japanese Female Facial Expression, is a database which is only made by samples from female Japanese subjects. All the samples are staged and produced in a lab. There are 213 samples of posed expressions from 10 different subjects. Each person has three or four images acting each one of the six basic facial expressions and one image with a neutral expression. Although this database produces models with good performances, the number of samples is too small and the lack of diversity makes the models not prepared for real-life situations.

FER2013 (Facial Expression Recognition 2013) Database

FER2013 [40] is a large-scale and unconstrained database collected automatically by the Google image search API. All the thirty-five thousand plus images are grey-scale and were registered and resized to 48 by 48 pixels after rejecting incorrectly labelled frames. This dataset contains 28,709 training images, 3,589 validation images and 3,589 test images with seven expression labels - six emotions + neutral state.

The FER-2013 dataset is more challenging than the other facial expression recognition datasets. Besides the intra-class variation of FER, another main challenge in this dataset is the imbalanced nature of different emotional classes. Some of the classes such as happiness and neutral have many more examples than others.

AFEW (Acted Facial Expressions In The Wild) and SFEW (Static Facial Expressions in the Wild) Databases

These two datasets, AFEW and SFEW [29], refer to the same data but in different states. The Acted Facial Expressions in the Wild (AFEW) database contains video clips collected from different movies with spontaneous expressions, various head poses, occlusions and illuminations. AFEW is a temporal and multimodal database that provides vastly different environmental conditions in both audio and video [57]. Static Acted Facial Expression (SFEW) is the result of selecting certain frames from the videos of AFEW. These two databases are very important since the samples are faces in “real" life scenarios. Although the facial expressions are “fake" and do not reflect a real emotion (because all the samples are from actors performing in movies), the conditions, like poses, angles and lighting are similar to a day-to-day situation.

AffectNet Database

The AffectNet Database [72] aims to offer a very large platform for robust affective computing research. It includes over one million face images retrieved online, of which more than half have annotation landmarks on seven discrete emotions. Furthermore, the samples have annotations about arousal and valence. It is a robust database with not only a lot of samples but also a lot of information about the intensity of the emotion which is being described.

FERG-2D (Facial Expression Research Group 2D) Database

FERG-2D Database [8] is a particular dataset. Opposite to other databases, this one is not made up from real human faces but is made from cartoon-like faces. The database contains 55773 different images distributed by six different stylized characters (3 males and 3 females) - Ray, Malcolm, Jules, Bonnie, Mery and Aia. The characters were produced using the MAYA software and rendered out in 2D to create the images. The images for each character are grouped into seven types of expressions - anger, disgust, fear, joy, neutral, sadness and surprise. This database can be very important to see how models trained with cartoonish samples can perform when faced with real human faces

Multi-PIE (CMU Multi-PIE) Database

The Multi-Pie Database [43] contains samples from 337 subjects recorded under 15 different viewpoints and 19 illumination conditions. It comprises more than 750 thousand samples labelled with the six basic emotions. It is a research database and was built under controlled laboratory circumstances and all the recorded facial expressions are posed. It is another useful database to reach good performing models, with the perk of comprising multiple angles which induces more variability and, consequently, will produce more robust approaches.

MMI Database

MMI Database [94] has both dynamic and static samples, and, like CK+, is also completely laboratory-controlled. Besides the six basic emotions, the dataset is also annotated for individual action units according to FACS guidelines. For experiments, the most common method is to choose the first frame (neutral face) and the three peak frames in each frontal sequence.

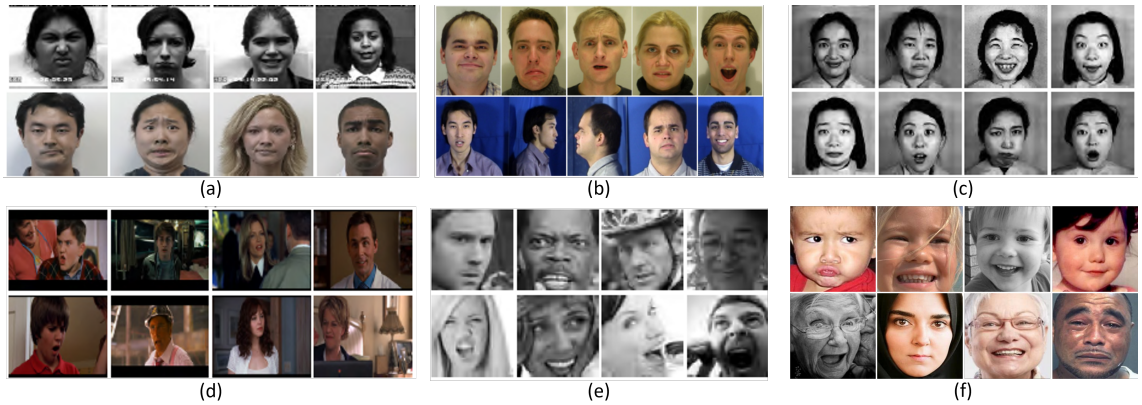


Figure 3.1: Images from six different datasets: (a) C+[65], (b) MMI[94]; (c) JAFFE[67]; (d) SFEW [29]; (e) FER-2013[40], (f) AffectNet[72]

Table 3.1: Summary of the used AFER Databases during the implementation of the models presented in this dissertation (L - Lab; D - Dynamic; St - Static; P - Posed; S - Spontaneous).

Database	Samples	Subjects	Time	Condition	Position	Expression
C+ [65]	593 images sequences	123	D	Lab	P & S	Seven basic expression plus contempt
JAFFE [67]	213 images	10	St	Lab	P & S	Seven basic expression
FER-2013 [40]	35,887 images	N/A	St	Web	P & S	Seven basic expression
AFEW [29]	1,809 videos	N/A	D	Movie	P & S	Seven basic expression
SFEW [29]	1,766 images	N/A	St	Movie	P & S	Seven basic expression
AffetcNet [72]	450,000 images	N/A	St	Internet	P & S	Seven basic expression
FERG [8]	55,773 images	6	St	Web	-	Seven basic expression
Multi-PIE [43]	755,370 images	337	St	Lab	P	Anger, Disgust, Neutral, Happy, Squint, Scream, Surprise
MMI [94]	740 images and 2,900 videos	25	D+St	Lab	P & S	Seven basic expression

3.3 Summary

In this Chapter, the different characteristics of the available facial expressions databases were highlighted and discussed. Aspects such as pose, recording conditions and even number of subjects are very important and can determine how good a database is suited for our problem.

As it was previously mentioned models are highly dependent on how good the data which they are trained with. Thus, to make models which are both robust and accurate for daily situations we need more images which can reliably describe more natural facial expressions. However, none of the databases is perfect and everyone has their limitations. For example, both AFEW and SFEW use movie samples in order to produce a more adaptable model, but due to having a high number of different posed samples, originate less accurate models; CK+ and JAFFE are the most used datasets and, though they are used to create the highest accurate models, they have fewer samples and diversity because of the reduced number of subjects.

The main conclusion we can take from this evaluation is that none of the presented databases fully satisfy the need for AFER systems. There is a need for vast databases that comprise both spontaneous and posed samples, not only in laboratory conditions but also in real-life scenarios. Nevertheless, the presented databases are the most used in the literature and it is important, not only, to replicate implemented methodology, but also to compare new results to the existing ones.

Chapter 4

State-Of-The-Art

While AFER algorithms can be constructed using multiple approaches, almost everyone follows the same global architecture which comprises three fundamental steps: Face Detection/Alignment, Feature Detection/Extraction and Emotion Classification (Figure 4.1). The next section dives deeper into each one of these stages.

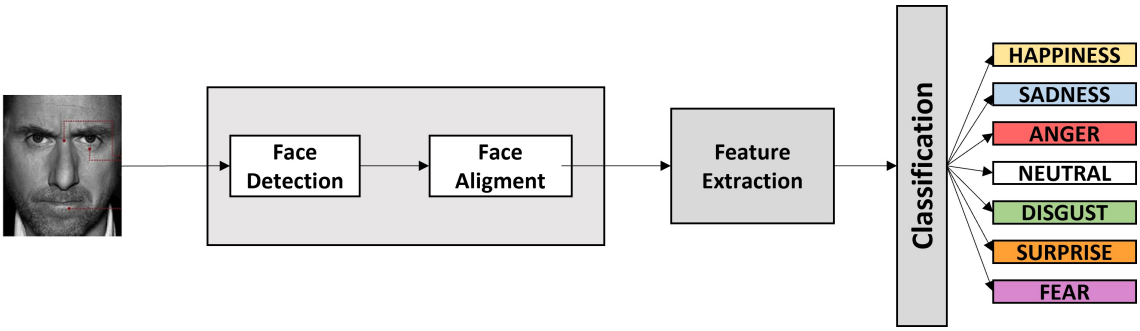


Figure 4.1: Block diagram of a typical AFER system.

4.1 Face Detection

Given an arbitrary image, the goal of face detection is to determine whether or not there are any faces in the image and, if present, return the image location and extent of each face. Naturally, automatic face localisation is the prerequisite step of facial image analysis for many applications, being AFER systems one of them. Therefore, the first step to determining which facial expression is being displayed is to find where the subject face is. Most of the time, in day-to-day situations the subject face will not be perfectly aligned nor in a perfect position for all essential features to be detected. Thus, there is a need to first track the face and, only then, continue the classification problem. The main objective of face detection is to detect the faces present on the face and return their location. Afterwards, the images are cropped to generate a new sample. This new image

will only have the subject face which will aid the model to find and extract the essential features to make an accurate prediction. Moreover, cropping the images can also help with tilted faces, because it reduces the information of the face to the vital information.

Face detection has been a prominent subject throughout the latest years, hence many different algorithms have been developed [105]. By 2001 one of the most successful algorithms and the one which had the most impact in the 2000s emerged, Viola&Jones [96]. This face detector contains three main ideas that make it possible to build and run in real-time: the integral image, classifier learning with AdaBoost, and the attentional cascade structure [104]. However, while fast, it has problems with occlusions and large pose variations. It remained state-of-the-art for many years and is still regarded as one of the best approaches for RGB images. Following the pioneering work of Viola-Jones face detector [96], most early works focus on designing robust features and training effective classifiers. These features could also be pixel-based [9], where they used the relationships between adjacent pixels to find the boundaries of faces; statistics-based features: Spatial Histogram (LBP-based) [107], a combination of Histogram oriented Gradients and Local Binary patterns (HOG-LBP) [97] or even Edge orientation [23]; finally, Haar-like features, like in Viola&Jones were also very prominent in the early works [58]. Nonetheless, these approaches depend on non-robust hand-crafted features and optimize each component separately, making the face detection pipeline sub-optimal.

In recent years, as it was previously mentioned, convolutional neural network (CNN) has achieved remarkable successes, ranging from image classification to object detection which also inspires face detection. CNNs are built to learn invariant representations of images, thus making them powerful weapons to solve some of the problems of traditional solutions, pose variations, light changes, viewing angles and even occlusions. Quickly, CNN-based models surpass the performance of traditional models and became state-of-the-art. Overall, these algorithms used the ability of convolutional layers to extract features without needing to manually define where to look. For example, in the work of Haoxiang *et al.*[56] a Cascade of CNNs is purposed, on the other hand, Najibi *et al.*[74] suggests a single-stage model which is able to detect faces directly from the early convolutional layers in a classification network.

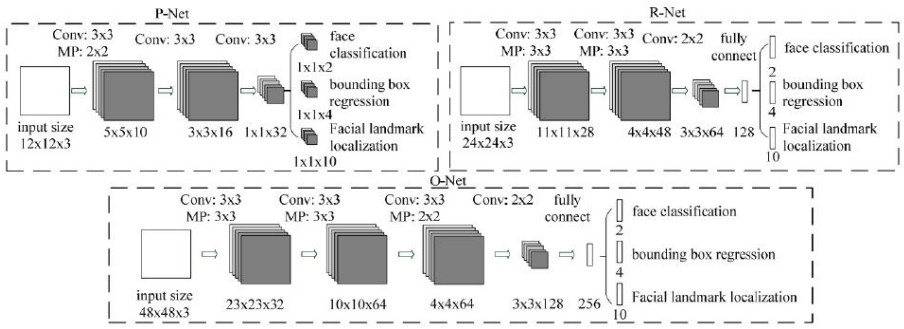


Figure 4.2: Architectures of P-Net, R-Net, and O-Net, where "MP" means max pooling and "Conv" means convolution. The step size in convolution and pooling is 1 and 2, respectively. Image from Kaipeng Zhang *et al.*[108].

Another state-of-the-art approach is MTCNN - ”Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks" [108]. These face detector model uses multiple CNNs in a cascade format to achieve unprecedented results. The model is divided into 3 different phases: (1) a fully CNN, called Proposal Network (P-Net), is used to obtain the candidate facial windows and their bounding box regression vectors. After, a non-maximum supression (NMS) is applied to merge highly overlapped candidates; (2) the second step consists in feeding the candidates into another CNN, refine network (R-Net), whose goal is to reject a large number of false candidates; (3) finally, stage three is very similar to the previous stage, but now we aim to identify face regions with more supervision. The overall architecture is described in Figure 4.2.

This model reaches a performance of 99.83% in LFW (Labeled Faces in the Wild) dataset [48] which even surpasses the human performance for the same dataset - 97.53%. For the purpose of this dissertation, MTCNN model will be used as the face detector. In Figure 4.3, it is shown the comparative performance of Viola&Jones models versus MTCNN face detector. Besides detecting more faces, MTCNN also makes fewer inaccurate predictions.

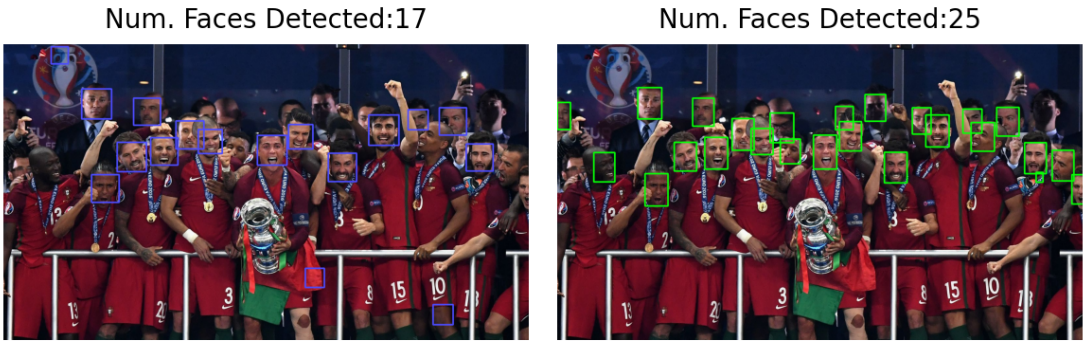


Figure 4.3: Face detection performed by Viola&Jones [96] and MTCNN [108] respectively.

4.2 Face Registration

After detecting the face, many AFER models depend on very specific information regarding our face. Thus, it is necessary to perform an intermediate process to make these points available to the model. The most common representation of face registration is the fiducial points. These key points represent very distinctive areas of our face and can be used to help the model extract features. In other words, during this phase, specific landmarks localization, for example, the eyes, the mouth or the nose are pinpointed. These points can be obtained by different processes ranging from more traditional approaches like scale-invariant feature transform (SIFT) [100] to CNNs [7]. Face registration methods only aim to locate and save the multiple existent landmarks, the ways those points can be used are shown and explained in Section 4.3.1.2.

4.3 Feature Detection/Extraction

Feature detection and extraction is the step where there is more diversity between different models. This stage is the core of all the processes of emotion recognition, if the features detected and extracted are good, the classification will probably be more accurate. Globally, feature detection and extraction can be spliced into two main categories: hand-crafted features and learned features. The first ones are associated with more traditional approaches and the second ones with deep learning methods. Also, the feature extraction algorithm can be divided into static or dynamic, depending on the type of data being used.

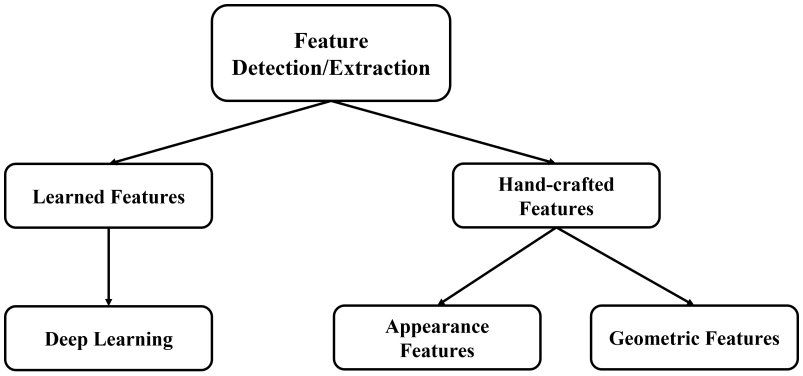


Figure 4.4: Diagram of the existent ramification for Feature Detection/Extraction stage.

4.3.1 Traditional Approaches

The majority of the AFER systems mentioned in the literature are ML (Machine Learning) based. The different approaches and techniques used mainly differ from each other during the feature extraction stage. The two main groups are Appearance Features and Geometrical Features. As it was stated before, the performance of the models created highly depends on the quality of the features chosen, the two aforementioned approaches focus on completely different types of features, though each one has its advantages and disadvantages.

As it was previously stated, hand-crafted features are associated with machine learning methods where the features are extracted transforming the original data into a new feature sub-space. The features gathered by these approaches can be divided into *Appearance Features* and *Geometrical Features*.

4.3.1.1 Appearance Features

Appearance-based methods rely on the principle that facial expressions involve a change in local texture, this change is used to create a pattern and understand which changes are associated with each facial expression [36]. Typically, a bank of filters, such as Local Binary Patterns (LBP) [86], Gabor Filters [60], Local Gabor Binary Patterns [6], Local Phase Quantization, Scale Invariant

Feature Transform (SIFT) [12] and Pyramids of Histograms of Gradients, are applied either to the whole face or specific face regions to encode the texture. (Figure 4.5)

Of all available image techniques, LBP's are probably the most used within appearance features for FER mechanism. Feng *et al.*[15] establish local LBP histograms and connect them to FER. Jabid and Chae (2010) [50] introduced an LBP-based Local Directional Pattern (LDP) algorithm, which is robust to illumination and has relatively low computational complexity. Chen *et. al* proposed a multi-model system combining audio and visual modalities. Regarding the face, an appearance descriptor named Histogram of Oriented Gradients from three orthogonal planes (HOG-TOP) was proposed to extract dynamic textures [18]. Continuing with LBP, Liu et. al[61] used both HOG and LBP features from salient areas and then used Principal Component Analysis to reduce those features to the essential. More recently, Ben Niu *et. al* (2021) used a combination of ORB and LBP descriptors in an attempt to better describe human face detection. Once again PCA was used to reduce the dimensions of the features of the fusion between both descriptors.

Moving on, Gabor wavelets are also very good feature extraction filters, though consume a lot of memory as they work on global features. Yu *et al.* [102] used linear and nonlinear synthesis of new algorithms based on Gabor feature.

Further, Kamarol *et al.* which proposed a spatiotemporal texture map (STTM) to capture subtle spatial a temporal variations [53]. Also, Karan Sikka *et al.* explored Bag of Words (BoW) as an approach for feature extraction, combining both highly discriminative Multi-Scale Dense SIFT (MSDF) features with spatial pyramid matching (SPM) [88].

Overall, appearance-based features can accurately describe human faces and comprise the needed information. However, the performance of these hand-crafted feature extraction methods decreases with illumination changes, noise variability, and changes in pose and expression conditions, making them less useful in real-life scenarios.

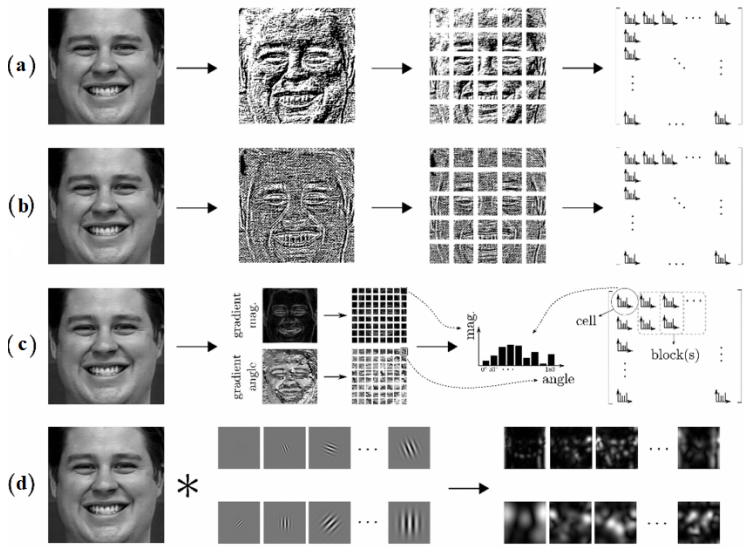


Figure 4.5: Appearance-based feature extraction examples: (a) LBP histograms, (b) LBQ histograms, (c) HOG, (d) BoW. Adapted from Sariyanidi *et al.*[85].

4.3.1.2 Geometrical Features

Geometric-based methods involve, in the first stage, the location of facial landmarks and/or some facial components (e.g., mouth, eyes, nose and eyebrows) and, then, the extraction of geometric correlations from those fiducial points. Geometric features attempt to measure distances, deformations, curvatures and other geometric properties to represent the face geometry [36]. Thus, most of the seen approaches use the active appearance model (AAM) to track the existent facial points.

One geometric approach, proposed by Ghimire *et al.*[38], defended that by dividing the whole face region into local specific regions geometric features could be detected and extracted. Using a landmark tracking sequence, they were able to record all the facial landmarks and compute the distances between them. The angle and the euclidean distances between each pair of landmarks were calculated and stored in a feature vector. Following the same thought process, Ozbey *et al.*[78] using dlib's facial landmark detector [7] detected the facial key-points and then extracted the needed measure between each landmark pair. In both papers, it is used feature selection algorithms like Sequential Forward Feature Selection (SFFS), Sequential Backward Feature Selection (SBFS) or even Principal Component Analysis (PCA) Figure 4.6 shows an example of how valuable information can be computed and extracted from a geometric point of view using only the landmark locations.

Globally, geometrical features can accurately describe faces through distances and shapes, though they lack depth and cannot be extracted from thermal data and are terrible at identifying small details like wrinkles, furrows and changes in skin texture [109].

