



Universidade Estadual de Campinas
Instituto de Computação



William Marques Dias

**Cross-dataset emotion recognition from facial
expressions through convolutional neural networks**

**Reconhecimento de emoções a partir de expressões
faciais em conjunto de dados cruzados através de
redes neurais convolucionais**

CAMPINAS
2020

William Marques Dias

**Cross-dataset emotion recognition from facial expressions
through convolutional neural networks**

**Reconhecimento de emoções a partir de expressões faciais em
conjunto de dados cruzados através de redes neurais
convolucionais**

Dissertação apresentada ao Instituto de Computação da Universidade Estadual de Campinas como parte dos requisitos para a obtenção do título de Mestre em Ciência da Computação.

Dissertation presented to the Institute of Computing of the University of Campinas in partial fulfillment of the requirements for the degree of Master in Computer Science.

Supervisor/Orientador: Prof. Dr. Anderson de Rezende Rocha

Este trabalho corresponde à versão final da Dissertação defendida por William Marques Dias e orientada pelo Prof. Dr. Anderson de Rezende Rocha.

CAMPINAS
2020

Ficha catalográfica
Universidade Estadual de Campinas
Biblioteca do Instituto de Matemática, Estatística e Computação Científica
Ana Regina Machado - CRB 8/5467

D543c Dias, William Marques, 1986-
Cross-dataset emotion recognition from facial expressions through convolutional neural networks / William Marques Dias. – Campinas, SP : [s.n.], 2020.

Orientador: Anderson de Rezende Rocha.
Dissertação (mestrado) – Universidade Estadual de Campinas, Instituto de Computação.

1. Aprendizado de máquina. 2. Visão por computador. 3. Redes neurais (Computação). 4. Reconhecimento de expressões faciais. 5. Reconhecimento de emoções. I. Rocha, Anderson de Rezende, 1980-. II. Universidade Estadual de Campinas. Instituto de Computação. III. Título.

Informações para Biblioteca Digital

Título em outro idioma: Reconhecimento de emoções a partir de expressões faciais em conjunto de dados cruzados através de redes neurais convolucionais

Palavras-chave em inglês:

Machine learning

Computer vision

Neural networks (Computer science)

Facial expressions recognition

Emotion recognition

Área de concentração: Ciência da Computação

Títuloção: Mestre em Ciência da Computação

Banca examinadora:

Anderson de Rezende Rocha [Orientador]

Paula Dornhofer Paro Costa

Teófilo Emidio de Campos

Data de defesa: 02-10-2020

Programa de Pós-Graduação: Ciência da Computação

Identificação e informações acadêmicas do(a) aluno(a)

- ORCID do autor: <https://orcid.org/0000-0003-1102-5484>

- Currículo Lattes do autor: <http://lattes.cnpq.br/0263640317775658>



Universidade Estadual de Campinas
Instituto de Computação



William Marques Dias

**Cross-dataset emotion recognition from facial expressions
through convolutional neural networks**

**Reconhecimento de emoções a partir de expressões faciais em
conjunto de dados cruzados através de redes neurais
convolucionais**

Banca Examinadora:

- Prof. Dr. Anderson de Rezende Rocha (Supervisor)
Instituto de Computação - UNICAMP
- Profa. Dra. Paula Dornhofer Paro Costa
Faculdade de Engenharia Elétrica e de Computação - UNICAMP
- Prof. Dr. Teófilo Emidio de Campos
Departamento de Ciência da Computação - UNB

A ata da defesa, assinada pelos membros da Comissão Examinadora, consta no SIGA/Sistema de Fluxo de Dissertação/Tese e na Secretaria do Programa da Unidade.

Campinas, 02 de outubro de 2020

Agradecimentos

Nosso caminho é construído aos poucos, bloco após bloco, à medida que por ele avançamos. Seria, porém, impraticável percorrê-lo sem que houvesse pessoas para ajudar-nos a assentar o piso, acompanhar-nos durante o percurso ou indicar-nos a direção precisa. Deste modo, gostaria, aqui, de agradecer àqueles que contribuíram para a realização deste trabalho.

Agradeço a meus pais por todo esforço dedicado para que eu e meu irmão tivéssemos acesso a uma boa educação. Sei que vocês abriram mão de muitas coisas ao priorizar nossa formação, e eu serei eternamente grato por isso. Este trabalho é uma pequena retribuição por tudo o que fizeram por mim. Sem o incentivo de vocês, decerto, ele não seria possível.

Agradeço a minha parceira de vida e melhor amiga, Ana Julia, com quem tenho a sorte de compartilhar meus dias. Obrigado pelo companheirismo, por apoiar-me em minhas escolhas e por fazer-me evoluir como pessoa. Percorrer esse caminho com você ao meu lado é muito mais divertido.

Agradeço a meu amigo de longa data e sócio de jornada empreendedora, Rodrigo. Obrigado por estar presente no decurso de todos esses anos e por confiar em meu trabalho. Seguimos em frente, encarando os alicives desse caminho que decidimos trilhar juntos.

Agradeço ao Prof. Anderson que, lá atrás, quando mostrei interesse por uma de suas disciplinas, incentivou-me a ingressar no mestrado e se dispôs a ser meu orientador. Obrigado por apontar-me a direção e ser paciente. Nestes tempos de desinformação e descrédito à ciência, seu papel é mais que imprescindível. Em um país tão doente de desigualdade social, a educação é o único remédio possível.

Agradeço aos colegas de RECOD, em especial: Fernanda, Gabriel, Rafael e Waldir. Obrigado pelas críticas, sugestões e pelo tempo que dispuseram para me ajudar. Foi um prazer ter a companhia de vocês durante este trabalho.

Por fim, agradeço à Motorola e à CAPES^{1,2} por financiaram esta pesquisa.

¹O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001

²This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001

Resumo

O rosto é a janela da alma. É o que pensava o médico francês do século XIX, Duchenne de Boulogne. Usando choques elétricos para estimular contrações musculares e induzir expressões de aparência assustadora e bizarra, ele queria entender como os músculos produzem expressões faciais e, assim, revelar as emoções mais ocultas do ser humano. Passados dois séculos, esse campo de pesquisa continua muito ativo, despertando o interesse de diversos segmentos da indústria. Vemos sistemas automáticos de reconhecimento de emoção e expressão facial sendo aplicados na medicina, em sistemas de segurança e vigilância, em propaganda e *marketing*, entre outros. Mas, apesar de sua ampla adoção, ainda existem questões fundamentais que os cientistas estão tentando responder quando analisamos o estado emocional de uma pessoa a partir de suas expressões faciais. É possível inferir, com segurança, o estado interno de alguém baseando-se apenas nos movimentos de seus músculos faciais? Existe uma configuração facial universal para expressar raiva, repulsa, medo, felicidade, tristeza e surpresa, comumente chamadas de emoções básicas? Nesta pesquisa, tentamos responder a essas questões explorando redes neurais convolucionais. Diferentemente da maioria dos estudos disponíveis na literatura, estamos particularmente interessados em examinar se as características aprendidas em um grupo de pessoas podem ser empregadas para prever, com sucesso, as emoções de outro. Nesse sentido, adotamos um protocolo de avaliação em conjunto de dados cruzados para mensurar o desempenho dos métodos propostos. Nosso método de base foi construído a partir do ajuste fino de um modelo originalmente empregado no problema de reconhecimento facial para o problema de categorização de emoções. Em seguida, aplicamos técnicas de visualização de dados para entender o que nossa rede de base havia aprendido para, então, derivarmos três outros métodos. O primeiro método visa direcionar a atenção da rede para regiões da face consideradas importantes na literatura, mas ignoradas pelo nosso modelo inicial, usando uma arquitetura multi-ramificada para uma abordagem baseada em partes. No segundo método, simplificamos essa arquitetura e trabalhamos nos dados de entrada, ocultando partes aleatórias da imagem facial, de modo que a rede pudesse aprender características discriminativas em diferentes regiões. No terceiro método, exploramos uma função de perda que gera representações de dados em espaços de alta dimensão, de forma que exemplos de uma mesma classe de emoção fiquem próximos e exemplos de classes diferentes fiquem distantes. Finalmente, investigamos a complementaridade entre dois de nossos métodos, propondo uma técnica de fusão tardia que combina seus resultados por meio da multiplicação de probabilidades. Para efeito de comparação de nossos resultados, compilamos uma extensa lista de trabalhos avaliados nos mesmos conjuntos de dados escolhidos. Em todos eles, quando comparados a trabalhos que seguiram um protocolo de avaliação em um único conjunto de dados, nossos métodos apresentam números competitivos. Já sob um protocolo de conjunto de dados cruzados, obtivemos resultados do estado da arte, superando até mesmo aplicações comerciais de grandes empresas de tecnologia.

Abstract

The face is the window to the soul. This is what the 19th-century French doctor, Duchenne de Boulogne, thought. Using electric shocks to stimulate muscular contractions and induce creepy and bizarre-looking expressions, he wanted to understand how muscles produce facial expressions and, thus, reveal the most hidden human emotions. Two centuries later, this research field remains very active, arousing the interest of several segments of the industry. We see automatic systems for recognizing emotion and facial expression being applied in medicine, in security and surveillance systems, in advertising and marketing, among others. But despite its widespread adoption, there are still fundamental questions that scientists are trying to answer when analyzing a person’s emotional state from their facial expressions. Is it possible to reliably infer someone’s internal state based only on the movements of their facial muscles? Is there a universal facial setting to express anger, disgust, fear, happiness, sadness, and surprise, commonly referred to as basic emotions? In this research, we seek to address some of these questions through convolutional neural networks. Unlike most studies in prior art, we are particularly interested in examining whether characteristics learned in one group of people can be generalized to successfully predict the emotions of another. In this sense, we adopted a cross-dataset evaluation protocol to assess the performance of the proposed methods. Our baseline method was created by custom-tailoring a model originally used in the problem of face recognition to the problem of emotion categorization. Next, we applied data visualization techniques to account for what our baseline model had learned in order to, then, derive three other methods. The first method aims to direct the network’s attention to regions of the face considered important in the literature but ignored by our baseline model, using a multi-branched network architecture for a parts-based approach. In the second method, we simplified this architecture and worked on the input data, hiding random parts of the facial image, so that the network could learn discriminative characteristics in different regions. In the third method, we explored a loss function that generates representations of data in high-dimensional spaces, so that examples of the same emotion class are close and examples of different classes are distant. Finally, we investigated the complementarity between two of our methods, proposing a late-fusion technique that combines their outputs through the multiplication of probabilities. To compare our results, we have compiled an extensive list of works evaluated in the same adopted datasets. In all of them, when compared to works that followed an intra-dataset protocol, our methods present competitive numbers. Under a cross-dataset protocol, we achieved state-of-the-art results, outperforming even commercial off-the-shelf solutions from well-known tech companies.

List of Figures

2.1	Examples of Action Units in the Facial Action Code	19
4.1	Examples from the Bosphorus dataset	26
4.2	Examples from the CFEE dataset	27
4.3	Example of sequence from the CK+ dataset	27
4.4	Examples from the KDEF dataset	28
4.5	Examples from the NVIE dataset	29
4.6	Example of sequence from the Oulu-CASIA dataset	30
4.7	Examples from the RaFD dataset	30
5.1	VGG-Face architecture	34
5.2	Pre-trained VGG-Face model as a feature extractor	36
5.3	Fine-tuning the pre-trained VGG-Face model	36
5.4	Regions of interest for Parts-based VGG-Face	41
5.5	Parts-based VGG-Face architecture	42
5.6	Random Patches batch example	43
5.7	Triplets categories	44
5.8	Triplet Loss architecture	45
6.1	Face landmarks	47
6.2	Face bounding box	47
6.3	Data augmentation operations	49
6.4	Baseline results: Fine-tuned VGG-Face model	51
6.5	Mean image for anger emotion from CAFE dataset	51
6.6	Mean images from NVIE dataset	52
6.7	Method 1 results: Parts-based VGG-Face	53
6.8	Method 2 results: Random Patches	55
6.9	Method 3 results: Triplet Loss	57
6.10	Methods comparison in terms of accuracy	58
6.11	Methods comparison in terms of class recall	58
6.12	Combination comparison in terms of accuracy	60
6.13	Combination comparison in terms of class recall	61
6.14	Distribution of signed-ranks sum W for $n = 5$	62

List of Tables

2.1	Organismic subsystems, functions and components of emotion	17
2.2	Examples of Action Units in the Facial Action Code	19
4.1	Datasets summary	31
4.2	Datasets breakdown	31
4.3	Train and validation datasets	32
4.4	Test datasets	32
5.1	Data visualization methods comparison	39
5.2	Mean images and mean emotion heatmaps	40
6.1	Baseline results: Fine-tuned VGG-Face model	50
6.2	Method 1 results: Parts-based VGG-Face	53
6.3	Method 2 results: Random Patches	54
6.4	Method 3 results: Triplet Loss	56
6.5	Methods comparison	57
6.6	Combination comparison	59
6.7	Combination results: Random Patches and Triplet Loss	60
6.8	Wilcoxon signed-rank test: Combination vs Baseline	62
6.9	Wilcoxon signed-rank test: Combination vs Method 2	63
6.10	Wilcoxon signed-rank test: Combination vs Method 3	63
6.11	Performance comparison for Bosphorus dataset	64
6.12	Performance comparison for CAFE dataset	64
6.13	Performance comparison for CFEE dataset	65
6.14	Performance comparison for KDEF dataset	65
6.15	Performance comparison for RaFD dataset	66
A.1	Emotions and Action Units	80

Contents

1	Introduction	12
1.1	Research questions	13
1.2	Contributions	14
1.3	Text organization	14
2	Theoretical overview	16
2.1	Emotion and other affective phenomena	16
2.2	Facial expressions and emotion	18
3	Related work	20
3.1	Handcrafted features	20
3.1.1	Shape features	20
3.1.2	Appearance features	21
3.1.3	Shape and appearance features combined	22
3.2	Data-driven features	22
3.2.1	CNN-derived features	23
3.2.2	Other data-driven derived features	24
4	Datasets	25
4.1	Bosphorus	25
4.2	CAFE	26
4.3	CFEE	26
4.4	CK+	27
4.5	KDEF	27
4.6	MUG	28
4.7	NVIE	28
4.8	Oulu-CASIA	29
4.9	RaFD	30
4.10	Datasets summary	30
5	Proposed methods	33
5.1	A pre-trained model as a baseline	33
5.2	Data visualization	37
5.3	Method 1: Parts-based VGG-Face	38
5.4	Method 2: Random Patches	41
5.5	Method 3: Triplet Loss	42

6	Experimental results	46
6.1	Data preparation	46
6.1.1	Data augmentation	47
6.2	Evaluation metrics	48
6.3	A pre-trained model as a baseline	50
6.4	Method 1: Parts-based VGG-Face	52
6.5	Method 2: Random Patches	54
6.6	Method 3: Triplet Loss	55
6.7	Methods comparison	56
6.8	Statistical test	61
6.9	Performance comparison with prior art	63
7	Conclusion and future work	67
	Bibliography	71
A	EMFACS	77

Chapter 1

Introduction

Emotion and facial expression recognition has gained increased attention lately. In the last decade, we have seen many companies focused on solving problems in the area popping up in the media. Emotient, FacioMetrics, Affectiva, Imotions, Face++, Eyeris, Sightcorp, Noldus, Kairos, and Nviso are some of the leading companies in the segment. This growing interest has also caught up the attention of some giants from Silicon Valley. Recent moves made by Apple [72], which acquired Emotient, and Facebook [73], which acquired FacioMetrics, expose the importance of understanding emotion and facial expressions to build systems that can effectively perceive and predict human needs.

Forecasts estimate that the emotion recognition market will grow from USD 21.6 Billion in 2019 to USD 56.0 Billion by 2024 [45]. This boost will be mainly due to the rising demands in sectors like marketing and advertising, healthcare, banking, defense, and commercial security, in which people have been trying to accomplish tasks such as (a) testing the impact and acceptance of content, product or service by analyzing facial responses of customers; (b) monitoring facial stress levels of individuals for security and safety purposes (e.g., the automotive industry using emotion recognition to improve car safety [1]); (c) detecting pain and recognizing depression in patients undergoing clinical treatments; and (d) measuring satisfaction levels of users while viewing a website, playing a game or using a software.

Despite all the progress in the field, some topics remain as open issues yet to be addressed. Head-pose variations, illumination changes, and the distinction between facial expressions of emotion and facial expressions caused by speech are some of the challenges of the overall pipeline in facial expression and emotion recognition systems. More specifically, these issues come into view when working in real-world environments, in which expressions occur in a spontaneous way and their intensities are typically low to moderate [56, 5]. In a world in which everything is mobile, connected and interactive, emotions and facial expressions analyses are expected to perform under daily uncontrolled situations in order to meet the current and future needs of people.

But if we take our attention only to these said practical challenges, we might miss the big picture. Take one step back and we will see that there are some fundamental questions whose answers scientists are still grasping with and seeking a consensus [4, 23]. Can we reliably judge how someone feels from their facial expressions? Do different cultures display facial expressions in the same manner to express the same emotions? In spite

of these open questions, we have seen a growing interest in artificial intelligence-based systems to automatically detect emotions from facial expressions. But, should we rely on such systems to solve a problem of this complexity?

With this backdrop, in this work, we rely upon deep learning techniques to analyze emotions from facial expressions, with a special interest in **cross-dataset scenarios**. As opposed to most existing solutions, which claim to have solved the problem by only evaluating their methods under an intra-dataset protocol, we want to know if characteristics learned from one group of people can be employed to successfully predict emotions in another one. By looking at a person’s portrait, our method should be able to decipher the presented facial expressions and reveal which of the seven basic emotions – anger, disgust, fear, happiness, neutral, sadness, and surprise – the person is more likely to be communicating. Particularly, it is desirable that our model shows an appropriate level of generalization, by adapting to new, previously unseen data (i.e. portraits of different people, collected under different conditions), as opposed to one used to create it.

Recent studies in the field [36], especially the ones from the winners of the Facial Expression Recognition and Analysis Challenge (FERA) [19] and the Emotion Recognition in the Wild Challenge (EmotiW) [17], show that deep features for automated facial expression and emotion recognition offer state-of-the-art results. Hence, in this work, rather than using handcrafted features (c.f., Section 3.1) to represent the captured information, we learn them directly from data through deep neural networks. First, we derive our baseline method from a network pre-trained for the face recognition task. Then, after applying accountable machine-learning techniques to interpret initial results, we introduced two methods that employ attention mechanisms in an attempt to learn alternative discriminative regions of the face, being the second method a simplification of the first in terms of computational costs. Finally, we proposed a third method, complementary to the first two, which learns data embeddings in a high dimensional vector space by the notion of similarity and dissimilarity.

1.1 Research questions

In order to guide our research, we formulate some investigative questions regarding emotion recognition from facial expressions and the quest to find a robust automated solution to accomplish the task in a cross-dataset scenario.

- Can characteristics learned from one group of people be employed to successfully predict emotions in another one?
- Is it possible to improve model generalization and reduce dataset bias while maintaining the same base solution?

1.2 Contributions

In summary, our research introduces the following contributions:

- We propose three data-driven methods, based on a convolutional neural network (CNN), for emotion recognition from facial expressions:
 - In the first method, we define the four most important regions of interest present in the face and, using a multi-branched CNN, drive the network’s attention to extract and combine features from these regions.
 - In the second method, we force a CNN to learn features from less obvious regions of the face by iteratively occluding some parts of the input images during the network training process.
 - The third method employs a metric learning approach which maps distances in a high dimensional space to emotion similarities, inducing proximity between examples from the same class of emotion while distancing examples with distinct emotion classes.
- We also propose to combine two of our methods, which seem to be complementary to each other, under a late-fusion approach.
- Differently from most of the works present in the literature, all proposed methods are evaluated in a **cross-dataset fashion**, a more challenging scenario than using the same dataset for both training and testing. The assembled collection of datasets used in this research provides multi-cultural data from individuals in a wide age range, captured in varied conditions of illumination and background information.
- Finally, we perform an extensive evaluation of the methods available in the literature in comparison to our own. Even though most of them achieve state-of-the-art results in the intra-dataset scenario, they often lack generalization when analyzing images with different conditions than those seen in training.

1.3 Text organization

We start by presenting a theoretical overview of emotion and facial expression from the perspective of neuroscience, psychology, and cognitive science in Chapter 2. There, we describe the differences between emotion and other affective phenomena and detail how facial expressions can relate to the perception of emotion. In Chapter 3, we compile a list of related works on automatic emotion recognition from facial expressions, giving particular attention to the different feature extraction methods and classification algorithms employed.

In Chapter 4, we detail the datasets used in this research, pointing out their main properties. Moving on to Chapter 5, we propose three methods based on convolutional neural networks towards an automated emotion recognition solution and explain the process that led us to them. The results of the three proposed methods, evaluated in a

cross-dataset setup, are reported and discussed in Chapter 6. There, we also compare our methods to existing approaches present in the literature. Concluding, in Chapter 7, we present our final remarks about the findings of this research and outline new directions to guide future work.

Chapter 2

Theoretical overview

In this chapter, we present an overview of emotions and facial expressions through the lens of neuroscience, psychology, and cognitive science. We introduce a formal definition of the term “emotion” so that we can differentiate it from other affective phenomena. Additionally, we discuss how facial expressions and emotions can be organized and how they relate to one another.

