Centro de Informática U·F·P·E

Pós-Graduação em Ciência da Computação

Pedro Diamel Marrero Fernández

# FERAtt: New Architecture Learning for Facial Expression Characterization

Universidade Federal de Pernambuco
posgraduacao@cin.ufpe.br
http://cin.ufpe.br/~posgraduacao

Recife
2019

Pedro Diamel Marrero Fernández

# FERAtt: New Architecture Learning for Facial Expression Characterization

Tese de Doutorado apresentada ao Programa de Pós-graduação em Ciência da Computação do Centro de Informática da Universidade Federal de Pernambuco, como requisito parcial para obtenção do título de Doutor em Ciência da Computação.

**Área de Concentração**: Inteligência Computacional
**Orientador**: Prof. Dr. Tsang Ing Ren

Recife
2019

Catalogação na fonte
Bibliotecária Monick Raquel Silvestre da S. Portes, CRB4-1217

F363f Fernández, Pedro Diamel Marrero

FERAtt: new architecture learning for facial expression characterization / Pedro Diamel Marrero Fernández. – 2019.

117 f.: il., fig., tab.

Orientador: Tsang Ing Ren.

Tese (Doutorado) – Universidade Federal de Pernambuco. CIn, Ciência da Computação, Recife, 2019.

Inclui referências.

1. Inteligência computacional. 2. Expressões faciais. 3. Aprendizagem profunda. I. Ren, Tsang Ing (orientador). II. Título.

006.3 CDD (23. ed.) UFPE - CCEN 2020-10

Pedro Diamel Marrero Fernández

**“FERAtt: New Architecture Learning for Facial Expression Characterization”**

Tese de Doutorado apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de Pernambuco, como requisito parcial para a obtenção do título de Doutor em Ciência da Computação.

Aprovado em: 12/08/2019.

**Orientador: Prof. Dr. Tsang Ing Ren**

**BANCA EXAMINADORA**

Prof. Dr. George Darmiton da Cunha Cavalcanti
Centro de Informática/UFPE

Prof. Dr. Cleber Zanchettin
Centro de Informática/UFPE

Prof. Dr. Leandro Maciel Almeida
Centro de Informática/UFPE

Prof. Dr. Alceu de Souza Britto Jr
Centro de Ciências Exatas e de Tecnologia/PUCPR

Prof. Dr. Luiz Eduardo Soares de Oliveira
Departamento de Informática/UFPR

*I dedicate this dissertation to my daughter, Sofia.*

# ABSTRACT

Affective computing is a branch of artificial intelligence responsible for the development of equipment and systems capable of interpreting, recognizing and processing human emotions. The automatic understanding of human behavior is of great interest since it allows the creation of new human-machine interfaces. Within this behavior, facial expressions are the most convenient because of the wide range of emotions that can be transmitted. The human face conveys a large part of our emotional behavior. We use facial expressions to demonstrate our emotional states and to communicate our interactions. In addition, we express and read emotions through the expressions of faces without effort. However, automatic understanding of facial expressions is a task not yet solved from the computational point of view, especially in the presence of highly variable expression, artifacts, and poses. Currently, obtaining a semantic representation of expressions is a challenge for the affective computing community. This work promotes the field of facial expression recognition by providing new tools for the representation analysis of expression in static images. First, we present an analysis of the methods of extracting characteristics and methods of combining classifiers based on sparse representation applied to the facial expression recognition problem. We propose a system of multi-classifiers based on trainable combination rules for this problem. Second, we present a study of the main deep neural networks architectures applied in this problem. A comparative analysis allows to determine the best models of deep learning for the classification of facial expressions. Third, we propose a new supervised and semi-supervised representation approach based on metric learning. This type of approach allows us to obtain semantic representations of the facial expressions that are evaluated in this work. We propose a new loss function that generates Gaussian structures in the embedded space of facial expressions. Lastly, we propose FERAtt, a new end-to-end network architecture for facial expression recognition with an attention model. The FERAtt neuralnet focuses attention in the human face and uses a Gaussian space representation for expression recognition. We devise this architecture based on two fundamental complementary components: (1) facial image correction and attention and (2) facial expression representation and classification.

**Keywords**: Facial Expression. Emotion Recognition. Attention Models. Deep Learning. Metric Learning.

# RESUMO

Computação afetiva é um ramo da inteligência artificial responsável pelo desenvolvimento de equipamentos e sistemas capazes de interpretar, reconhecer e processar emoções humanas. A compreensão automática do comportamento humano é de grande interesse, já que permitiria a criação de novas interfaces homem-máquina. O rosto humano transmite uma grande parte do nosso comportamento emocional. Usamos expressões faciais para demonstrar emoções e para melhorar nossas interações sem esforço, devido a que as expressões são um reflexo incorporado a nosso mecanismo de comunicação. No entanto, a compreensão automática das expressões faciais é uma tarefa ainda não solucionada do ponto de vista computacional, especialmente na presença de expressão altamente variável, artefatos e poses. Atualmente, obter uma representação semântica de expressões faciais é um desafio para a comunidade de computação afetiva. Este trabalho promove o campo do reconhecimento da expressão facial, fornecendo novas ferramentas para a análise de expressão em imagens estáticas a partir do estudo da representação no espaço de características. Em primeiro lugar, apresentamos uma revisão dos principais métodos de extração de características e dos métodos de combinação de classificadores com base em representação escassa que são aplicadas aos problemas de reconhecimento de expressão facial. Propomos um sistema de multi-classificadores baseado em regras de combinação treináveis para a classificação das expressões faciais. Em segundo lugar, apresentamos um estudo das principais arquiteturas de redes neurais profundas aplicadas neste problema. Uma análise comparativa nos permite determinar os melhores modelos de aprendizagem profunda para a classificação das expressões. Em terceiro lugar, propomos uma nova abordagem supervisionada e semi-supervisionada de representação baseada na aprendizagem por métrica. Este tipo de abordagem nos permite obter representações semânticas das expressões faciais que são avaliadas neste trabalho. Propomos uma nova função de perda que geram estruturas Gaussianas no espaço de representação. Finalmente, propomos FERAtt, uma nova arquitetura de rede ponta-a-ponta para o reconhecimento de expressões faciais com um modelo de atenção. A rede FERAtt, concentra a atenção no rostro humano e usa uma representação do espaço Gaussiano para reconhecimento de expressão. Concebemos essa arquitetura com base em dois componentes fundamentais: (1) correção e atenção à imagem facial; e (2) representação e classificação da expressão facial.