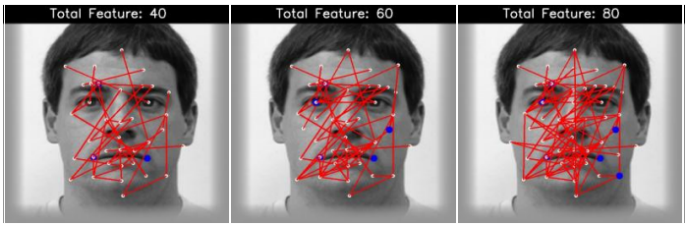


Figure 4.6: Representation of a geometric feature extraction. The blue dots indicates a feature vector extracted only from that landmark tracking result; whereas a red line connecting two landmarks indicates the feature vector extracted from that pair of landmarks tracking result, as the expression evolves over time. Adapted from Ghimire *et al.*[38].

4.3.2 Deep Learning Approaches

Traditional ML-based approaches are, as today, still more frequent and common among the literature. However, with the growth of deep learning methods in the last years, AFER systems developed using deep approaches became increasingly popular and are now the state-of-the-art for producing reliable Automated Facial Emotion Recognition algorithms.

Deep learning approaches are framed as "Learned Features" because ideally, the developed models can gather all the necessary information to make accurate predictions without needing any prior bias. In contrast to the Appearance and Geometric-based models, deep learning methods

have a way of extracting the needed features even if we don´t know what to look for. Deep learning models can range from Convolutional Neural Network (CNNs), Long Short-Term Memory (LSTM), Attentions Networks or even Generative Adversarial Networks (GANs).

The most common deep learning approach within AFER systems is, by far the CNN. In A.Lopes *et al.* a simple CNN is built by combining two convolutional layers, two sub-sampling layers and one fully connected layer. This model performs the three learning stages in just one classifier and achieves very good results [62]. Also, Deepak Jain *et al.* presents a model based on a residual network. This network uses two residual blocks each one with four convolutional layers and achieves high-performing results in multiple datasets [51].

Yu *et al.* (2018) proposed a model which comprised spatio-temporal convolutional features paired up with an LSTM classifier. Taking advantage of emotions being a dynamic process, this approach tries to extract the spational features directly by using a 3DCNN - 3D Convolutional Neural Network. These models can receive three-dimensional samples and are followed by a T-LSTM to preserve the temporal dynamic and then the convolutional C-LSTM for modelling the multi-level features [103].

Further, in 2021, S. Minaee *et al.* described an Attentional Convolutional Network based on a grid generator produced by using a Localization Network, known as a spatial transformer. This transformer, made by two convolutinal layers and two fully-connected layers tries to focus on the most relevant parts of the input image estimating a sample over the region of interest [70]. On the same note, Y.Gan *et al.* implemented a multi attention network. The model uses a region-aware sub-net (RASnet) which learns binary masks for locating expression-related critical regions. Embedded in the convolutional layers, the Multi-Attention (MA) block fuses diversified attention using the learned masks from the RASnet.

Finally, Generative Adversarial Networks (GANs) are also starting to be used in AFER systems. Yang *et al.* utilizes GAN potential and employs them to train the conditional generative models, and each of them is designed to generate one of the prototypic facial expression images. Then, a CNN is used for classification purposes [99]. On the contrary, Zhang *et al.* uses a GAN as a end-to-end model. This model incorporates a generator, two discriminators and a classifier. Using this structure, the encoder-decoder is able to learn to identify representation for face images which is then demonstratively specified through the expression and pose codes. Finally, the model can enlarge the FER training set by automatically generating face images with arbitrary expressions and head-poses [106].

One of the bigger problems that all these approaches have in common is they are highly dependent on the quality of the data given to the model: if the data is extensive and diverse the model will gather enough information to create a pattern and correctly classify emotions in multiple situations if not, the model can fail in finding the correlation between data and labels. Therefore, even if the model is well built and fine-tuned, if the data which is fed to him is poor, the model will hardly succeed.

4.4 Emotion Classification

After the right set of features has been extracted from the image they are fed to a classifier. This last stage of the pipeline performs the classification task, where the module receives the input (image or a set of features) and outputs a final prediction. Emotion Classification can be performed by a multitude of classifiers like Support Vector Machine (SVM) [76][50][78], Random Forest (RF) [73], k-Nearest Neighbor (kNN) [90] or even, deep learning-based ones, like Multi-Layer Perceptrons (MLP).

Finally, unlike traditional machine learning approaches, most, deep learning models rely on themselves to make the final classification. As it is explained in Section 4.3.2, most models already incorporate a final stage for prediction, accumulating both feature extraction and classification stages.

4.5 Summary

All in all, AFER algorithms follow a standard pipeline made by 3 fundamental steps: Face Detection/Registration, Feature Extraction and Emotion Classification. To achieve a good-performing Facial Expression classification system, we need good-performing models in each one of the stages. Face Detection/Registration models are well-established and deep learning models, like MTCNN [108], dominated the state-of-the-art with outstanding results. The Feature Extraction step portrays a fundamental role and most of the time defines how well the overall model performs. This stage was dominated by image processing approaches like LBPs, HOG or Gabor filters. However, with the development of better hardware and breakthrough within deep learning models, they lost their prominent role. In the latest years, pure deep learning approaches started to dominate the literature performing both Feature Extraction and Emotion classification tasks, using LSTMs, CNNs and ACNNs. At last, the Emotion Classification stage comprises machine learning classifiers like SVMs, RF or kNN. Also, MLPs and other deep learning classifiers are now widely used for this task. A summary of traditional and deep learning approaches is presented in Table 4.1 and 4.2.

Table 4.1: Summary of Traditional Approaches.

		Feature Extraction			
Database	**Year**	**Apperance**	**Geometric**	**Classifer**	**Performance(%)**
CK+	2021	[76]LBP		SVM	97.3
	2021	[61]LBP-HOG		SVM	98.3
	2006	[102]Gabor Wavelets		SVM	80.95
	2010	[50]LDP		SVM	96.40
	2015	[59]HOG		AdaBoost	86.00
	2016	[53]STTM		SVM	97.70
	2018		[73]Landmarks	RF	90.0
	2021		[78]Landmarks - SFFS	SVM	89.5
	2021		[78]Landmarks - SBFS	SVM	93.5
JAFFE	2021	[76]LBP		SVM	86.7
	2021	[61]LBP-HOG		SVM	89.6
	2010	[50]LDP		SVM	85.4
	2015	[59]HOG		SVM	89.5
	2005	[10]Gabor		SVM	85.8
	2007		[90]Landmarks	kNN	90.76

Table 4.2: Summary of Deep Learning Approaches.

Database	**Year**	**Approach**	**Performance(%)**
CK+	2017	[62]CNN	96.76
	2019	[51]CNN	93.24
	2018	[103]STC-NLSTM	99.8
	2021	[70]Spatial Transformer	98.0
	2020	[37]Multi-Attention	96.28
	2018	[99] GAN	96.57
JAFFE	2017	[62]CNN	82.10
	2019	[51]CNN	95.23
	2021	[70]Spatial Transformer	92.8
SFEW	2015	[55]CNN	54.56
	2017	[71]CNN	47.7
	2018	[106]GAN (End-to-End)	26.58

Chapter 5

Methodology

The work developed for this dissertation is separated into two categories: the first experiments dwell in more traditional approaches based on machine learning feature extracting and classification methods; the second experiments move to deep learning methods. Within the deep learning methods, the experimental work is once again divided by the type of data used to train them.

Summarily, the first section is a detailed description of the multiple alterations made to the original databases. These changes range from face detection models to the process of creating image samples focusing on facial fiducial points, also known as landmarks (Section 5.2). Then, a Support Vector Machine (SVM) and a Multi-layer Perceptron (MLP) were trained with multiple handcrafted features (Section 5.3). Additionally, Convolutional Neural Networks (CNNs) build from scratch and based on transfer learning were developed (Section 5.4.1 and Section 5.4.2). Finally, a deep learning model based on an attention mechanism was built (Section 5.5). It should be noted that the deep learning approaches are not presented in chronological, as the important evaluation of the models is made during the next chapter. All the methodology described is summarized in the following diagram, Figure 5.1.

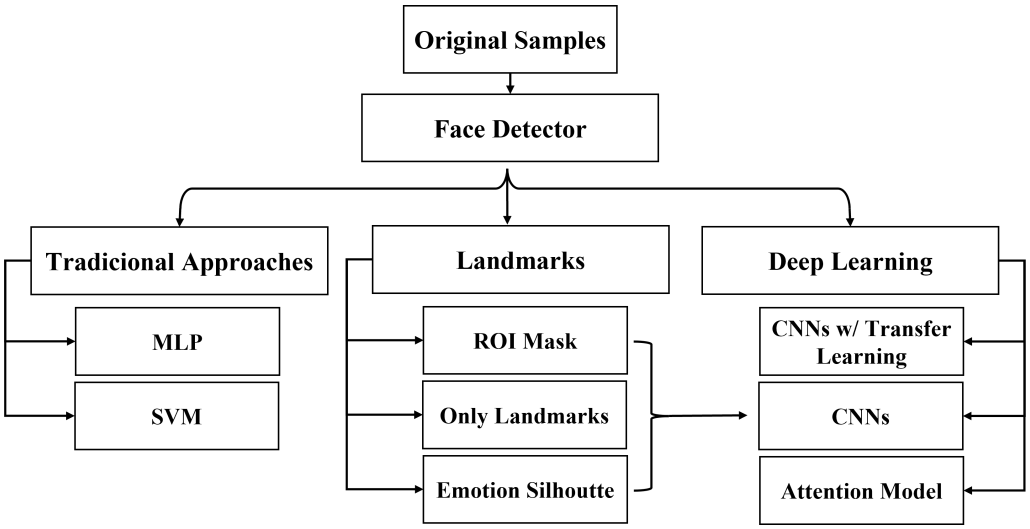


Figure 5.1: Summary of all developed approaches.

5.1 Evaluation Framework

This section describes the evaluation framework defined to train and evaluate the created models. Taking into consideration the different nature of the developed models, we used two different frameworks. The training phase consists in feeding a fraction of the available data to the model in order to train it. This data, in our case, corresponds to a combination of images samples and the corresponding label. The goal is to make the model robust and familiarized with the data to be able to make a prediction of the labels regarding similar samples. At the end of each training iteration, a validation step is performed. During this phase, a small set of unseen data is fed to the model in order to evaluate its performance. The validation phase is often used to optimize the models and to choose the parameters which produce the best model. For the purpose of this dissertation, the validation phase is performed differently for different approaches. For the approaches regarding deep learning methods, a cross-validation algorithm was chosen. In this validation method, the training and validation sets are defined accordingly with the number of folds chosen. Figure 5.2 shows an example of a Cross-Validation with 3 folds, in each iteration, two-thirds of the data are used for training and one-third for validation. On the other hand, for traditional machine learning classifiers, we chose a nested cross-validation algorithm which is used to optimize the parameters of the models. This technique is a combination of two loops: outer loop and inner loop. The first loop (outer loop) is similar to standard cross-validation and the second loop happens inside each iteration of the first loop. The second loop is used to find the best parameters for the model being trained. This inner loop also divides the data into folds and uses a portion for training and another for validation. The inner loop is complete when all the combinations of parameters are tested and the best model is found for that specific data (Figure 5.3).

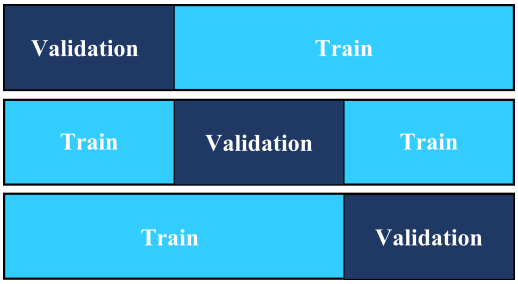


Figure 5.2: Example of a Cross-Validation with 3 folds.

Finally, the testing phase is the same for every approach. During this phase, a set of data, which has not been shown to the algorithms during the training or the validation processes, is used to evaluate the performance of the models.

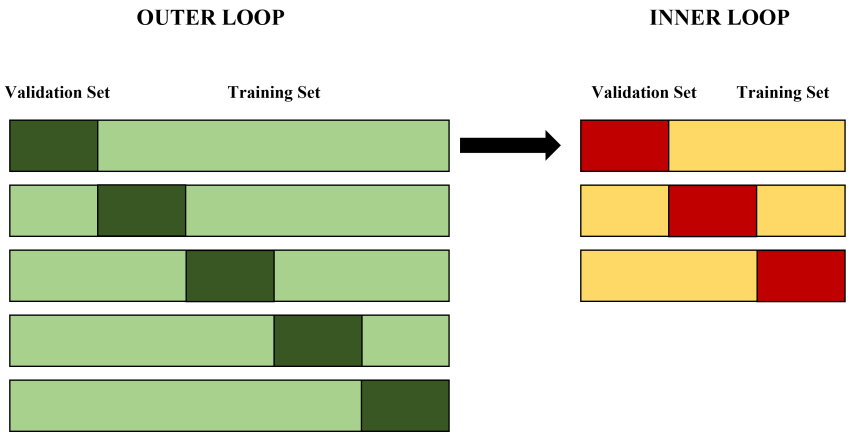


Figure 5.3: Example of a Nested Cross-Validation with 5 folds in the outer loop and 3 folds in the inner loop.

5.2 Dataset Preparation

5.2.1 Face Detector

As was highlighted in Section 4.3, face detection is the first step of pre-processing. The face detection algorithm used was created based on the MTCNN face detector [108]. Using this algorithm we were able to see with how much precision the model detected the existent faces in the samples and then we were able to crop the samples down to the essential. As it is possible to see in Figure 5.4, the pre-processing reduces the image transforming our samples into cropped versions of themselves. It is important to mention that this method is meant to improve our results helping the models ignore non-essential information like the background surrounding the human faces. However, the algorithm was not able to detect faces in every sample. In the first two databases - CK+ [65] and JAFFE [67] - every sample had a face detected and was cropped, but in SFEW database the number of samples decreased. Only 1322 from 1766 samples had a face detected and therefore these were samples used for the rest of the experimental work. All the cropped samples were resized to $224 \times 224 \times 3$ images.

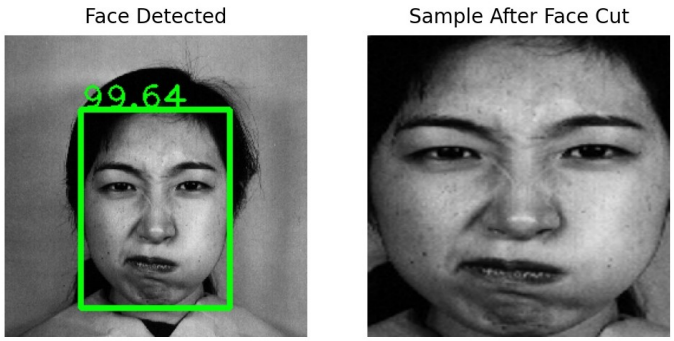
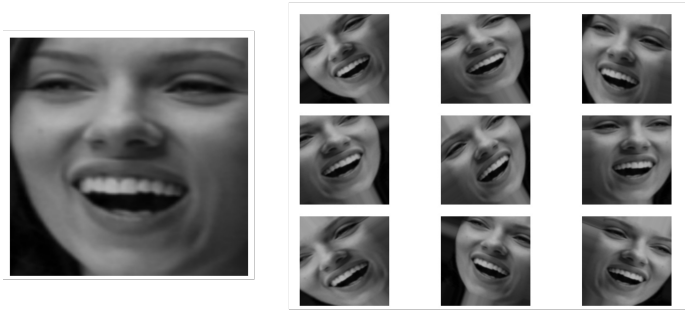


Figure 5.4: Face detector computation. Example of a JAFFE Database [67] sample detected and cropped.

5.2.2 Data Augmentation

Additionally, for some of the data used on deep learning models, a data augmentation technique was implemented. Using height and height shifts, zoom in and out, a shear factor and even rotations up to a 40 degree angle, several new samples were forged. This process induces sample variability, making our dataset more diverse and able to produce more robust and accurate models. Figure 5.5 illustrates one example of a sample from SFEW Database [29], a single image can be transformed and originates, in this case, 9 seemingly different samples.



(a) (b)

Figure 5.5: Data augmmentation example: (a) original image from SFEW Database [29]; (b) 9 different examples produced from the original sample from SFEW Database [29].

5.2.3 Landmarks

Exploring the premise that landmarks play an essential role in emotion classification, stated in Section 2.1, we developed a new set of samples based on the original ones but focused on the landmark key points. Inspired by existent literature papers [7][82][34], we developed a facial landmarks detection function based on Google's MediaPipe Python package [66][4]. This package comprises multiple functionalities ranging from simple Face Detection to Iris Detection to even Pose Estimation. However, we choose this package because of its uncanny ability to detect 468 different landmarks. Besides, the framework which outputs 68 landmarks [7] had already been used for Automated Facial Expression Recognition (AFER), but this fairly recent package from Google has not yet been incorporated in AFER models.

To develop all the necessary functions to develop landmark-oriented samples, first, we needed to comprehend how the Face Mesh was designed and how the key points were outputted. For each individual image, the Face Landmark Model detects if a human face is present in the image and then outputs 468 relative coordinates, one for each key point. Although the model outputs a constant number of coordinates, assuming that a face is detected, the indexation of those coordinates does not follow any standard order, like left-to-right or up-and-down. The way the key points are indexed is consistent throughout samples, which means, for example, that the coordinate for the tip of the nose will always be associated with the same index number in different images. However, if the order of indexation is not known it is impossible to associate each coordinate to a specific

facial area, and consequently is impossible to work with particular landmarks. To solve this problem, an automatic labelling mechanism was made. Using a random sample as ground truth, the index number of each of the 468 key points was drawn on the picture. This solution gave us the possibility to visually evaluate which was the order of indexation. As it is normal, 468 numbers written in a very limited space would be limited to draw any kind of conclusion, thus the coordinates were divided in sets of 10 and then were saved in 47 different images, Figure 5.6 shows some of those images. This way it was possible to know which was the index number of the key point allocated to a certain region of the face. Finally, we conclude that the indexation order was area oriented, but did not respect any standard continuity, for example, first was the nose bridge then the left eye and then the upper right lip and so on.

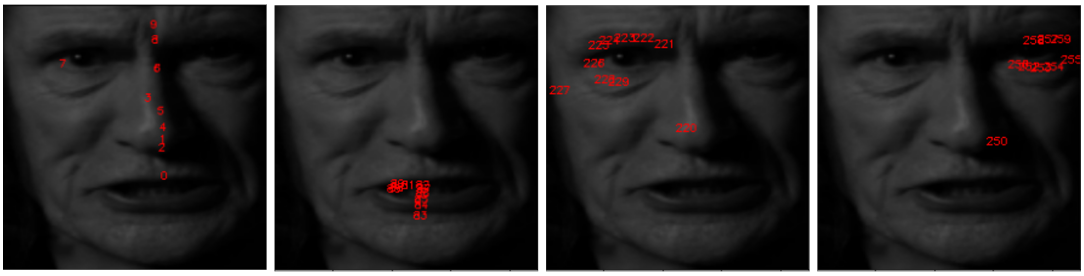


Figure 5.6: Landmark Numbering examples: (a) [0-9], (b) [80-89], (c) [220-229], (d) [250-259].

Using the landmarks key points and, now knowing how to choose a specific landmark, we were able to develop three different approaches: *Region Of Interest Mask*, *Only Landmarks* and *Emotion Silhouette*.

5.2.3.1 Region Of Interest Mask

The *Region Of Interest* (ROI) *Mask* was our first approach, our main goal using the available key points was to make our samples more landmark oriented, but influence the information that the image could give the model. As was previously stated, standard AFER models tend to focus on specific regions like the eyes, mouth or even nose to make a more accurate classification. Therefore, our objective was to isolate those areas using a binary mask.

The mask was custom-built for each one of the samples to make it more personalized and invariant to pose shifts. Also, the developed mask was a combination of two orthogonal rectangles, one vertical from the eyebrow until the mouth and the other horizontal covering the region of the eyes. These rectangles were drawn depending on the relative location of 5 different key points.

The process of developing the mask and applying it to the image was divided into 3 main stages:

1. Verify if a face is detected and collect all 468 key-point coordinates
2. Create a binary mask
3. Merge original sample and binary mask

Table 5.1: Landmark Indexes for *Region Of Interest Mask*.

Region Of Interest Mask	**Landmark Indexes**
Horizonal Rectangle	[71, 68, 346]
Vertical Rectangle	[66, 431]

The key step is the creation of the custom-built binary mask for each image. After collecting all the coordinates for the facial landmarks we had to choose which key points to use to grasp all the fundamental areas. Using multiple random samples from different datasets, the coordinates with the index 66, 68, 71, 346 and 431 were chosen to defined the binary mask (Table 5.1 and Figure 5.7).



Figure 5.7: Three examples from CK+ Database with the chosen landmarks for the Mask.

The rectangles are defined by the upper right corner and the lower-left coordinates, thus our two perpendicular quadrilaterals were defined by 4 tuples of relative coordinates. For the horizontal rectangle, the right corner was defined by the x value of the 71st landmark and y value of the 68th landmark; the lower-left corner was defined by the x and y value of the 346th landmark. Regarding the vertical rectangle, the upper right corner was defined by the tuple from the 66th landmark and the other corner was defined by both values from the 431st landmark. The rectangle was drawn and filled using white, RGB (255,255,255), colour and, finally, the whole mask was normalized, being the final result a 224×224 image made by 0 and 1 pixels.

The last step was merging the original image and the binary mask. To match the same image size as the original image ($224 \times 224 \times 3$), each binary mask was stacked with two copies of itself, resulting in a three-channel image, changing the mask size from 224×224 to ($224 \times 224 \times 3$).

The described method is summarized in Figure 5.8, illustrating the process from the original image to collecting all the landmarks and creating the binary mask, until the final image.

5.2.3.2 Only Landmarks

The first approach was based on the assumption that to make a good prediction the model does not need the complete image of the face. In fact, reducing the available information to the minimum could improve the chances of the model capturing the essential portions of the image and lead to better results. The *Only Landmarks* (OL) approach has somewhat the same principle but is applied in a very different way. Instead of reducing the quantity of information given to the model by cropping the image using a mask, this method uses all the available landmarks and draws



Figure 5.8: *Region Of Interest Mask* approach: (a) Original Image; (b) Indexes of Landmarks (c) Binary Mask; (d) Final Image.

small circles in each one of them. Creating this way a facial representation using only the facial landmarks. The goal is to try to create a representation of the facial expression using only circles placed in specific places.

The final result is a $224 \times 224 \times 3$ image having white circles drawn in each one of the key-point coordinates, Figure 5.9.