2.1 Emotion and other affective phenomena

There is a wide diversity of definitions for the concept of emotion. Kleinginna and Kleinginna [32] compiled an extensive list of 92 definitions and 9 skeptical statements from a variety of sources in the literature, evincing the lack of consensus in the scientific community. On the basis of this study, they proposed a broad definition of the term:

“Emotion is a complex set of interactions among subjective and objective factors, mediated by neural/hormonal systems, which can (a) give rise to affective experiences such as feelings of arousal, pleasure/displeasure; (b) generate cognitive processes such as emotionally relevant perceptual effects, appraisals, labeling processes; (c) activate widespread physiological adjustments to the arousing conditions; and (d) lead to behavior that is often, but not always, expressive, goal-directed, and adaptive”.

Scherer [60, 61] stated that emotions are normally triggered by events that are relevant to the organism. These events can be external, such as the behavior of others, a change in a current situation or an encounter with novel stimuli, or internal, such as thoughts, memories, or sensations. Emotion episodes are intertwined, synchronized changes in the states of all or most of the five organismic subsystems, that last for a certain duration and then fade away with decreasing intensity. The state of each subsystem is a component of the emotion episode. Table 2.1 shows the suggested correspondence between organismic subsystems, functions, and components of emotion.

Dolan [10] argues that emotions, as psychological experiences, have unique qualities, when distinguishing it from other affective phenomena. He affirms that, unlike other psychological states, emotions are manifested after stereotyped patterns of facial expressions

in a uniquely recognizable way, that they are less susceptible to our intentions, and that they are more capable of influencing other aspects of cognition. In recent years, however, these assumptions were put into question by the scientific community.

Barrett et al. [4] state that there seems to be a more common facial configuration to express some emotions, but it is not forcibly a rigid one-to-one mapping between facial expressions and emotion categories. In fact, facial configurations can vary substantially across cultures and situations and are more context-dependent than previously thought. Furthermore, similar facial configurations can possibly express instances of different emotion categories.

Organismic subsystem	Emotion component	Emotion function
Information processing	Cognitive component (appraisal)	Evaluation of objects and events
Support	Neurophysiological component (bodily symptoms)	System regulation
Executive	Motivational component (action tendencies)	Preparation and direction of action
Action	Motor expression component (facial and vocal expression)	Communication of reaction and behavioral intention
Monitor	Subjective feeling component (emotional experience)	Monitoring of internal state and organism–environment interaction

Table 2.1: Relationship between organismic subsystems, functions, and components of emotion. Extracted from [61].

Treating emotion and feeling as synonyms is a frequent source of confusion, says Scherer [61]. As shown in Table 2.1, extracted from his work, while emotion is the whole phenomenon, consisting of a multi-modal component process, feeling is the subjective emotional experience component of emotion. Damasio [9] affirms that emotional events lead to rapid and automatic responses that contrast with more long-term modulatory behavioral influences mediated by feeling states. He even suggests that humans have distinct brain systems to support emotional perception and feeling states. Sentiment, as proposed by Gordon [22], on the other hand, “is a socially constructed pattern of sensations, expressive gestures and cultural meanings, organized around a relationship to a social object, usually another person”. Grief, love, envy, and hatred, for instance, persist beyond the duration of bodily changes. They are social rather than organic states.

2.2 Facial expressions and emotion

As pointed out in the previous section, facial expressions are one of the possible physiological responses to emotion. Ekman and Friesen, in their study [15], hypothesized that the relationship between distinctive patterns of facial muscles and particular emotions are universal, although cultural differences would be seen in some of the stimuli due to social settings influence. They conducted an experiment in which they showed still photographs of faces to people from different cultures (Brazil, United States, Argentina, Chile, and Japan) and asked them to identify the emotion conveyed on each photograph. To overcome the probable bias created by the exposure of these aforementioned cultures to mass media presentations of facial expressions, they also extended the study to members of isolated communities in New Guinea.

Despite the failure of New Guinean people to discriminate fear from surprise, the results corroborate the proposed hypothesis. Ekman [13, 14] believes that it is reasonable to say that the connection between particular facial configurations and specific emotions are universal. Nevertheless, that does not mean that expressions will always occur when emotions are experienced, nor does it mean that emotions will always occur when a facial expression is shown (we are capable of inhibiting and fabricating expressions). Anger, contempt, disgust, fear, happiness, sadness, and surprise are believed to have a universal facial configuration and are known as basic emotions. This discrete perception of facial expressions of emotion, in which there is a finite set of predefined classes, is known as the categorical model.

Apart from studies on emotion characterization, authors have also targeted on the examination of facial expressions in a more physiological way. Facial expressions can be described in terms of coding schemes, in which each facial configuration receives a unique parametrization. The most explored coding scheme is the Facial Action Code System (FACS), a framework for measuring visibly different facial movements, developed by Ekman and Friesen [16]. FACS can be used to describe any facial movement in terms of anatomically based Action Units (AU). The authors developed a comprehensive system by associating how each muscle of the face acts to change visible appearance and defined Action Units that all facial expressions could be broken down into. Table 2.2 shows some examples of Action Units in the Facial Action Code. Figure 2.1 depicts the relationship between facial muscles and Action Units.

According to Ekman [12], facial expressions are considered to be organized in families. He argues that a variety of related but visually different expressions can occur as a response to an emotion. Variations within a family of facial expressions reflect the intensity of the emotion, whether the emotion is controlled, simulated, or spontaneous. The Emotion FACS (EMFACS), developed Friesen and Ekman [20], scores facial actions in terms of the latent emotions or affects that are believed to generate them. EMFACS is a selective application of FACS scoring in which only certain combinations of AUs, that is, the combinations that have found to suggest emotions, are coded. Appendix A contains a detailed table showing the combination of Action Units present in each basic emotion.

Although the literature may provide other emotion characterization models, in this reasearch, we assume the categorical model. As we are interested in pointing out which of

AU	FAC Name	Muscular Basis
1	Inner Brow Raiser	Frontalis, Pars Medialis
2	Outer Brow Raiser	Frontalis, Pars Medialis
4	Brow Lowerer	Depressor Glabellae; Depressor Supercilli; Corrugator
5	Upper Lid Raiser	Levator Palpebrae Superioris
6	Cheek Raiser	Orbicularis Oculi, Pars Orbitalis
7	Lid Tightener	Orbicularis Oculi, Pars Palebralis

Table 2.2: Examples of Action Units in the Facial Action Code. Extracted from [16].

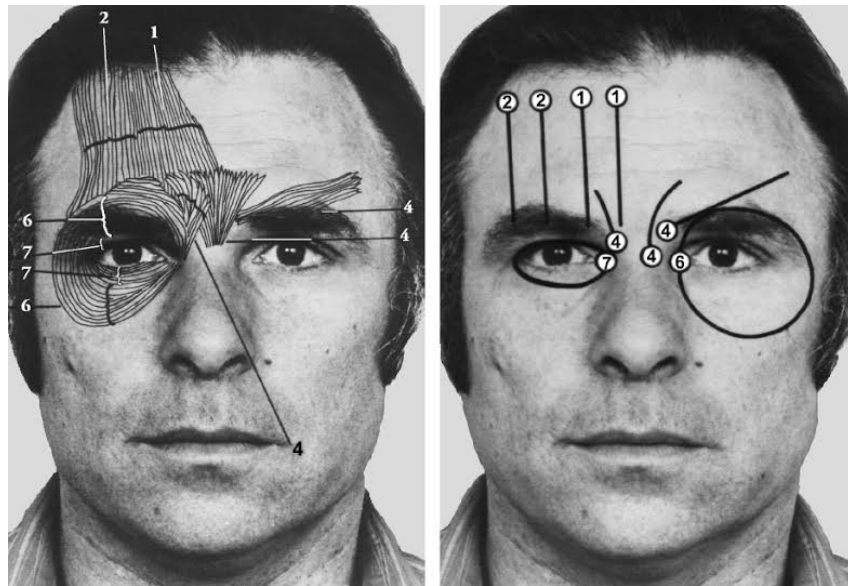


Figure 2.1: Examples of Action Units in the Facial Action Code. Extracted from [37].

the seven basic emotions – anger, disgust, fear, happiness, neutral, sadness, and surprise – the person is more likely to be communicating, the categorical model can provide us an easy way to map facial expressions into emotion classes through the aid of the EMFACS frameworks.

Chapter 3

Related work

Over the last 20 years, an extensive body of work on facial expression and emotion recognition has been proposed. Considering the *categorical model* and the EMFACS framework, while some studies focus on detecting Action Units (AUs) to, subsequently, find the corresponding emotion, others attempt to recognize the basic emotion directly. But, despite the differences, there exists a consensus in the literature on a pipeline for automatic emotion recognition from facial expressions, as suggested by Sariyanidi et al. [56] and Corneanu et al. [5] in their surveys. This pipeline can be decomposed into four fundamental components: face localization, face registration, feature extraction and emotion recognition.

Face localization and registration have been vastly explored by the scientific community, as they serve as foundation to a myriad of face-related problems. Therefore, in this study, we will focus on the last two steps of the pipeline: the feature extraction and the emotion recognition components. Hence, in this chapter, we will discuss in more depth common approaches that address these problems within the context of the categorical model, in which emotions are discretely classified into one of the seven basic categories – anger, disgust, fear, happiness, neutral, sadness, and surprise.

3.1 Handcrafted features

Handcrafted features are manually engineered data representations whose creation involves a human design through the use of domain knowledge. These features are then fed to a machine-learning algorithm to perform typical classification or regression tasks. Handcrafted features, in the context of emotion recognition, is generally divided into two groups: shape and appearance features. However, it is also possible to combine both and benefit from their positive aspects concurrently.

3.1.1 Shape features

Shape features are obtained by transforming face information into geometric properties, such as point coordinates, distances and angles. To accomplish that, distinctive regions of the face, such as eyes, mouth, nose and chin, are located and their positions are registered

as landmarks. From these registered points, we extract the desired descriptors to train a machine-learning model.

Zhang et al. [85], for instance, benefited from a Microsoft Kinect sensor to track 121 fiducial points and construct a mesh representation. Features are generated from facial points displacements between expressive and neutral faces and only the most informative ones are selected. The selection process employed is both manual, using domain knowledge, and automatic, based on minimal-redundancy-maximal-relevance criterion (mRMR). Finally, a Neural Network (NN) and a Support Vector Regressor (SVR) estimate the intensity of targeted AUs.

Ahmed et al. [2], in turn, divided the face into three regions of interest (eyes region, mouth regions and a variable auxiliary region) and tracked fiducial points from these regions to generate shape features based on the Euclidean distance between each pair of points. Emotion recognition was carried out by an SVM classifier. Similarly, Zheng et al. [86], in their study about adopting existing methods tuned for adults to recognize facial expressions in children, extracted 68 facial landmarks and used an SVM to categorize the facial expressions based on their positions.

Though vastly explored, mainly in the past, shape features are known for their limitations. Their performance is highly sensitive to errors in the registration phase, especially if there are variations in head tilt, and is subject to identity bias due to subjects' unique facial structure [56, 5]. Common approaches to mitigate these issues include data normalization and the use of the face in a neutral state to reduce user specificities.

3.1.2 Appearance features

Appearance features, on the other hand, use texture information by considering the intensity values of pixels. Over the years, a series of different algorithms that use pixel intensity levels to extract visual descriptors have been proposed in the literature. Among them, it is possible to find methods using histogram representations to encode information, methods based on kernel convolution that look for specific frequency content in the image, and methods based on the detection of robust local features.

Silva and Pedrini [6], for example, employed Local Binary Patterns (LBP), Histogram of Oriented Gradients (HOG) and Gabor filters as features to describe the facial expressions for six different basic emotions. For comparison's sake, they applied the Principal Component Analysis (PCA) to the LBP and the Gabor filters feature vectors in order to reduce their dimensionality to the same size of the HOG feature vector. The evaluation procedure was carried out by three classification algorithms: SVM, NN and K-Nearest Neighbors (KNN).

Savran et al. [58] used Gabor wavelets to analyze faces for an automatic AU recognition. Feature selection is performed by the AdaBoost algorithm, based on Gabor magnitude responses for each AU. Then, an SVM classifier was trained on top of the most discriminative features to automatically detect AUs. Nagpal et al. [47] compared the use of Dense Scale-Invariant Feature Transform (DSIFT) feature detection algorithm along with a Random Forest (RF) classifier to categorize emotions against several other methods. One of them, which also employed an RF as a classifier, used a PCA as a feature

extractor, applying it directly to the holistic representation of the face.

While appearance features are usually robust against registration errors and geometric transformations, identity bias remains a major issue for this kind of descriptor [56]. Furthermore, as the list of possible appearance features is quite extensive, finding the most adapted descriptor for our problem is nearly an impractical challenge.

3.1.3 Shape and appearance features combined

To take advantage of the best of both worlds, some studies combine shape and appearance features when extracting information from facial expressions.

Benitez-Quiroz et al. [18] explored distances between landmarks and angles defined by Delaunay triangles emanating from these landmarks in conjunction with Gabor filters. The combined feature vector was then used in a Kernel Subclass Discriminant Analysis (KSDA) to recognize AUs and their intensities. Du et al. [11] also combined shape and appearance features, but for the categorization of the six basic emotions. Features generated by Gabor filters and pairwise distance between facial landmarks were merged into a single vector to train a KSDA and an SVM classifiers.

Lucey et al. [41] extracted Similarity Normalized Shape Features (SPTS), which refers to the 68 facial landmarks tracked by an Active Appearance Model (AAM), resulting in a feature vector of size 136. They also extracted Canonical Normalized Appearance Features (CAPP) by applying a piece-wise affine warp on each Delaunay triangle so that the source image aligns with the base face shape. These combined features were then used to train an SVM to detect AUs.

In this research, however, due to the intrinsic limitation of both shape and appearance features, we opted to learn descriptors directly from data. Our choice is backed up by recent results in emotion recognition competitions, such as FERA [19] and EmotiW [17], whose winners use deep features in their solutions.

3.2 Data-driven features

Unlike its handcrafted counterpart, data-driven features are generated by algorithms in an automated fashion, following an optimization process to best suit the task we are trying to accomplish. According to Li and Deng [36], for a variety of applications, including emotion recognition, features learned through deep learning techniques have been shown to achieve state-of-the-art performances.

Inspired by the hierarchical architecture of the human biological neural system, deep learning architectures attempt to capture high-level abstractions through multiple stacked layers of learning nodes. As data flows through the network, each layer transforms the output of the previous level into a slightly more abstract representation and before passing it to the next level.

In computer vision problems, such as emotion recognition from facial images, a particular class of deep architecture called convolutional neural networks has been extensively used with great success [64]. In its core building blocks lies the convolutional layer, which consists of a set of learnable filters that convolves across the whole input image, computing

a dot product to produce 2-dimensional activation feature maps. These convolutional operations learn correlations among neighboring pixels in such a way the filters get activated when some specific type of feature is detected.

3.2.1 CNN-derived features

There is a series of considerations one can make when building a CNN based solution. To begin with, it is necessary to decide upon the network architecture, the types of layers it should have, their quantities and how they connect to each other. Also, another frequent concern has to do with how data is inputted to the network. We must choose the kind of image to use and determine if some sort of preprocessing is required.

Shin et al. [65] proposed a baseline CNN structure for facial expression recognition by analyzing four different architectures. These networks were trained with five types of input data: raw, histogram equalization, isotropic smoothing, diffusion-based normalization and difference of Gaussian. Comparing the results, the best accuracy was obtained by a three-level structure, each level consisting of a convolutional and a max-pooling layer, fed with the histogram equalized image.

Yet, in this research, instead of assessing the performance of different network architectures for emotion categorization, we picked just a single model and explored different ways to adapt it to our problem. Our selection is quite popular and known for its competitive results in face recognition.

One of the paths we explored was to analyze how input data can be manipulated to help us improve results. We took the same direction of Ruiz-Garcia et al. [53], who compared two different approaches to classify images into seven emotions using CNNs. In the first one, they examined an architecture with a reduced number of convolutional layers, while in the second approach, they horizontally split the input image into two parts based on eyes and mouth positions.

Li et al. [35] presented a Deep Fusion Convolution Neural Network (DF-CNN) for multimodal 2D and 3D facial expression recognition. DF-CNN is composed by a feature extraction subnet, a feature fusion subnet, and a softmax layer. Textured 3D face scans are decomposed into six 2D maps (geometry, normal-x, normal-y, normal-z, curvature and texture maps) which are then fed to the DF-CNN for feature extraction and fusion, generating a 32-dimensional feature vector. Predictions for one of the six base emotions are made by both a softmax layer and a linear SVM.

The choice for CNN-derived features, however, brings us some substantial challenges. Training a CNN from scratch, for instance, is not always possible or desirable, especially if we do not have sufficient data. In fact, the optimization process, in which the weights of a deep architecture are learned, is slow and cumbersome, involving a multitude of mathematical operations. Having good initial values for the network weights can help us speed up that process and achieve better results.

Ruiz-Garcia et al. [54] analyzed the effects of different weight initializations of a CNN for emotion recognition using facial expression images. They compared the performance of a CNN when its weights are randomly initialized to the performance when pre-training each layer of the network as an Auto-Encoder in an unsupervised fashion as a Stacked

Convolutional Auto-Encoder (SCAE).

An alternative approach regarding weight initialization would be to explore pre-trained models. In this technique, we transfer the knowledge acquired in one domain to another one by either using the pre-model as a feature extractor or by fine-tuning it. The closer the source task is to the target task, the better the results [81]. In our case, the pre-trained model we selected was trained on face images, the same kind of images used to recognize emotions from facial expressions.

Zhou and Shi [87] used the first five convolutional layers of a network pre-trained on a visual object recognition dataset as a feature extractor. It generates 256 feature maps that are forwarded to three different subsystems. The first uses all 256 feature maps. The second uses a deconvolutional neural network to filter out feature maps responding primarily to patterns in the background. The third determines the extent to which each feature map is selective for AUs. All three subsystems have fully-connected and softmax layers on their ends and were fine-tuned to recognize facial expression. Zavares et al. [82] investigated the influence of fine-tuning a CNN in a cross-dataset approach to recognize facial expressions. Using a total of seven datasets, the authors employed the leave-one-out approach when fine-tuning a model pre-trained for the face recognition problem.

Mavani et al. [46] fine-tuned an existing CNN model pre-trained for the object recognition problem to predict seven basic emotions from facial expressions. The fine-tuned model was used to compute visual saliency maps for each dataset image. The network was then trained again using the product between the original images and their saliency maps. Visual saliency maps are intensity maps that highlight areas of an image that most attracted the attention of the network [66]. We explored saliency maps with the intent of understanding what our model has learned during training. This valuable information helped us to find alternative approaches and to improve our results.

3.2.2 Other data-driven derived features

Although CNN-based models are the most commonly used to extract features in computer vision problems, the literature also provides other data-driven methods. Here we cite a couple of them used to classify emotions from facial expressions.

Nagpal et al. [47] propose a Mean Supervised Deep Boltzmann Machine (msDBM) as a feature extractor for the seven basic emotions recognition problem. The supervised phase is incorporated in the loss function, by maximizing inter-class and minimizing intra-class variations through the use of the distances between learned features and mean features of a particular class. Learned features are used to train a Random Forest classifier.

Sun et al. [68] employ Multi-scale Dense LBP (MDLBP) to extract descriptors in different resolutions. These descriptors are concatenated into a single vector and sent to a Stacked Binarized Auto-encoder for feature learning in an unsupervised fashion. Then a Binarized Neural Network (BNN) is trained on the learned features to predict emotions labels.

Chapter 4

Datasets

In deep neural networks, as with all supervised learning algorithms, data plays a key role. Often, a huge quantity of data is required to capture variability and generate models less prone to bias. A representative training data set should include subjects of different ethnic groups, pictured under different conditions: illumination, pose, image quality, etc.

Here, we describe the details of the datasets used in this research. We underline information about their subjects, the data acquisition process, the captured media and their properties, and the annotation procedure that attributed class labels to each example. All datasets were made available to the scientific community by their authors. We have not collected any new data nor built any new dataset during this research.

We organized this chapter by reserving one section for each dataset. In Section 4.10, we present a summary of the datasets and explains how they were split into training, validation and test sets.

4.1 Bosphorus

The Bosphorus dataset¹ [57, 59], a 3D/2D database of FACS annotated facial expressions, consists of 105 subjects (61 men and 44 women) and 4666 faces in various poses and occlusion conditions. The majority of the subjects are Caucasian, aged between 25 and 35. There are 29 professional actors/actresses subjects incorporated in the database. The images have variable dimensions but are in high resolution.

In this study, we used only 2D images annotated with one of the seven basic emotions (anger, disgust, fear, happiness, neutral, sadness and surprise). The selected images do not contain occlusions of any type and head pose does not vary, as faces are always in frontal position. Figure 4.1 reproduces some examples from the dataset.

¹Following the Bosphorus dataset license agreement, only images from subjects marked as publishable were used in this document.



Figure 4.1: Examples from subject 11 of the Bosphorus dataset (anger, disgust, fear, happiness, sadness and surprise).

4.2 CAFE

The Child Affective Facial Expression dataset² [38, 39] comprises photographs of a racially and ethnically diverse group of 2 to 8-year-old children posing for six emotional facial expressions – anger, disgust, fear, happiness, sadness and surprise – and a neutral face.

The full set features 90 female models and 64 male models with no prior training on how to pose to the photos. A professional photographer elicited naturalistic expressions by engaging each child in unscripted play based on each emotion. The photographer attempted to obtain all FACS codes related to them, although not all children were able to successfully pose for all seven basic emotions. Images have 2530×2530 pixels resolution.

To conform to the EMFACS framework, images captured when children were deliberately prompted to pose for the expression with their mouths open were not considered. The only exception is the surprise emotion, which naturally occurs with an open mouth. We also discarded the disgust emotion examples performed with a tongue protrusion.

4.3 CFEE

The Ohio State University Compound Facial Expressions of Emotion database³ [11] comprehends images of facial expressions of emotion of 230 subjects in 22 different categories. All images have 1000×750 pixels resolution and were annotated following the EMFACS protocol.

Besides the neutral state and the six basic emotions (anger, disgust, fear, happiness, sadness and surprise), the dataset also contains posed compound emotions, that means, emotions that can be constructed by combining basic component categories to create new ones. For instance, happily surprised and angrily surprised are two distinct compound emotion categories. For this work, however, these compound emotions were not taken into consideration. Figure 4.2 shows some examples from the dataset.

²Images from the CAFE dataset are copyright protected and cannot be reproduced.

³The CFEE dataset terms of use do not restrict the use of its images in scientific publications.



Figure 4.2: Examples from subject 24 of the CFEE dataset (anger, disgust, fear, happiness, sadness and surprise).

4.4 CK+

The Extended Cohn-Kanade dataset⁴ [29, 41] contains 123 different subjects in 593 sequences of images, with variable duration (i.e. 10 to 60 frames), which goes from the neutral state to peak formation of the facial expressions. Most of the expressions are posed, but a few are spontaneous. The majority of the sequences were recorded in 8-bit gray-scale in 640×490 pixels resolution. Only 327 of the 593 sequences fit the prototypic definition given by EMFACS for one of the basic emotions: anger, contempt, disgust, fear, happiness, sadness and surprise.

We manually looked through each of the 327 sequences of images and selected the first frame to the neutral emotion set and every other frame with a clear manifestation of the facial expression to the correspondent emotion set, except for the contempt emotion, which is not used. Figure 4.3 depicts an example of this procedure.



Figure 4.3: Surprise emotion sequence from subject 55 of the CK+ dataset. The first frame was selected to the neutral emotion set while all frames from the bottom row were selected to the surprise emotion set.

4.5 KDEF

The Karolinska Directed Emotional Faces dataset⁵ [42] encompasses 4900 pictures of human facial expressions. The set contains 70 amateur actors (35 females and 35 males) displaying 7 different emotional expressions (anger, disgust, fear, happiness, neutral, sadness and surprise). Each expression was photographed twice from 5 different angles, although in this research we have only used images with frontal face position. Images have a resolution of 562×762 pixels.

⁴Following the CK+ dataset guidelines, only images from subjects with consent for publication were reproduced in this document.

⁵We followed the KDEF dataset guidelines when reproducing its images in this document.

All subjects received written instructions in advance about the seven different expressions that they were to pose during the photo session. It was emphasized that they should evoke the emotion in a way that felt natural to them while, at the same time, trying to make the expression strong and clear. Although the dataset authors do not mention the protocol used to annotate images, they seem to follow the prototypic definition given by EMFACS, if we refer to Appendix A. Figure 4.4 illustrates some examples from the dataset.



Figure 4.4: Examples from subject 1 of the KDEF dataset (anger, disgust, fear, happiness, sadness and surprise). Files: AF01ANS, AF01DIS, AF01AFS, AF01HAS, AF01SAS and AF01SUS, respectively.

4.6 MUG

The MUG facial expression database⁶ [3] consists of image sequences of 52 subjects performing posed facial expressions. They were captured in a controlled laboratory environment with a resolution of 896×896 pixels and no occlusions. Each image sequence contains 50 to 160 images. Six basic emotions (anger, disgust, fear, happiness, sadness and surprise) plus the neutral state were captured. The number of sequences for each emotion and subject is variable.

The subjects were given a tutorial on how to perform the six facial expressions according to the emotion prototypes defined in EMFACS. The image sequences start and end at the neutral state and follow the onset, apex, offset temporal pattern.

We manually looked through each of the 986 sequences of images and selected the first and last frames to the neutral emotion set and every other frame with a clear manifestation of the facial expression to the correspondent emotion set. The neutral emotion set is also composed of images from neutral emotion sequences.

4.7 NVIE

The Natural Visible and Infrared facial expression database⁷ [43, 74, 75, 76] contains both spontaneous and posed expressions, recorded simultaneously by a regular and an infrared thermal camera. Here, we are only interested in the posed dataset captured by

⁶The use of MUG images in publications is only permitted upon explicit grant from subjects.

⁷The NVIE dataset release agreement does not restrict the use of subject images in scientific publications.

the regular camera, which includes the apex expressional images of 107 subjects, both with and without glasses. Each emotion (anger, disgust, fear, happiness, neutral, sadness and surprise) was recorded under three different illumination conditions: left, front and right illumination. Images have a resolution of 704×480 pixels.

It is worth noting that, even though images were annotated with one of the basic emotions, there is no information about how this process was conducted. Put in other words, we do not know if the facial expressions were evaluated according to the FAC system. Figure 4.5 depicts some examples from the dataset.



Figure 4.5: Examples from subject 4 of the NVIE dataset (anger, disgust, fear, happiness, sadness and surprise).

4.8 Oulu-CASIA

The Oulu-CASIA NIR&VIS facial expression database⁸ contains videos with the six typical expressions: anger, disgust, fear, happiness, sadness and surprise, from 80 subjects captured with two imaging systems, NIR (Near Infrared) and VIS (Visible light), under three different illumination conditions: normal indoor illumination, weak illumination (only computer display is on) and dark illumination (all lights are off). The database includes two parts, one was taken by the Machine Vision Group of the University of Oulu, consisting of 50 subjects and most of them are Finnish people. The other was taken by the National Laboratory of Pattern Recognition, Chinese Academy of Sciences, consisting of 30 subjects and all of them are Chinese people.

Subjects were asked to make a facial expression according to an expression example shown in picture sequences. The videos were recorded at 25 frames per second with an image resolution of 320×240 pixels. There is no information about how images were annotated, although they seem to conform to the EMFACS prototypic definition present in Appendix A. Only images captured with visible light in normal indoor illumination were used in this study.

We manually looked through each of the 480 sequences of images and selected the first frame to the neutral emotion set and every other frame with a clear manifestation of the facial expression to the correspondent emotion set. Figure 4.6 shows an example of this procedure.

⁸The Oulu-CASIA dataset guidelines do not restrict the use of subject images in scientific publications.

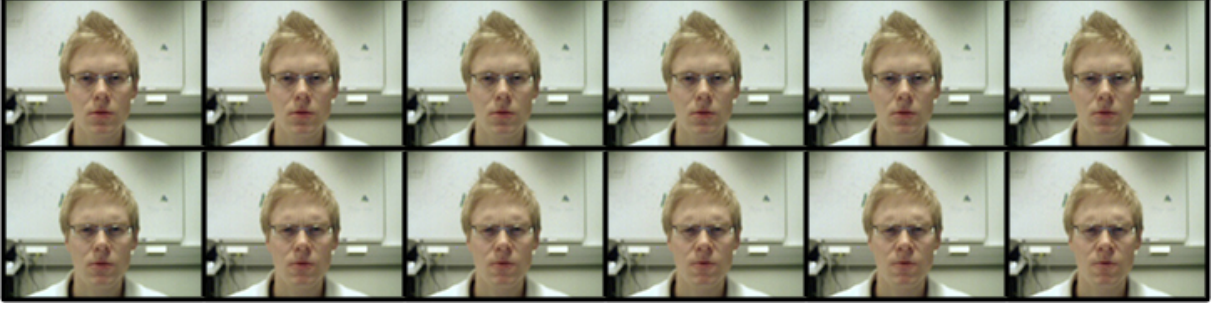


Figure 4.6: Anger emotion sequence from subject 1 of the Oulu-CASIA dataset. The first frame was selected to the neutral emotion set while the last 4 frames from the bottom row were selected to the anger emotion set.

4.9 RaFD

The Radboud Faces Database⁹ [34] is a set of pictures of 67 models (including Caucasian males and females, Caucasian children, both boys and girls, and Moroccan Dutch males) displaying 8 emotional expressions (anger, contempt, disgust, fear, happiness, neutral, sadness and surprise) according to the FAC system. Each emotion was shown with three different gaze directions and all pictures were taken from five camera angles simultaneously. Images have 681×1024 pixels resolution.

Here, we used images of all basic emotions but contempt. We selected only images with frontal face position and all three gaze directions: looking left, looking frontal and looking right. Figure 4.7 reproduces some examples from the dataset.

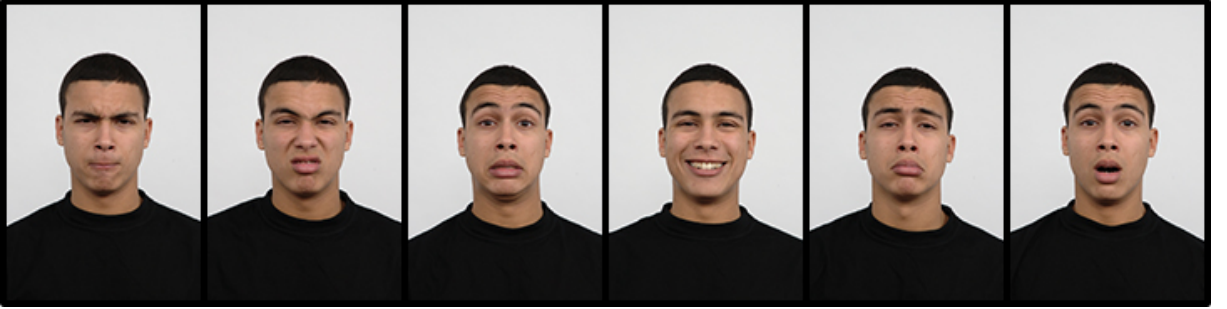


Figure 4.7: Examples from subject 69 of the RaFD dataset (anger, disgust, fear, happiness, sadness and surprise).

4.10 Datasets summary

For a better visualization of the datasets and their properties, in this section, we summarize all relevant information related to them. For each dataset, Table 4.1 specifies the number of subjects, the number of samples, the image resolution, how the facial expression was elicited, the available classes of emotion, if the number of samples is balanced among these classes and which protocol was used when annotating images. Table 4.2 gives

⁹The RaFD dataset guidelines state that there is no restriction on using its images in scientific publications.

more information on the subjects who participated in the composition of the datasets. It provides details such as age, gender and ethnicity.

	Subj.	Samples	Image resolution	Elicitation	Classes of emotion	Bal.	Ann.
Bosphorus	105	4,666 images	Variable (HI-RES)	Posed	6 basic + neutral	No	FACS
CAFE	154	1,192 images	2530×2530 pixels	Posed	6 basic + neutral	No	FACS
CFEE	230	5,060 images	1000×750 pixels	Posed	6 basic + neutral + 15 compound	Yes	FACS
CK+	123	593 image sequences	640×490 pixels	Posed + Spontaneous	6 basic + neutral + contempt	No	FACS
KDEF	70	4,900 images	562×762 pixels	Posed	6 basic + neutral	Yes	N/D
MUG	52	986 image sequences	896×896 pixels	Posed	6 basic + neutral	No	FACS
NVIE	107	3,960 images	704×480 pixels	Posed	6 basic + neutral	No	N/D
Oulu-CASIA	80	2,880 image sequences	320×240 pixels	Posed	6 basic + neutral	No	N/D
RaFD	67	1,608 images	681×1024 pixels	Posed	6 basic + neutral + contempt	Yes	FACS

Table 4.1: Datasets summary: Number of subjects (Subj.), Number of samples (Samples), Image resolution, Facial expression elicitation (Elicitation), Classes of emotion, Balanced classes (Bal.) and Annotation method (Ann.).

	Age	Gender	Ethnicity
Bosphorus	Adults	58% male, 42% female	Majority Caucasian
CAFE	Children	42% male, 58% female	African-american, Asian, Caucasian, Hispanic, South-asian
CFEE	Adults	43% male, 57% female	African-american, Asian, Caucasian, Hispanic
CK+	Adults	31% male, 69% female	African-american, Caucasian, others
KDEF	Adults	50% male, 50% female	Majority Caucasian
MUG	Adults	59% male, 41% female	Majority Caucasian
NVIE	Adults	N/A	Asian
Oulu-CASIA	Adults	74% male, 26% female	Asian, Caucasian
RaFD	Adults + Children	63% male, 37% female	Caucasian, Moroccan

Table 4.2: Datasets breakdown by Age, Gender, and Ethnicity.

Given that CK+, MUG and Oulu-CASIA datasets were manually prepared by looking through each sequence of images and selecting only samples with a clear manifestation of the emotion, they were picked for training and validation steps. The remaining datasets

(Bosphorus, CAFE, CFEE, KDEF, NVIE and RaFD), as they have a ground-truth label for every single image, were used in the test phase. Thereby, we will be able to compare our methods to the literature in a more honest and reliable way.

Table 4.3 and Table 4.4 show the number of examples for each of the seven basic emotions (anger, disgust, fear, happiness, neutral, sadness and surprise) in training and validation datasets and in test datasets, respectively. These values correspond to the number of samples available after filtering out undesired images. Data preparation processes, including data augmentation operations used to increase the diversity of training data, are discussed in Chapter 6.

	Emotion							Total
	Ang	Dis	Fea	Hap	Neu	Sad	Sur	
CK+	463	419	285	644	593	270	534	3,208
MUG	5,582	4,833	4,442	6,032	5,100	5,467	5,713	37,169
Oulu-CASIA	638	613	626	614	480	608	622	4,201
Total	6,683	5,865	5,353	7,290	6,173	6,345	6,869	44,578
CK+*	18,520	16,760	11,400	25,760	23,720	10,800	21,360	128,320
MUG*	22,328	19,332	17,768	24,128	20,400	21,868	22,852	148,676
Oulu-CASIA*	25,520	24,520	25,040	24,560	19,200	24,320	24,880	168,040
Total*	66,368	60,612	54,208	74,448	63,320	56,988	69,092	445,036

* After data augmentation. For details about data augmentation, we refer the reader to Subsection 6.1.1.

Table 4.3: The number of examples for each emotion (Anger, Disgust, Fear, Happiness, Neutral, Sadness and Surprise), before and after data augmentation, in datasets used for both **training** and **validation**.

	Emotion							Total
	Ang	Dis	Fea	Hap	Neu	Sad	Sur	
Bosphorus	71	69	70	106	299	66	71	752
CAFE	121	96	79	120	129	62	103	710
CFEE	230	230	230	230	230	230	230	1,610
KDEF	140	140	140	140	140	140	140	980
NVIE	628	619	629	633	208	618	625	3,960
RaFD	201	201	201	201	201	201	201	1,407
Total	1,391	1,355	1,349	1,430	1,207	1,317	1,370	9,419

Table 4.4: The number of examples for each emotion (Anger, Disgust, Fear, Happiness, Neutral, Sadness and Surprise) in datasets used for **testing**.

Chapter 5

Proposed methods

In this chapter, we present our approaches towards an automated method for emotion recognition from facial expressions. All methods described here use a convolutional neural network (CNN) and are based on the VGG-Face model from the Visual Geometry Group at the University of Oxford [50]. Here, we report the process that led us to the proposed methods and explain the rationale behind our choices in view of a cross-dataset scenario. Throughout the text, we consider the network in a top-down representation when referring to its architecture.

In Section 5.1, we describe how we benefit from the pre-trained VGG-Face model to create our baseline solution. Then, we discuss data visualization and how it can be used to interpret and improve our results in Section 5.2. Our first method is explained in Section 5.3 and is based on a multi-branch CNN to handle different regions of interest present in the face. In Section 5.4, we suggest a less resource-demanding approach, compared to the previous one, that uses patches to occlude random regions of the input image during training. Finally, in Section 5.5, the third explores the triplet loss function as an alternative way to achieve model generalization.

5.1 A pre-trained model as a baseline

The VGG-Face [50] is a deep architecture trained on a dataset containing 2,622 subjects and over 2.6M images for face recognition. The input of the network is a three-channel image of 224×224 pixels resolution. Its architecture comprises a long sequence of convolutional layers, each followed by one or more non-linear operators such as rectified linear unit (ReLU) and max pooling, and three fully-connected layers at the end of the network. The last fully-connected layer is adjusted to an output of 2,622 values, the number of classes of the problem. A softmax log loss function is used to predict the subject's identity. The model was evaluated on the Labeled Faces in the Wild [25] and the YouTube Faces [79] datasets, obtaining classification accuracy above 97% in both. Figure 5.1 shows the network architecture.

To overcome the lack of data to train a convolutional neural network from scratch for the emotion recognition problem, we leverage the knowledge the pre-trained VGG-Face model has for face recognition and adapt it to our task, applying a concept called Trans-

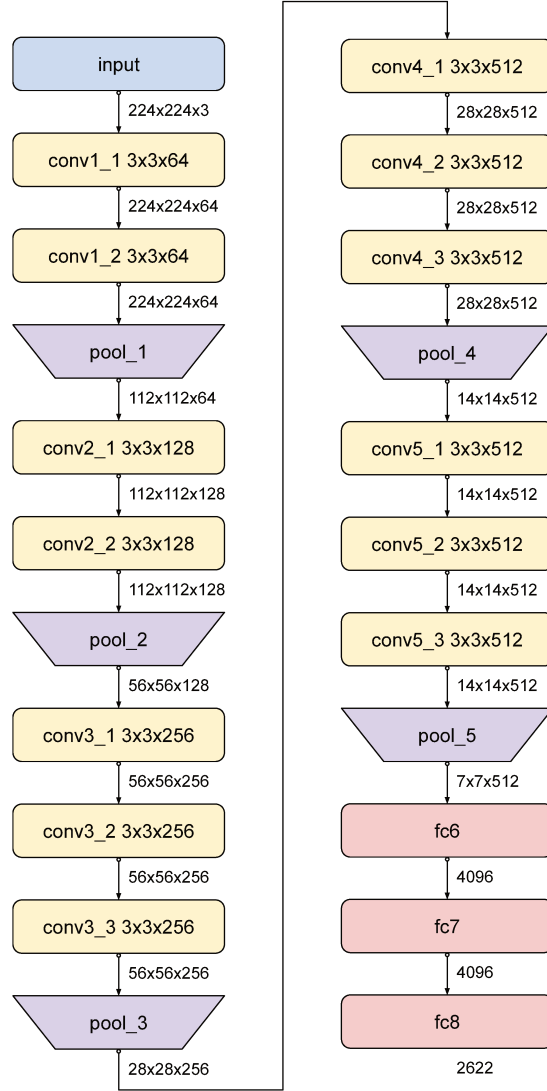


Figure 5.1: VGG-Face architecture. Each convolutional layer is annotated with a tuple $W \times H \times N$, where W is the filter width, H the filter height, and N the number of filters. The tuple $W \times H \times N$ below each layer are the output dimensions. W is the activation map width, H the activation map height, and N the number of activation maps. For the fully-connected layers, the output is flattened into a 1-D array.

fer Learning. Both classification problems have similar domains (human faces and human faces conveying facial expressions) but different tasks. Pan and Yang [49] present a framework for understanding Transfer Learning in terms of domain, probability distributions, and tasks. A formal definition is as follows:

Definition 5.1. A domain D consists of two components, a feature space \mathcal{X} and a marginal probability distribution $P(X)$, where $X = \{x_1, \dots, x_n\} \in \mathcal{X}$.

Definition 5.2. Given a specific domain $D = \{\mathcal{X}, P(X)\}$, a task consists of two components, a label \mathcal{Y} and a predictive function $f(\cdot)$, and is denoted by $T = \{\mathcal{Y}, f(\cdot)\}$. The task

T can be learned from training data consisting of pairs $\{x_i, y_i\}$, where $x_i \in \mathcal{X}$ and $y_i \in \mathcal{Y}$, and consequently $f(\cdot)$ can be used to predict the corresponding label of a new instance x .

Definition 5.3. *Given a source domain D_s and a learning task T_s , a target domain D_t and a learning task T_t , transfer learning aims to help improve the learning of the target predictive function $f_T(\cdot)$ in D_t using the knowledge in D_s and T_s , where $D_s \neq D_t$ and $T_s \neq T_t$.*

For deep learning based methods, Tan et al. [71] list possible categories of what they call Deep Transfer Learning. In our case, we are mainly interested in Network-Based Deep Transfer Learning, in which a part of the network pre-trained in the source domain is reused in the target domain. This approach is possible owing to the fact that neural networks tend to learn very similar features in their first layers when trained on images [81]. Their mechanism is an iterative and continuous abstraction process, roughly comparable to the way the human brain works.

In this work, we experimented with two different Network-Based Deep Transfer Learning approaches: pre-trained VGG-Face model as a feature extractor and fine-tuning the pre-trained VGG-Face model.

Pre-trained VGG-Face model as a feature extractor

When training a deep learning model, each layer of its architecture learns a different set of features that will be passed on as inputs to subsequent layers until the final layer is reached and the final output yielded. Any of these layers can be used as a fixed feature extractor. Training data from the target domain are passed through the network until they reach the selected layer and the extracted features can be then used to train any classifier for the target task.

Figure 5.2 illustrates the procedure. In this example, we cut the network after a given layer, e.g. the *pool5* layer, and used its output as features to train a Support Vector Machine (SVM) classifier for emotion recognition.

Fine-tuning the pre-trained VGG-Face model

It is also possible to selectively retrain some of the layers of the network so that it is adapted for the target task. In this fine-tuning process, considering a top-down network representation, we may want to freeze the weights of some top layers, as they capture generic features, while task-specific bottom layers are updated in an optimization process. It is also possible to modify or replace some layers to make the network best fit our needs.

One common adjustment is to replace the last fully connected layer with a new one whose output matches the number of classes of our problem. We carried out this procedure as shown in 5.3.