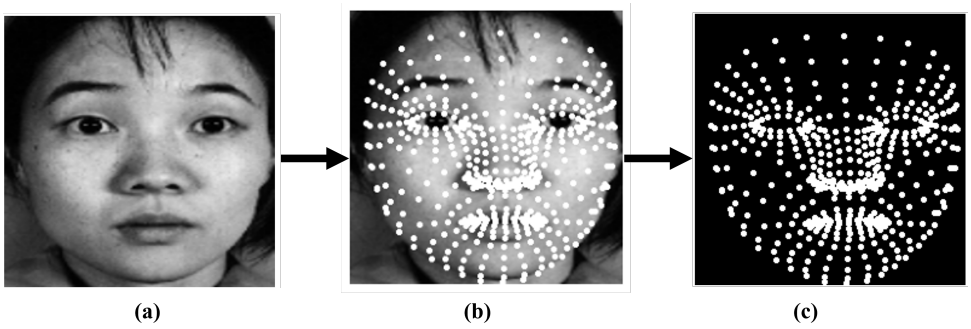


Figure 5.9: *Only Landmarks* approach: (a) Original Image; (b) Detection of the Landmarks; (c) Final Image.

5.2.3.3 Emotion Silhouette

The last approach, *Emotion Silhouette* (ES) is an attempt to mix both previous premises and create images that can, with the bare minimum information, accurately represent a specific emotion. The goal is to create a representation which grasps the most fundamental points of the face without needing the actual image information. This portrayal, for the purpose of this dissertation and throughout this document, is denominated as the *Emotion Silhouette*. The common definition of silhouette does not entirely match our created samples, however, in order to simplify communication, this was the denomination chosen.

The process of developing the *Emotion Silhouette* and applying to the image was divided in 3 main stages:

1. Verify if a face is detected and collect all 468 key-point coordinates
2. Choose specific key-points

3. Create a new image and design the silhouette

As it happens in the binary mask method, the key step is choosing which of the 468 key points should be used. Using the standards from the 2011 ISO/IEC *Information Technology - Biometric Data interchange formats - Part 5: Face Image Data* [49], we were able to sort out which are the most prominent landmarks for defining a human face (Figure 5.10).

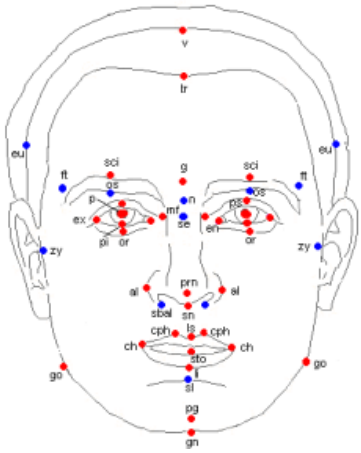


Figure 5.10: Anthropometric landmarks defined by ISO/IEC 19794-5 [49].

According to the ISO/IEC 19794-5 documentation, we need a minimum of 29 fiducial points to define a human face, however, these points would not be enough to capture the details inherent to facial expressions. Hence, for this approach, a higher number, 70, of landmarks were chosen (Table 5.2). Nonetheless, the key points were picked taking into consideration their similarity to the ones displayed in ISO example. Moreover, and in order to induce a more fluid and natural appearance, we decided to pick a higher number of points and connected them in certain regions, like the eyes and mouth. Crossing all the available information, we were able to choose the best indexes for the creation of our silhouette. The indexes of the landmarks used are shown in Figure 5.11 and Table 5.3.

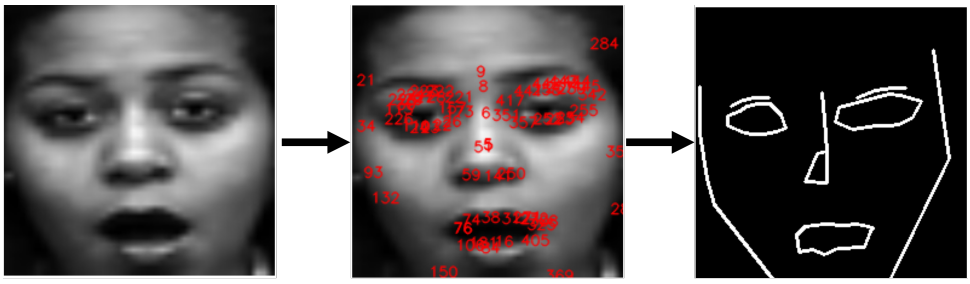
Table 5.2: Number of key-points used in ISO Documentation and in *Emotion Silhouette* images.

Facial Areas	**Number Points ISO**	**Number Points Silhouette**
Nose Bridge	1	4
Left Eyebrow	1	5
Right Eyebrow	1	5
Left Eye	6	13
Right Eye	6	13
Nose	4	5
Jaw Line	3	11
Mouth	7	14
Total	**29**	**70**

Table 5.3: Landmark Indexes used for each Facial Area in *Emotion Silhouette* images.

Facial Areas	**Landmark Indexes**
Nose Bridge	[5,6,8,9]
Left Eyebrow	[225, 224, 223, 222, 221]
Right Eyebrow	[441, 442, 443, 444, 445]
Left Eye	[28, 27, 29, 30, 113, 226, 110, 24, 23, 22, 26, 157, 173]
Right Eye	[417, 351, 357, 252, 253, 254, 255, 342, 259, 257, 258]
Nose	[5, 51, 59, 141, 250]
Jaw Line	[21, 34, 93, 132, 150, 152, 175, 369, 352, 284, 288]
Mouth	[76, 106, 181, 84, 16, 405, 325, 308, 272, 271, 312, 38, 74, 76]

After having the specific indexes, the final step was to build our image. After a black image is created, a combination of lines are drawn respecting the landmarks chosen. The nose, eyes and mouth are closed in an attempt to make the image more ”human like. Finally, it is important to mention that each one of the images is custom built and dependent on its facial landmarks, however, the indexes chosen are the same for every sample. Thus, some of the created samples could have a more accurate representation than others.



(a) **(b)** **(c)**

Figure 5.11: *Emotion Silhouette*: (a) Original Image; (b) Landmarks chosen to delineate the silhouette; (c) Final Image.

The final result of the *Emotion Silhouette* is a $224 \times 224 \times 3$ binary representation of a human face and respective expression (Figure 5.11)

5.3 Handcrafted Features

In this section, the proposed methodologies for handcrafted features are described in detail. Local Binary Patterns (LBP) and Histogram Oriented Gradients (HOG) were chosen as our image descriptors. Both textural descriptors are fully explained in Chapter 2.

5.3.1 Local Binary Pattern

During the feature extraction step, every sample is transformed from a 224×224 cropped image, resulting from the explained in Section 5.2, into a feature array. In order to produce better results the Local Binary Pattern(LBP) descriptor is created from a solo patch and then a final vector is made from all the LBP histograms, one from each patch.

The feature extraction method is divided into several phases. The first step was to resize the images from 224×224 to 128×128. This change is necessary to make the division of the image into patches easier. The following step was to divide the image into 8 by 8 patches, giving 16 symmetrical squares per row, hence 256 equal segments in total. Henceforth, an LBP with a radius equal to 1 and a neighbourhood made by the surrounding 8 pixels was applied to each individual patch. Using these specifications, the LBP applied to each 16×16 region resulted in a 10 bin histogram for each one of the 256 patches. At last, the array for the training phase is constructed, and the 256 normalized histograms are chained together giving origin to a 2560 array. Each array is combined with the respective label, numbers from 0 to 6, corresponding to the facial expression being displayed in the sample.

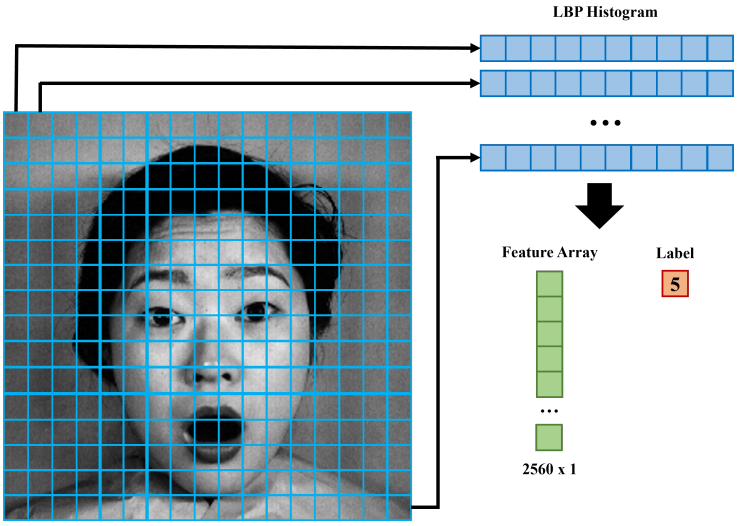


Figure 5.12: Example of a SFEW [29] sample transformed into a LBP Histogram.

5.3.2 Histogram Oriented Gradient

The feature extraction process for Histogram Oriented Gradient (HOG) is slightly different from the previous one. In this case, the HOG descriptor is applied directly using the specifications

shown in Table 5.4. Initially, the gradient is obtained by combining magnitude and angle from the image considering a block of 3×3 pixels, first Gx and Gy are calculated for each pixel. After obtaining the gradient of each pixel the gradient matrices (magnitude and angle matrix) are divided into 8×8 cells to form a block. For each block, a 9-point histogram is calculated. The final result is a 54756 by 1 array representing the HOG descriptor of one sample.

Table 5.4: HOG Descriptor Parameters.

Parameter	Value
Orientations	9
Pixels Per Cell	(8,8)
Cells Per Block	(3,3)

5.3.3 Classifiers

After manually extracting LBP and HOG features from the images, two different classifiers were trained with them: SVM and MLP. (Figure 5.13) These models were chosen to evaluate the performance of machine learning classifiers versus deep learning ones using the same input data.

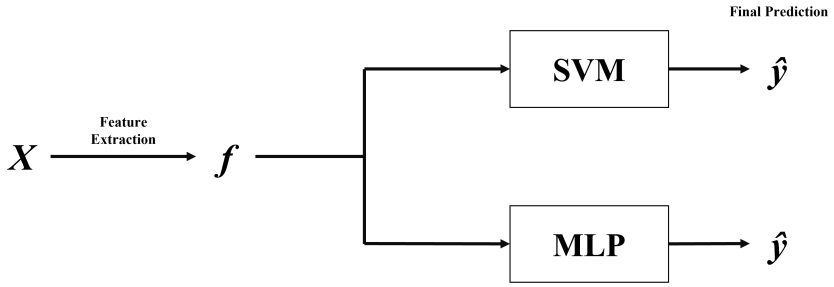


Figure 5.13: Diagram of MLP and SVM classifier. X represents an image, f the corresponding handcrafted features and $\hat{y}$ the predicted class.

- **Support Vector Machine** The SVM classifier, whose functioning is fully dissected in Chapter 2, is one of the most common classifiers in the literature. Moreover, these classifiers has been proven to be one of the most versatile machine learning classifier. Therefore, it was the obvious choice to be trained with our handcrafted features. The input of the SVM classifier is a combination of the feature array and respective classification label. The model hyperparameters were chosen using nested cross-validation, 10-fold cross-validation for the outer loop and 3-fold for the inner loop. The combination of parameter values which led to the best accuracy were the ones then used to re-train the model and, finally, test it. For the SVM classifier, the parameters considered in the optimization process are shown in Table 5.5.

- **Multi-layer Perceptron** The other classifier used for handcrafted features was a Multi-layer Perceptron. Just like SVM, MLP underwent a process of optimization, the type of architecture remained the same, which means that every MLP tested had 3 hidden layers and

Table 5.5: SVM parameters used during optimization step.

Parameter	Value
Kernel	"rbf", "poly, "sigmoid"
Regularization parameter	0.1, 1, 10, 100
Gamma	1, 0.1, 0.01, 0.001, 0.0001

produced a 7-class output. However, the size of hidden layers was changed to find which combination would produce the best results. Also, the type of optimizer, activation function and alpha, which defines the strength of the L2 regularization term, were also parameters which were changed through the optimization process, the learning rate was set at 0.001 (Table 5.6). In contrast with SVM, MLP outputs a probability for each one of the classes, the final prediction is the one with the highest probability. In the end, the most suitable MLP for each data feature (LBP and HOG) and for each database was selected.

Table 5.6: MLP parameters used during optimization step.

Parameter	Value
Hidden Layer Sizes	(150,100,50), (120,80,40), (100,50,30)
Activation	"tanh", "relu"
Optimizer	"Adam", "SGD"
Alpha	0.01, 0.001, 0.0001

5.4 Deep Learning - Convolutional Neural Network

In this section, the potential of deep learning methods as both feature extractors and classifiers is explored. For the purpose of this dissertation, deep learning approaches can be divided into two different stages: CNNs with and without transfer learning. To facilitate the creation of the models and keep all the methods less ambiguous, all the developed CNNs have input data images with shape $224 \times 224 \times 3$.

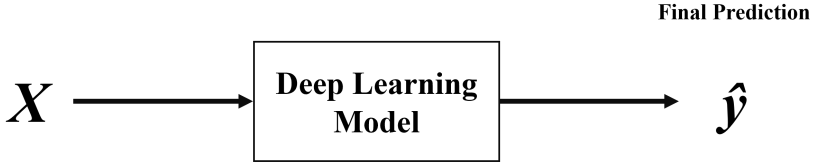


Figure 5.14: Diagram of generic deep learning approach. X represents an image and $\hat{y}$ the predicted class.

At last, it is necessary to mention, that all the developed networks, TransVGG16, myCNN and myVGG, were trained using all the developed types of samples. Besides being trained with the original samples which were submitted to a face detector, the aforementioned models were

developed using the following samples: *Region Of Interest* (ROI) *Mask*, *Only Landmarks* (OL) and *Emotion Silhouette* (ES).

5.4.1 Convolutional Neural Network with Transfer Learning

The first attempt at working with pure deep learning methods was made by Transfer Learning. In a few words, Transfer Learning is a machine learning technique where a model trained on one task is re-purposed on a second related task. This process enables the used of huge and difficult to train deep learning models like VGGs [89], ResNets [47] and even InceptionNets [92]. These deep networks are pre-trained on ImageNet Scale Visual Recognition Challenge dataset [26], an enormous image database, and are suitable for image classification tasks.

This approach's objective was to use some of the most well-performing models trained with ImageNet weights and adapt them to output a seven-value vector, as is portrayed in Figure 5.15.

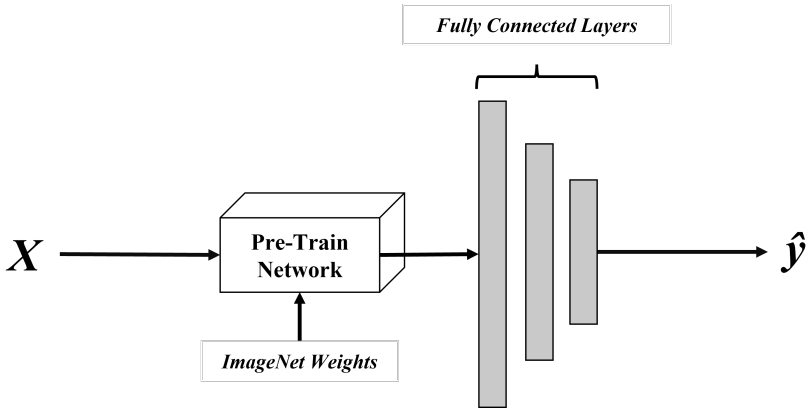


Figure 5.15: Diagram of generic transfer learning approach. X represents an image and $\hat{y}$ the predicted class.

The first pure deep learning approach was made by using the VGG16 network. Very Deep Convolutional Networks or VGGs [89] are famous for having very good performance in image classification, thus being a prime candidate for AFER systems.

VGG16 networks comprise 16 weighted layers, 13 convolutional plus 3 fully connected layers, and were originally built to output 1000 values. Hence, the last original 3 layers were removed and 2 new fully connected layers plus a global average pooling were added to adapt it to our problem. In order to preserve the weights of the already trained VGG, all the layers were frozen and only the fully connected layers were trainable during the training phase. The final model is VGG16 trained with ImageNet database with frozen weight plus 2 fully connected layers - [512,7]. Dropout regularization with a value of 0.4 was also added after the first fully connected layer.

The developed network architecture is illustrated in Figure 5.16 and described in Table 5.7. Table 5.7 also comprises all training parameters used for these method.

Table 5.7: TransVGG16 architecture details and training parameters.

Architecture	**Values**
Number Convolutional Layers	13
Convolution Filter	[64,64,128,128,256,256,256, 512,512,512,512,512,512]
Convolutional Kernel	[3,3,3,3,3,3,3,3,3,3,3,3,3]
Number Max Pooling Layers	5
Max Pooling	kernel=(2x2); stride 2
Fully Connected Layers	[512,7]

Training Parameters	**Values**
Number of Epochs	100
Batch Size	32
Loss Function	Sparse Categorical Cross Entropy
Optimizer	Adam
Learning rate	10^{-4}
Weight Decay	10^{-4}

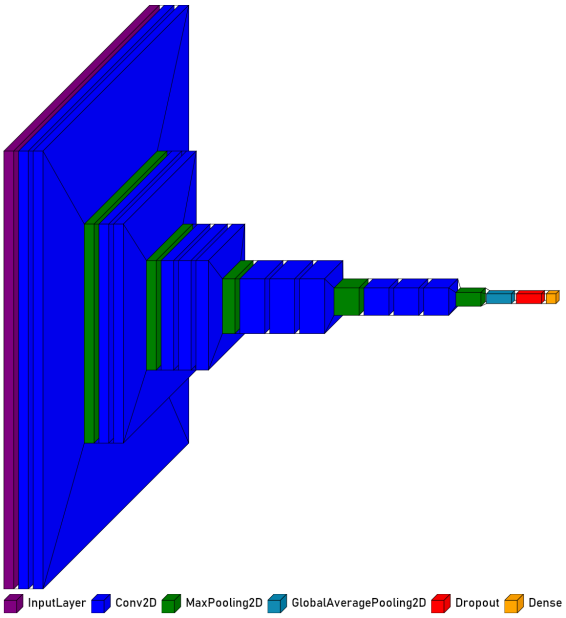


Figure 5.16: Illustration of the TransVGG16 Architecture.

5.4.2 Convolutional Neural Network

5.4.2.1 First Convolutional Neural Network - myCNN

The major objective of the first pure deep learning method, First Convolutional Neural Network (myCNN), is to test the potential of a simple combination of convolutional and pooling layers trained from scratch. The architecture of this CNN is schematized in Figure 5.17 and displayed in Table 5.8. Also, its training parameters are summarized in Table 5.8.

Table 5.8: CNN Baseline architecture details and training parameters.

Architecture	**Values**
Number Convolutional Layers	5
Convolution Filter	[32,32,64,128,256]
Convolutional Kernel	[3,3,3,3,3]
Number Max Pooling Layers	4
Max Pooling	kernel=(2x2); stride 2
Fully Connected Layers	[1024,7]

Training Parameters	**Values**
Number of Epochs	150
Batch Size	32
Loss Function	Sparse Categorical Cross Entropy
Optimizer	Adam
Learning rate	10^{-3}
Weight Decay	10^{-4}

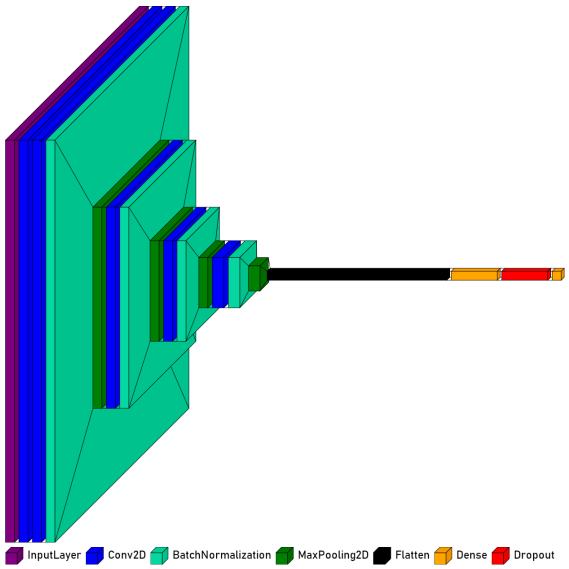


Figure 5.17: Illustration of the myCNN Architecture.

5.4.2.2 Second Convolutional Neural Network - myVGG

In the second approach regarding convolutional layers trained from scratch, we try to develop a deeper network compared to the first attempt to see if more layers would lead to better performance. Motivated by the results from the transfer learning models mixed with knowing that a network trained from scratch would be more adaptable to our data, a VGG-inspired network was built.

However, due to hardware limitations, the developed CNN is a smaller and personalized version of the traditional VGG19, the more complex VGG available. Instead of starting with 64

convolutional layers, the first two layers are convolutional with filters of size 32. Then, instead of the 16 convolution layers, we used the first 8 and the last 4 convolutional layers. Also, 5 max-pooling layers and 2 fully connected layers were added, being the last one adapted for our 7 classes. Despite the original VGG19 not having batch normalization layers, we added 4 of them, one after each one of the max-pooling layers. Also, before the first fully connected layer a flatten operator was used. The complete characteristics of this model, as well as its training parameters, are summarized in Table 5.9.

Table 5.9: myVGG architecture details and training parameters.

Architecture	**Values**
Number Convolutional Layers	14
Convolution Filter	[32,32,64,64,128,128,256,256, 256,256,512,512,512,512]
Convolutional Kernel	[3,3,3,3,3,3,3,3,3,3]
Number Max Pooling Layers	5
Max Pooling	kernel=(2x2); stride 2
Fully Connected Layers	[2048,7]

Training Parameters	**Values**
Number of Epochs	100
Batch Size	32
Loss Function	Sparse Categorical Cross Entropy
Optimizer	Adam
Learning rate	10^{-4}
Weight Decay	10^{-4}

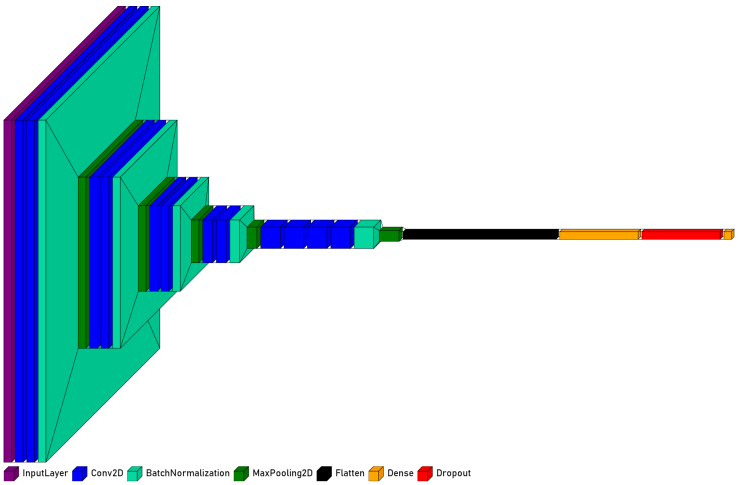


Figure 5.18: Illustration of the myVGG Architecture.