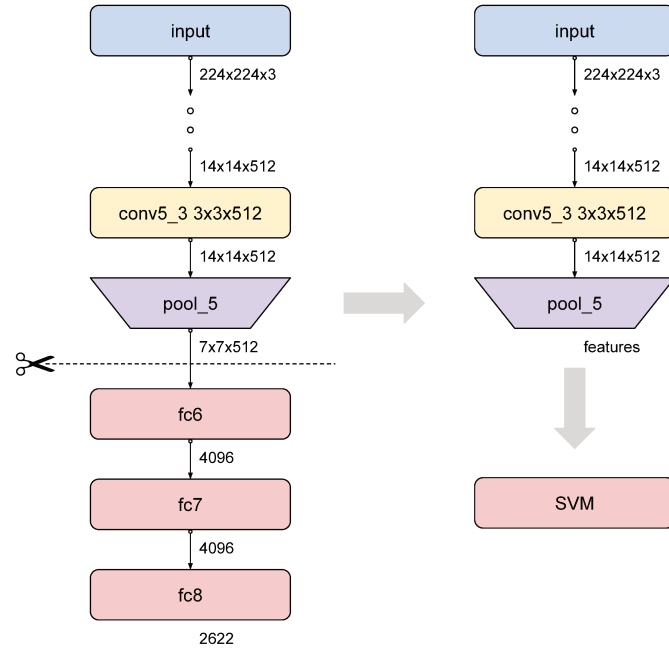


Figure 5.2: Pre-trained VGG-Face model as a feature extractor. Extracted features are used to train the SVM classifier.

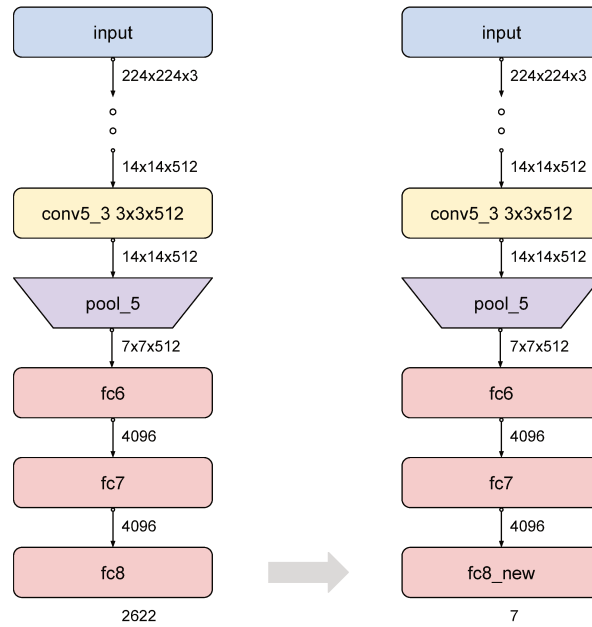


Figure 5.3: Fine-tuning the pre-trained VGG-Face model. The last fully-connected was replaced to a new one with a different output size and the whole network is trained for the target task.

5.2 Data visualization

In the last decade, we have seen the adoption of deep neural networks in many different areas, including ones with high-stakes decisions. If, on one hand, it is now the state-of-the-art technique to many different applications, on the other hand, its complex and non-linear structure, at times seen as a black box, imposes an obstacle to human interpretability. Understanding what a deep neural network has learned during training not only can increase transparency and accountability in the decision process, but also provides explanations to help us improve its accuracy.

For that reason, the scientific community has proposed plenty of methods that try to connect the dots between data inputs, learned inner parameters, and network outputs. For image data, for example, one commonly wants to know which regions of the image are the most significant when the network attributes a certain class to it, in a sort of map representation.

Zeiler and Fergus [83] propose to monitor the classification confidence while systematically occluding different parts of the input image. A heatmap can be created by analyzing the confidence variation for each occluded part of the image. Sundararajan et al. [69] suggested the Integrated Gradients, a method in which a sequence of images, interpolating in increasing intensity from a baseline to the actual image, is passed through the network and the gradients of the outputs with respect to the inputs are integrated to a final map. Springenberg et al. [67] introduced the Guided Backpropagation method, which computes the gradient of the output with respect to the input, backpropagating through the ReLU layers only values that were not negative in both forward and backward passes. The resulting image shows the most discriminative parts.

We tested these three methods, but, eventually, opted for another for our work: the Gradient-weighted Class Activation Mapping (Grad-CAM) [63]. The Grad-CAM uses class-specific gradient information which is passed back through the final convolutional layer to create a heatmap that highlights important regions in the image for class discrimination. The method is applicable to a variety of CNNs, including the VGG-Face, without the need of modifying its architecture. Next, we present a formal definition of the method.

Consider a Grad-CAM heatmap $L_{Grad-CAM}^c \in \mathbb{R}^{u \times v}$ of width u and height v with respect to class c . To generate it, we first compute the gradients of the score y^c for class c with respect to feature maps A^k of the selected convolutional layer. Then we perform a global-average-pooling operation on these gradients to obtain weights α_k^c that capture the importance of each feature map k for class c :

$$\alpha_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k} . \quad (5.1)$$

The heatmap can be then calculated by a linear combination of α_k^c weights and A^k feature maps, following a ReLU function:

$$L_{Grad-CAM}^c = ReLU\left(\sum_k \alpha_k^c A^k\right) . \quad (5.2)$$

Note that the generated heatmap has the same dimensions of the selected feature map. In our case, we used the *conv5_3* layer, which is 14×14 , as per Figure 5.1. By upscaling the heatmap, it can be used to confront the input image. Table 5.1 compares the heatmaps generated from each one of the aforementioned methods: Occlusions, Integrated Gradients, Guided Backpropagation, and Grad-CAM.

Although the Occlusions method works in a pretty intuitive way, the heatmap generation process is very expensive once we have to forward pass the input image through the network several times. Moreover, it is not clear what color and size the patches used to occlude the parts of the input should have and how such choices may impact the final heatmaps. The Guided Backpropagation method, in turn, can give fine-grained visualizations at the pixel level but is not class-discriminative. As we can see from Table 5.1, it highlights similar regions for both happiness and sadness classes. Finally, for Integrated Gradients, which is also a computationally expensive method due to multiple feed-forwards passes, it is difficult to assess how different baseline image hyperparameter choices affect the resulting heatmaps. For those reasons, we picked the Grad-CAM method.

5.3 Method 1: Parts-based VGG-Face

When evaluating our baseline model, the fine-tuned VGG-Face for emotion recognition, we generated Grad-CAM heatmaps for each example in the training dataset to visualize what the network had learned and to gain insights about how to improve initial results. For each one of the six basic emotions (anger, disgust, fear, happiness, sadness, and surprise), we calculated the mean heatmap and applied it to the respective mean image of that class. This procedure was executed for all three training datasets: CK+, MUG, and Oulu-CASIA. Table 5.2 presents mean images and heatmaps side by side for each emotion and each dataset.

Comparing the heatmaps from Table 5.2 with the combination of Action Units required to convey a basic emotion as suggested in the literature and available in Table A.1, it is possible to see that there is not a complete correspondence between highlighted regions and the areas where Action Units occur. We suppose that, once the network has found enough discriminative regions, it stops looking for new ones, as if it was trapped in local minima. All observed differences are listed below.

- **Anger:** frontal, supralabial, and nasolabial fold regions emphasized; eyelids (AU5 and AU7) and lips (AU23) regions possibly minimized.
- **Disgust:** periorbital, perinasal, and mental regions emphasized; lips (AU16) region possibly minimized.
- **Fear:** periorbital, nasal, and supralabial regions emphasized; glabellar (AU4) and mouth (AU20 and AU26) regions possibly minimized.
- **Happiness:** nasal and supralabial regions emphasized; infraorbital (AU6) and lip corners (AU12) regions possibly minimized.

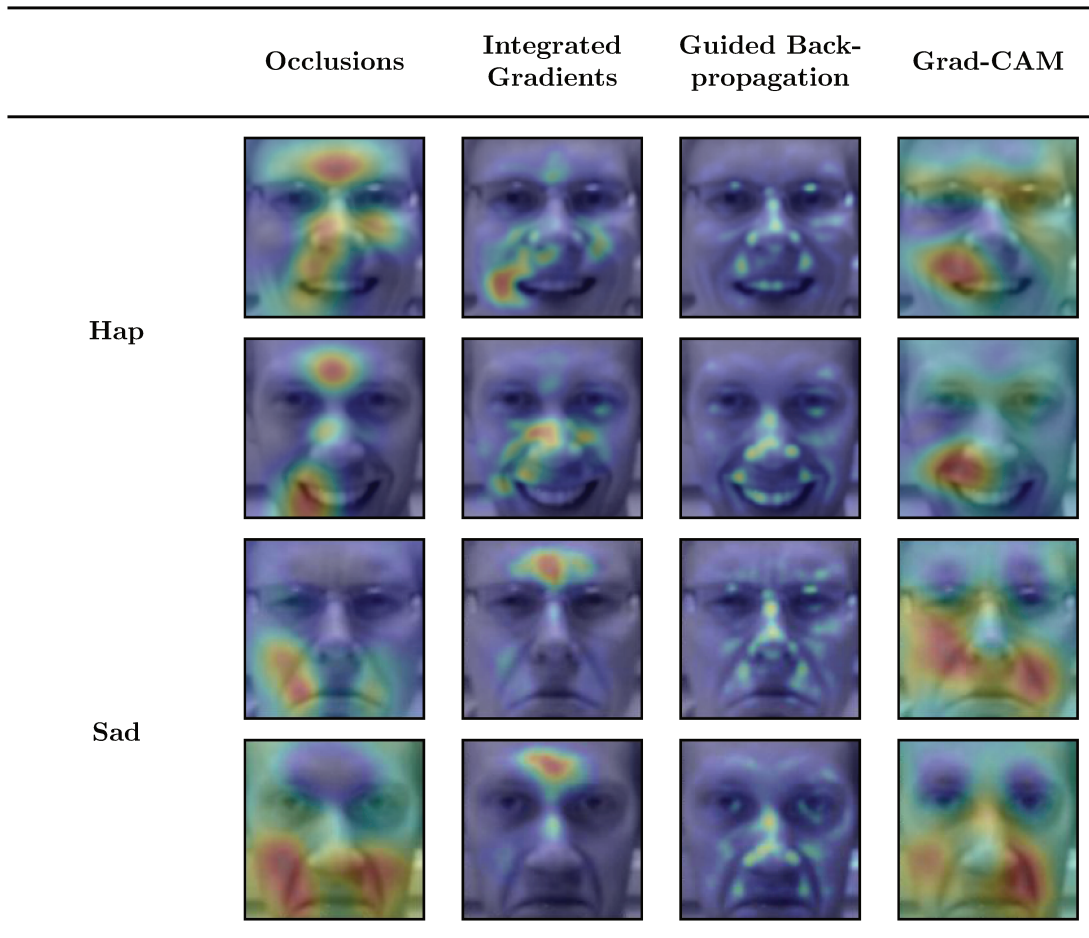


Table 5.1: Heatmaps generated from different data visualization methods for happiness and sadness classes: Occlusions (Zeiler and Fergus [83]), Integrated Gradients (Sundararajan et al. [69]), Guided Backpropagation (Springenberg et al. [67]) and Grad-CAM (Selvaraju et al. [63]). The less important regions are in cold colors (bluish), while the more important regions are in hot colors (reddish). Samples extracted from the Oulu-CASIA dataset.

- **Sadness:** perinasal and lips regions emphasized; supraorbital (AU1), glabellar (AU4) and lip corners (AU15) regions possibly minimized.
- **Surprise:** periorbital, glabellar, and lip corners region emphasized; mouth (AU 26) region possibly minimized.

To make the network also take those minimized regions into account during training, we modified its architecture and added an attention mechanism. In object detection problems, for instance, a common approach is to work with region proposals. Selected regions can be extracted from the input image to feed a CNN [21] or be learned directly from feature maps by an auxiliary Region Proposal Network to accelerate the process [52]. The idea behind it is to reduce the search space and look only into the most promising areas of the image.

For fine-grained categorization problems, in contrast, we want to differentiate between hard-to-distinguish object classes and, therefore, drive the network’s attention to the most




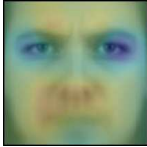













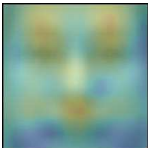


















	CK+		MUG		Oulu-CASIA	
Ang						
Dis						
Fea						
Hap						
Sad						
Sur						

Table 5.2: Mean image and mean heatmap generated from Grad-CAM (Selvaraju et al. [63]) for each emotion (Anger, Disgust, Fear, Happiness, Sadness, and Surprise) and train dataset.

discriminative regions of the image. In the case of emotion categorization, we leverage the Facial Action Code System to propose the positions of such regions of interest.

Regions of interest were created by looking where Action Units occur. We grouped them into a total of four candidate regions: forehead, eyes, nose, and mouth. For each region, a correspondent binary mask was applied to the input image in order to hide out the rest of the face. Figure 5.4 shows the resulting images. They were used to train a multi-branch CNN in which each branch is responsible for learning attributes for a specific region of interest.

In the multi-branch CNN architecture, each one of the four branches is formed by all VGG-Face layers until *pool5*. These branches output a feature map of size $7 \times 7 \times 512$ each, which are then concatenated to a $7 \times 7 \times 2048$ tensor. A new block of convolutional layers is plugged into its bottom to reduce the number of parameters before we get to



Figure 5.4: The original and resulting masked input images, one for each region of interest: forehead, eyes, nose, and mouth. Example extracted from subject 74 of the CK+ dataset.

the fully-connected layers. Figure 5.5 shows the Parts-based VGG-Face final architecture. Note that branches do not share weights, as to allow the network to specialize in each particular region, and that the model size is around four times greater than the original VGG-Face.

5.4 Method 2: Random Patches

If the original VGG-Face model can be considered a bit challenging to handle owing to its large number of parameters, the Parts-based method proposed in the previous section is a heavy load. It either requires a lot of GPU resources in terms of memory or demands us to juggle with batch sizes during the training process.

In light of that, in this second method, our goal is to direct the network’s attention to complementary regions of the face while keeping the final architecture much more compact. To this regard, starting from the concept of adversarial erasing strategy [77], which iteratively occludes the most discriminative regions of the image to force the network to discover new and complement object regions, originally proposed for the object segmentation, we apply the same idea to the emotion recognition problem.

By occluding some regions of the face during training, the learning process tends to be harder as the network cannot rely on the holistic representation of the face. As a consequence, it is forced to find out new features that would not probably be used otherwise to distinguish between emotions. We presume those less obvious features will decrease bias and help the model generalize better.

But what regions should we occlude? Padilha et al. [48] showed that the use of random patches to occlude parts of the input image produced results as good as more sophisticated methods, such as the ones in which occlusions are guided by the most activated regions of an image. With that in mind, we decided on the random patches approach.

During the training process, we divided the batch into two halves. The first one contains the original input images while the other contains those same images but with a

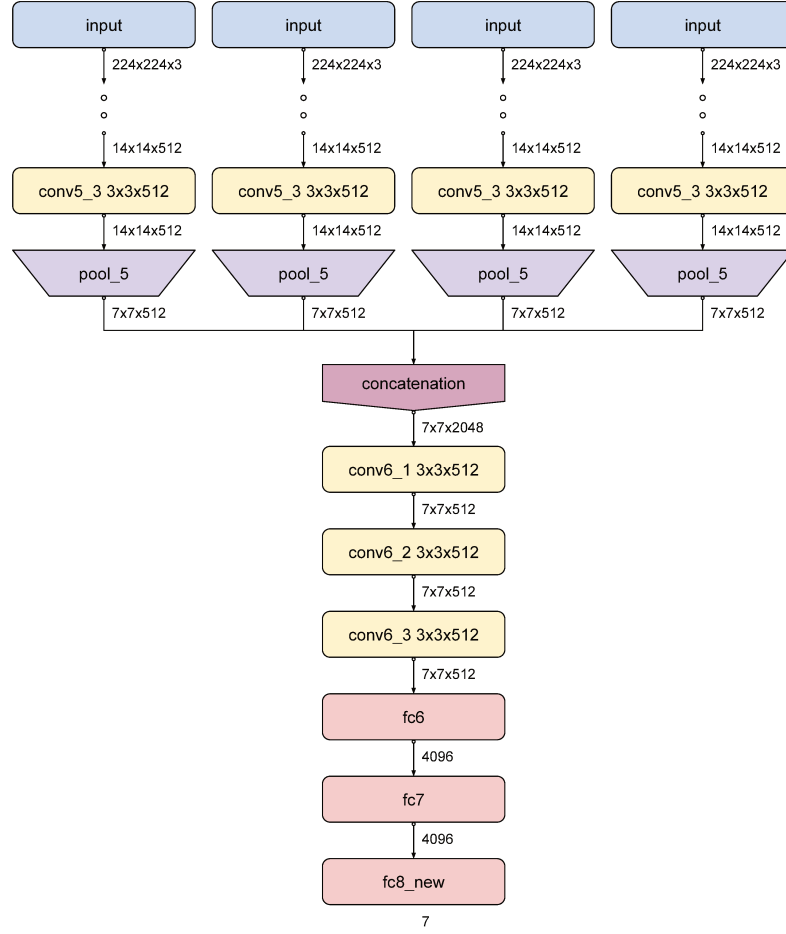


Figure 5.5: Parts-based VGG-Face architecture.

random patch occluding some region of the face. Figure 5.6 shows an example batch.

5.5 Method 3: Triplet Loss

In our third method, we explore another direction in an effort to improve cross-dataset accuracy for emotion recognition. Here, we try to learn good embeddings by the notion of similarity and dissimilarity of facial expressions. Feature vectors are transported to a hyperspace where the distance between examples of the same class is small, whereas the distance between examples of different classes is large. This is accomplished by employing a triplet loss function [62] during model training.

Every feature vector $f(x) \in \mathbb{R}^d$ of a given image x is contained in a d -dimensional hypersphere, with $\|f(x)\|_2 = 1$. A triplet is formed by an anchor image x^a , a positive image x^p of the same class as the anchor, and a negative image x^n of a different class. The relation between the components of each triplet is given by:

$$\|f(x^a) - f(x^p)\|_2^2 + \alpha < \|f(x^a) - f(x^n)\|_2^2 \quad (5.3)$$

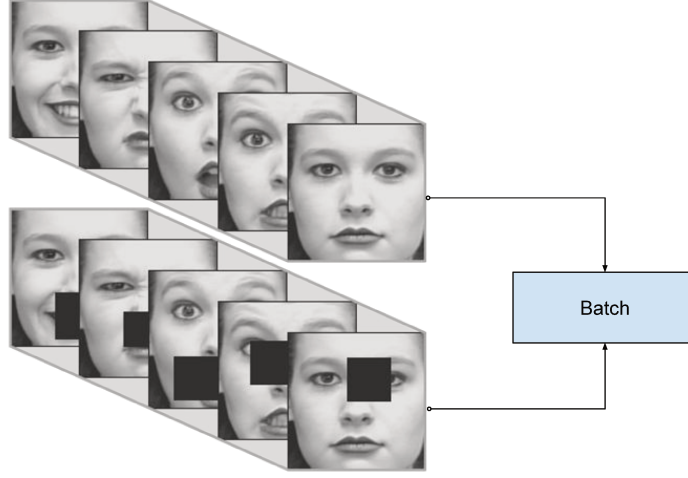


Figure 5.6: Random Patches batch example. Original input images and images with occluding random patches. Examples extracted from subject 74 of the CK+ dataset.

where α is a margin between positive and negative pairs. For a training process containing N triplets, the loss function that we want to minimize is:

$$L = \left[\sum_i^N \|f(x_i^a) - f(x_i^p)\|_2^2 - \|f(x_i^a) - f(x_i^n)\|_2^2 + \alpha \right]_+ \quad (5.4)$$

so that the distance between the anchor and the positive example, $dist(a, p)$, is pushed to zero and the distance between the anchor and their negative example, $dist(a, n)$, is greater than $dist(a, p)$ plus the margin α . Based on the loss definition, triplets can be classified into three groups:

- **Easy triplets:** are triplets in which the positive example is closer to the anchor than the negative example by the margin α , i.e. $dist(a, p) + \alpha < dist(a, n)$.
- **Semi-hard triplets:** are triplets in which the negative example is farther to the anchor than the positive example but not by the margin α , i.e. $dist(a, p) < dist(a, n) < dist(a, p) + \alpha$.
- **Hard triplet:** are triplets in which the negative example is closer to the anchor than the positive example, i.e. $dist(a, n) < dist(a, p)$.

Note that, when the loss is computed for an easy triplet, its value is zero. For semi-hard and hard triplets, the loss has a positive value. In our work, we filter all candidate triplets and select only the semi-hard for the training process. The inclusion of hard triplets led us to a collapsed model, in other words, the result is a model that classifies every example into the same class. Schroff et al. [62] state that this behavior occurs due to bad local minima along the optimization process. Figure 5.7 shows an example of each of the three triplets categories in the hyperspace.

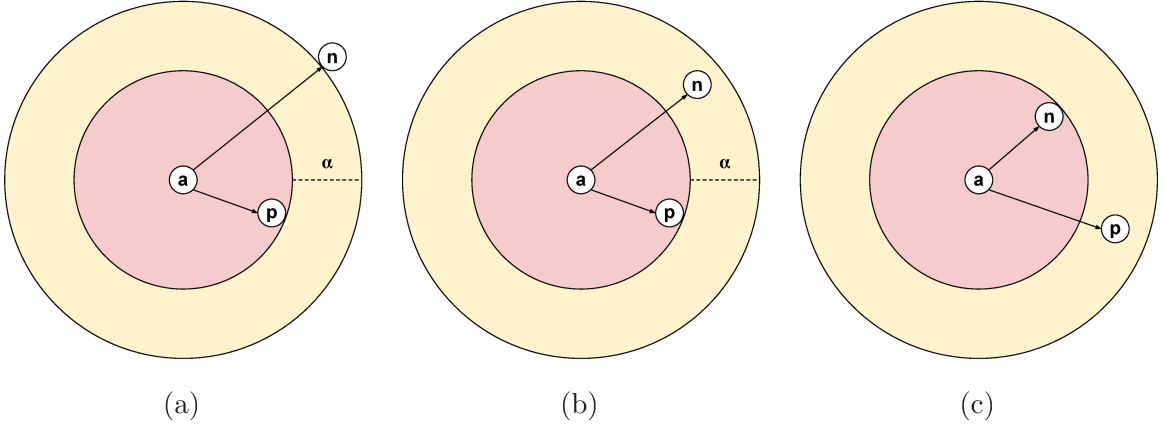


Figure 5.7: Triplets categories: (a) easy, (b) semi-hard and (c) hard.