5.5 Deep Learning - Attention Mechanism

Attention mechanisms, as described in Section 2.2.3.3, are a fairly recent emerging type of Recurrent Neural Network (RNN). For our purpose, a specific type of model based on attention, called Vision Transformer, was developed.

The Vision Transformer (Vit) developed was composed by 8 attention modules followed by a layer normalization step, then a flatten layer with a dropout of 0.5. This transformer outputs for each image the respective tensor - [*batch_size*, *num_patches*, *projection_dim*]. Taking this tensor as input, an MLP head, made by two hidden layers with sizes 2048 and 1024 and 0.5 of dropout rate, and a final dense layer, with the same size as the number of classes, with "relu" as activation function, are used to output the final result.

Regarding the attention modules, which are the crucial step of the developed model, they are based on a Multi-Head Attention layer with 4 heads. The attention mechanism performed by these modules is described as being self-attention, as both inputs of this layer are the encoded patches. At last, it is crucial to mention that before being fed to the attention model each image was resized and divided into patches. To find the optimal model several patch sizes were tested (Table 5.11). Starting with resizing the image from a 224×224 to a 72×72 and having patches with size 6; to then try patches with sizes 12, 16 and 32. Then the variable number of patches were flattened and projected onto a vector of size 64. Further, a positional encoding was performed and a learnable positional was added to each one of the projected vectors, Figure 5.19.

Both ViT and Attention Model architecture are shown in Figure 5.20. Also, all the parameters used are displayed in Table 5.10.

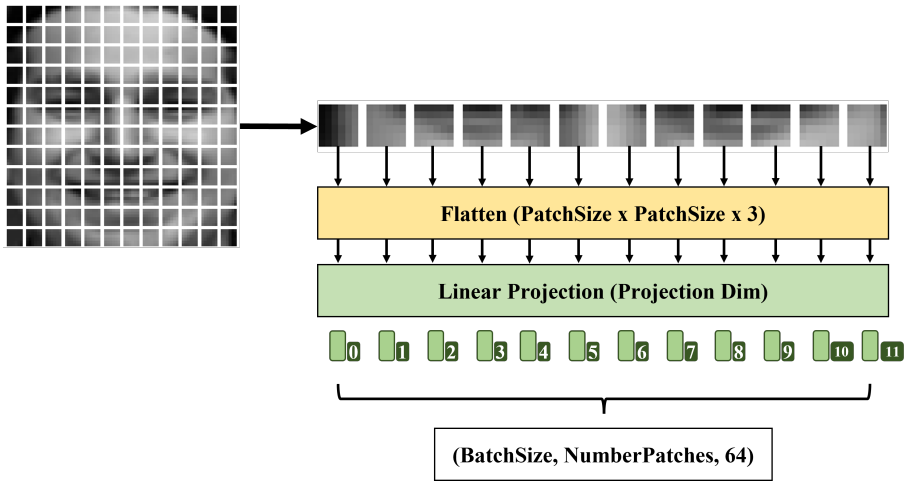


Figure 5.19: Illustration of Positional Encoding.

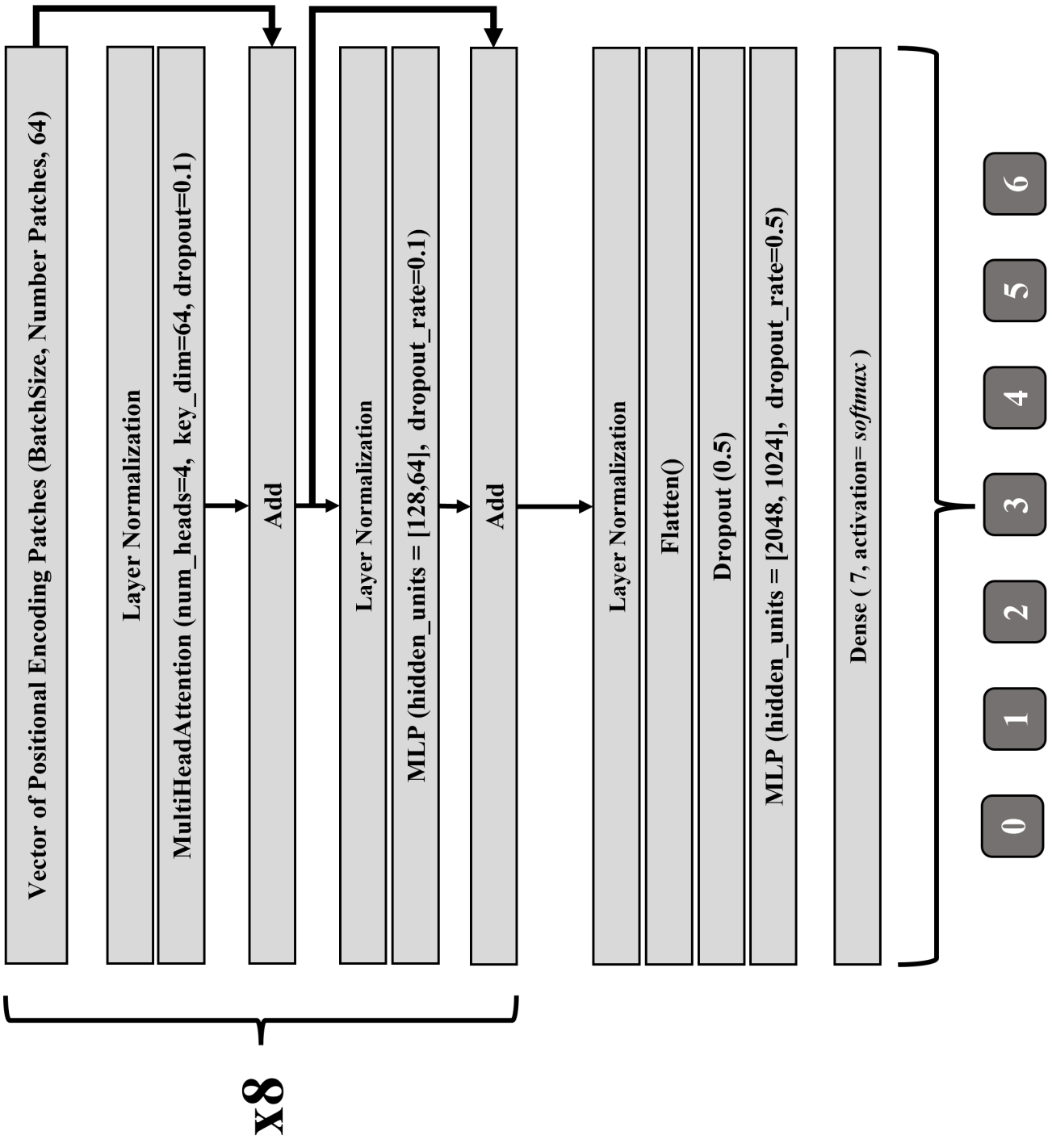
Table 5.10: Attention Mechanism Model architecture details and training parameters.

Architecture	**Values**
ViT Transformer Layers	8
ViT MLP Hidden Layers	[128, 64]
Projection Dim.	64
Head MLP Hidden Layers	[2048, 1024]
Fully Connected Layers	2

Training Parameters	**Values**
Number of Epochs	150
Batch Size	64
Loss Function	Sparse Categorical Cross Entropy
Optimizer	Adam
Learning rate	10^{-3}
Weight Decay	10^{-4}

Table 5.11: Attention Model parameters to optimize

Parameter	**Values**
Image Size	[72, 224]
Patch Size	[6, 12, 16, 32]


Vector of Positional Encoding Patches (BatchSize, Number Patches, 64)
Layer Normalization
MultiHeadAttention (num_heads=4, key_dim=64, dropout=0.1)
Add
Layer Normalization
MLP (hidden_units = [128,64], dropout_rate=0.1)
Add
x8
Layer Normalization
Flatten()
Dropout (0.5)
MLP (hidden_units = [2048, 1024], dropout_rate=0.5)
Dense (7, activation= softmax)
0
1
2
3
4
5
6

5.6 Summary

The methodology presented in this dissertation addresses traditional image processing techniques allied with machine learning classifiers, proven good performing deep learning approaches and, finally, more innovative and experimental landmark based approaches.

First, all data preparation methods are described. These methods range from simple face detecting algorithms to more complex landmark-based alterations. After those specifications, the first approaches are presented. They were based on handcrafted features paired up with both, traditional machine and deep learning classifiers, in the form of an SVM and an MLP.

Regarding the field of pure deep learning, we developed one CNN based on transfer learning and more two CNNs developed from scratch.

These models were trained using both original images and landmark-oriented images, in order to evaluate the impact of certain facial regions in emotion classification.

Finally, and continuing to follow the idea that different facial regions have a different impact on the final result, a model based on attention mechanism is presented.

All in all, the methods proposed in this dissertation perform a comprehensive study on AFER systems through the recreation of established methods and then move towards understanding the impact of facial regions using emerging approaches, regarding both classifiers but also data modification tasks.

Chapter 6

Experimental Work Results

6.1 Datasets descriptions

The different methods described in Chapter 5 were evaluated in the following databases: CK+ citec+, SFEW [29] and JAFFE [67]. These databases were chosen as they are the most used in the current literature and give the possibility to perform more direct comparisons to previous works. Also, these databases gather different types of samples. JAFFE [67] and CK+ [65] are mainly composed of samples from controlled environments and posed subjects and SFEW [29] give us samples from more “natural" situations with spontaneous images in more uncontrolled circumstances. Overall, these different databases enable us to make comparisons between the models and speculated about their performances in more ideal situations or more real-life scenarios.

For the experiments made within this dissertation, the available data was divided using an 80/20 partition - 80% of the data was used for training and validation and 20% was used for testing. Moreover, as it was explained in Section 5.1 cross-validation with 10 folds and a nested 3-folds cross-validation were the chosen techniques for traditional machine learning classifiers. Regarding all deep learning training, 3-fold cross-validation was the preferred approach. To ensure that the split of Train/Test was balanced, a bar chart was plotted for each of the used datasets (Figure 6.1). For CK+ [65] and SFEW [29] databases there is some imbalance between classes within the same database, but the class distribution remains consistent across the train and test sets, thus the split is suitable for the user databases. The same split was applied to JAFFE [67], however because of the reduced number of samples, this database showed a higher degree of imbalance between train and test sets - Figure 6.1 (b)).

Finally, due to changes performed to the images during the approaches described in Section 5.2.3, the number of samples differs from method to method. The number of samples of each dataset for each method is shown in Table 6.1.

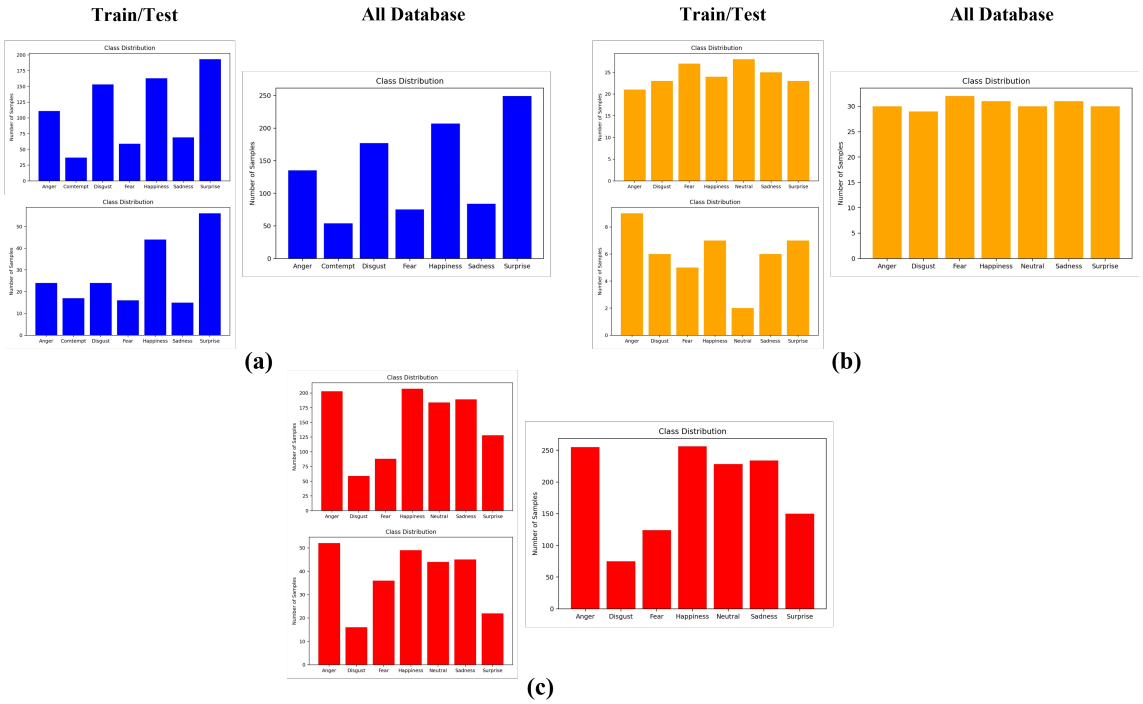


Figure 6.1: Class distribution for: (a) CK+ [65], (b)JAFFE [67] and (c) SFEW [29].

6.2 Experiments Description

For better comprehension of all the results obtained by the models described in Chapter 5, the results and respective discussion are divided into four different experiments. Labelled from A to D, these were the following experiments:

- **Experiment A - Handcrafted:** This first experiment comprises the evaluation of two different classifiers (SVM and MLP) trained with two different handcrafted features (LBP and HOG). The main goal was to see how traditional machine learning methods would perform.

- **Experiment B - CNNs:** This experiment comprised the evaluation of three different networks trained and evaluated only with original samples. First, TransVGG16 was developed to see if with frozen weight a model trained in a completely different set of samples could have a good performance. Second, myCNN served as an initial approach to developing a network end-to-end. Finally, myVGG is an attempt to merge the last two models, trying to have a trained from the scratch network but having more and deeper layers. All the developed models were evaluated using both original and landmark-created images.

- **Experiment C - Minimalist Approach:** The ultimate goal of this experiment matches the second objective of this dissertation: to discover if using minimalist representations of human facial expressions we can reach the same type of results as using original samples. Experiment C is divided into the different developed samples, in other words: *ROI Mask*, *Only Landmarks* and *Emotion Silhouette*.

Table 6.1: Number of samples for each method

Train/Validation Samples

Method					
		Original	**LBP**	**HOG**	**Landmarks Approaches**
Database	**CK+**[65]	785	785	785	738
	SFEW[29]	1058	1058	1058	769
	JAFFE[67]	171	171	171	171

Test Samples

Method					
		Original	**LBP**	**HOG**	**Landmarks Approaches**
Database	**CK+**[65]	196	196	196	182
	SFEW[29]	264	264	264	196
	JAFFE[67]	42	42	42	42

Alternatively to the previous experiment (Section 6.4.2), during this experiment the performance of the models *per se* will not be discussed. Instead, we will see if each one of the landmark transformations was successful and how the models change their performance with completely different samples. We will take the values of the models as a baseline and evaluate the three approaches by direct comparison to the same models trained with the original samples.

In addition, it is imperative to mention that landmark approaches are not new to the literature, but in the majority of cases the relative positions of the landmarks are used to make those models. In our case, we wanted to see and test if minimalist representations could lead to similar results as a full human face. We wanted to see if the models with a smaller set of visual information could reach the same final results.

- **Experiment D - Attention Model:** The final experiment was the evaluation of the Attention Model (AM). This model was built on the premise that different areas have a different value for the model and so dividing our samples and feeding them as positional patches to the model would benefit the final results.

6.3 Implementation Details

This section aims to describe the technical details of the experimental work from Chapter 5. First of all, every single experiment was developed in Python 3.8.13 in the Anaconda Scientific Python Development Environment (Spyder).

Regarding handcrafted features, both LBP and HOG descriptors were computed using *scikit-image* library, which is a collection of algorithms for image processing. The classifiers, SVM and MLP, and respective optimization using GridSearchCV, were developed using python *sklearn* library. Further, for deep learning methods *tensorflow keras* libraries was used.

Regarding the evaluation of the results, *seaborn*, a python statistical data library, was chosen to facilitate the visualization of the metrics and *visualkeras* was used to build graphic representations of the developed deep learning models.

At last, all the models were trained with the help of Graphic Processing Units (GPUs), specifically a NVIDIA GeForce RTX 2060. Python libraries like *numpy*, *matplotlib*, *pickle*, *cv2* and *pandas* were also used along this dissertation.

6.4 Discussion and Results

In this section, the results from the experiments described in Chapter 5 are presented, as well as the respective discussion of each method. To simplify the analysis, all the results have been presented in tables (Table 6.2 to 6.6). Also, accuracy was the chosen metric to evaluate and compare the experiment. Being the most used and intuitive metric, accuracy was chosen to ease the direct comparison between models. Nonetheless, the confusion matrices for each exmperiment were also evaluated.

6.4.1 Experiment A - Handcrafted

As it was described in Chapter 5, the first approaches were based on handcrafted features. Using LBP and HOG descriptors 4 different models for 3 different databases were computed. Besides comparing which descriptor achieve better results, we also wanted to compare which classifier provided the best classification. From Table 6.2, the main conclusions are the following:

Table 6.2: Handcrafted (A) - Results accuracy depicted in percentage(%).

		CK+	**JAFFE**	**SFEW**
SVM	**LBP**	95.91%	59.52%	35.98%
	HOG	99.48%	73.80%	-
MLP	**LBP**	92.34%	59.52%	37.12%
	HOG	**100.0%**	**80.95%**	-

- Regarding the descriptor HOG, it showed a better performance in both databases and independently of the classifier used. This was expected, because of the discrepancy of elements between the used feature vector. Even though LBP was performed in patches and not directly on all images, HOG produced a much heavier feature vector, transforming each image into arrays with much more elements than LBP. Moreover, HOG took much more time to be computed and even limited the use of SFEW as a viable database, because of the time it would take to train those samples as HOG features.

- Regarding the classifiers, it is clear that MLP has a better performance. Mainly on JAFFE database and for both Hog and LBP, MLP was able to produce a much better performance than its counterpart. SVM reached very satisfying results with both descriptors for CK+ database, but lost by a large margin concerning JAFFE database.

Table 6.3: CNNs (B) - Results accuracy depicted in percentage(%).

	CK+	JAFFE	SFEW
TransVGG16	92.85%	**61.90%**	45.08%
myCNN	92.34%	54.76%	**45.45%**
myVGG	**98.46%**	30.95%	36.36%

- Comparing the results to the state-of-the-art, all the combinations regarding the CK+ dataset are considered very good and can be placed among the best literature results, showing values above 90% for every combination of descriptor and classifier. On the other hand, JAFFE results are a bit below the ones shown in the literature, produced by identical methods. Primarily, LBP descriptor could not deliver performance close to state-of-the-art values, being almost 20% below. Handcrafted models are not usually implemented in SFEW, though we reach an interesting result very close to the challenge baseline performance, 35.96% [28].

6.4.2 Experiment B - CNNs

During Experiment B we evaluated the built CNNs using both original and landmark-orientated images. The following section is divided by the different databases to ease the evaluation of each one of the developed models. The results obtained from these models are shown in Table 6.3.

CK+

The three developed models all had the expected results. Having accuracies above 92% and could be framed in state-of-the-art results. The worst performing model with an accuracy of 92.34% in the testing set was myCNN. Although, it had a lot of parameters the number of convolutional layers was reduced, which might difficult for the model to go deeper and make a good generalization.

Regarding the TransVGG16 it achieves very close results with myCNN, having an accuracy of 92.85% during the testing phase. The frozen weights still delivered a very satisfying result and enabled a more fast and lighter training phase. Only having the two last fully connected layers being trainable this model was able to gather the necessary information to make a good prediction in the testing set. This shows that models built with frozen layers previously trained with ”ImageNet" samples can also be an option to develop satisfying results.

The last model, myVGG achieve very promising results, being able to reach an almost perfect score even during the training phase. However, not only, these models took much more time to be trained and validated but also required a higher computational power. As it was expected, myVGG being inspired by a more deep network (VGG19) was able to reach a better understating of the data and consequently better accuracy during the testing phase. The corresponding confusion matrix, Figure 6.2, shows that the model got an accuracy different from 100% only in two classes, Fear and Anger, miss-classifying 3 samples from a total of 196.

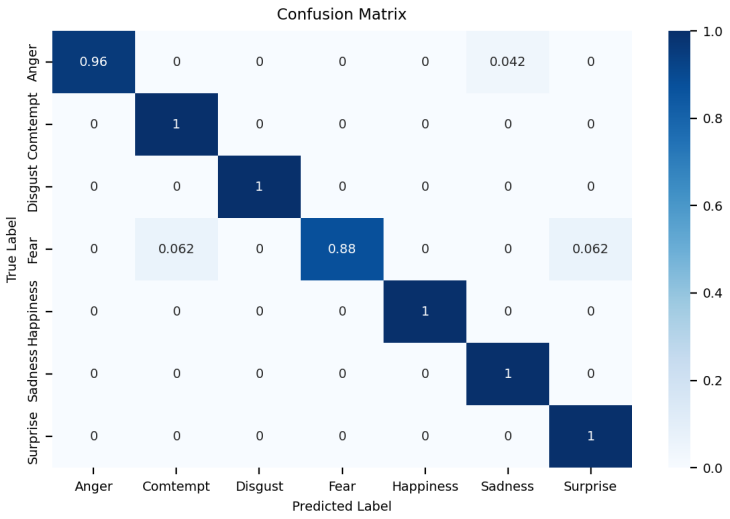


Figure 6.2: Confusion matrix for myVGG.

JAFFE

The results for JAFFE were very underwhelming, all the three models performed compared with CK+ database and with state-of-the-art approaches. After, analysing the results obtained from JAFFE database it was clear that the models were overfitting during the training phase. The accuracy during this phase rapidly rises to 100%, yet the value for the validation subset reaches values around 20% and stays in a plateau for the rest of the epochs (Figure 6.3). For overcoming this problem, a data augmentation phase was implemented in order to see if increasing the number of samples and inducing more variability the results would improve. Even with data alterations the model still continued to overfit and produced even more inaccurate results. As it is possible to see in Figure 6.4, the accuracy values during training did not stabilise and were very unstable until the last epoch; during the validation phase the accuracy value did stabilise in the first epochs but the value was very continued very low compared to the training phase.

Table 6.4 show that even with data augmentation, the models could grasp the necessary aspects of JAFFE samples leading even worst and disappointing results. Better model regularization and more data augmentation could improve those results

Nonetheless, the model with the best performance was TransVGG16 with 61.90%. TransVGG16 was not the only one with the best performance but also had a smaller discrepancy

Table 6.4: Results from methods with and without data augmentation.

No Data Augmentation	**TransVGG16**	61.90%	1.09
	myCNN	54.76%	2.03
	myVGG	30.95%	2.85
Data Augmentation	**TransVGG16**	16.66%	1.98
	myCNN	14.28%	5.84
	myVGG	16.66%	17.72

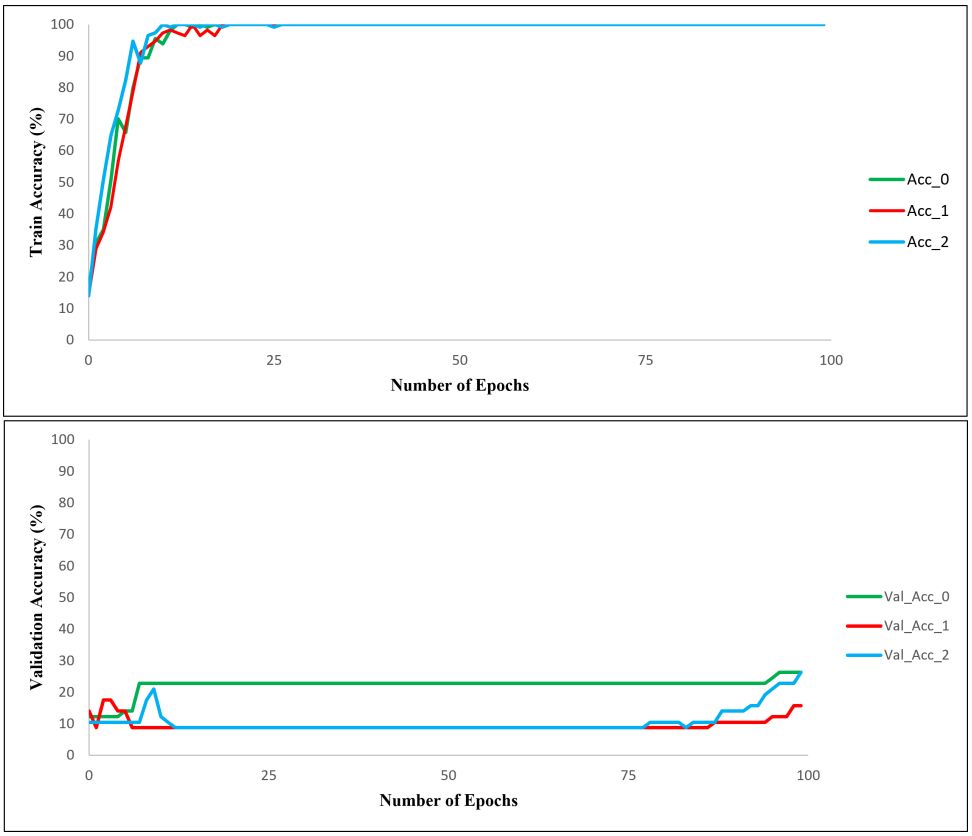


Figure 6.3: Evolution of the myVGG without data augmentation accuracy for training and validation phases.

between the training and validation phase. Not reaching as good results during the training phase, but having a much better evolution during validation. It is possible to project with more epochs and a better regularization a state-of-the-art result could be achieved. Moreover, the success of this model and the failure of the other two can be explained by the fact that JAFFE is the database with the least amount of samples. Saying this, this database could benefit from a model which was pre-trained with a much larger dataset. Instead of using a deeper network to understand the necessary details, a more generalised network could be the solution to deal with smaller datasets. Both myCNN and myVGG had very poor performance with 54.76% and 30.95% respectively.

SFEW

Regarding the SFEW database, the results were accordingly to the expected. None of the models had an outstanding performance, but all have better results than the baseline proposed by the database itself, 35.96% [28]. Compared to more ambitious values, two of the models achieve similar accuracies to the one proposed by Mohammadpour *et al.*[55], which reached a performance of 47.7%.

In this instance, myCNN beat TransVGG16 by a close margin, 45.45% against 45.08%, as the best-performing model. As it happens with JAFFE data, all the models were slightly overfitted

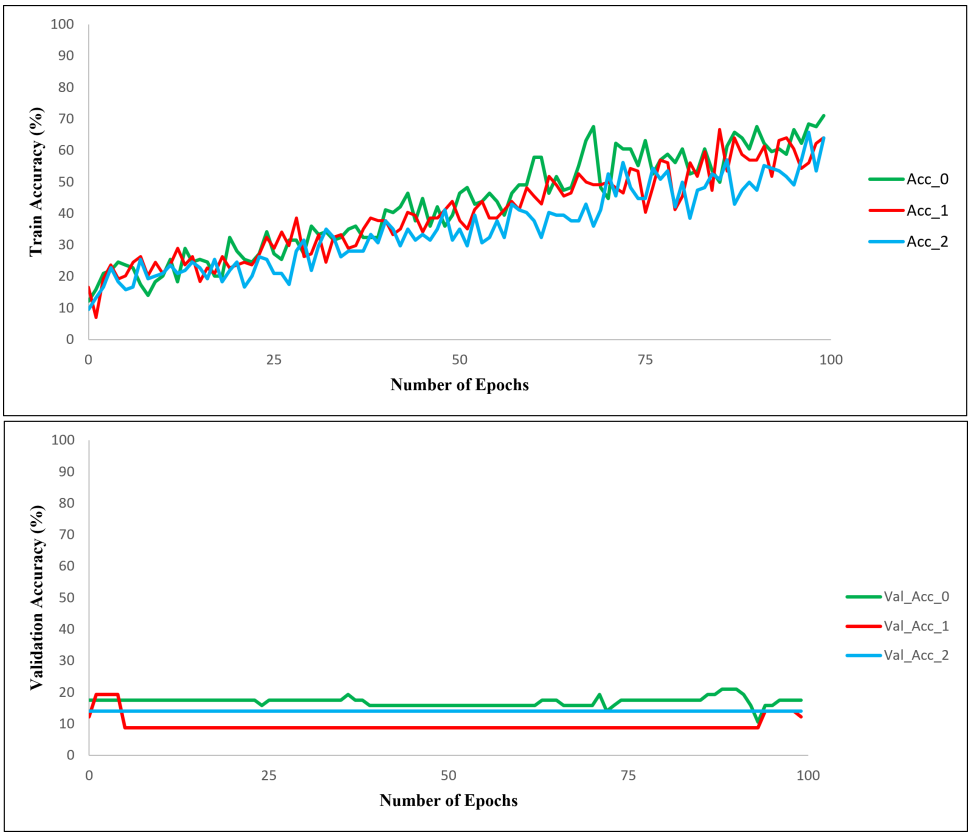


Figure 6.4: Evolution of the myVGG with data augmentation accuracy for training and validation phases.

during the training phase. However, in this case, the models reach satisfying values when faced with testing samples. Once more, the model built from frozen layers had a better performance than one of the models trained from scratch, myVGG. This means that applying transfer learning leads to good results when applied to fewer balance datasets like SFEW. A more fine-tuned model, with having some of the original layers unfrozen and trainable, could lead to a more adaptable network and could be an interesting approach for any future work.

The myVGG network, even though being the deeper and more complex model, could not get good results missing out by almost 10%. The model was overfitted during the training phase and even during the validation phase the results were not better than 40%, Figure 6.6. Values going as much as 95% were obtained during the training phase, however when faced when unknown samples the models could not deliver the same values. Thus, a thorough regularization process and even the addition of more dropout layers would be the next step to see if the model could reach better accuracy values.

Finally, scrutinizing our best performing model with SFEW samples, myCNN, confusion matrix, Figure 6.7, it is possible to see that the classes with the least samples, Disgust and Fear, lead to the worst results. Happiness, on the other hand, is the class where the model excelled and made better predictions. We can conclude that with a more balanced dataset better results could have been achieved.

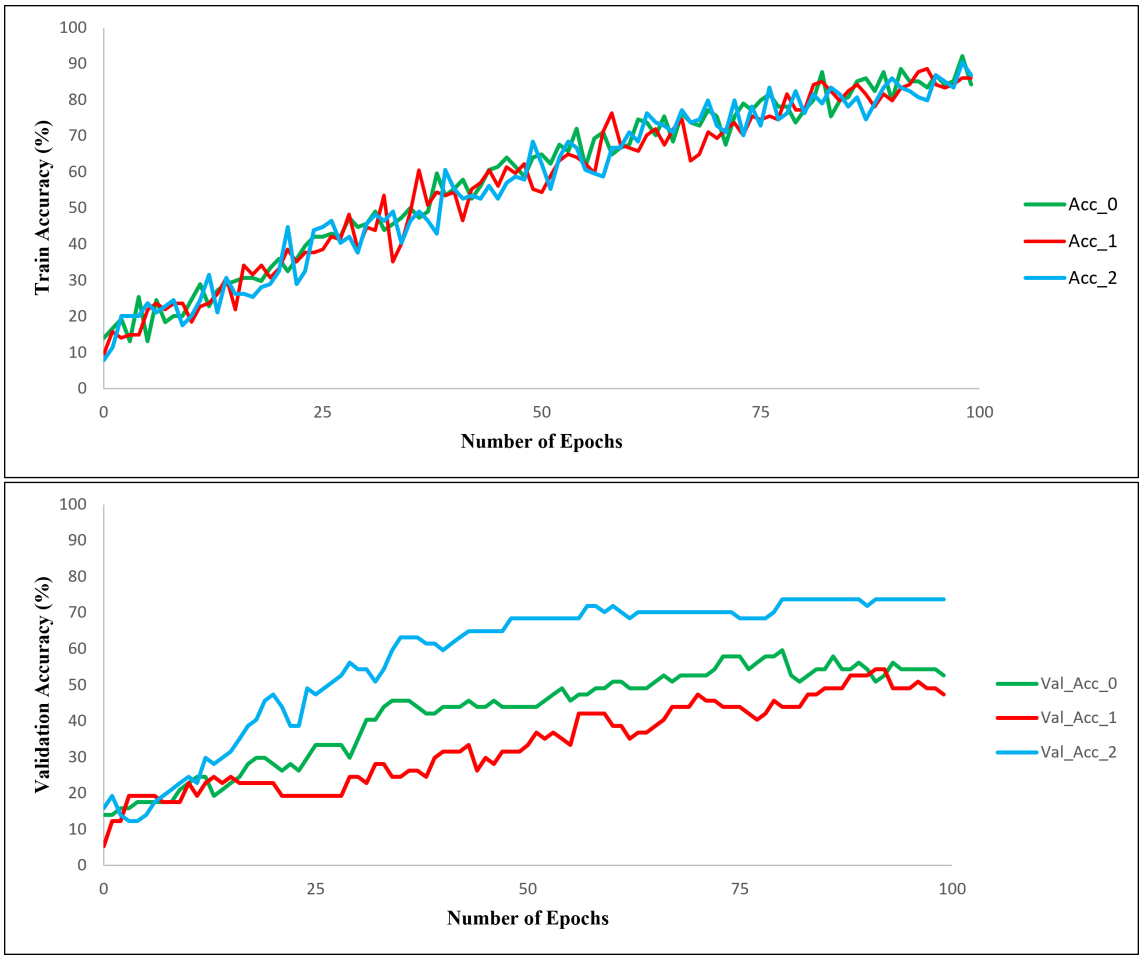


Figure 6.5: Evolution of the TransVGG16 accuracy for training and validation phases.

6.4.3 Experiment C - Minimalist Approach

Experiment C aims to evaluate the developed landmarks-based images and compare the performance of the same models trained with original samples and with our landmark-based samples. Also, the analysis will be made by comparing the success of each of the landmark modifications for each used database. All the gathered results are presented in Table 6.4.3.

Region Of Interest Mask

The main premise behind the development of the Region of Interest (ROI) Mask was the fact that most of the action units are related to certain specific facial regions. Thus, forcing the models onto those areas the model should produce equal or even better results.

Overall, the samples originated by this approach met the initial goal, all the pre-determined key areas (eyes, eyebrow, nose, and mouth) were correctly enclosed. However, because of the original positioning of the faces, a diverse range of regions of interest masks originated.

Starting by CK+ database, all the results are very satisfactory, meaning none of the relevant areas was blocked or cut from the image. As it is possible to see in Figure 6.8 all the important

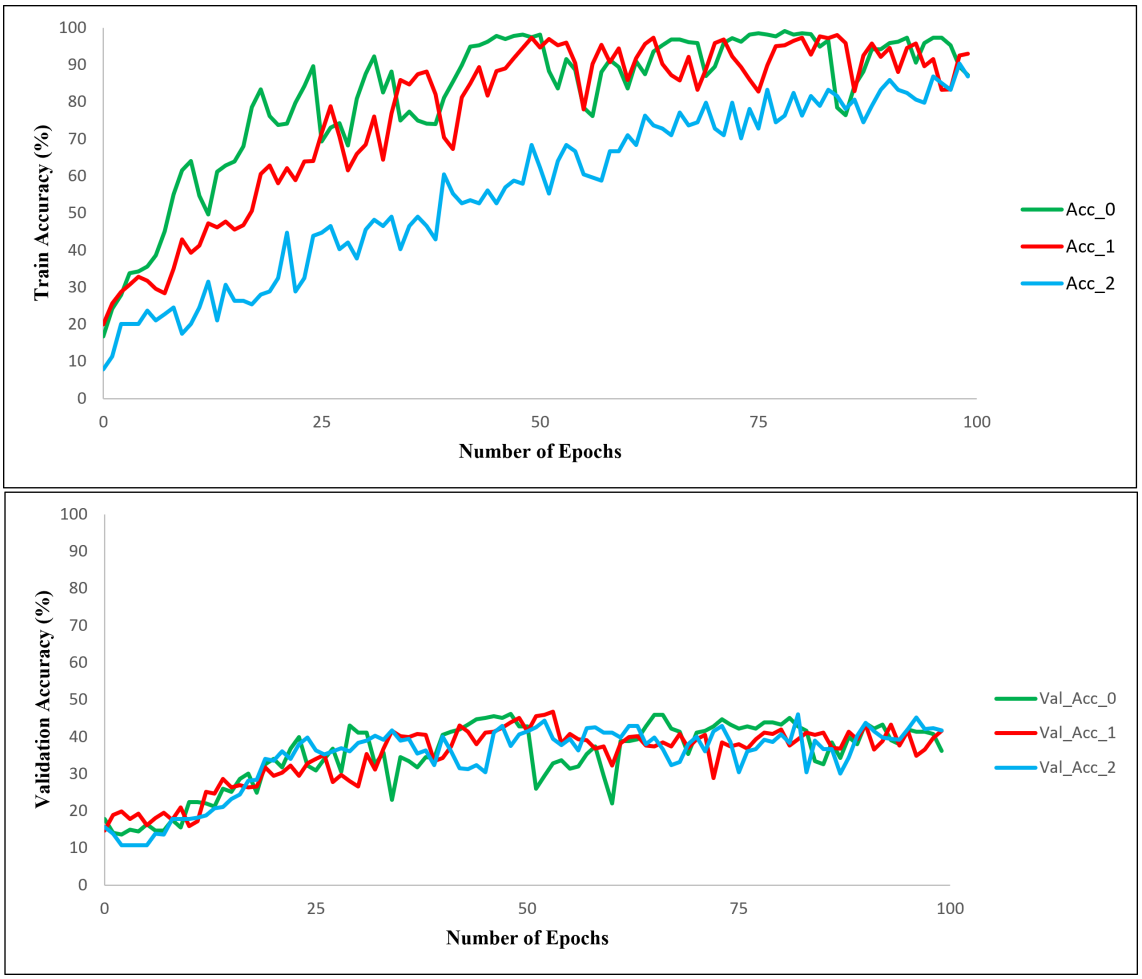


Figure 6.6: Evolution of the myCNN accuracy for training and validation phases.

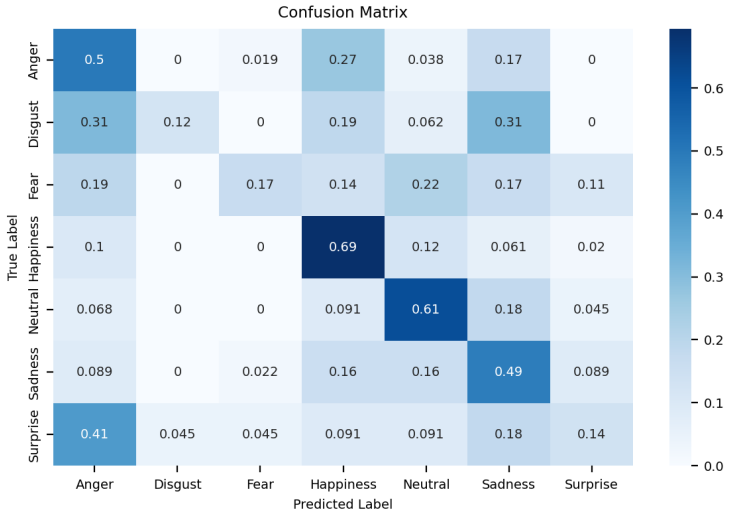


Figure 6.7: Confusion matrix for myCNN.

Table 6.5: Handcrafted (C) - Results accuracy depicted in percentage(%).

		CK+	JAFFE	SFEW
Original	**TransVGG16**	92.85%	**61.90%**	45.08%
	myCNN	92.34%	54.76%	**45.45%**
	myVGG	**98.46%**	30.95%	36.36%
ROI Mask	**TransVGG16**	60.98%	**35.71%**	**35.51%**
	myCNN	**91.75%**	26.19%	26.53%
	myVGG	84.07%	14.28%	30.61%
OL	**TransVGG16**	**82.42%**	47.62%	**39.28%**
	myCNN	76.69%	**54.76%**	33.67%
	myVGG	70.97%	0.05%	32.14%
ES	**TransVGG16**	72.52%	**52.38%**	**36.73%**
	myCNN	**79.12%**	50.0%	35.20%
	myVGG	74.72%	19.05%	34.70%

regions necessary for a prediction are highlighted. However, the produced result shows a discrepancy in accuracy between some classes. From analysing the confusion matrices of the three developed models (Figure 6.9), we notice that three classes stand out with much better results. Classes Happiness, Surprise and Disgust, independently from the model use showed performances well above any other class with Happiness reaching 95% in every model. This can be explained by the fact that these classes are defined by very distinctive features in comparison with the rest. Opening the mouth for Surprise or frowning the eyebrows for Disgust is very easy to associate with specific emotional states, and because of this approach is highlighted even more.

On the other hand, Fear and Contempt are the least successful classes. As is depicted by Figure 6.10, the crop was still well performed as the same areas as in the previous examples are emphasized, but these emotions required another sort of information that these regions could not deliver.

All in all, the models trained with ROI Mask had the worst performance than the original, though the decrease was not consistent along with the three models. Both models trained from scratch had only minor decreases with myCNN having an overall accuracy of 91.75% which is less than 1% below the performance of the original samples. Showing that, even though reducing the images to these key areas does not improve the score and consequently slightly deteriorates the model's accuracy, the majority of emotions rely on the information present in these regions. In contrast, TransVGG16 had a major drop from 92.85% to 60.98%. We can assume that changing the nature of the images to this can be explained by the fact that the nature of these new images is very different from the image the frozen layers were trained with, leading to less accurate results.

Concerning the results produced using the JAFFE database, the delimitation of the interest areas was as satisfactory as it was for CK+ database. Figure 6.11 shows some examples where we can see that once again the eyebrows, the eyes, the nose and the mouth are correctly isolated. However, neither one of the trained models could reach values close to the ones existent in the literature nor to the ones produced with original samples. JAFFE emotions are very well portrayed,

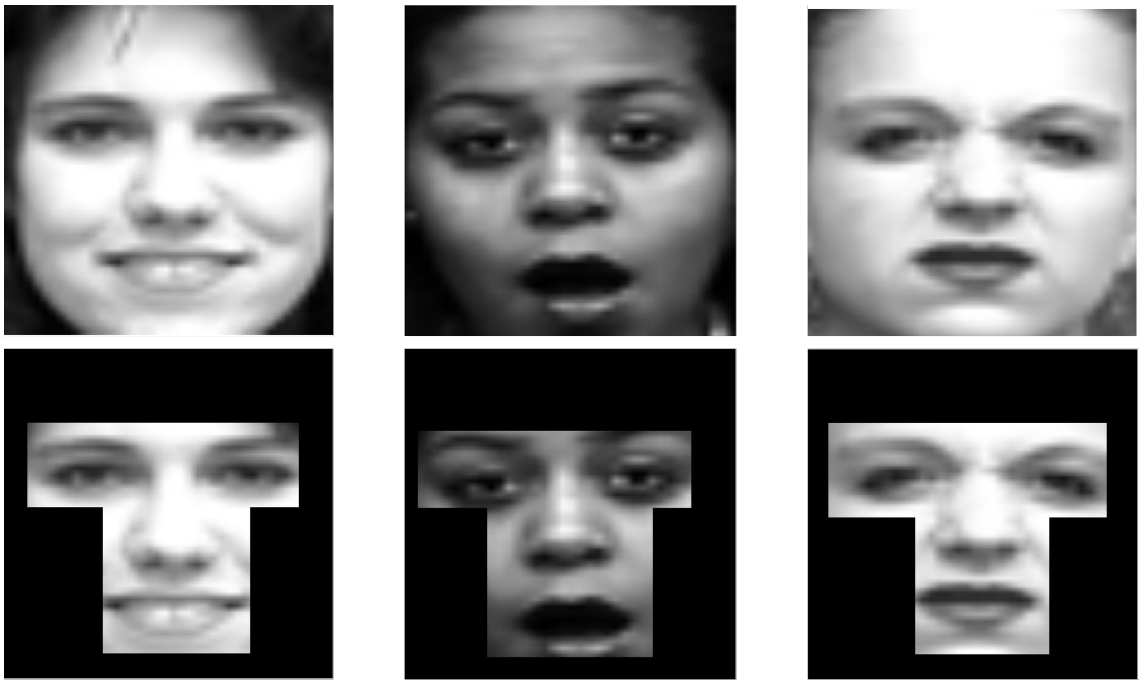


Figure 6.8: ROI Mask samples of CK+ Database for three different emotions: (a) Happiness; (b) Surprise; (c) Disgust.

but are not as intense and emphasized as the ones depicted in CK+ databases. For example, Surprise is not defined by an exaggerated open mouth or Anger which is not defined by narrow eyes. These small details, which sometimes give a more natural look to the expression, make the model less dependable on specific facial traits and make our ROI Mask less useful.