Triplets are generated in an online manner for each batch of the training process. Given a balanced training batch, with the same number of examples for each class of our problem and a sufficient number of examples for each class, we use the following protocol to generate triplets. Considering that each input image x is converted to a feature vector by applying $f(x)$:

1. Compute the Euclidean distance between each pair of feature vectors in the batch;
2. For each feature vector as an anchor, split positive and negative examples into two sets;
3. Create triplets by performing a 2-combination of one element of the positive set and one element of the negative set;
4. Select only semi-hard triplets, i.e. $dist(a, p) < dist(a, n) < dist(a, p) + \alpha$;
5. Calculate the loss function L using only the selected semi-hard triplets;

We then divide L by the number of semi-hard triplets selected in each batch to find L_{mean} . This final value is used in the optimization process during the backpropagation algorithm.

Since we would like an end-to-end classification process, the triplet loss function is linearly combined with a cross-entropy loss so that our model can also measure class probabilities. In the final architecture, after passing the *pool5* layer, the network is split into two branches. One branch contains the remainder fully-connected layers, while the other flattens the data into a 1-D array and then applies a normalization to place the feature vector inside a hypersphere of radius 1. Figure 5.8 shows the final network architecture. In the test phase, only the layers over the gray background are used.

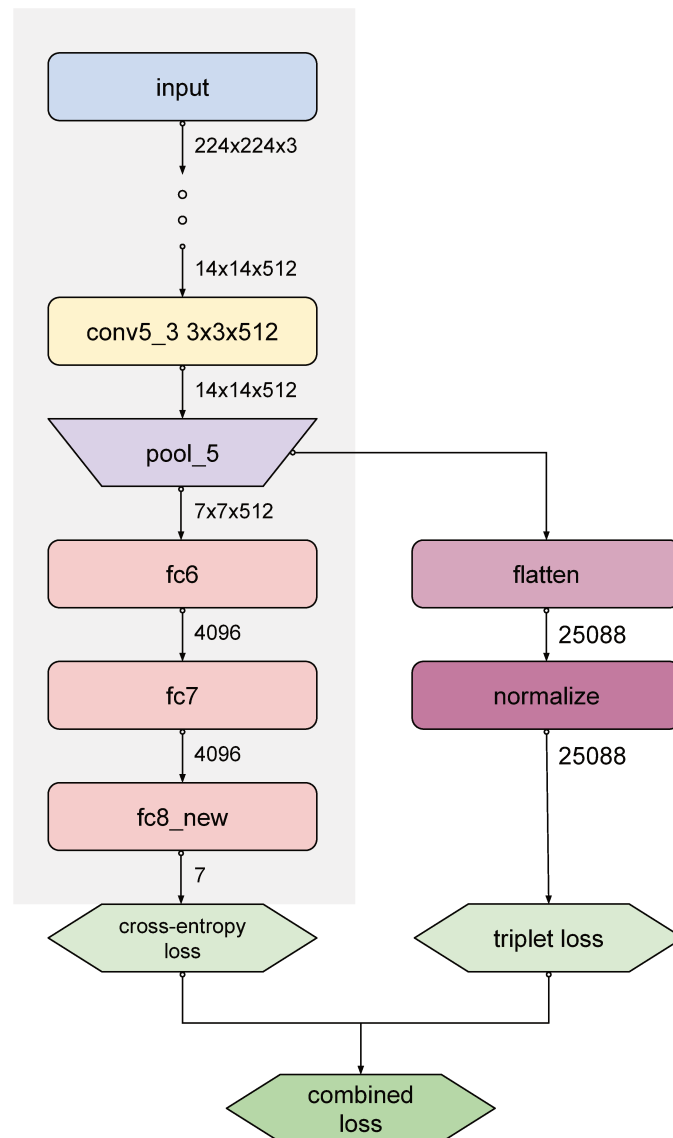


Figure 5.8: Triplet Loss architecture. The layers over the gray background are the only ones considered during the test phase.

Chapter 6

Experimental results

In this chapter, we evaluate the efficacy of the proposed methods by carrying out experiments with the datasets described in Chapter 4. As state there, we used CK+, MUG and Oulu-CASIA datasets in the training and validation phases, while Bosphorus, CAFE, CFEE, KDEF, NVIE, and RaFD datasets were used in the test phase.

We organized this chapter as follows. In Section 6.1, we explain the data preparation process, including data augmentation operations performed in training datasets. Section 6.2 lists the evaluation metrics used in this research. For each method proposed in Chapter 5, a section was reserved to present and discuss their results individually. Then, we proceed to compare our methods in Section 6.7. In Section 6.8, we apply a statistical test to verify if our methods have led us to significant improvements. Finally, in Section 6.9, we compare our methods performance with works present in the literature as well as with commercial applications.

6.1 Data preparation

The first step in data preparation was to detect facial landmarks. We used the Dlib C++ Library [31] to find 68 points of interest, which include points on the corners of the mouth, along the eyebrows, on the eyes, and so forth. Dlib face detector was created using Histogram of Oriented Gradients (HOG) [8], combined with a linear classifier, an image pyramid, and a sliding window detection scheme. The pose estimator was implemented following Kazemi and Sullivan paper [30] and the dataset detection model was trained on the iBUG 300-W face landmark dataset [55].

After finding the landmarks, we used the six points around each eye to calculate its center. Figure 6.1 shows an example face with 68 landmarks and eyes center points. Next, we performed an in-plane rotation to have the eyes horizontally aligned and then computed the bounding box used to crop a squared region of the face. For this last task, we defined three ratios:

$$d_1 = 0.333, \quad d_2 = 0.277 \quad \text{and} \quad d_3 = 0.446$$

where d_1 is the height fraction of the face corresponding to the distance between the eyes and upper boundary, d_2 the width fraction of the face corresponding to the distance



Figure 6.1: On the left, the 68 landmarks detected by Dlib. On the right, the calculated eyes center points. Sample extracted from subject 74 of the CK+ dataset.

between the eye center and the nearest side boundary and d_3 the width fraction of the face corresponding to the distance between the two eyes centers. Figure 6.2 shows the computed bounding box after face alignment and the cropped region.



Figure 6.2: Face bounding box after alignment and cropped region. Sample extracted from subject 74 of the CK+ dataset.

Finally, given that some datasets contain images in grayscale, all cropped images were converted to grayscale and pixel values replicated to form a three-channel image. They were also resized to 224×224 pixels to match our model input size, before being used in both training and evaluation processes.

6.1.1 Data augmentation

The lack of a sufficient amount of training data is a constant problem in machine learning. The more variable the data to which an algorithm has access during training, the more

it can learn to generalize and be effective. Data augmentation overcomes (at least partially) this issue by artificially generating new samples through transformations applied to existing ones. In this work, data augmentation was carried out statically, that is, all data were generated previous to the training phase, using the following operations:

- **Translation:** the bounding box was shifted left, right, up, and down by 10% of the distance value between the eyes centers before cropping the region of interest.
- **Mirror:** the cropped image was flipped horizontally.
- **Contrast Limited Adaptive Histogram Equalization (CLAHE):** an image contrast enhancement algorithm, which divides the images into regions and performs local adaptative histogram equalization with a limited contrast amplification to reduce noise [88]. CLAHE was employed with two different kernel sizes (8×8 and 16×16) on the cropped images.
- **Gaussian Blur:** the cropped image was blurred using a Gaussian function with a random standard deviation value between 0.05 and 2.3.
- **Gamma Correction:** is used to correct the differences between the way a camera captures content and the way our visual system processes light, defining the relationship between a pixel's numerical value and its actual luminance. It transforms the cropped image in a pixel-wise manner.
- **Multi-Scale Retinex with Color Restoration (MSRCR):** is an algorithm that tries to transform digital images into renditions that approach the realism of direct scene observation [51]. It is applied to enhance local contrast/lightness and color constancy, being the former our point of interest, since we use only grayscale images.

Figure 6.3 presents the results of the aforementioned operations applied to a sample image. In the data augmentation process, we oftentimes combined these operations intending to generate more data.

6.2 Evaluation metrics

We adopted four different metrics as a means to evaluate our trained models. They were chosen with a multiclass classification problem in mind and help us compare the efficacy of the trained models for each class of emotion as well as for each dataset as a whole.

- **Recall:** computes the fraction of correct predictions for one class over the number of actual examples of that class. It is given by Equation 6.1, where \hat{y}_i^c is the predicted value of the i -th sample from class c , y_i^c the corresponding true value for class c , and n the total number of samples in class c :

$$recall(y^c, \hat{y}^c) = \frac{1}{n} \sum_{i=1}^n \begin{cases} 1, & \text{if } y_i^c = \hat{y}_i^c \\ 0, & \text{otherwise} \end{cases} . \quad (6.1)$$

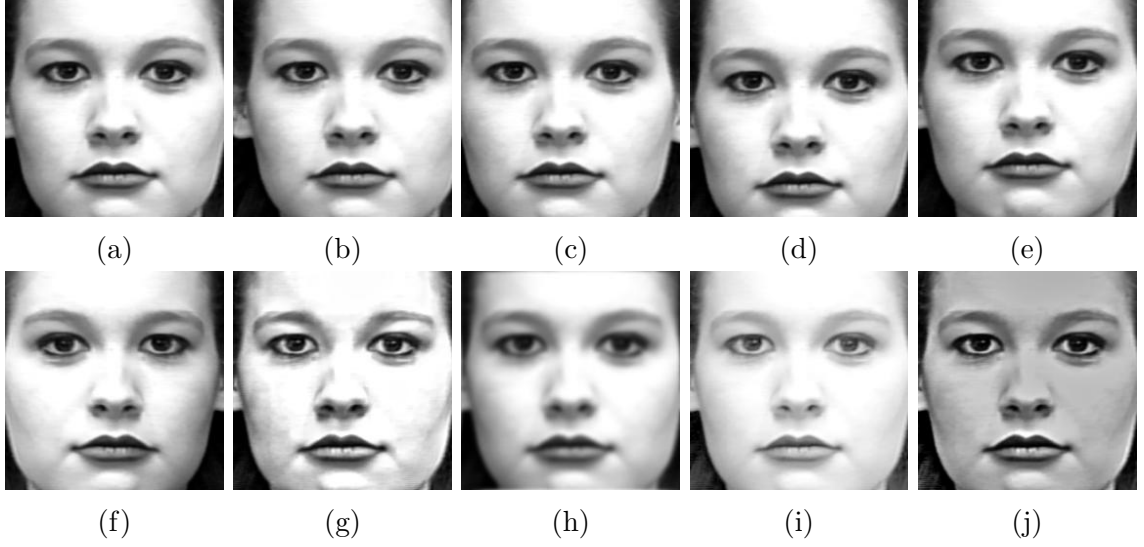


Figure 6.3: The operations used to augment data: (a) the original cropped image, (b) translation left, (c) translation right, (d) translation up, (e) translation down, (f) mirror, (g) CLAHE, (h) Gaussian Blur, (i) Gamma Correction and (j) MSRCR. Sample extracted from subject 74 of the CK+ dataset.

Recall is also known as True Positive Rate (TPR) in a binary classification setup. It is described in Equation 6.2, where TP is the number of true positives and FN is the number of false negatives:

$$TPR = \frac{TP}{TP + FN} . \quad (6.2)$$

- **Normalized accuracy:** computes the average recall for all classes of the dataset. It is given by Equation 6.3, where r_c is the recall for the c -th class of the dataset and n the total number of classes. The normalization gives equal weights for each class when computing the accuracy if the number of examples is different amongst classes:

$$norm_acc(r) = \frac{1}{n} \sum_{c=1}^n r_c . \quad (6.3)$$

- **Accuracy:** computes the fraction of correct predictions for one dataset. It is given by Equation 6.4, where \hat{y}_i is the predicted value of the i -th sample, y_i the corresponding true value, and n the total number of samples in the dataset:

$$acc(y, \hat{y}) = \frac{1}{n} \sum_{i=1}^n \begin{cases} 1, & \text{if } y_i = \hat{y}_i \\ 0, & \text{otherwise} \end{cases} . \quad (6.4)$$

- **Number of parameters:** is directly linked to the quantity of memory the model consumes during training and testing, and the amount of disk space it uses when persisted.

6.3 A pre-trained model as a baseline

Our baseline method consists of fine-tuning the pre-trained VGG-Face model [50] for the emotion recognition problem, as described in Section 6.3. Using the Deep Transfer Learning concept, we depart from the original VGG-Face network with its weights optimized for the face recognition problem, replace the last fully-connect layer to match the number of classes of our task, and restart the network optimization process feeding it with examples from the seven basic emotions. In our experiments, we tried different approaches towards which layers should have their weights frozen and which should not. The best result was obtained when all layers were left available for optimization, which means that none of them were frozen.

We also tried to use the pre-trained VGG-Face model as a feature extractor with a multiclass SVM classifier on top of it. In these experiments, we extracted features from different network layers, some of them yielding huge feature vectors (the *pool4* layer, for instance, generates a vector of size greater than 100K features) that had to have its dimensions reduced through the use of the Principal Component Analysis (PCA) before being served as input data to the SVM. None of these tests, however, outperformed the results we had by fine-tuning the pre-trained VGG model. Also, it is usually preferable, in terms of simplicity, to have an end-to-end architecture than to cherry-pick the feature extractor layer and attach an external classifier on top of it.

Table 6.1 and Figure 6.4 show the results obtained with our baseline method. These results will be used as a basis for comparison with our subsequent methods.

	Recall (%)							Acc (%)	Norm. Acc (%)
	Ang	Dis	Fea	Hap	Neu	Sad	Sur		
Boshorus	81.69	50.72	50.00	99.06	81.94	51.52	73.24	75.00	69.74
CAFE	2.48	84.38	59.49	100.00	86.05	41.94	100.00	69.15	67.76
CFEE	77.39	83.91	71.74	97.83	78.70	73.04	88.26	81.55	81.55
KDEF	72.14	90.00	80.71	99.29	91.43	77.14	77.14	83.98	83.98
NVIE	27.71	25.04	11.29	62.24	88.94	7.28	20.80	29.14	34.76
RaFD	99.00	99.00	89.55	98.51	91.54	88.56	100.00	95.17	95.17
Mean	60.07	72.18	60.46	92.82	86.43	56.58	76.57	72.33	72.16
Mean*	66.54	81.60	70.30	98.94	85.93	66.44	87.73	80.97	79.64

* Without the NVIE dataset.

Table 6.1: Baseline results obtained by fine-tuning the pre-trained VGG-Face model for the emotion recognition problem.

By looking at Table 6.1 and Figure 6.4, two points call our attention instantly. The first is the low accuracy obtained for the anger emotion in the CAFE dataset. To track down the cause of such low accuracy, we looked through the dataset and generate the mean image for the emotion in question. Figure 6.5 presents the resulting mean image.

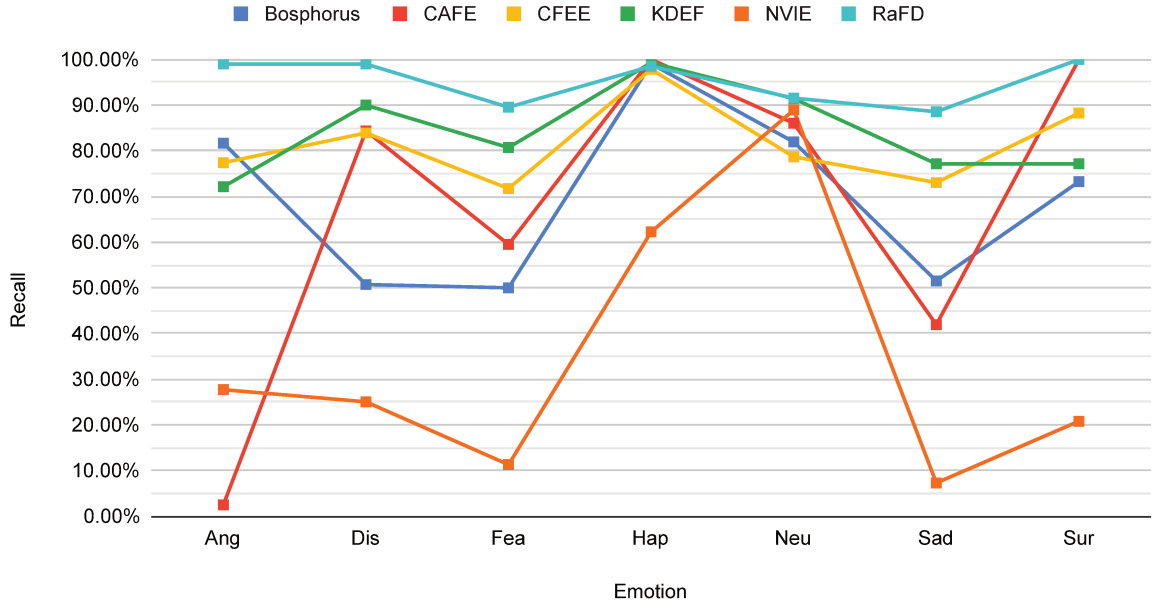


Figure 6.4: Baseline results obtained by fine-tuning the pre-trained VGG-Face model for the emotion recognition problem.



Figure 6.5: Mean image for anger emotion from CAFE dataset.

Appendix A reveals that the Action Unit 23, Lip Tightener, is expected to be present in the facial expression of the anger emotion. But, compared to the generated mean image, we can see that there is a movement in the sense of showing the subjects' teeth, which is characteristic of the disgusted emotion and the intensified form of the happiness emotion. As a matter of fact, our results corroborate these observations. For the CAFE dataset, 60.33% of the examples in the anger emotion set were misclassified as disgust and 28.93% as happiness.

The second point is the low accuracy obtained in the anger, disgust, fear, sadness and surprise emotions for the NVIE dataset. We followed the same approach as with the CAFE dataset and generated the mean image for each one of those emotions. Figure 6.6 depicts the resulting mean images.

All mean images look very alike and it is hard to distinguish which emotion is being conveyed in each one of them. Actually, one could say that they all represent the neutral state, as facial expressions are very latent. That is exactly what our results assert. For our baseline model, 47.93% of the examples in the NVIE anger emotion set were misclassified

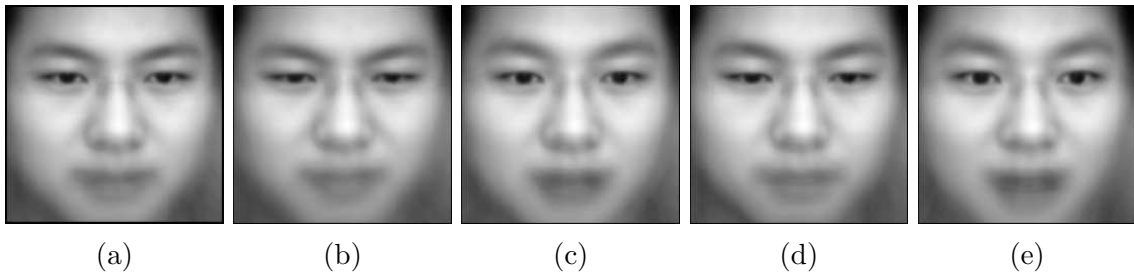


Figure 6.6: Mean images from NVIE dataset: (a) anger, (b) disgust, (c) fear, (d) sadness and (e) surprise.

as neutral. The same occurs with the other mentioned emotions: 30.21% for the disgust emotion, 50.72% for the fear emotion, 53.24% for the sadness emotion, and 45.92% for the surprise emotion.

These findings show us that some images may not follow the EMFACS framework. It does not mean, however, that the emotion is not present nor that it was incorrectly portrayed. What it suggests is that, at times, facial expressions can be too subtle or simply do not match the prototypic definition which our model relies on to categorize emotions.

Particularly, the NVIE dataset does not provide any information about how images were annotated or which reference system was used. However, it seems that either Asian people tend to express their emotions in a much more subtle way than Western people, or that is a particularity of this dataset. For that reason, we calculate two separate means: one taking into account all datasets and one ignoring the NVIE dataset. We will use the latter to compare the results reported in this chapter.

Hence, our baseline method achieves a mean accuracy of 80.97% and a normalized mean accuracy of 79.64%. Its network architecture, which is the original VGG-Face architecture, has a total of 134,289,223 parameters.

6.4 Method 1: Parts-based VGG-Face

The Parts-based VGG-Face explores the idea of driving the networks attention to specific regions of the face. That is accomplished by creating a multi-branch network architecture, in which each branch is responsible for learning features from each region of interest. Each branch has the exactly same layers of the VGG-Face original architecture, except the last fully-connected layers. Consequently, the Parts-based VGG-Face model is 43% bigger than the original architecture, with a total of 192,590,599 parameters.

Our architecture defines four network inputs, one for each face region: forehead, eyes, nose, and mouth. In our experiments, we tried two different approaches regarding the input images. In the first one, we cropped the regions of interest and resized them to match the input size of 224×244 pixels. The second approach, in turn, consisted of using binary masks to hide out all regions of the face but the ones in which we were interested in. The former approach deforms the regions of the face to fit them into the network's input layers, whereas the latter keeps their original aspect ratio. We believe this could be

one possible explanation to the fact that the best results were obtained with the second approach, which was thus used with Method 1. Table 6.2 and Figure 6.7 show the results of the Parts-based VGG-Face method.