Moreover, having the same overfitting problems as happened with original samples, the main conclusion from this approach is that if the models could not grasp the necessary information when faced with the whole face, it gets even worse when the face is cropped. Still, we considered TransVGG16 as the best performing with 35.71%, having a drop of almost 25% in comparison with the same model trained with original samples.

Finally, ROI Mask samples for SFEW database were more challenging. Because they are based on movie frames, SFEW database is comprised of many faces which are not aligned like CK+ or JAFFE. Thus, the landmark locations were not as consistent throughout the samples as in the previous examples. In Figure 6.12 we can visualize some less perfect examples of created samples based on SFEW samples. For the first time, we can see some misaligned rectangles forming irregular shapes, similar to a cross. Due to the existence of different poses and the inconsistency of the positioning of the faces, the developed binary mask were much more diverse. Narrower horizontal rectangles and longer vertical rectangles were more common. In order to solve the existing problems, we would need a more personalized system. Instead of applying the same index for every image, we should analyse the positioning of the face and only crop out non-essential regions. Nonetheless, even if without the same precision as in CK+ and JAFFE databases, the

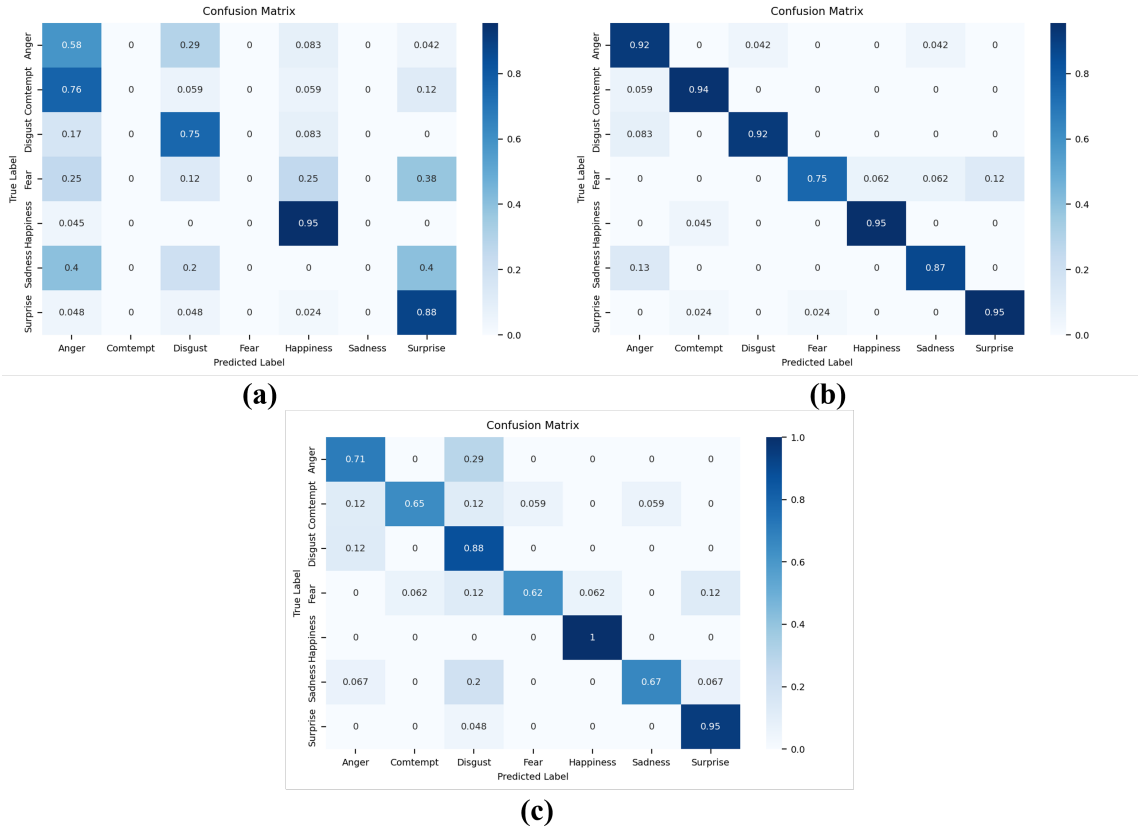


Figure 6.9: Confusion matrices of the models trained with ROI Mask samples: (a) TransVGG16; (b) myCNN; (c) myVGG.

majority of the samples created achieve its purpose and successfully bonded the chosen regions (Figure 6.13).

Regarding model performances, all three models had slight drops in their accuracy scores, showing that using databases with more variability, standardization of certain regions can lead to more inaccurate results. The models TransVGG16 and myVGG had the best results, having a drop of 10% and 6% respectively.

Only Landmarks

Continuing to focus on the use of facial landmarks, the second approach was an attempt to prove if the locations of the 468 key points would be enough to accurately describe a human expression. Based on the assumption that every fundamental area is described by each one of the available landmark coordinates, we developed images which only used those points to recreate a human face. The implementation of this method was the most straightforward one, as the creation of our sample depended exclusively if the landmarks were detected. Upon confirming the existence of a usable human face a small circle was drawn in each location.

As it was seen in the ROI Mask method, for CK+ samples, the same three classes stood out as having the best performance: Happiness, Surprise and Disgust. The same distinctive feature was

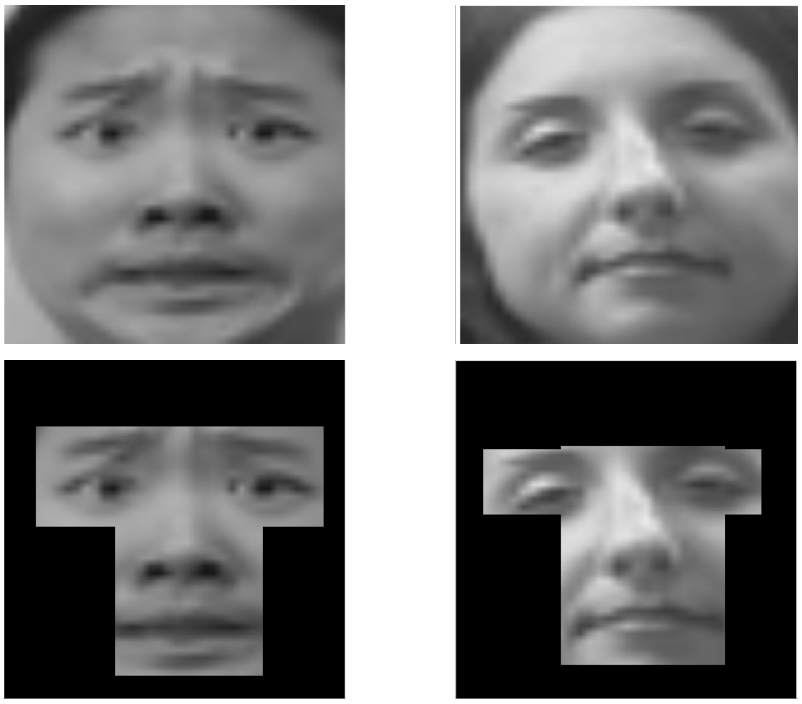


Figure 6.10: ROI Mask samples of CK+ Database of two different emotions: (a) Fear; (b) Contempt.



Figure 6.11: ROI Mask samples of JAFFE Database for of different emotions: (a) Anger; (b) Surprise; (c) Disgust; (d) Happiness.

once again captured by the landmarks circles. As it is possible to see in Figure 6.14 (a), the open mouth is well characterized without needing any texture representation. Likewise, in Figure 6.14



Figure 6.12: ROI Mask less accurate samples of SFEW Database of four different emotions: (a) Fear; (b) Happiness; (c) Anger; (d) Surprise.

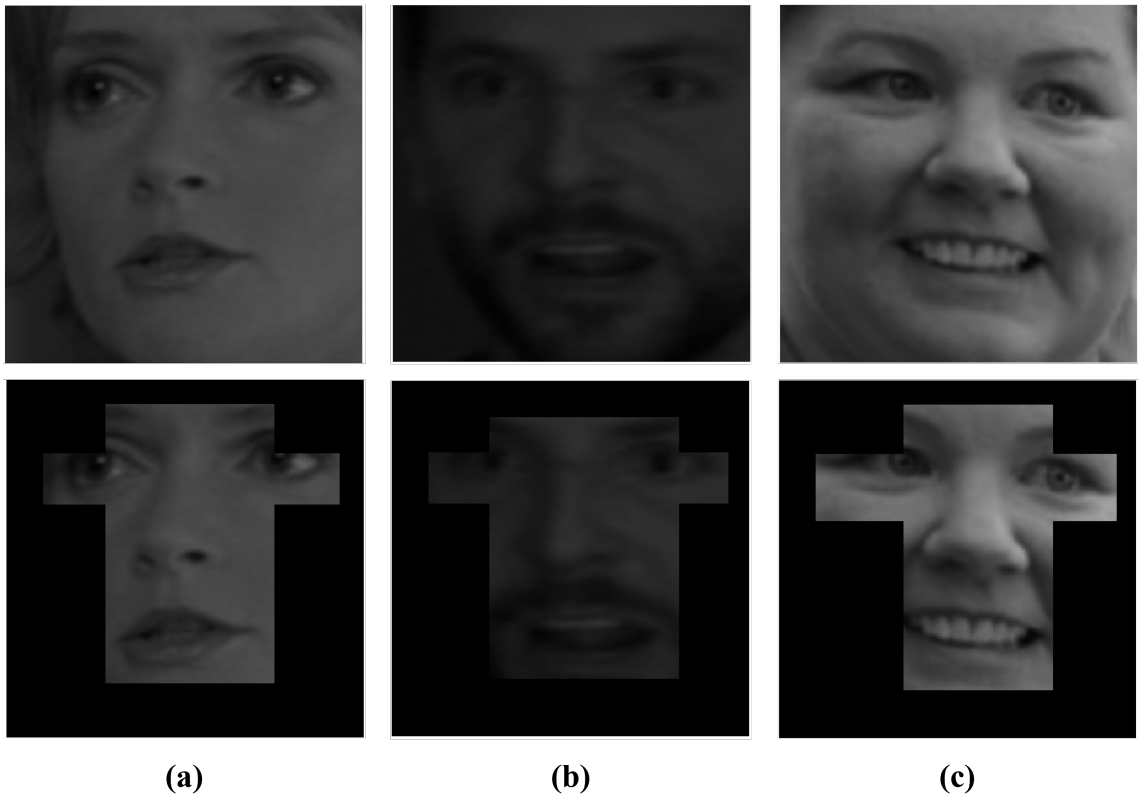


Figure 6.13: ROI Mask more accurate samples of SFEW Database of three different emotions: (a) Sadness; (b) Surprise; (c) Happiness.

(b) the smile is correctly portrayed giving the models enough formations to associate the sample with Happiness. Both the presented images are correctly predicted by all three developed models.

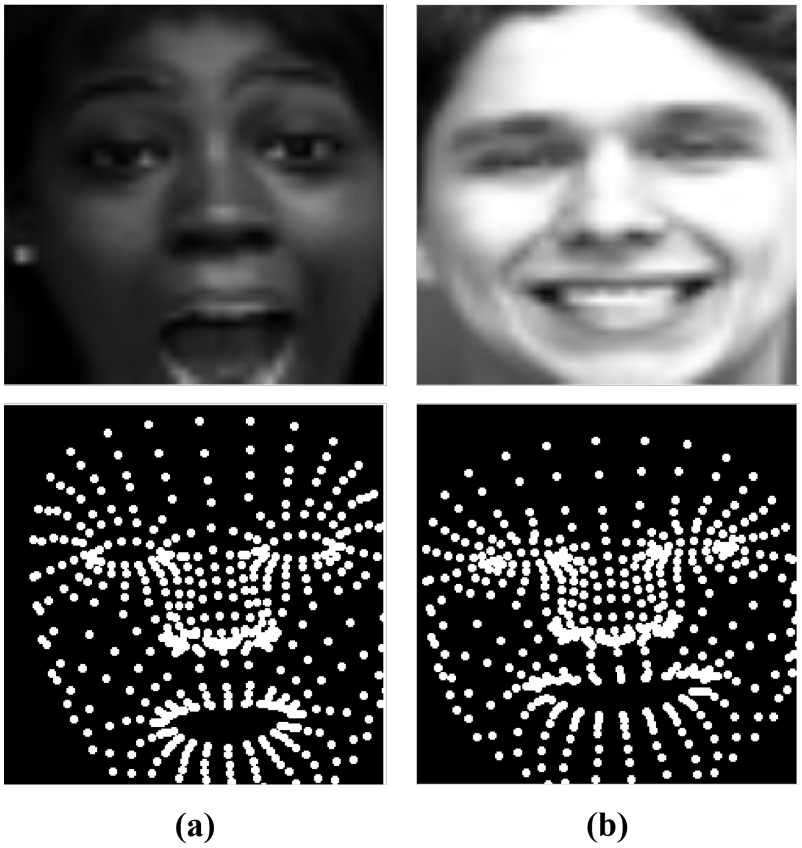


Figure 6.14: Only Landmarks samples of CK+ Database of two different emotions: (a) Surprise; (b) Happiness.

In contrast, emotions like Anger and Contempt have worse results because they are easily confused. Figure 6.15 shows how similar these three emotions are portrayed using only landmarks, the key information to differentiate them is the small details like the movement of the eyebrow and the curvature of the lips. However, the landmarks are not able to capture those details and subsequently originate inaccurate representations. The main conclusion is that this approach is very suitable for expressions which are dependent on the mouth and the wideness of the eyes. The landmarks capture really well the overall essence of a human face but fail to grasp the important details to capture an expression.

The performance of the models was inferior in comparison when compared whit models trained with original samples. Still, the combined average performance of the three models was around 76% which only represents an average decrease of around 14%. Additionally, the classes Happiness, Surprise and Disgust continue to produce scores as high as 95%. From the developed models TransVGG16 was the one with the best performance with 82.42%

Moving to the next database, Figure 6.16 shows some of the developed samples using JAFFE samples. As it is possible to see and as it was highlighted before, JAFFE expressions appear to be

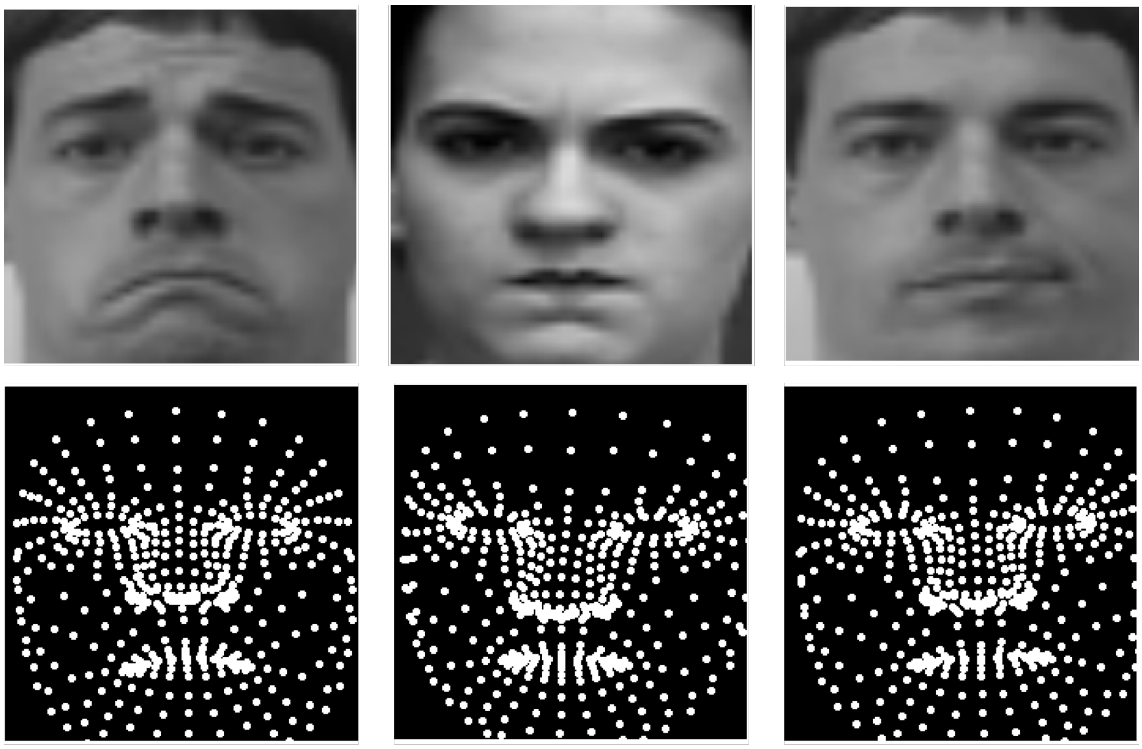


Figure 6.15: Only Landmarks samples of CK+ Database of three different emotions: (a) Sadness; (b) Anger; (c) Comtempt.

much less intense and more naturalistic. Thus the respective OL representation seems to be much more generic throughout the different emotions. Nevertheless, the emotions which rely more on the mouth and eyes related features, like Happiness and Surprise, where the slope of the lips or the elevation of the eyebrow can be fundamental, stood out, delivered a better result and even propel the global performance of the models to values which rival even the models trained with original samples. TransVGG16 had the worst performance than with the original samples, much better than with ROI Mask samples; myCNN trained with OL samples achieved even more impressive values, having exactly the same accuracy than trained with original samples, 54.76%.

Analysing the model's performance it is possible to conclude that for CK+ the Only Landmarks approach resulted in a worst approach than ROI Mask. However, for SFEW database, the OL method leads to better results. Instead of cutting some of the necessary traits like in ROI Mask, OL offered a very good description of the all face. For the other databases, this was a disadvantage because the visual information was better than only having the landmarks circles, but in SFEW samples this approach showed to be much more adaptable to the high variability of poses and tilted faces leading to better performance. In Figure 6.17 the same samples are shown in their ROI Mask and OL format, it is easy to understand why the second method lead to an improvement. The developed images completely defined the human face and ultimately give the models more information only by using white dots, than ROI Mask approach gave in total. What was a crucial

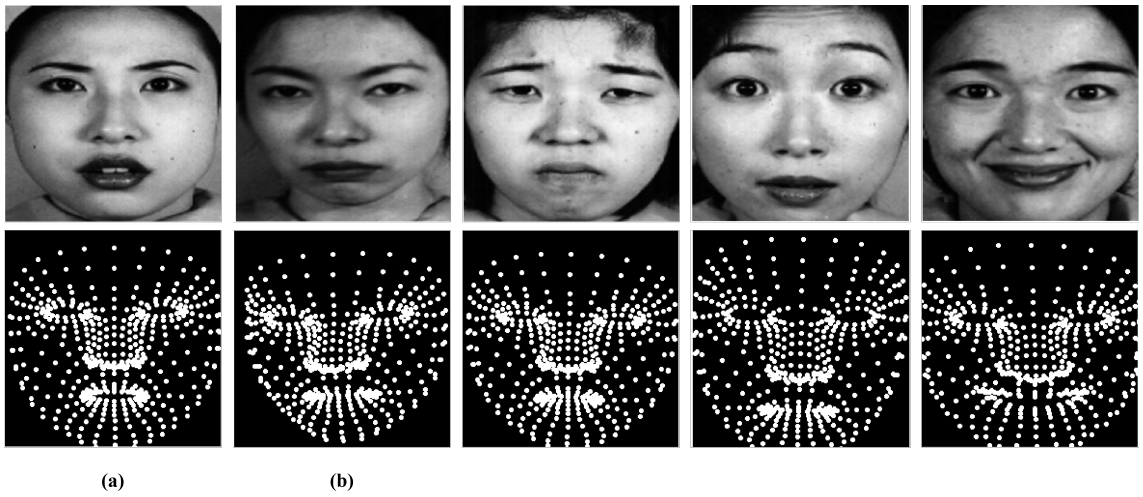


Figure 6.16: Only Landmarks samples of JAFFE Database of five different emotions: (a) Fear; (b) Anger; (c) Sadness; (d) Surprise; (e) Happiness.

disadvantage of SFEW database during ROI Mask method, was easily overcome by OL.

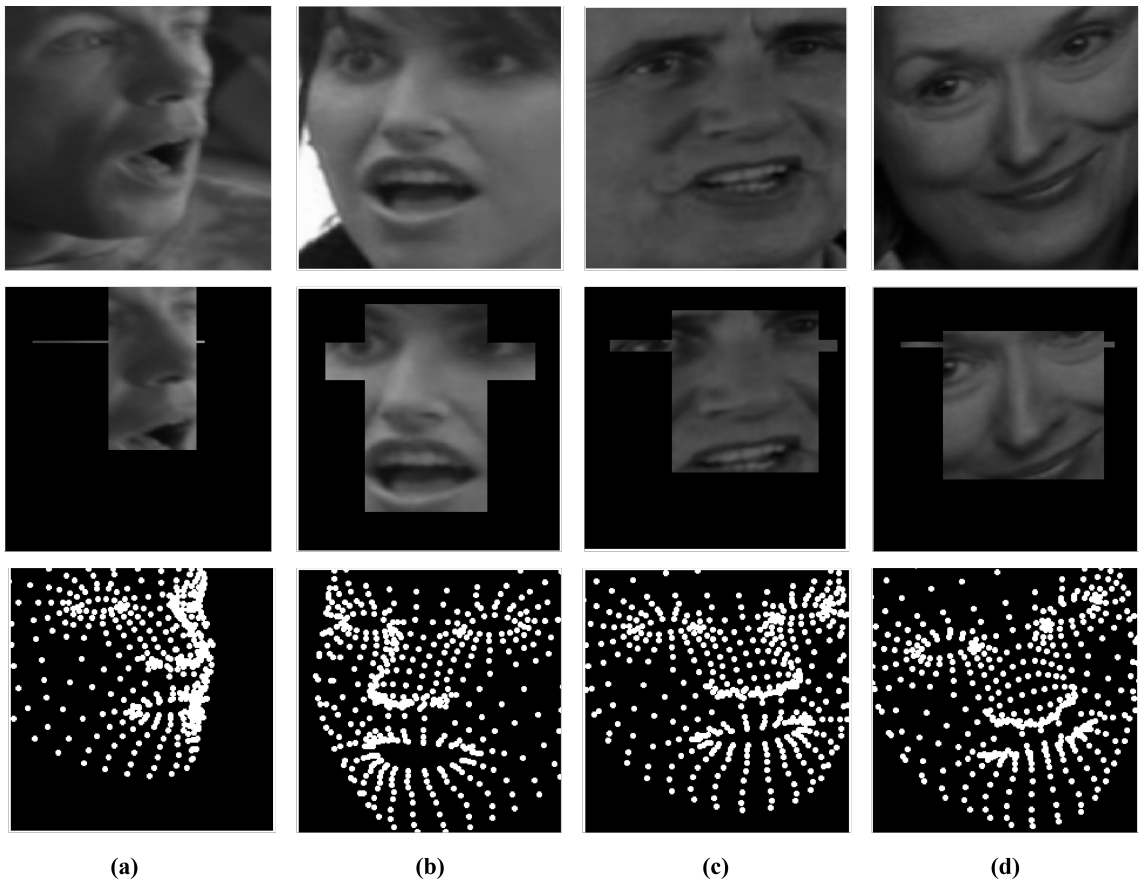
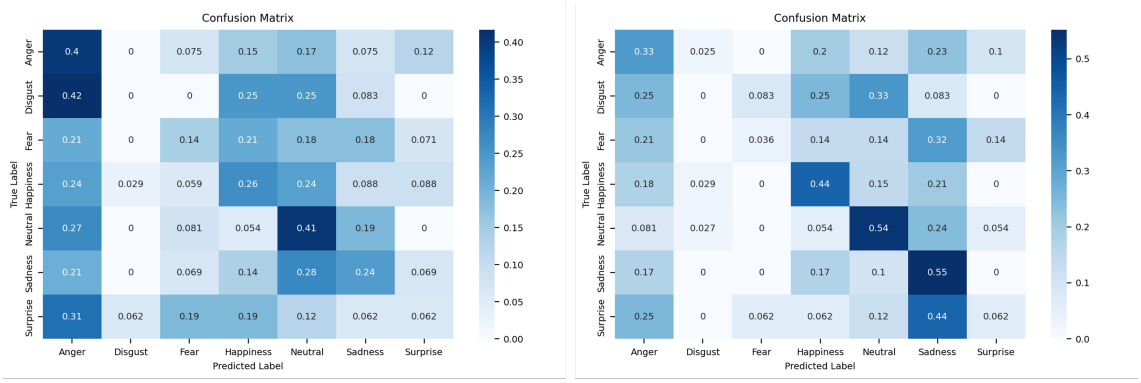


Figure 6.17: Only Landmarks and ROI Mask samples of SFEW Database of four different emotions: (a) Fear; (b) Surprise; (c) Anger; (d) Happiness.

All the three models registered an accuracy improvement, with myCNN having an improvement of 10% . Figure 6.18 shows the confusion matrix for the same models trained with ROI Mask samples (a) and OL samples (b), the main difference can be seen in Happiness and Sadness. We can assume that because of bad cropping explained in the previous Section 6.4.3, the models were deprived of the vital information and now, with OL samples, were able to make better predictions. OL samples, as shown in Figure 6.17, even though do not gather all the necessary feature to describe an emotion, are able to adapt to more challenging images and correctly portray a human face.

Anyhow, and even though the created images made a much more accurate global representation, the models still underperformed in comparison with the same models trained with original samples



(a) **(b)**

Figure 6.18: Confusion matrices of myCNN trained with : (a) ROI Mask Samples; (b) OL samples.

Overall, Only Landmarks approach was a very good method to describe a human face but lacked the necessary details to fully grasp what defined facial expressions. Having such a high number of landmark coordinates was an advantage, still, in the regions of the eyes and eyebrows, which play a fundamental role in most facial expressions, the positioning of the key points was not good enough to capture the essence of what is an emotion.

Emotion Silhouette

Emotion Silhouette was the last approach and aimed to produce images which were simultaneously dependent on reducing the number of facial features, like ROI Mask, and is only represented by minimalist drawings, like OL. Also, from all the developed methods, ES was the most challenging one due to being the one with more specifications. On top of it all. it was the method which gave less information to the model, thus making even more important how well the facial attributes are portrayed in our representation.

One of the problems which make this approach even more difficult was the fact that the drawing needed to be respected in order to make the facial features more natural and fluid. Saying this,

a slight change in the positioning of any of the chosen key points leads to a more distorted representation. The developed methods comprised what OL approach had excelled in, highlighting the changes in the mouth and eyes region, and solved one of its biggest problems, the lack of differentiation on the eyebrows. ES samples due to using fewer points and connecting them to produce a continuous line were able to show the dynamic present in certain areas like the eyebrows. On the other hand, the depiction of the lips was very hard, making it super intuitive to detect emotions like surprise (mouth wide open), but very difficult to describe the slope for Sadness, for example. The representation of the nose was also very hard, the nose was easily drawn, but it was impossible to capture the nuances of frowning making it really hard to characterise emotions like Disgust. Some examples of Emotion Silhouette samples of CK+ Database are displayed in Figure 6.19.

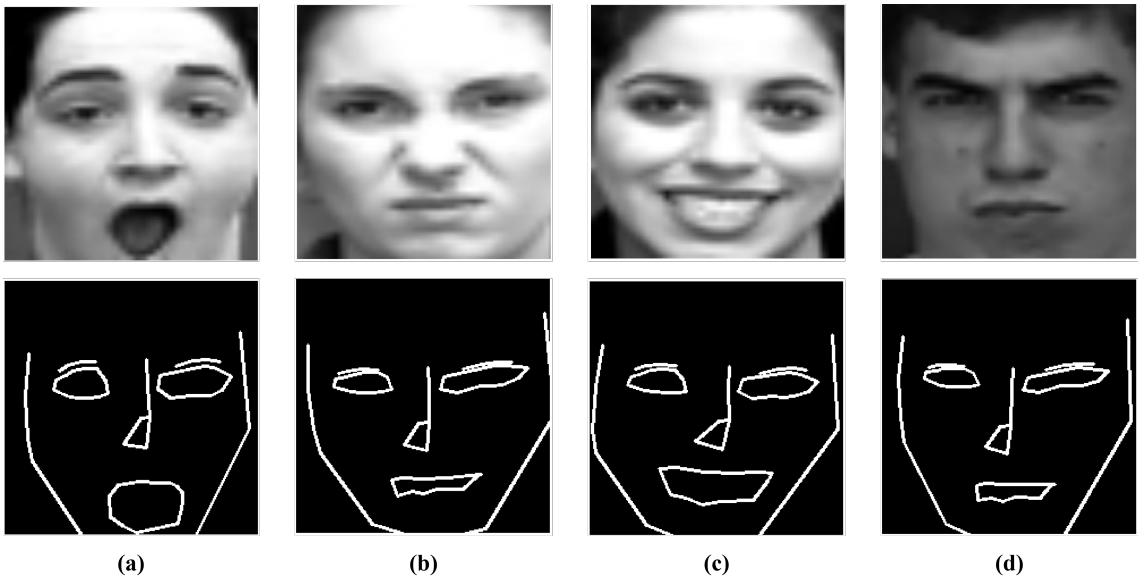


Figure 6.19: Emotion Silhouette samples of CK+ Database of four different emotions: (a) Surprise; (b) Disgust; (c) Happiness; (d) Anger.

Yet again, the same emotions as in the previous methods stood out with better results. Happiness and Surprise were the best performing classes reaching an average of 90% accuracy. This is a direct consequence of the capability of this method being able to portray very well the relationship between the eyes and mouth which are exaggerated during those emotions. Moreover, Anger class was the biggest success in comparison with the previous methods. As it was said previously, and as it is possible to see by Figure 6.19 (d), the slight slope of the eyebrows gives an edge to this approach in comparison to OL. Even though, reaching performance values were almost as good as with the original samples, in this case, the class Anger was predicted much more time giving a better accuracy score but also miss classifying more Disgust and Contempt as Anger. All these classes have in common the importance of the eyebrow for the overall classification. The confusion matrix from myCNN displayed in Figure 6.20, shows the aforementioned emotions still deliver very good results, but mainly because of Contempt and Sadness, the overall score was lower. Still, myCNN and myVGG achieve better performances than OL and are only 10% below

the results obtained for ROI Mask. Proving without a doubt, that even with some limitations, ES is a viable approach for CK+ samples.

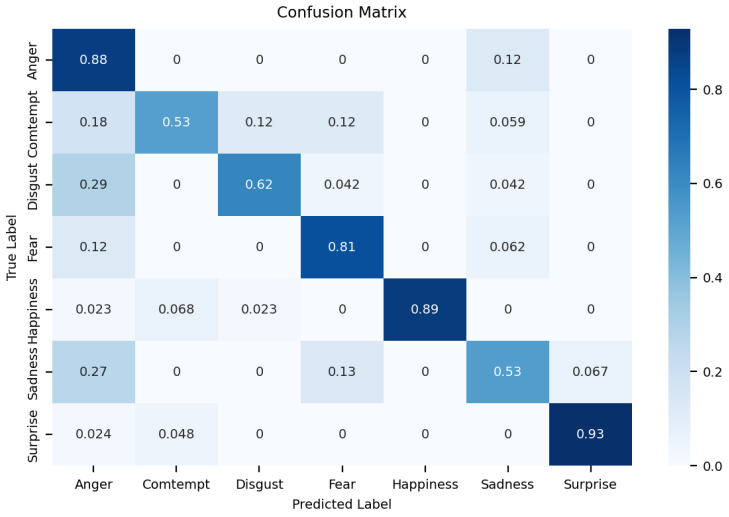


Figure 6.20: Confusion matrices of myCNN trained with Emotion Silhouette samples.

In regards to JAFFE database the obtained results are displayed in Figure 6.21. The majority of the samples presented the same problems like the ones developed from CK+, more intense emotions result in a better representation and the ”mouth-based" emotions were much better classified than the remaining, but JAFFE samples demonstrated to be even more generic. The relation between the eyes and the eyebrows was not well accomplished because the chosen key points were selected using primarily CK+ images, thus using the same landmarks indexes for JAFFE resulted in a poor differentiating mainly in the eyebrows and the lips. As it is possible to see in Figure 6.21 the second row, the chosen landmarks for the eye regions are too compacted removing the necessary flexibility and preventing the optimal display of these facial traits.

Performing a direct comparison, ES samples are able to portray a human reaction having much less with only 15% of the available key points. Selecting only a few landmarks and connecting the points creates a more authentic and unique look to each sample. Giving the already best-developed samples like Happiness and Surprise, an even more distinctive look. Still, not all emotions achieve the same results. Even with this force variability, Disgust and Sadness are very similar due to being dependent on frowning and the lips, which both fail for the ES representation. Just like OS, ES samples lack the necessary details to fully grasp all the emotions, primarily failing in the nose and lips.

The models trained with JAFFE ES samples had performance very similar to the ones trained with OL samples and much better than the ones trained with ROI Mask samples. We can assume that both OL and ES samples were able to capture small essential details, thus reaching higher accuracies values.

SFEW ES developed samples, just like the OL samples, overcome all the issues regarding the multitude of facial positions. Every one of the developed images was able to correctly portray

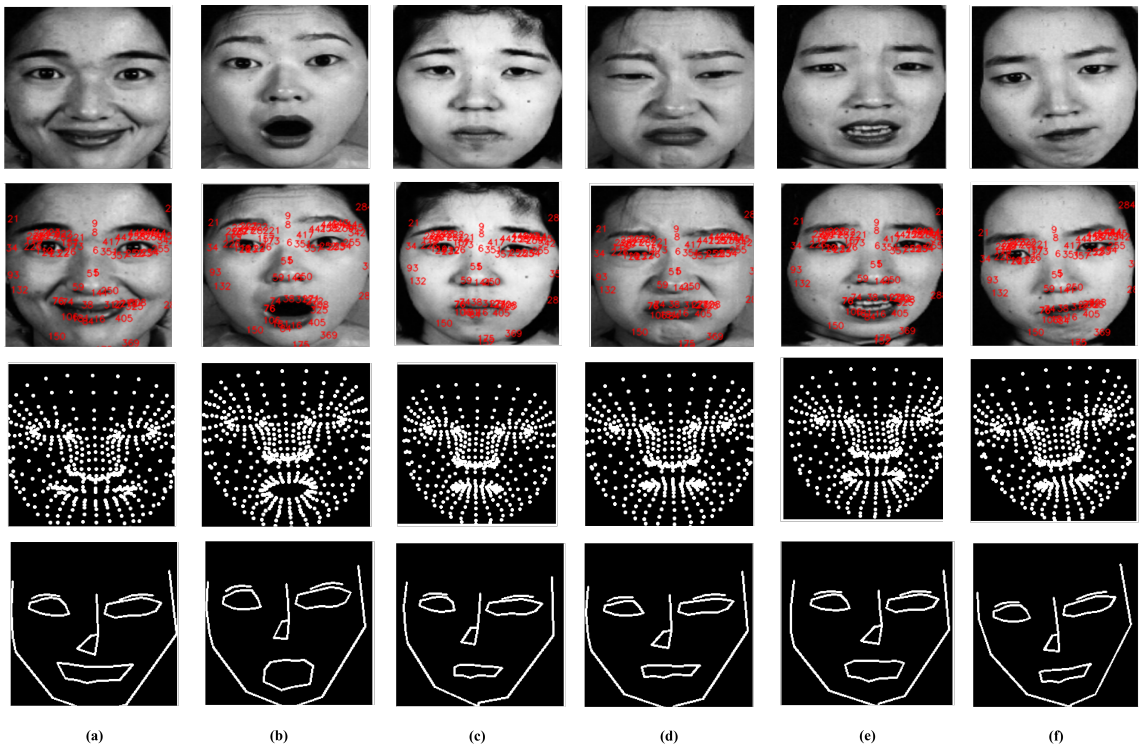


Figure 6.21: Emotion Silhouette and Only Landmarks samples of JAFFE Database of four different emotions: (a) Happiness; (b) Surprise; (c) Sadness; (d) Disgust; (e) Fear; (f) Contempt.

a human face unlike what happened for *ROI Mask* method. But just like for other databases, some expressions were more suited for our methods. As it is possible to see from Figure 6.22, Surprise and Happiness reached much better representations. The accuracy scores from the models trained with these samples reached very analogous values to OL method. ES method produced more balance results between the developed models, having only a max 2% deviation between TransVGG16, myCNN and myVGG. The best performing model with 36.73% was TransVGG16.

6.4.4 Experiment D - Attention Model

Table 6.6: Attention Model (D) - Results accuracy depicted in percentage(%).

	CK+	**JAFFE**	**SFEW**
AM - (16x16)	28.57%	42.85%	18.56%
AM - (32x32)	93.87%	50.0%	17.04%
AM - (6x6)	**96.42%**	54.76%	26.89%
AM - (12x12)	95.41%	**66.66%**	**35.98%**

Experiment D is the evaluation of the developed Attention Model based on the Visual Transformer. In order to achieve the best results, the number of patches produced in each image was changed. The results of the four developed models are displayed in Table 6.6.



(a) (b) (c) (d)

Figure 6.22: Emotion Silhouette samples of SFEW Database of six different emotions: (a) Anger; (b) Fear; (c) Happiness; (d) Surprise.

CK+

Regarding CK+ the achieved results were very good apart from the model built with patches 16x16 from a non-resized image of 224 by 224. This specific model was not able to grasp the essential details from the samples and accuracy values from the testing set were very low compared to the other models. However, this value is not a result of overfitting, as during the training and validation phase the AM-(16x16) model was the only one with poor performance. The other three models show excellent performance with values above 90% which is within the results shown in the literature. From these three models, the models trained with resized images show the best results and from those two models, the one with smaller patches leads to slightly higher accuracy. This was expected as more and smaller patches should give the model more detail to make better predictions. Overall, for CK+ the AM-(6x6) model with 96.42% accuracy was the best models. As it seen in Figure 6.23, AM-(6x6) had more than 95% of accuracy in 3 of the classes and even 100% regarding Happiness. Anger and Sadness were the most challenging emotions, this can be explained by the fact that these classes had fewer samples and the model could not make a good make generalization for unseen images.

JAFFE

In comparison with CK+, Attention Model could not produce good results when trained with JAFFE images. Every single model independently from the patch size used unperformed compared to the state-of-the-art values. Still, the best performance models were once again the ones trained with images which were rezided at the beginning of the training. However, this time the

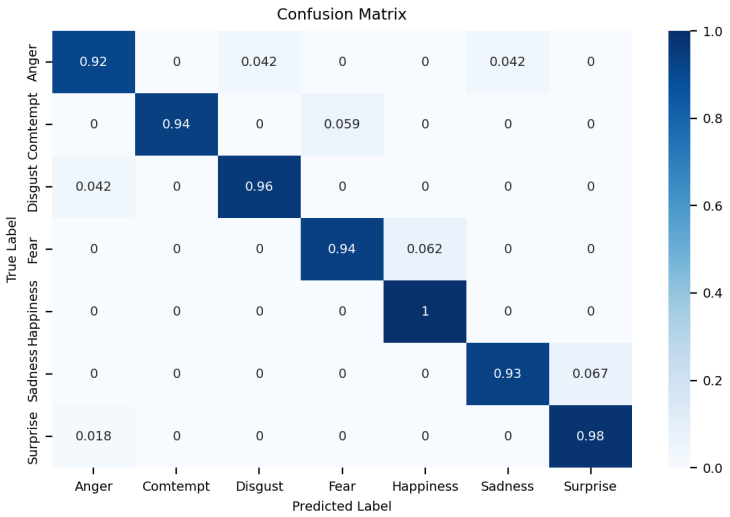


Figure 6.23: Confusion matrix for AM-(6x6).

smaller patch did not achieve the best performance. This can mean that when using JAFFE samples bigger patches are needed to capture essential features in comparison with CK+. Also, the models produced 90 plus accuracy during the training phase, but fail to reach those values during validation. Using AM-(12x12), the best performing model, as an example, we can see from training and validation accuracy evolution (Figure 6.24) that the model was able to correctly classify the samples with which it was trained, but did not gather enough information for a good generalization. This can be explained by the fact that JAFFE database was the one with fewer samples, and perhaps with a more diverse dataset training-like value could be reached during the training phase.

SFEW

About SFEW it was expected that the results would be the lowest. As it was mentioned for every previous approach, because of its more challenging nature, SFEW database produced the worst absolute results. However, if we frame the accuracy value in the overall context AM models are not far off from some of the literature performances. Once more, the smaller patches from a resized 72 by 72 image achieve better results, in this case, the AM-(12x12) was the best performing model. The maximum value reached was of 35.98% which is still better than the challenge baseline performance is 35.96% [28] but not good enough to be comparable with some of the best results. The main factor in the underachievement of the Attention Model could be the existence of multiple poses. As the model uses the information patch by patch the more “natural" faces that this database gives made the generalization more difficult resulting in poor accuracy. By analysing its confusion matrix, Figure 6.25, it is possible to see that a class like Happiness, which have the most number of samples reached a much more satisfying and well above state-of-the-art result than Disgust, for example. The more imbalanced classes once again coincide with the worst predictions, showing how important is the quality and diversity of the dataset for the model’s success.

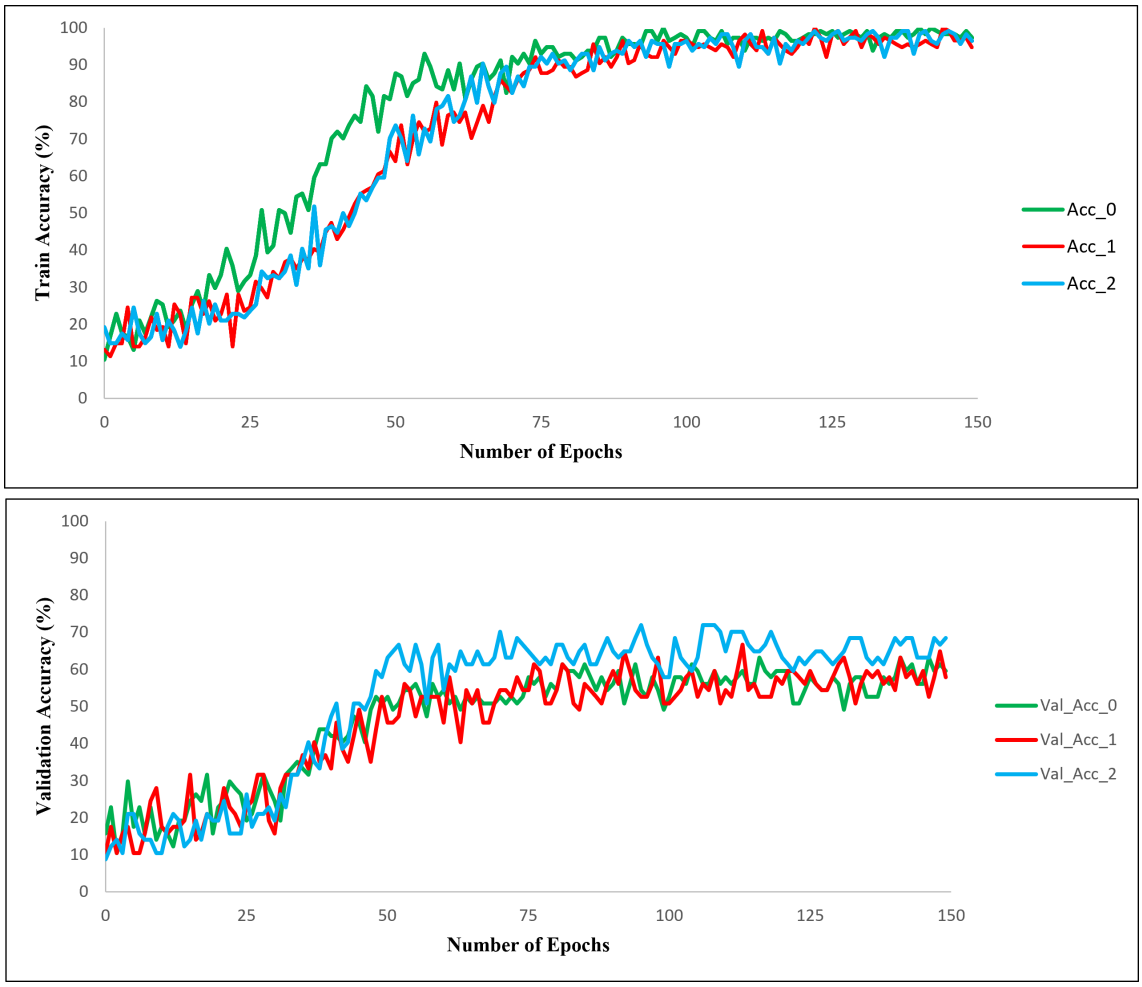


Figure 6.24: Evolution of the AM-(12x12) accuracy for training and validation phases.