	Recall (%)							Acc (%)	Norm. Acc (%)
	Ang	Dis	Fea	Hap	Neu	Sad	Sur		
Bosporus	74.65	46.38	37.14	99.06	90.97	46.97	81.69	76.73	68.12
CAFE	3.31	83.33	73.42	99.17	96.90	50.00	98.06	72.96	72.03
CFEE	77.83	79.13	69.13	97.83	93.04	81.74	90.43	84.16	84.16
KDEF	69.29	89.29	73.57	98.57	94.29	87.14	84.29	85.20	85.20
NVIE	16.56	17.61	6.36	66.35	93.27	4.05	27.68	26.89	33.12
RaFD	97.51	100.00	94.53	98.51	95.02	97.01	100.00	97.51	97.51
Mean	56.52	69.29	59.02	93.25	93.92	61.15	80.36	73.91	73.36
Mean*	64.52	79.63	69.56	98.63	94.04	72.57	90.89	83.31	81.41

* Without the NVIE dataset.

Table 6.2: Results obtained by employing Method 1: Parts-based VGG-Face.

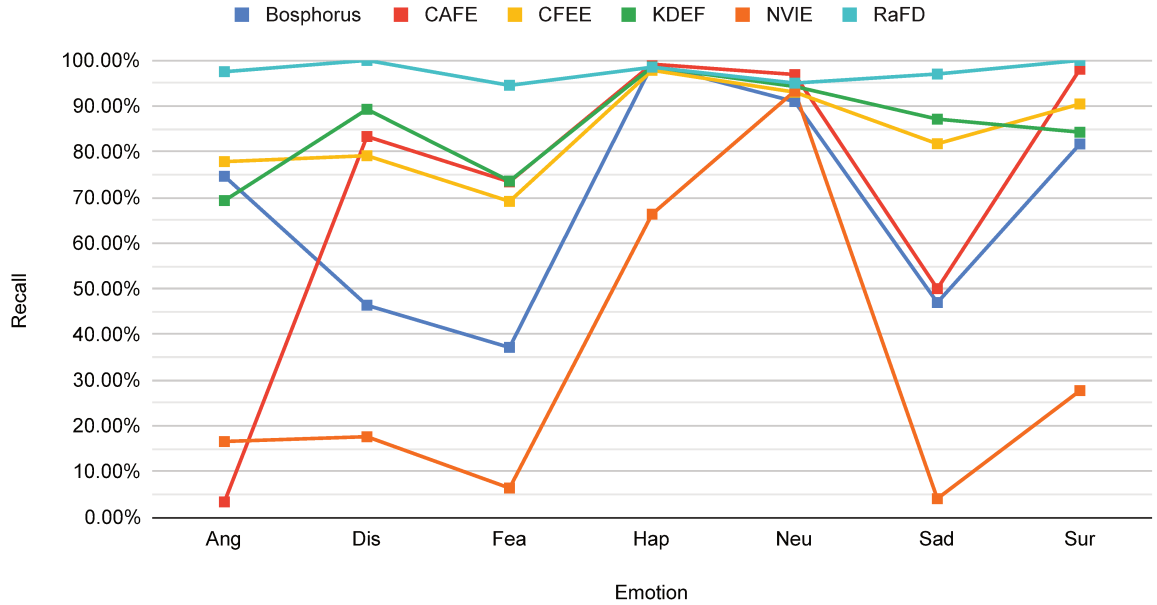


Figure 6.7: Results obtained by employing Method 1: Parts-based VGG-Face.

Apart from the NVIE dataset, the Parts-based VGG-Face achieved higher accuracy in every test dataset when compared to the baseline. We can see gains of 1.73 percentage points for the Bosphorus dataset, 3.80 for the CAFE dataset, 2.61 for the CFEE dataset, 1.22 for the KDEF dataset, and 2.35 for the RaFD dataset. As a result, the mean classification accuracy, disregarding the NVIE dataset, went from 80.97% to 83.31%. Looking at the emotion class level, we observe improvements in the neutral, sadness and surprise emotions.

6.5 Method 2: Random Patches

The Random Patches method is an attempt to implement the same rationale behind the Parts-based VGG-Face method, but with a much simpler network architecture. We force the network to learn less obvious features to draw a distinction between emotions by occluding some regions of the face during training. In fact, the network architecture used here is the original VGG-Face architecture, without any modifications. Therefore, the total number of parameters present in the network is 134,289,223.

During the experiments with this method, we tested whether or not the original face image should be used alongside the modified face image, with random patches, when training the network. We found out that, when the original face image is left out of the training process, the classification accuracy drops consistently. Supposedly, it is important for the network to be exposed to the whole face representation during training, since the same whole face will be used in the test phase. We also tested different sizes of patches to hide out parts of the face. We observed better performances when patches are randomly created with sizes between 10% and 40% of the input size and equally distributed in this interval. In our case, patches vary from squares of 24×24 pixels to squares of 88×88 pixels. Table 6.3 and Figure 6.8 show the results of this second method.

	Recall (%)							Acc (%)	Norm. Acc (%)
	Ang	Dis	Fea	Hap	Neu	Sad	Sur		
Bosphorus	70.42	56.52	52.86	100.00	91.64	53.03	76.06	79.12	71.50
CAFE	3.31	84.38	64.56	100.00	93.02	46.77	97.09	71.13	69.87
CFEE	66.96	84.78	71.74	98.70	93.91	82.17	93.48	84.53	84.53
KDEF	60.71	94.29	74.29	100.00	95.00	84.29	80.71	84.18	84.18
NVIE	14.49	13.25	9.54	65.40	94.71	4.85	20.96	25.38	31.89
RaFD	95.52	100.00	89.05	99.50	96.52	96.02	100.00	96.66	96.66
Mean	51.90	72.20	60.34	93.93	94.13	61.19	78.05	73.50	73.11
Mean*	59.38	83.99	70.50	99.64	94.02	72.46	89.47	83.13	81.35

* Without the NVIE dataset.

Table 6.3: Results obtained by employing Method 2: Random Patches.

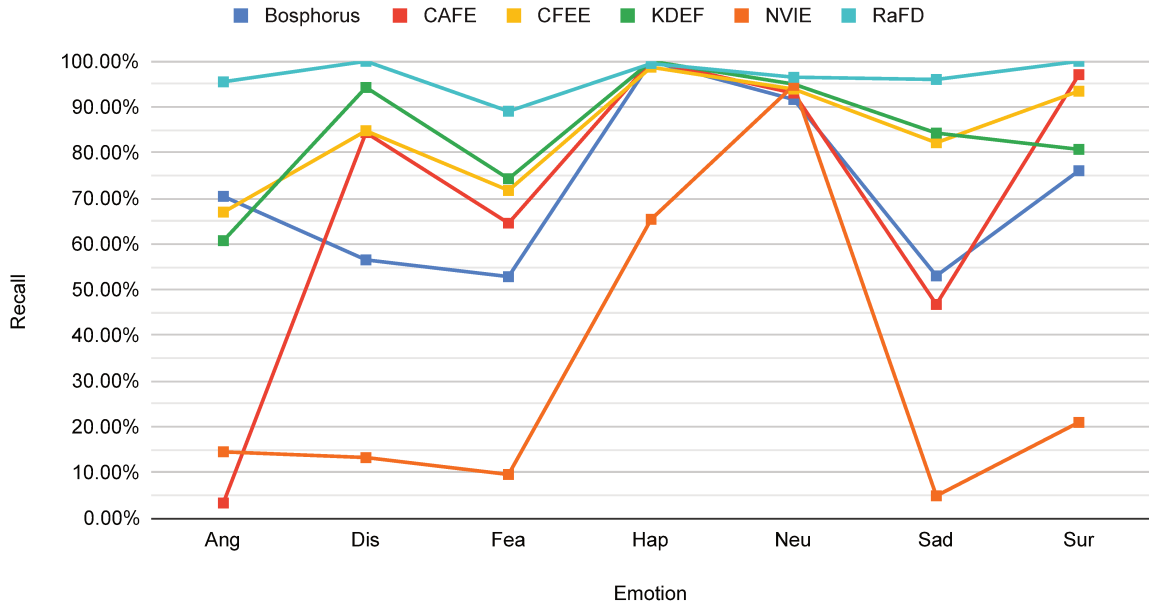


Figure 6.8: Results obtained by employing Method 2: Random Patches.

Leaving the NVIE dataset out, the mean classification accuracy obtained with the Random Patches method is 83.13%, which is very close to the 83.31% of Method 1, but with the advantage of using a much more compact network architecture. Comparing to the baseline results, we can see gains of 4.12 percentage points for the Bosphorus dataset, 1.97 for the CAFE dataset, 2.98 for the CFEE dataset, 0.20 for the KDEF dataset, and 1.49 for the RaFD dataset. At the emotion class level, we can see improvements in every emotion, except for the class anger.

Another benefit of this method is the possibility of replacing the underlying network architecture without great efforts. One can easily unplug the VGG-Face architecture and use an architecture more adapted to their needs in its place. If we want a lightweight network, for instance, we could try SqueezeNet [26] or MobileNet [24]. It is important to mention, though, that we did not perform such kind of tests as they are out of the scope of this work.

6.6 Method 3: Triplet Loss

If with the Random Patches method we worked on the top layers of the network (where data comes in), with the Triplet Loss method, we operate on its bottom. We combined the cross-entropy loss, responsible for giving us class probabilities for each input image, with the triplet loss, a cost function that tries to place examples of the same class close together and examples of different classes far apart in a given hyperspace.

Special care was taken when choosing the weights for each loss function. We wanted both of them to equally contribute to the final combination, without one loss dominating the other. In our case, after checking individual values, we noticed that the cross-entropy

loss was four times higher than the triplet loss in magnitude, so we adjusted the weights of the linear combination accordingly to balance them out.

As triplets are generated in an online manner within each batch of the training process, we defined our batches to have 12 examples of each class of emotion, totaling 84 images per batch. This guarantees a sufficient number of examples from each class to form triplets.

The method’s final model maintains the original VGG-Face architecture and adds a new layer to the bottom of the network to compute the triplet loss function. Nevertheless, the number of parameters in the network is kept unchanged: 134,289,223 in total. Table 6.4 and Figure 6.9 display the results of Method 3.

	Recall (%)							Acc (%)	Norm. Acc (%)
	Ang	Dis	Fea	Hap	Neu	Sad	Sur		
Boshorus	77.46	59.42	48.57	100.00	89.97	40.91	76.06	77.93	70.34
CAFE	3.31	85.42	82.28	97.50	91.47	46.77	98.06	72.68	72.12
CFEE	78.70	84.35	76.09	95.65	85.65	71.30	89.13	82.98	82.98
KDEF	70.71	91.43	72.86	97.14	93.57	68.57	85.00	82.76	82.76
NVIE	16.56	20.84	8.43	57.98	91.83	3.07	33.92	27.15	33.23
RaFD	97.51	99.00	92.54	98.51	94.03	90.05	100.00	95.95	95.95
Mean	57.38	73.41	63.46	91.13	91.09	53.45	80.36	73.24	72.90
Mean*	65.54	83.92	74.47	97.76	90.94	63.52	89.65	82.46	80.83

* Without the NVIE dataset.

Table 6.4: Results obtained by employing Method 3: Triplet Loss.

The Triplet Loss method achieves a mean classification accuracy of 82.46% when the NVIE dataset is left out, compared to the 80.97% of our baseline method. That indicates an increase of 2.93 percentage points for the Bosphorus dataset, 3.52 for the CAFE dataset, 1.43 for the CFEE dataset and 0.78 for the RaFD dataset. For the KDEF dataset, however, the result is 1.22 percentage points worse than the baseline. Also, it is possible to see improvements for disgust, fear, neutral and surprise emotions.

Just like with the Random Patches method, one can easily replace the backbone network architecture to another one that best fits their needs. Just plug the triplet loss layer to the desired architecture before training it to the emotion recognition problem.

6.7 Methods comparison

For each proposed method, we computed the average recall for every class of emotion with respect to all test datasets, as well as the average accuracy and average normalized accuracy. The results are shown in Table 6.5 as well as in Figures 6.10 and 6.11.

When comparing all three proposed methods to the baseline, we see improvements

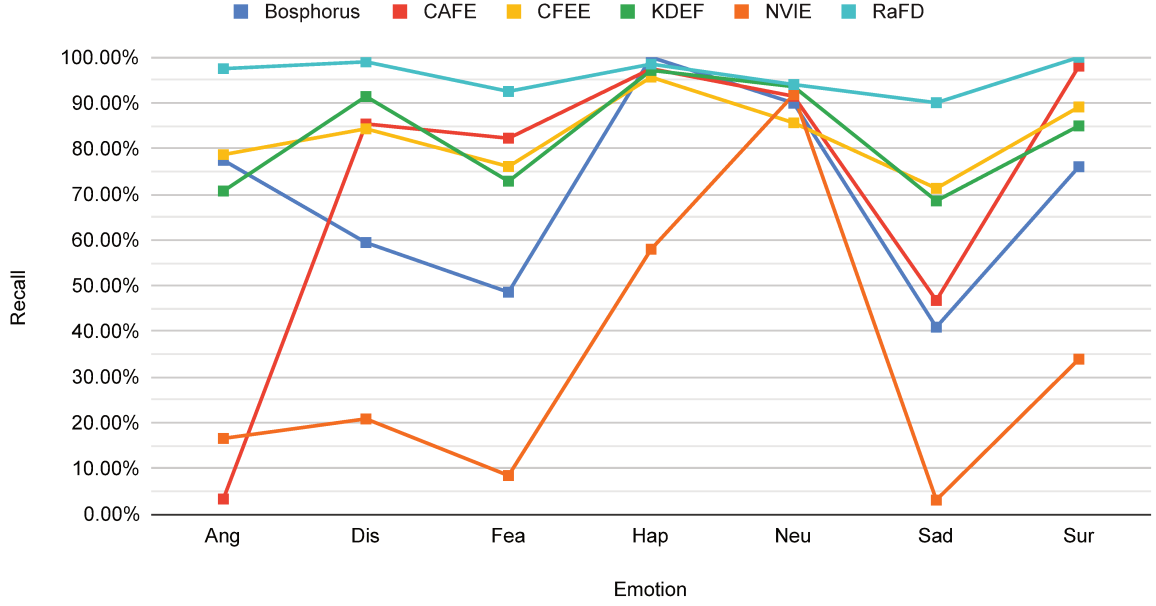


Figure 6.9: Results obtained by employing Method 3: Triplet Loss.

	Recall (%)							Acc (%)	Norm. Acc (%)
	Ang	Dis	Fea	Hap	Neu	Sad	Sur		
Baseline	60.07	72.18	60.46	92.82	86.43	56.58	76.57	72.33	72.16
Method 1	56.52	69.29	59.02	93.25	93.92	61.15	80.36	73.91	73.36
Method 2	51.90	72.20	60.34	93.93	94.13	61.19	78.05	73.50	73.11
Method 3	57.38	73.41	63.46	91.13	91.09	53.45	80.36	73.24	72.90
Baseline*	66.54	81.60	70.30	98.94	85.93	66.44	87.73	80.97	79.64
Method 1*	64.52	79.63	69.56	98.63	94.04	72.57	90.89	83.31	81.41
Method 2*	59.38	83.99	70.50	99.64	94.02	72.46	89.47	83.13	81.35
Method 3*	65.54	83.92	74.47	97.76	90.94	63.52	89.65	82.46	80.83

* Without the NVIE dataset.

Table 6.5: Comparison among baseline and all three proposed methods: Parts-based VGG-Face (Method 1), Random Patches (Method 2) and Triplet Loss (Method 3).

in the mean accuracy. Besides that, what also stands out when we look at the numbers is the complementarity between Random Patches (Method 2) and Triplet Loss (Method 3). If we take Figure 6.11, we can see that, while Method 2 obtained the best results for happiness, neutral and sadness classes of emotion, Method 3 achieved the best results for anger and fear classes of emotion. Hence, combining both methods is a straightforward attempt towards improving results.

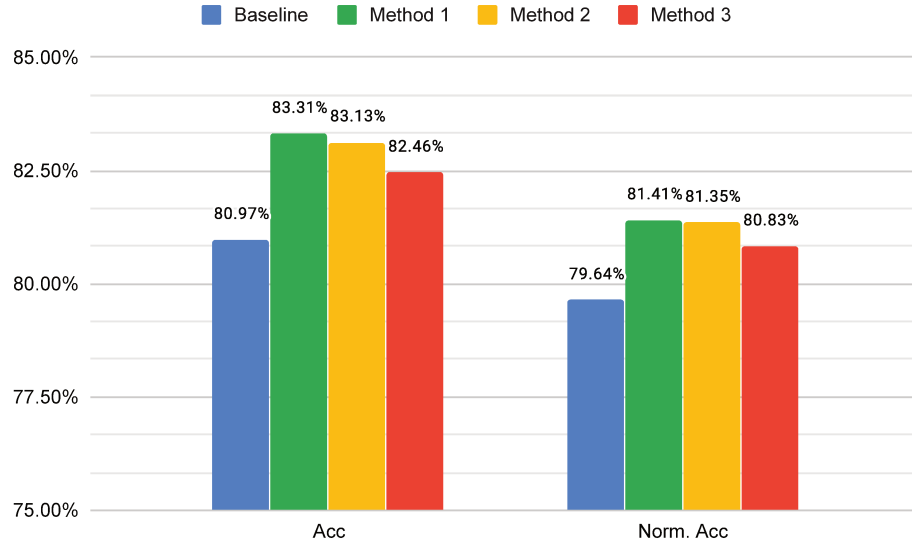


Figure 6.10: Comparison in terms of classification accuracy and normalized accuracy, disregarding the NVIE dataset, among baseline and proposed methods: Parts-based VGG-Face (Method 1), Random Patches (Method 2) and Triplet Loss (Method 3).

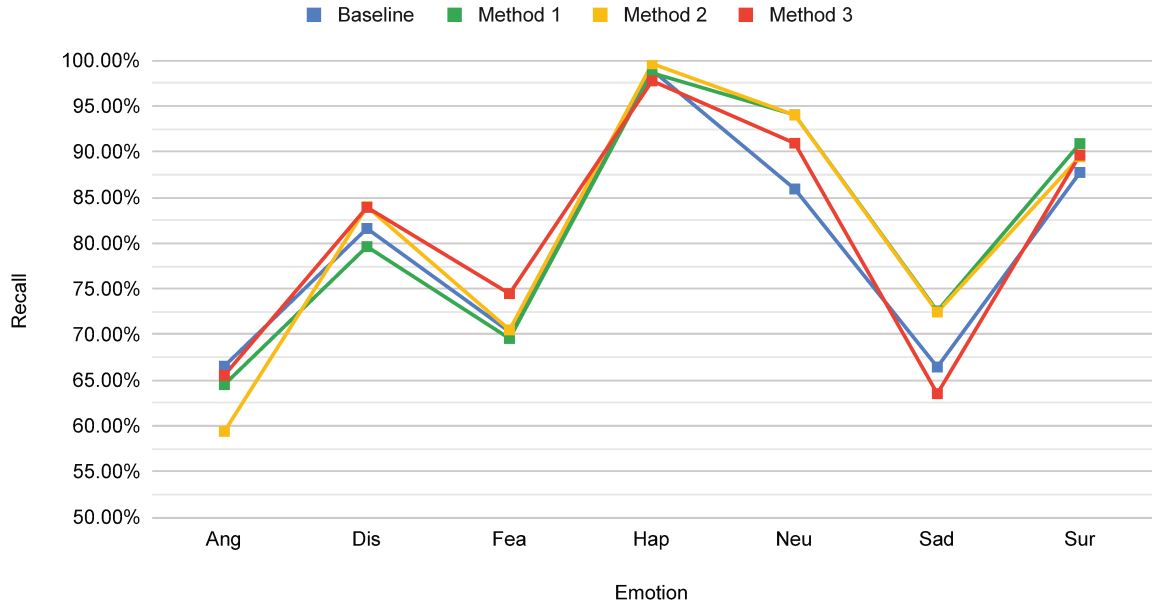


Figure 6.11: Comparison in terms of recall for each class of emotion, disregarding the NVIE dataset, among baseline and proposed methods: Parts-based VGG-Face (Method 1), Random Patches (Method 2) and Triplet Loss (Method 3).

The same complementarity can be seen between Parts-based VGG-Face (Method 1) and Triplet Loss (Method 3). But, as previously discussed, the Parts-based VGG-Face is 43% bigger than the original VGG-Face model and it does not have the property of being easily replaced by another network architecture that Methods 2 and 3 have. For that reason, we will put it aside in our attempts to combine the proposed methods.

There are different ways to combine or fuse methods. In an early fusion approach, feature vectors are combined to create a joint representation of the input data. A single model is then trained in order to learn the domain characteristics. Particularly, we could try to combine both Random Patches and Triplet Loss methods into a single model.

In a late fusion, we combine decision values from multiple models to make a final prediction. It is an easier approach once we already have multiple models trained, though usually less effective. For simplicity’s sake, in this work, we use late fusion to combine the results of Methods 2 and 3. We took the probability vectors of both methods and combined them under three different strategies: max probability, sum of probabilities and multiplication of probabilities. Table 6.6 and Figure 6.12 show the results for each combination strategy.

	Recall (%)							Acc (%)	Norm. Acc (%)
	Ang	Dis	Fea	Hap	Neu	Sad	Sur		
Method 2	59.38	83.99	70.50	99.64	94.02	72.46	89.47	83.13	81.35
Method 3	65.54	83.92	74.47	97.76	90.94	63.52	89.65	82.46	80.83
Comb. (MAX)	60.46	83.82	71.82	99.55	93.99	71.59	89.65	83.33	81.55
Comb. (SUM)	60.30	84.11	71.59	99.47	94.36	71.31	89.65	83.38	81.54
Comb. (MUL)	60.64	83.94	72.51	99.47	94.34	70.89	90.12	83.60	81.70

Table 6.6: Results of combining Random Patches (Method 2) and Triplet Loss (Method 3) under different strategies: max probability (MAX), sum of probabilities (SUM) and multiplication of probabilities (MUL), disregarding the NVIE dataset.

All combination strategies outperformed the results of Methods 2 and 3 alone. Furthermore, they also outperformed Method 1, our best model in terms of accuracy so far. The data also shows us that the multiplication of probabilities strategy provides the best result in terms of classification accuracy, which means it is important that both models have a minimal level of agreement on the predicted class. In other words, very low probabilities tend to heavily penalize high probabilities when computing the final class probability. The complete results of the combination of Methods 2 and 3 through the multiplication of probabilities are shown in Table 6.7.

The combination of Methods 2 and 3 through the multiplication of probabilities achieves a mean accuracy of 83.60% in contrast to 83.13% of Method 2 and 82.46% of Method 3. Except for Method 2 performance in the Bosphorus dataset and Method 3 performance in the CAFE dataset, the fusion of methods yields better results when compared to the ones obtained with each method individually.

In Figure 6.13, we compare the results at the emotion class level. It is possible to see that, for disgust, happiness, neutral and surprise classes of emotion, the combination of methods leads to the best of individual values. For anger, fear and sadness classes, however, its values are placed halfway between individual results of Methods 2 and 3.

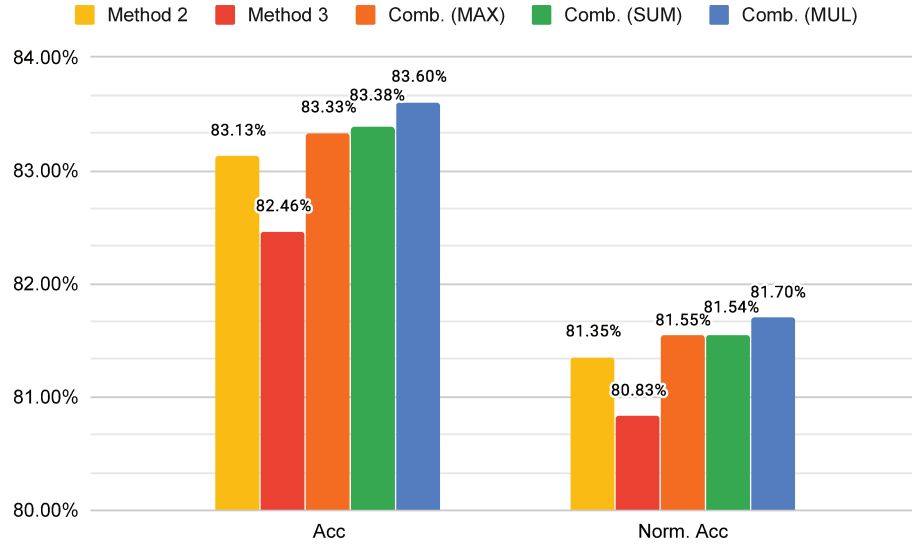


Figure 6.12: Comparison in terms of classification accuracy and normalized accuracy among Random Patches (Method 2), Triplet Loss (Method 3) and their combinations under different strategies: max probability (MAX), sum of probabilities (SUM) and multiplication of probabilities (MUL), disregarding the NVIE dataset.

	Recall (%)							Acc (%)	Norm. Acc (%)
	Ang	Dis	Fea	Hap	Neu	Sad	Sur		
Boshorus	66.20	56.52	52.86	100.00	92.64	50.00	76.06	78.86	70.61
CAFE	3.31	84.38	68.35	100.00	93.80	45.16	98.06	71.69	70.44
CFEE	72.17	85.22	74.35	97.83	93.04	80.87	94.35	85.40	85.40
KDEF	65.00	93.57	76.43	100.00	95.71	81.43	82.14	84.90	84.90
NVIE	15.29	15.19	10.33	63.67	95.19	3.72	23.36	25.88	32.39
RaFD	96.52	100.00	90.55	99.50	96.52	97.01	100.00	97.16	97.16
Mean	53.08	72.48	62.14	93.50	94.48	59.70	78.99	73.98	73.48
Mean*	60.64	83.94	72.51	99.47	94.34	70.89	90.12	83.60	81.70

* Without the NVIE dataset.

Table 6.7: Results of combining Random Patches (Method 2) and Triplet Loss (Method 3) through the multiplication of probabilities (MUL).

The results obtained with these simple late fusion approaches provide us a potential track to be explored in the future. By applying more sophisticated fusion techniques, we may have a good chance of achieving even better results.

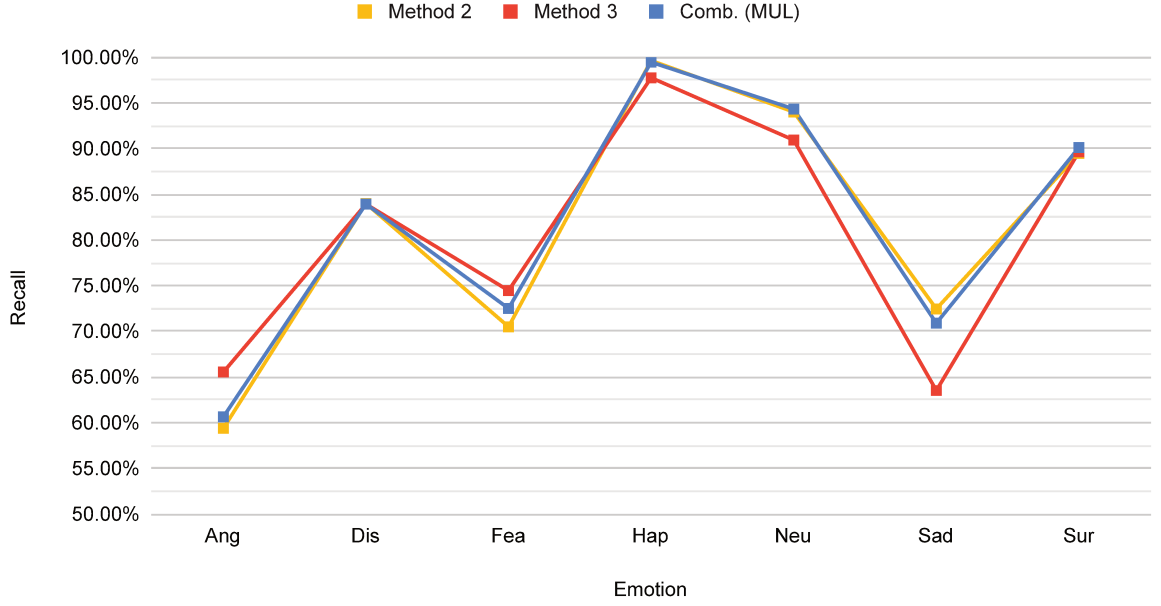


Figure 6.13: Comparison in terms of recall for each class of emotion among Random Patches (Method 2), Triplet Loss (Method 3) and their combination through the multiplication of probabilities (MUL), disregarding the NVIE dataset.

6.8 Statistical test

In this section, we apply the Wilcoxon signed-rank test to verify if our best method, the combination of Methods 2 and 3 through the multiplication of probabilities, has a statistically different performance when compared to our baseline method. We evaluate their classification accuracy on all test datasets, except the NVIE, which has been shown problematic.

The Wilcoxon signed-rank test is a paired non-parametric test. Paired means that it should be used to compare the same test data (e.g., compare the accuracy of two different models on the same dataset). Non-parametric means that it does not presume data to follow any particular statistical distribution [40].

We start by formulating two hypothesis:

$$H_0 : m_B = m_A$$

$$H_1 : m_B \neq m_A$$

where H_0 is our null hypothesis, and states that the population median of scores of method B is equal to the population median of scores of method A , and H_1 is our alternative hypothesis, and states that the population median of scores of method B is different than the population median of scores of method A .

Next, for each test dataset, we compute the difference between scores from methods B and A , the sign of difference score, and its absolute value. Then, we assign ranks to the absolute difference scores, starting from the smallest value, and compute the sum of signed ranks W . Table 6.8 depicts this procedure.

	A	B	$B - A$	Sign	$ B - A $	Signed-rank
Boshorus	75.00	78.86	3.86	+	3.86	5
CAFE	69.15	71.69	2.54	+	2.54	3
CFEE	81.55	85.40	3.85	+	3.85	4
KDEF	83.98	84.90	0.92	+	0.92	1
RaFD	95.17	97.16	1.99	+	1.99	2
						$W = 15$

Table 6.8: Wilcoxon signed-rank test: (A) Our baseline method; (B) The combination of Methods 2 and 3 through the multiplication of probabilities.

For a population of size $n = 5$, there are 32 possible combinations for the signed-rank values. The distribution of signed-ranks sum W is shown in Figure 6.14. According to it, the probability value of obtaining W as large as ± 15 under the null hypothesis is 0.0625. Therefore, we can reject the null hypothesis H_0 with 90% of confidence in favor of the alternative hypothesis H_1 and state that the combination of Methods 2 and 3 through the multiplication of probabilities yields a statistically different distribution from the baseline method.

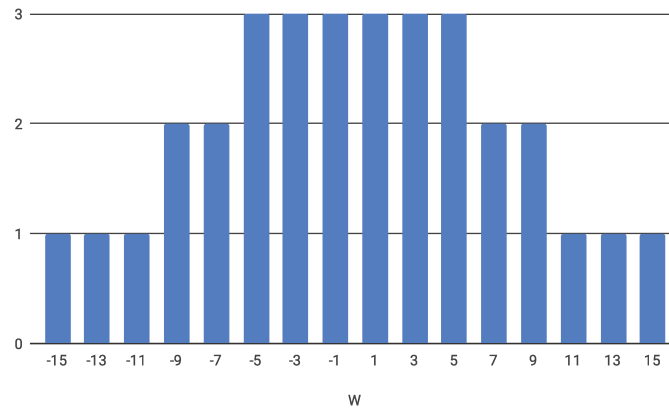


Figure 6.14: Distribution of signed-ranks sum W for a population of size $n = 5$.

Likewise, we also applied the Wilcoxon signed-rank test to the combination of Methods 2 and 3 through the multiplication of probabilities against Method 2 and Method 3. Tables 6.9 and 6.10 show the computed values.

	A	B	$B - A$	Sign	$ B - A $	Signed-rank
Boshorus	79.12	78.86	-0.26	-	0.26	-1
CAFE	71.13	71.69	0.56	+	0.56	3
CFEE	84.53	85.40	0.87	+	0.87	5
KDEF	84.18	84.90	0.72	+	0.72	4
RaFD	96.66	97.16	0.50	+	0.50	2
						$W = 13$

Table 6.9: Wilcoxon signed-rank test: (A) Method 2; (B) The combination of Methods 2 and 3 through the multiplication of probabilities.

When testing the combination of Methods 2 and 3 through the multiplication of probabilities against the Method 2, we obtained $W = 13$. By looking at Figure 6.14, we can see that the probability of obtaining W as large as ± 13 under the null hypothesis is 0.125. In the same way, with Method 3, we observe $W = 11$ and $p\text{-value} = 0.1875$. Hence, for these two tests, we cannot reject the null hypothesis H_0 and, consequently, we can state that the combination of Methods 2 and 3 through the multiplication of probabilities is not statistically different from Method 2 or Method 3 in isolation at a significance level of 0.1.

	A	B	$B - A$	Sign	$ B - A $	Signed-rank
Boshorus	77.93	78.86	0.93	+	0.93	1
CAFE	72.68	71.69	-0.99	-	0.99	-2
CFEE	82.98	85.40	2.42	+	2.42	5
KDEF	82.76	84.90	2.14	+	2.14	4
RaFD	95.95	97.16	1.21	+	1.21	3
						$W = 11$

Table 6.10: Wilcoxon signed-rank test: (A) Method 3; (B) The combination of Methods 2 and 3 through the multiplication of probabilities.

6.9 Performance comparison with prior art

In this section, we benchmark the results obtained with our methods and compare them to results of existing works and of popular commercial applications for emotion recognition from facial expressions. This benchmark was developed on a per-dataset basis, as most methods described in the literature are evaluated on a single dataset.

In the following tables, we detail the methods used, the evaluation protocols, the number of classes of emotion, and accuracy obtained for the most relevant works found in the literature. It is worth mentioning that the majority of them follow an **intra-dataset protocol**, which means that the same dataset was used in both training and testing phases. Only a few adopt a cross-dataset evaluation protocol.

Bosphorus (intra-dataset protocol)				
Work	Method	Protocol	Classes	Acc (%)
Zhang et al. [85]	Shape features + mRMR + SVR + SVM ensemble	5-fold cross-validation	6	92.20
Taha and Hatzinakos [70]	CNN	5-fold cross-validation	6	88.20
Li et al. [35]	DF-CNN + SVM	10-fold cross-validation	6	80.28
Ahmed et al. [2]	Shape features + SVM	80/20 holdout	7	78.38
Silva and Pedrini [6]	Gabor filters + NN	80/20 holdout	6	73.13
Bosphorus (cross-dataset protocol)				
Work	Method	Protocol	Classes	Acc (%)
Ours	Method 2	cross-dataset (CK+, MUG, Oulu-CASIA)	7	79.12
Silva and Pedrini [6]	HOG + SVM	cross-dataset (MUG)	6	53.80
Silva and Pedrini [6]	HOG + SVM	cross-dataset (CK+)	6	43.00

Table 6.11: Performance comparison for Bosphorus dataset.

CAFE (intra-dataset protocol)				
Work	Method	Protocol	Classes	Acc (%)
Zheng et al. [86]	Shape features + SVM	70/30 holdout	7	77.40
Witherow et al. [78]	CNN	10-fold cross-validation	7	76.03
Nagpal et al. [47]	msDBM + RF	5 x 30/70 sub-sampling validation	7	48.00
CAFE (cross-dataset protocol)				
Work	Method	Protocol	Classes	Acc (%)
Ours	Method 1	cross-dataset (CK+, MUG, Oulu-CASIA)	7	72.96
Zheng et al. [86]	Shape features + SVM	cross-dataset (CK+, CFEE, Multi-PIE)	7	64.70
Witherow et al. [78]	CNN	cross-dataset (CK+)	7	46.50

Table 6.12: Performance comparison for CAFE dataset.

CFEE (intra-dataset protocol)				
Work	Method	Protocol	Classes	Acc (%)
Du et al. [11]	Shape features + Gabor filters + KSDA	10-fold cross-validation	7	96.86
Koujan et al. [33]	CNN + SVM	10-fold cross-validation	7	96.43
Neto et al. [7]	CNN	10-fold cross-validation	7	82.54
Mavani et al. [46]	CNN	85/15 holdout	7	74.79
CFEE (cross-dataset protocol)				
Work	Method	Protocol	Classes	Acc (%)
Ours	Methods 2 and 3 combined	cross-dataset (CK+, MUG, Oulu-CASIA)	7	85.40
Zeng et al. [84]	LTNet	cross-dataset (AffectNet, RAF)	7	76.02

Table 6.13: Performance comparison for CFEE dataset.

KDEF (intra-dataset protocol)				
Work	Method	Protocol	Classes	Acc (%)
Ruiz-Garcia et al. [53]	CNN	70/30 holdout	7	96.93
Ruiz-Garcia et al. [54]	SCAE	70/30 holdout	7	92.52
Koujan et al. [33]	CNN + SVM	10-fold cross validation	7	92.24
Islam et al. [28]	Gabor filter + EML	10-fold cross validation	7	86.84
Lucey et al. [41]	SPTS + CAPP + SVM	leave-one-subject-out	7	82.86
KDEF (cross-dataset protocol)				
Work	Method	Protocol	Classes	Acc (%)
Ours	Method 1	cross-dataset (CK+, MUG, Oulu-CASIA)	7	85.20
Magyar et al. [44]	Face++ (commercial)	N/A	7	77.08
Magyar et al. [44]	Microsoft Face (commercial)	N/A	7	75.33
Zavares et al. [82]	CNN	cross-dataset (CK+, JAFFE, MMI, RaFD, BU3DFE, ARFace)	7	72.55
Magyar et al. [44]	F.A.C.E. (commercial)	N/A	7	63.21

Table 6.14: Performance comparison for KDEF dataset.

RaFD (intra-dataset protocol)				
Work	Method	Protocol	Classes	Acc (%)
Koujan et al. [33]	CNN + SVM	10-fold cross validation	7	97.65
Yaddaden et al. [80]	CNN	10-fold cross validation	7	97.57
Islam et al. [28]	Gabor filter + EML	10-fold cross validation	7	96.94
RaFD (cross-dataset protocol)				
Work	Method	Protocol	Classes	Acc (%)
Ours	Method 1	cross-dataset (CK+, MUG, Oulu-CASIA)	7	97.51
Zavares et al. [82]	CNN	cross-dataset (CK+, JAFFE, MMI, RaFD, BU3DFE, ARFace)	7	85.97
Mavani et al. [46]	CNN	cross-dataset (CFEE)	7	77.19
Magyar et al. [44]	Microsoft Face (commercial)	N/A	7	76.24
Magyar et al. [44]	Sighthound (commercial)	N/A	7	72.33
Magyar et al. [44]	Face++ (commercial)	N/A	7	71.33

Table 6.15: Performance comparison for RaFD dataset.

Even though there are few works following a cross-dataset protocol to compare to, our methods achieve the best results in all datasets in this scenario. In addition to that, it is possible to say that the performance of our methods are significantly competitive when compared to works that adopted the intra-class evaluation protocol. From the presented results, we can conclude that the most challenging datasets are Bosphorus and CAFE.

One curious observation concerns the results obtained by commercial applications. Magyar et al. [44] tested several of the most popular solutions available in the market, including the ones from giant tech companies such as Microsoft and Google, using the KDEF and RaFD datasets. Interestingly, our methods outperformed all of them in both datasets. Although we are not aware of the methods and datasets used to build such solutions, we imagine they have access to a much larger amount of data than we do when training our models.

Chapter 7

Conclusion and future work

Recognizing emotions from facial expressions, even for humans, is not an easy task. As discussed in Section 2.1, emotions are complex phenomena, which are comprised of events in several organismic subsystems, and the manifestation of facial expressions is only one of possible reactions the body can communicate [32, 61]. Although some authors state that emotions lead to automatic and stereotyped body responses, including facial expressions [10, 9], they seem to ignore the fact that it is possible for someone to experience an emotion without moving their facial muscles and to also fake an emotion by mimicking the suitable facial expression. In fact, that is the case with all the datasets we used in this work, as images were captured from subjects in posed, not spontaneous, expressions.

Basic emotions – anger, disgust, fear, happiness, sadness, and surprise – are believed to have a universal facial configuration among different cultures and populations across the globe [13, 14]. These configurations can be described in terms of different visibly facial movements, known as Action Units, and the resulting coding scheme can be used to map facial expressions into emotion categories, following a prototypic definition. In this research, however, we found examples that contradict these claims.

In the CAFE dataset, for example, composed exclusively by children subjects, the mean image for the anger emotion does not match the typical facial expressions suggested in the literature. Likewise, in the NVIE dataset, formed by only Asiatic people, it is practically impossible to tell which emotion is being displayed in each mean image as facial expressions are subtle. Hence, even though there seems to be a more frequent facial configuration for each basic emotion, we cannot take its universality across different populations and cultures for granted. Facial expressions may vary in both format and intensity depending on the subject, even when they are asked to perform posed expressions.

These examples also evince that the EMFACS framework, which we adopted in this research, has its own limitations. If, on one hand, the framework specifies a precise prototypic definition for each emotion in terms of facial movements, on the other, it narrows down the margins for detecting variations such as observed in CAFE and NVIE datasets.

For all those reasons, we do not believe it is possible to reliably infer someone’s internal emotional state from their facial expressions, an external easily manipulable tiny portion of the whole emotion phenomenon. Ideally, a reliable solution should also take into consideration other organismic subsystems involved in the emotion episode, analyz-

ing distinct and possibly complementary body information. That being said, the most we can postulate by simply looking at someone’s facial expressions is which emotion is more likely to be represented according to the prototypic definitions. We believe this is a fairer description of what we sought to accomplish in this study.

Throughout our investigation, we proposed three different methods to categorize emotions from facial expressions, all three derived from the same baseline solution. Accompanying the winners of the latest competitions in the field, we opted for a CNN-based approach to accomplish the task. The selected architecture was the VGG-Face, a popular network for the face recognition problem.

We started by fine-tuning the pre-trained VGG-Face model for our problem, being the result of this process considered our baseline method. Then, with the help of the Grad-CAM algorithm, we generated activation maps to visualize what the network had learned from its initial training process. The heatmaps showed us that some regions of the face considered important in the literature were not taken into account. Consequently, our first proposed method is an attempt to drive the network’s attention to specific regions of the face. It is a parts-based architecture derived from the original VGG-Face model, with four different convolutional branches, one for each region of interest: forehead, eyes, nose, and mouth. The second proposed method follows the same rationale, but instead of using a heavy multi-branched model, we preserved the original VGG-Face architecture and only modified the images inputted to the network. We applied random patches to occlude some portions of the face, as a means to force the network to learn alternative features. Finally, in the third proposed method, we employed a triplet loss function with the objective of learning good face embeddings for emotion categorization. The idea is that, when transported to a new hyperspace, the embeddings will place emotions of the same class close together while keeping emotions of different classes apart.