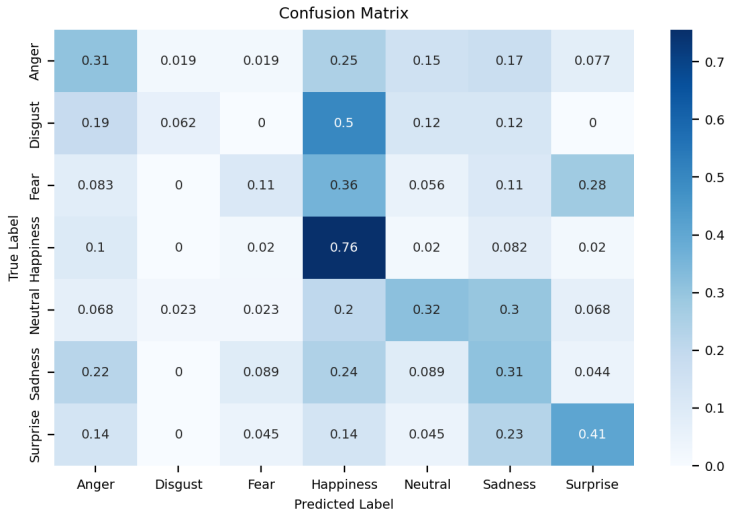


Figure 6.25: Confusion matrix for AM-(12x12).

Chapter 7

Conclusions

7.1 Summary and Conclusions

The presented work has brought a comprehensive and vast study on the different approaches regarding facial expression recognition methods. Exploring new paths and showing new ways of facing the same problems that have yet been fully developed in the state of the art.

Considering the traditional machine learning techniques, the main conclusion revolves around the fact that, mainly for smaller and more imbalanced databases like JAFFE, these approaches are still valid and achieve state-of-the-art performances. Inclusive, HOG-MLP, from all the developed models, produced the best accuracy for JAFFE samples. In contrast, these models require more pre-processing and fail to deliver the same type of results when face with more challenging databases.

Moving to the development of CNN the first and more intuitive conclusion is that the same model leads to absolutely different results when trained with different samples. This supports what was previously mention, a model is as good as the data with which it was trained.

Concerning the overall performance of the three developed models, myCNN was the one which achieved the most consistent results throughout the used database. This can be explained by the fact that this model was an intermediate step between the deeper and more complex model in myVGG and the less personalized model in TransVGG. Despite being the steady in terms of

Table 7.1: Review of the best results presented in Experiment from a to D - Results accuracy depicted in percentage(%)

	CK+	**JAFFE**	**SFEW**
HOG-MLP	100.0%	80.95%	-
CNN - Original	98.46%	61.90%	45.45%
CNN - ROI Mask	91.75%	35.71%	35.51%
CNN - OL	82.42%	54.76%	39.28%
CNN - ES	79.12%	52.38%	36.73%
AM-12	95.41%	66.66%	35.98%

accuracy, myCNN did not achieve any outstanding results, having good predictions for CK+ and SFEW and below average predictions for JAFFE.

TransVGG16 similarly with myCNN also had a very consistent performance, not excelling at CK+ like myVGG, but producing good results for SFEW and the least underwhelming values for JAFFE. This is a proof that transfer learning combined with fined tuning and with some crucial layers being unfrozen can lead to state-of-the-art accuracies. TransVGG16 proved that it was the only valid approach for the JAFFE database. As it was said this database had fewer elements than the remaining and led to very overffited myCNN and myVGG models. Due to having most of its parameters frozen, TransVGG16 was the most prepared model and achieved better scores.

Finally, myVGG had exceptional results for CK+ initially proven that a more complex and VGG-inspired model could delivered very good performance for AFER problems. However, when faced with more challenging databases, the model could not excel. Still having a considerably good performance for SFEW, myVGG had dissapointing values for the JAFFE database, having a decrease of almost 60% from one database to another. As it was discussed, the overfitting during the training phase, that could not be solve with data augmentation techniques, led to very bad predictions. More fully connected layers with dropout, combined with a better regularization and more augmented data would be the next steps in order to take this model to more acceptable results.

Regarding the landmarks-based approaches, first of all the results obtained from the three different databases were not equally conclusive. As it happened during the training phase with the original samples, Experiment B (6.4.2), the models trained mainly with JAFFE samples and in some cases with SFEW samples were higly overfitted. Thus, the comparison between the performance with original samples and with landmarks samples was not as relevant in those case than it was with CK+ database. Saying this, the overall results regarding CK+ and at some extend SFEW were much more prominent in evaluating each approach.

Secondly, all types of developed images were able to represent what is a human emotion. *ROI Mask* proved that, although, the region of the eyebrow, eyes, nose and mouth are much more prominent to the final classification, the rest of the face is also needed to achieve state-of-the-art. *Only Landmarks* showed the capability of a human face being represent only by its landmarks. Finally, *Emotion Silhouette* shown that only by using some basic shapes and lines, we are still able to achieve a pretty accurate representation of what is human facial expression.

Overall, the three approaches but mainly *Emotion Silhouette* proved our initial premise. A human face can be accurately represented using only a small number of facial landmarks. Inevitably there was a substantial decrease on the global accuracy score, but comparing the quantity of information given to the models, *Emotion Silhouette* shows that it is possible to reach very good performances when using the correct bare minimum. Giving the model successively less information starting with original samples and ending up with only few shapes and lines, the models had only a maximum drop of 17% for CK+ and only 5% for SFEW, showing empirically the partial success of these approaches (Table 7).

Finally, regarding the developed Attention Model, our final approach showed the potential of the self-attention mechanisms and the ViT. Once again, we proven that different facial regions

contribute differently to the final output. AM-12 model was able to achieve very consistent values independently from the database use, having the second best accuracy for both CK+ (95.41%) and even JAFFE (66.66%) databases (Table 7).

All in all, as it was expected in every experiment some expression revealed to be more challenging that others, this is a direct consequence of certain emotions being dependent on classic facial traits and others being more unique. Above all, the most important conclusion is that emotions are dynamic processes. The secret is, not only in the individual details of each facial landmark, but also in the way those area relate to each other. In order to find the optimal essential portrayal of what is a human facial expression, we need to know how to grasp each individual detail intrinsic to an emotion.

7.2 Future work

Future work based on this dissertation must, first of all extend the produced work to new and more natural databases. There are many datasets for facial expression recognition which have not been used during this dissertation and might give new information about the developed approaches.

This dissertation opens the doors showing that minimalist representation can indeed lead to very promising results. However, the optimal solution is yet to be find. Future works should continue to experiment and find out which combination of landmarks lead to a better and more consistent minimalist design.

Finally, the results of the developed attention model showed very promising results. Saying this, training the developed attention model with landmark orientated samples would be the next natural step for future work. As both approaches complement each other and revolve around the same premise.

References

[1] Facial expression recognition (face recognition techniques) part 1. URL http://what-when-how.com/face-recognition/facial-expression-recognition-face-recognition-techniques-part-1/.

[2] Types of neural networks activation functions. URL https://www.v7labs.com/blog/neural-networks-activation-functions.

[3] Emotion | definition of emotion by oxford dictionary on lexico.com. URL https://www.lexico.com/definition/emotion.

[4] Mediapipe google´s python api. URL https://google.github.io/mediapipe/.

[5] Facial action coding system, Jan 2020. URL https://www.paulekman.com/facial-action-coding-system/.

[6] T. R. Almaev and M. F. Valstar. Local gabor binary patterns from three orthogonal planes for automatic facial expression recognition. In *2013 Humaine association conference on affective computing and intelligent interaction*, pages 356–361. IEEE, 2013.

[7] G. Amato, F. Falchi, C. Gennaro, and C. Vairo. A comparison of face verification with facial landmarks and deep features. 04 2018.

[8] D. Aneja, A. Colburn, G. Faigin, L. Shapiro, and B. Mones. Modeling stylized character expressions via deep learning. In *Asian Conference on Computer Vision*, pages 136–153. Springer, 2016.

[9] S. Baluja, M. Sahami, and H. A. Rowley. Efficient face orientation discrimination. In *2004 International Conference on Image Processing, 2004. ICIP'04.*, volume 1, pages 589–592. IEEE, 2004.

[10] M. S. Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. Fasel, and J. Movellan. 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, 2005.

[11] S. E. Bekhouche, A. Ouafi, A. Benlamoudi, A. taleb ahmed, and A. Hadid. Automatic age estimation and gender classification in the wild. 11 2015. doi: 10.13140/RG.2.1.1893.8323.

[12] S. Berretti, A. Del Bimbo, P. Pala, B. B. Amor, and M. Daoudi. A set of selected sift features for 3d facial expression recognition. In *2010 20th International Conference on Pattern Recognition*, pages 4125–4128. IEEE, 2010.

[13] P. Bota, C. Wang, A. Fred, and H. Silva. Emotion assessment using feature fusion and decision fusion classification based on physiological data: Are we there yet? *Sensors*, 20 (17):4723, 2020.

[14] G. Campbell. URL https://www.youtube.com/watch?v=huIeAoj6XNc.

[15] W.-L. Chao, J.-J. Ding, and J.-Z. Liu. Facial expression recognition based on improved local binary pattern and class-regularized locality preserving projection. *Signal Processing*, 117:1–10, 2015.

[16] A. Chastel et al. *Leonardo on Art and the Artist*. Courier Corporation, 2002.

[17] V. K. Chauhan, K. Dahiya, and A. Sharma. Problem formulations and solvers in linear svm: a review. *Artificial Intelligence Review*, 52(2):803–855, 2019.

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

[19] C. Clavel, I. Vasilescu, L. Devillers, G. Richard, and T. Ehrette. Fear-type emotion recognition for future audio-based surveillance systems. *Speech Communication*, 50(6):487–503, 2008.

[20] C. A. Corneanu, M. O. Simón, J. F. Cohn, and S. E. Guerrero. 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, 2016.

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

[22] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, volume 1, pages 886–893. Ieee, 2005.

[23] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, volume 1, pages 886–893. Ieee, 2005.

[24] A. Damasio. *Feeling*. Knopf Doubleday Publishing Group, 2021.

[25] C. Darwin and P. Ekman. *Charles Darwin: The expression of the emotions in man and animals*. Harper Collins Publishers, 1999.

[26] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

[27] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[28] A. Dhall. URL https://cs.anu.edu.au/few/emotiw2015.html.

[29] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon. Collecting large, richly annotated facial-expression databases from movies. *IEEE multimedia*, 19(03):34–41, 2012.

[30] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[31] G.-B. Duchenne and G.-B. D. de Boulogne. *The mechanism of human facial expression*. Cambridge university press, 1990.

[32] P. Ekman. *Universal and cultural differences in facial expression of emotion*, volume 19. 1971.

[33] P. Ekman and W. V. Friesen. Facial action coding system. *Environmental Psychology & Nonverbal Behavior*, 1978.

[34] S. Eppel. Setting an attention region for convolutional neural networks using region selective features, for recognition of materials within glass vessels. *arXiv preprint arXiv:1708.08711*, 2017.

[35] T. Fadrus and O. Spindler. The grimace project. URL http://www.datadeluge.com/2013/03/the-grimace-project.html.

[36] P. M. Ferreira, F. Marques, J. S. Cardoso, and A. Rebelo. Physiological inspired deep neural networks for emotion recognition. *IEEE Access*, 6:53930–53943, 2018. doi: 10.1109/ACCESS.2018.2870063.

[37] Y. Gan, J. Chen, Z. Yang, and L. Xu. Multiple attention network for facial expression recognition. *IEEE Access*, 8:7383–7393, 2020.

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

[39] J. M. Girard, J. F. Cohn, M. H. Mahoor, S. M. Mavadati, Z. Hammal, and D. P. Rosenwald. Nonverbal social withdrawal in depression: Evidence from manual and automatic analyses. *Image and vision computing*, 32(10):641–647, 2014.

[40] I. J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, D.-H. Lee, et al. Challenges in representation learning: A report on three machine learning contests. In *International conference on neural information processing*, pages 117–124. Springer, 2013.

[41] M. R. Graver. Cicero on the emotions: Tusculan disputations 3 and 4. 2002.

[42] S. Greenblatt. Toward a universal language of motion: Reflections on a seventeenth century muscle man. *LiNQ*, 21(2):56–62, 1994.

[43] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker. Multi-pie. In *2008 8th IEEE International Conference on Automatic Face & Gesture Recognition*, pages 1–8, 2008. doi: 10.1109/AFGR.2008.4813399.

[44] K. Han, D. Yu, and I. Tashev. Speech emotion recognition using deep neural network and extreme learning machine. In *Interspeech 2014*, 2014.

[45] Y. N. Harari. *Sapiens: A brief history of humankind*. Harper Perennial, 2018.

[46] K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015.

[47] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition, 2015. URL https://arxiv.org/abs/1512.03385.

[48] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, October 2007.

[49] ISO/IEC 19794-5:2000(E). Information Technology - Biometric Data interchange formats - Part 5: Face Image Data. International standard, International Organization for Standardization, Geneva, CH, Nov. 2011.

[50] T. Jabid, M. H. Kabir, and O. Chae. Robust facial expression recognition based on local directional pattern. *ETRI journal*, 32(5):784–794, 2010.

[51] D. K. Jain, P. Shamsolmoali, and P. Sehdev. Extended deep neural network for facial emotion recognition. *Pattern Recognition Letters*, 120:69–74, 2019.

[52] Q. Ji, P. Lan, and C. Looney. A probabilistic framework for modeling and real-time monitoring human fatigue. *IEEE Transactions on systems, man, and cybernetics-Part A: Systems and humans*, 36(5):862–875, 2006.

[53] S. K. A. Kamarol, M. H. Jaward, J. Parkkinen, and R. Parthiban. Spatiotemporal feature extraction for facial expression recognition. *IET Image Processing*, 10(7):534–541, 2016.

[54] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

[55] G. Levi and T. Hassner. 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*, pages 503–510, 2015.

[56] H. Li, Z. Lin, X. Shen, J. Brandt, and G. Hua. A convolutional neural network cascade for face detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5325–5334, 2015.

[57] S. Li and W. Deng. Deep facial expression recognition: A survey. *IEEE transactions on affective computing*, 2020.

[58] R. Lienhart and J. Maydt. An extended set of haar-like features for rapid object detection. In *Proceedings. international conference on image processing*, volume 1, pages I–I. IEEE, 2002.

[59] C. F. Liew and T. Yairi. Facial expression recognition and analysis: a comparison study of feature descriptors. *IPSJ transactions on computer vision and applications*, 7:104–120, 2015.

[60] C. Liu and H. Wechsler. Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition. *IEEE Transactions on Image processing*, 11(4): 467–476, 2002.

[61] Y. Liu, Y. Li, X. Ma, and R. Song. Facial expression recognition with fusion features extracted from salient facial areas. *Sensors*, 17(4):712, 2017.

[62] A. T. Lopes, E. De Aguiar, A. F. De Souza, and T. Oliveira-Santos. Facial expression recognition with convolutional neural networks: coping with few data and the training sample order. *Pattern recognition*, 61:610–628, 2017.

[63] C. Lord et al. *Aristotle's "Politics"*. University of Chicago Press, 2013.

[64] D. G. Lowe. Object recognition from local scale-invariant features. In *Proceedings of the seventh IEEE international conference on computer vision*, volume 2, pages 1150–1157. Ieee, 1999.

[65] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *2010 IEEE computer society conference on computer vision and pattern recognition-workshops*, pages 94–101. IEEE, 2010.

[66] C. Lugaresi, J. Tang, H. Nash, C. McClanahan, E. Uboweja, M. Hays, F. Zhang, C.-L. Chang, M. G. Yong, J. Lee, W.-T. Chang, W. Hua, M. Georg, and M. Grundmann. Mediapipe: A framework for building perception pipelines, 2019. URL https://arxiv.org/abs/1906.08172.

[67] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba. Coding facial expressions with gabor wavelets. In *Proceedings Third IEEE international conference on automatic face and gesture recognition*, pages 200–205. IEEE, 1998.

[68] D. Matsumoto. More evidence for the universality of a contempt expression. *Motivation and Emotion*, 16(4):363–368, 1992.

[69] A. Mehrabian. Communication without words. In *Communication theory*, pages 193–200. Routledge, 2017.

[70] S. Minaee, M. Minaei, and A. Abdolrashidi. Deep-emotion: Facial expression recognition using attentional convolutional network. *Sensors*, 21(9):3046, 2021.

[71] M. Mohammadpour, H. Khaliliardali, S. M. R. Hashemi, and M. M. AlyanNezhadi. Facial emotion recognition using deep convolutional networks. In *2017 IEEE 4th international conference on knowledge-based engineering and innovation (KBEI)*, pages 0017–0021. IEEE, 2017.

[72] A. Mollahosseini, B. Hasani, and M. H. Mahoor. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, PP(99):1–1, 2017.

[73] M. Munasinghe. 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, 2018.

[74] M. Najibi, P. Samangouei, R. Chellappa, and L. S. Davis. Ssh: Single stage headless face detector. In *Proceedings of the IEEE international conference on computer vision*, pages 4875–4884, 2017.

[75] Neeraj, A. Rosebrock, Arun, Charles, Michael, Sveder, Abu, Suyuancheng, Abu, Dharmendra, and et al. Local binary patterns with python & opencv, Apr 2021. URL https://pyimagesearch.com/2015/12/07/local-binary-patterns-with-python-opencv/.

[76] B. Niu, Z. Gao, and B. Guo. Facial expression recognition with lbp and orb features. *Computational Intelligence and Neuroscience*, 2021, 2021.

[77] T. Ojala, M. Pietikainen, and T. Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on pattern analysis and machine intelligence*, 24(7):971–987, 2002.

[78] N. Özbey and M. B. Gülmezoğlu. Facial expression recognition using proposed geometric features. Technical report, EasyChair, 2021.

[79] L. Perez and J. Wang. The effectiveness of data augmentation in image classification using deep learning. *arXiv preprint arXiv:1712.04621*, 2017.

[80] P. C. Petrantonakis and L. J. Hadjileontiadis. Emotion recognition from eeg using higher order crossings. *IEEE Transactions on information Technology in Biomedicine*, 14(2):186–197, 2009.

[81] A. Proffitt. Market for emotion recognition projected to grow as some question science, Jun 2021. URL https://www.aitrends.com/emotion-recognition/market-for-emotion-recognition-projected-to-grow-as-some-question-science/.

[82] Y. Qiu and Y. Wan. Facial expression recognition based on landmarks. In *2019 IEEE 4th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC)*, volume 1, pages 1356–1360. IEEE, 2019.

[83] S. Ranganathan, K. Nakai, and C. Schonbach. *Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics*. Elsevier, 2018.

[84] F. Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386, 1958.

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

[86] C. Shan, S. Gong, and P. W. McOwan. Facial expression recognition based on local binary patterns: A comprehensive study. *Image and vision Computing*, 27(6):803–816, 2009.

[87] C. Shorten and T. M. Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of big data*, 6(1):1–48, 2019.

[88] K. Sikka, T. Wu, J. Susskind, and M. Bartlett. Exploring bag of words architectures in the facial expression domain. In *European Conference on Computer Vision*, pages 250–259. Springer, 2012.

[89] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[90] A. S. M. Sohail and P. Bhattacharya. Classification of facial expressions using k-nearest neighbor classifier. In *International Conference on Computer Vision/Computer Graphics Collaboration Techniques and Applications*, pages 555–566. Springer, 2007.

[91] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision, 2015. URL https://arxiv.org/abs/1512.00567.

[92] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning, 2016. URL https://arxiv.org/abs/1602.07261.

[93] Y.-l. Tian, T. Kanade, and J. F. Cohn. Recognizing lower face action units for facial expression analysis. In *Proceedings Fourth IEEE International Conference on Automatic Face and Gesture Recognition (Cat. No. PR00580)*, pages 484–490. IEEE, 2000.

[94] M. Valstar, M. Pantic, et al. Induced disgust, happiness and surprise: an addition to the mmi facial expression database. In *Proc. 3rd Intern. Workshop on EMOTION (satellite of LREC): Corpora for Research on Emotion and Affect*, page 65. Paris, France., 2010.

[95] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[96] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition. CVPR 2001*, volume 1, pages I–I. IEEE, 2001.

[97] X. Wang, T. X. Han, and S. Yan. An hog-lbp human detector with partial occlusion handling. In *2009 IEEE 12th international conference on computer vision*, pages 32–39. IEEE, 2009.

[98] C.-H. Wu, Z.-J. Chuang, and Y.-C. Lin. Emotion recognition from text using semantic labels and separable mixture models. *ACM transactions on Asian language information processing (TALIP)*, 5(2):165–183, 2006.

[99] H. Yang, Z. Zhang, and L. Yin. Identity-adaptive facial expression recognition through expression regeneration using conditional generative adversarial networks. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 294–301. IEEE, 2018.

[100] J. Yang, S. Liao, and S. Z. Li. Automatic partial face alignment in nir video sequences. In *International Conference on Biometrics*, pages 249–258. Springer, 2009.

[101] H. Yingge, I. Ali, and K.-Y. Lee. Deep neural networks on chip - a survey. pages 589–592, 02 2020. doi: 10.1109/BigComp48618.2020.00016.

[102] J. Yu and B. Bhanu. Evolutionary feature synthesis for facial expression recognition. *Pattern Recognition Letters*, 27(11):1289–1298, 2006.

[103] Z. Yu, G. Liu, Q. Liu, and J. Deng. Spatio-temporal convolutional features with nested lstm for facial expression recognition. *Neurocomputing*, 317:50–57, 2018. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2018.07.028. URL https://www.sciencedirect.com/science/article/pii/S0925231218308634.

[104] S. Zafeiriou, C. Zhang, and Z. Zhang. A survey on face detection in the wild: Past, present and future. *Computer Vision and Image Understanding*, 138:1–24, 2015. ISSN 1077-3142. doi: https://doi.org/10.1016/j.cviu.2015.03.015. URL https://www.sciencedirect.com/science/article/pii/S1077314215000727.

[105] C. Zhang and Z. Zhang. A survey of recent advances in face detection. 2010.

[106] F. Zhang, T. Zhang, Q. Mao, and C. Xu. Joint pose and expression modeling for facial expression recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3359–3368, 2018.

[107] H. Zhang, W. Gao, X. Chen, and D. Zhao. Object detection using spatial histogram features. *Image and Vision Computing*, 24(4):327–341, 2006.

[108] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao. Joint face detection and alignment using multi-task cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, 2016.

[109] Z. Zhang, M. Lyons, M. Schuster, and S. Akamatsu. Comparison between geometry-based and gabor-wavelets-based facial expression recognition using multi-layer perceptron. In *Proceedings Third IEEE International Conference on Automatic face and gesture recognition*, pages 454–459. IEEE, 1998.