Our proposed methods were evaluated following a cross-dataset protocol, in which the datasets used to train and validate our models are different from the datasets used to test their performance. With this approach, we are interested to know if characteristics learned from one group of people can be employed to successfully categorize emotions from facial expressions in another one. While we achieved good results in test datasets annotated after the EMFACS framework, we saw a drastic performance drop in the NVIE dataset, which provides no information about how its annotation process took place. Therefore, we can conclude that model generalization is unlikely to occur unless the targeted dataset follows a similar prototypic definition of emotion of the training datasets. For that reason, we advocate for the adoption of a standard annotation protocol based on objective traits, such as facial muscle movements, as seen in FACS. Additionally, deciding for a more flexible framework, capable of addressing possible variations of each facial representation of emotion, can also be of help.

Disregarding the NVIE dataset, we started from a baseline method with a mean classification accuracy among test datasets of 80.97% to a mean classification accuracy of 83.31% with Method 1. Then, with Method 2 and Method 3, we achieved classification accuracies of 83.13% and 82.46%, respectively. We also explored a late fusion technique to verify the complementarity between methods. By combining Methods 2 and 3 through the multiplication of probabilities, we obtained 83.60% of accuracy.

Although there is not a standard protocol adopted by the scientific community to compare results in this area, we compiled an extensive list of related work to assess our methods. We tried to make the comparison fair by underlining the evaluation protocol used in each work. From there, it is possible to see that, under a cross-dataset protocol, our methods achieved state-of-the-art results in all tested datasets, outperforming even commercial applications from popular tech companies. In addition, they also present competitive numbers when compared to works that followed an intra-dataset protocol, arguably an easier validation protocol. While we acknowledge that the use of different datasets to train a model will probably lead to different results, by examining common methodologies employed in the literature, we can point out the practices we believe have contributed the most to provide results superior to prior works.

To begin with, data plays a major role in deep convolutional neural networks. It is usually required a huge amount of data to achieve specific performance goals. Besides, data variability is also a critical point when our objective is to generate unbiased models. If we do not have enough data to train a model from scratch, we can leverage techniques such as transfer learning and data augmentation. With transfer learning, the closer the task of the pre-trained model is to our problem, the better the results are expected to be. In our case, we adapted a model originally trained for face recognition to a problem of emotion categorization from facial expressions. In both cases, we are dealing with the same kind of data: facial images. With data augmentation, it is important that the operations used to increase the sets ensure a correspondent level of diversity that we expect to see when evaluating the models. Images under different illumination conditions, blurred, and not totally aligned are some of the examples we employed in this work.

Another key point in this research was the application of activation maps to analyze what regions of the face the network learned as the most important to categorize emotions. We noticed a mismatch between the heatmaps generated by the Grad-CAM method and the prototypic definition of each emotion presented in the literature. This finding gave us room to explore attention mechanisms with the purpose of making the network aware of these formerly ignored regions. As a result of this investigation, we developed both Methods 1 and 2. The data visualization techniques also helped us figure out issues with the CAFE and the NVIE datasets, as we discussed earlier.

Moving on from the network input to its output, Method 3 evinces that our solution can benefit from a metric learning approach to achieve better generalization, when combined to a predictive function. We believe the way clusters are formed by the triplet loss function, when data embeddings are transported to a high dimensional feature space during the optimization process, is less biased towards training datasets.

Despite the good results reported in this research, it is clear that the problem is far from being solved, especially when models are evaluated under a cross-dataset protocol, which is closer to real-world scenarios. Therefore, as future work, there are some directions we would like to explore. Firstly, as we have already seen good evidence of complementarity between Methods 2 and 3, we could explore a more elaborated early fusion approach by proposing a single end-to-end trainable network. The network would be fed with batches containing the original face image and its version with occluded regions during the training phase to an optimization process lead by a combination between a triplet loss function

and a categorical loss function. Other loss functions other than these two could also be experimented with.

Additionally, we would like to explore the use of activation maps as a way to guide the network's attention during training. As the optimization process progresses, online generated maps could be used to mask the most discriminative parts of the face, hoping that the network would look for alternate regions to recognize emotions. As an output, we expect a more generalizable model. Furthermore, we could also verify how other popular network architectures, particularly the ones trained for face recognition, would adapt to our problem.

At last, considering the problem as a whole, it would be great to extend this study to datasets in which facial expressions are captured spontaneously, under uncontrolled situations. Since this is a very complex subject, involving different areas of science, we recommend future studies to be carried out by a multidisciplinary team, composed of psychologists, neuroscientists, and computer scientists.

Bibliography

- [1] Affectiva. Driver emotion recognition and real time facial analysis for the automotive industry. Retrieved August 7, 2020, from <https://blog.affectiva.com/driver-emotion-recognition-and-real-time-facial-analysis-for-the-automotive-industry>.
- [2] Haval A. Ahmed, Tarik A. Rashid, and Ahmed T. Sidiq. Face behavior recognition through support vector machines. *International Journal of Advanced Computer Science and Applications*, 7(1):101–108, 2016.
- [3] Niki Aifanti, Christos Papachristou, and Anastasios Delopoulos. The mug facial expression database. In *11th International Workshop on Image Analysis for Multimedia Interactive Services WIAMIS 10*, pages 1–4, 2010.
- [4] Lisa Feldman Barrett, Ralph Adolphs, Stacy Marsella, Aleix M. Martinez, and Seth D. Pollak. Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements. *Psychological Science in the Public Interest*, 20(1):1–68, 2019.
- [5] Ciprian A. Corneanu, Marc O. Simón, Jeffrey F. Cohn, and Sergio 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.
- [6] Flávio Altinier Maximiano da Silva and Helio Pedrini. Effects of cultural characteristics on building an emotion classifier through facial expression analysis. *Journal of Electronic Imaging*, 24(2):023015, 2015.
- [7] Humberto da Silva Neto, Clebeson Canuto, Mariana Rampinelli, and Jorge Samatelo. Transfer learning for facial emotion recognition. In *XIV Workshop de Visão Computacional*, 11 2018.
- [8] Navneet Dalal and Bill 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 vol. 1, 2005.
- [9] Antonio R. Damasio. *The Feeling of What Happens: Body and Emotion in the Making of Consciousness*. Harcourt Brace and Co, 1999.
- [10] Raymond J. Dolan. Emotion, cognition, and behavior. *Science*, 298(5596):1191–1194, 2002.
- [11] Shichuan Du, Yong Tao, and Aleix M. Martinez. Compound facial expressions of emotion. *Proceedings of the National Academy of Sciences*, 111(15):E1454–E1462, 2014.
- [12] Paul Ekman. Facial expression and emotion. *American psychologist*, 48(4):384–392, 1993.
- [13] Paul Ekman. Facial expressions. In Tim Dalgleish and Mick Power, editors, *Handbook of cognition and emotion*, chapter 16, pages 301–320. John Wiley & Sons Ltd., 1999.
- [14] Paul Ekman and Daniel Cordaro. What is meant by calling emotions basic. *Emotion Review*, 3(4):364–370, 2011.
- [15] Paul Ekman and Wallace V. Friesen. Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology*, 17(2):124, 1971.

- [16] Paul Ekman and Wallace V. Friesen. Measuring facial movement. *Environmental Psychology and Nonverbal Behavior*, 1(1):56–75, 1976.
- [17] EmotiW. Emotion recognition in the wild challenge. Retrieved March 23, 2020, from <https://sites.google.com/view/emotiw2020>.
- [18] C. Fabian Benitez-Quiroz, Ramprakash Srinivasan, and Aleix M. Martinez. Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5562–5570, 2016.
- [19] FERA. Facial expression recognition and analysis challenge. Retrieved March 23, 2020, from <http://www.fg2017.org>.
- [20] Wallace V. Friesen and Paul Ekman. Emfacs-7: Emotional facial action coding system. *Unpublished manuscript, University of California at San Francisco*, 2(36):1, 1983.
- [21] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 580–587, 2014.
- [22] Steven. L. Gordon. The sociology of sentiments and emotion. In Morris Rosenberg and Ralph H. Turner, editors, *Social Psychology: Sociological Perspectives*, chapter 18, pages 562–592. Transaction Publishers, 1990.
- [23] Douglas Heaven. Why faces don’t always tell the truth about feelings. *Nature*, 578:502–504, 02 2020.
- [24] Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.
- [25] Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik 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.
- [26] Forrest N. Iandola, Song Han, Matthew W. Moskewicz, Khalid Ashraf, William J. Dally, and Kurt Keutzer. Squeezenet: Alexnet-level accuracy with 50x fewer parameters and < 0.5 mb model size. *arXiv preprint arXiv:1602.07360*, 2016.
- [27] Imotions. Facial action coding system. Retrieved March 23, 2020, from <https://imotions.com/blog/facial-action-coding-system>.
- [28] Bayezid Islam, Firoz Mahmud, and Arfat Hossain. Facial region segmentation based emotion recognition using extreme learning machine. In *2018 International Conference on Advancement in Electrical and Electronic Engineering (ICAEEE)*, pages 1–4. IEEE, 2018.
- [29] Takeo Kanade, Jeffrey F. Cohn, and Yingli Tian. Comprehensive database for facial expression analysis. In *IEEE Intl. Conference on Automatic Face and Gesture Recognition*, pages 46–53, 2000.
- [30] Vahid Kazemi and Josephine Sullivan. One millisecond face alignment with an ensemble of regression trees. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1867–1874, 2014.
- [31] Davis E. King. Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, 10:1755–1758, 2009.
- [32] Paul R. Kleinginna and Anne M. Kleinginna. A categorized list of emotion definitions, with suggestions for a consensual definition. *Motivation and Emotion*, 5(4):345–379, 1981.
- [33] Mohammad Rami Koujan, Luma Alharbawee, Giorgos Giannakakis, Nicolas Pugeault, and Anastasios Roussos. Real-time facial expression recognition in the wild by disentangling 3d expression from identity. *arXiv preprint arXiv:2005.05509*, 2020.

- [34] Oliver Langner, Ron Dotsch, Gijsbert Bijlstra, Daniel Wigboldus, Skyler Hawk, and Ad Knippenberg. Presentation and validation of the radboud face database. *Cognition & Emotion*, 24:1377–1388, 12 2010.
- [35] Huibin Li, Jian Sun, Zongben Xu, and Liming Chen. Multimodal 2d+ 3d facial expression recognition with deep fusion convolutional neural network. *IEEE Transactions on Multimedia*, 19(12):2816–2831, 2017.
- [36] Shan Li and Weihong Deng. Deep facial expression recognition: A survey. *IEEE Transactions on Affective Computing*, 2020.
- [37] Paul Ekman Group LLC. Facial action coding system. Retrieved March 23, 2020, from <https://www.paulekman.com/facial-action-coding-system/>.
- [38] Vanessa LoBue and Cat Thrasher. The child affective facial expression (cafe) set. databrary. Retrieved April 5, 2020 from <http://doi.org/10.17910/B7301K>.
- [39] Vanessa LoBue and Cat Thrasher. The child affective facial expression (cafe) set: Validity and reliability from untrained adults. *Frontiers in psychology*, 5:1532, 01 2014.
- [40] Richard Lowre. Concepts & applications of inferential statistics. Vassar College. Retrieved June 16, 2020, from <http://vassarstats.net/textbook/>.
- [41] Patrick Lucey, Jeffrey F. Cohn, Takeo Kanade, Jason Saragih, Zara Ambadar, and Iain Matthews. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 94–101, 2010.
- [42] Daniel Lundqvist, Andreas Flykt, and Arne Ohman. The karolinska directed emotional faces - kdef. *Psychology Section, Department of Clinical Neuroscience, Karolinska Institutet*, 1998. ISBN: 91-630-7164-9.
- [43] Yanpeng Lv, Shangfei Wang, and Peijia Shen. A real-time attitude recognition by eye-tracking. In *Proceedings of the Third International Conference on Internet Multimedia Computing and Service, ICIMCS 2011*, page 170–173, 2011.
- [44] Ján Magyar, Gergely Magyar, and Peter Sincak. A cloud-based voting system for emotion recognition in human-computer interaction. In *2018 World Symposium on Digital Intelligence for Systems and Machines (DISA)*, pages 109–114. IEEE, 2018.
- [45] MarketsandMarkets. Emotion detection and recognition market. Retrieved March 23, 2020, from <https://www.marketsandmarkets.com/Market-Reports/emotion-detection-recognition-market-23376176.html>.
- [46] Viraj Mavani, Shanmuganathan Raman, and Krishna P. Miyapuram. Facial expression recognition using visual saliency and deep learning. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 2783–2788, 2017.
- [47] Shruti Nagpal, Maneet Singh, Mayank Vatsa, Richa Singh, and Afzel Noore. Expression classification in children using mean supervised deep boltzmann machine. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019.
- [48] Rafael Padilha, Fernanda Andaló, and Anderson Rocha. Improving the chronological sorting of images through occlusion: A study on the notre-dame cathedral fire. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 01 2020.
- [49] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2009.
- [50] Omkar M. Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep face recognition. In *British Machine Vision Conference*, 2015.
- [51] Zia-ur Rahman, Daniel Jobson, and Glenn Woodell. Retinex processing for automatic image enhancement. *J. Electronic Imaging*, 13:100–110, 01 2004.









- [52] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(6):1137–1149, June 2017.
- [53] Ariel Ruiz-Garcia, Mark Elshaw, Abdulrahman Altahhan, and Vasile Palade. Deep learning for emotion recognition in faces. In *International Conference on Artificial Neural Networks*, pages 38–46. Springer, 2016.
- [54] Ariel Ruiz-Garcia, Mark Elshaw, Abdulrahman Altahhan, and Vasile Palade. Stacked deep convolutional auto-encoders for emotion recognition from facial expressions. In *2017 International Joint Conference on Neural Networks (IJCNN)*, pages 1586–1593. IEEE, 2017.
- [55] Christos Sagonas, Epameinondas Antonakos, Georgios Tzimiropoulos, Stefanos Zafeiriou, and Maja Pantic. 300 faces in-the-wild challenge: database and results. *Image and Vision Computing*, 47:3 – 18, 2016. 300-W, The First Automatic Facial Landmark Detection In-The-Wild Challenge.
- [56] Evangelos Sariyanidi, Hatice Gunes, and Andrea 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, 2015.
- [57] Arman Savran, Neşe Alyüz, Hamdi Dibeklioglu, Oya Çeliktutan, Berk Gökberk, Bülent Sankur, and Lale Akarun. Bosphorus database for 3d face analysis. In Ben Schouten, Niels Christian Juul, Andrzej Drygajlo, and Massimo Tistarelli, editors, *Biometrics and Identity Management*, pages 47–56, Berlin, Heidelberg, 2008. Springer Berlin Heidelberg.
- [58] Arman Savran, Bülent Sankur, and M. Taha Bilge. Facial action unit detection: 3d versus 2d modality. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops*, pages 71–78. IEEE, 2010.
- [59] Arman Savran, Bulent Sankur, and M. Taha Bilge. Regression-based intensity estimation of facial action units. *Image Vision Comput.*, 30(10):774–784, October 2012.
- [60] Klaus R. Scherer. Psychological models of emotion. In Joan C. Borod, editor, *The Neuropsychology of Emotion*, Series in Affective Science, chapter 6, pages 137–162. Oxford University Press, 2000.
- [61] Klaus R. Scherer. What are emotions? and how can they be measured? *Social Science Information*, 44(4):695–729, 2005.
- [62] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2015.
- [63] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.
- [64] Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. Cnn features off-the-shelf: an astounding baseline for recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 806–813, 2014.
- [65] Minchul Shin, Munsang Kim, and Dong-Soo Kwon. Baseline cnn structure analysis for facial expression recognition. In *2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pages 724–729. IEEE, 2016.
- [66] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.
- [67] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806*, 2014.

- [68] Wenyun Sun, Haitao Zhao, and Zhong Jin. An efficient unconstrained facial expression recognition algorithm based on stack binarized auto-encoders and binarized neural networks. *Neurocomputing*, 267:385–395, 2017.
- [69] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 3319–3328. JMLR.org, 2017.
- [70] Bilal Taha and Dimitrios Hatzinakos. Emotion recognition from 2d facial expressions. In *2019 IEEE Canadian Conference of Electrical and Computer Engineering (CCECE)*, pages 1–4. IEEE, 2019.
- [71] Chuanqi Tan, Fuchun Sun, Tao Kong, Wenchang Zhang, Chao Yang, and Chunfang Liu. A survey on deep transfer learning. In Věra Kůrková, Yannis Manolopoulos, Barbara Hammer, Lazaros Iliadis, and Ilias Maglogiannis, editors, *Artificial Neural Networks and Machine Learning – ICANN 2018*, pages 270–279, Cham, 2018. Springer International Publishing.
- [72] TechCrunch. Apple dives deeper into artificial intelligence by acquiring emotient. Retrieved March 3, 2017, from <https://techcrunch.com/2016/01/07/apple-dives-deeper-into-artificial-intelligence-by-acquiring-emotient>.
- [73] TechCrunch. Like by smiling? facebook acquires emotion detection startup faciometrics. Retrieved March 3, 2017, from <https://techcrunch.com/2016/11/16/facial-gesture-controls>.
- [74] Yan Tong, Yang Wang, Zhiwei Zhu, and Qiang Ji. Robust facial feature tracking under varying face pose and facial expression. *Pattern Recognition*, 40(11):3195 – 3208, 2007.
- [75] Shangfei Wang, Zhilei Liu, Siliang Lv, Yanpeng Lv, Guobing Wu, Peng Peng, Fei Chen, and Xufa Wang. A natural visible and infrared facial expression database for expression recognition and emotion inference. *IEEE Transactions on Multimedia*, 12(7):682–691, 2010.
- [76] Shangfei Wang, Zhilei Liu, Zhaoyu Wang, Guobing Wu, Peijia Shen, Shan He, and Xufa Wang. Analyses of a multimodal spontaneous facial expression database. *IEEE Transactions on Affective Computing*, 4(1):34–46, 2013.
- [77] Yunchao Wei, Jiashi Feng, Xiaodan Liang, Ming-Ming Cheng, Yao Zhao, and Shuicheng Yan. Object region mining with adversarial erasing: A simple classification to semantic segmentation approach. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1568–1576, 2017.
- [78] Megan A. Witherow, Manar D. Samad, and Khan M. Iftekharruddin. Transfer learning approach to multiclass classification of child facial expressions. In *Applications of Machine Learning*, volume 11139, page 1113911. International Society for Optics and Photonics, 2019.
- [79] Lior Wolf, Tal Hassner, and Itay Maoz. Face recognition in unconstrained videos with matched background similarity. In *CVPR 2011*, pages 529–534, 2011.
- [80] Yacine Yaddaden, Mehdi Adda, Abdenour Bouzouane, Sebastien Gaboury, and Bruno Bouchard. User action and facial expression recognition for error detection system in an ambient assisted environment. *Expert Systems with Applications*, 112:173–189, 2018.
- [81] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 3320–3328. Curran Associates, Inc., 2014.
- [82] Marcus Vinicius Zavarez, Rodrigo F Berriel, and Thiago Oliveira-Santos. Cross-database facial expression recognition based on fine-tuned deep convolutional network. In *2017 30th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, pages 405–412. IEEE, 2017.
- [83] Matthew D. Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014.

- [84] Jiabei Zeng, Shiguang Shan, and Xilin Chen. Facial expression recognition with inconsistently annotated datasets. In *Proceedings of the European conference on computer vision (ECCV)*, pages 222–237, 2018.
- [85] Yang Zhang, Li Zhang, and M. Alamgir Hossain. Adaptive 3d facial action intensity estimation and emotion recognition. *Expert systems with applications*, 42(3):1446–1464, 2015.
- [86] Zhi Zheng, Xingliang Li, Jaclyn Barnes, Chung-Hyuk Park, and Myounghoon Jeon. Facial expression recognition for children: Can existing methods tuned for adults be adopted for children? In *International Conference on Human-Computer Interaction*, pages 201–211. Springer, 2019.
- [87] Yuqian Zhou and Bertram E. Shi. Action unit selective feature maps in deep networks for facial expression recognition. In *2017 International Joint Conference on Neural Networks (IJCNN)*, pages 2031–2038. IEEE, 2017.
- [88] Karel Zuiderveld. *Contrast Limited Adaptive Histogram Equalization*, page 474–485. Academic Press Professional, Inc., USA, 1994.
















Appendix A

EMFACS

AU	FAC Name	Neutral	Apex
Anger			
4	Brow Lowerer		
5	Upper Lid Raiser		
7	Lid Tightener		
23	Lip Tightener		















Continues on next page

Continued from previous page

AU	FAC Name	Neutral	Apex
Disgust			
9	Nose Wrinkler		
15	Lip Corner Depressor		
16	Lower Lip Depressor		
Fear			
1	Inner Brow Raiser		
2	Outer Brow Raiser		
4	Brow Lowerer		
5	Upper Lid Raiser		
7	Lid Tightener		

Continues on next page

Continued from previous page

AU	FAC Name	Neutral	Apex
20	Lip Stretcher		
26	Jaw Drop		
Happiness			
6	Cheek Raiser		
12	Lip Corner Puller		
Sadness			
1	Inner Brow Raiser		
4	Brow Lowerer		
15	Lip Corner Depressor		

Continues on next page

Continued from previous page









AU	FAC Name	Neutral	Apex
Surprise			
1	Inner Brow Raiser		
2	Outer Brow Raiser		
5	Upper Lid Raiser		
26	Jaw Drop		

Table A.1: The combination of Action Units present in each basic emotion - anger, disgust, fear, happiness, sadness and surprise. Here we exhibit the face in its initial neutral state and then facial expression in its apex. Extracted from [27].