**Palavras-chaves**: Expressões Faciais. Reconhecimento da Emoção. Aprendizagem Profunda. Modelos de Atenção. Aprendizado por Métrica.

# LIST OF FIGURES

## LIST OF TABLES

# LIST OF ABBREVIATIONS AND ACRONYMS

| | |
|---|---|
| **AAM** | Active Appearance Model |
| **ADs** | Action Descriptors |
| **AFER** | Automatic Facial Expressions Recognition |
| **AffectNet** | Affect from the InterNet |
| **AN** | Anger |
| **ANOVA** | Analysis of Variance |
| **ASD** | Autism Spectrum Disorder |
| **AU** | Action Units |
| **BPD** | Borderline Person-Ality Disorder |
| **BU-3DFE** | Binghamton University 3D Facial Expression |
| **CelebA** | CelebA |
| **CK+** | Extended Cohn-Kanade |
| **CNN** | Convolutional Neural Netword |
| **CO** | Contempt |
| **CycleGAN** | Cycle-Consistent Adversarial Networks |
| **DCNN** | Deep Convolutional Neural Network |
| **DFE** | Deep Feature Embedding |
| **DI** | Disgust |
| **DML** | Deep Metric Learning |
| **DMSL** | Deep Metric Structured Learning |
| **DP** | Decision Profile |
| **DSMMSE** | Deep Structure Metric MSE |
| **DSMP** | Deep Structure Metric Probability |
| **EM** | Expectation Maximization Algorithm |
| **EmotiW** | Emotion Recognition in the Wild |
| **FACS** | Facial Action Coding System |
| **FE** | Facial Expressions |
| **FER** | Facial Expressions Recognition |
| **FER+** | Extended Emotion FER |
| **FERAtt** | Facial Expression Recognition with Attention Net |

| | |
|---|---|
| **FFL** | Focal Features Learning |
| **FI** | Fuzzy Integral |
| **FR** | Fear |
| **GAN** | Generative Adversarial Network |
| **GEO** | Geometric Features |
| **GMM** | Gaussian Mixture Model |
| **GW** | Gabor Wavelets |
| **HA** | Happiness |
| **HCI** | Human-Computer Interaction |
| **HOG** | Histogram of Oriented Gradients |
| **ISA** | Independent Subspace Analysis |
| **JAFFE** | Japanese Female Facial Expression |
| **KNN** | K Nearest Neighbour |
| **LBP** | Local Binary Patterns |
| **LDA** | Linear Discriminant Analysis |
| **LOP** | Linear Opinion Pools |
| **LOSO** | Leave-One-Subject-Out |
| **LPP** | Locality Preserving Projections |
| **LPQ** | Local Phase Quantization |
| **MB** | Bayes Model |
| **MCS** | Multiple Classifier Systems |
| **MNIST** | Modified National Institute of Standards and Technology |
| **MRL** | Multiple Representation Learning |
| **MSE** | Mean Square Error |
| **MSVML** | SVM Linear Kernel Model |
| **MSVMP** | SVM Polynomial Kernel Model |
| **MV** | Majority Vote |
| **NB** | Naive Bayes Rule |
| **NE** | Neutral |
| **NMI** | Normalized Mutual Information |
| **PCA** | Principal Component Analysis |
| **PLS** | Partial Least Squares |
| **PSF** | Point Spread Function |

| | |
|---|---|
| **RAW** | Image Raw |
| **RBM** | Restricted Boltzmann Machines |
| **REC** | Recall Rule |
| **RMD** | Median Rule |
| **RMI** | Min Rule |
| **RMX** | Max Rule |
| **RP** | Product Rule |
| **RQ** | Research Questions |
| **RS** | Sum Rule |
| **SA** | Sadness |
| **SGD** | Stochastic Gradient Descent |
| **SLPP** | Semi Supervised Locality Preserving Projections |
| **SLR** | Systematic Literature Review |
| **SR** | Sparse Representation |
| **SRC** | Sparse Representation based Classifier |
| **SRFC** | Sparse Representation Fusion Classification |
| **ST** | Spatio Temporal |
| **STM** | Spatio Temporal Manifold |
| **SU** | Surprise |
| **SVM** | Support Vector Machine |
| **TCR** | Trainable Combination Rules |
| **TDLC** | Temporal Dynamics of Learning Center |
| **UI** | User Interface |
| **UMM** | Universal Manifold Model |
| **WMV** | Weighted Majority Vote Rule |
| **ZSL** | Zero Shot Learning |

## CONTENTS

# 1 INTRODUCTION

Computers are quickly becoming an ubiquitous part of our lives. We spend a great amount of time interacting with computers of one type or another. At the moment, the devices we use are indifferent to our affective states, since they are emotionally blind. However, successful human-human communication relies on the ability to read affective and emotional signals. Human-Computer Interaction (HCI) which does not consider the affective states of its users loses a large part of the information available in the interaction.

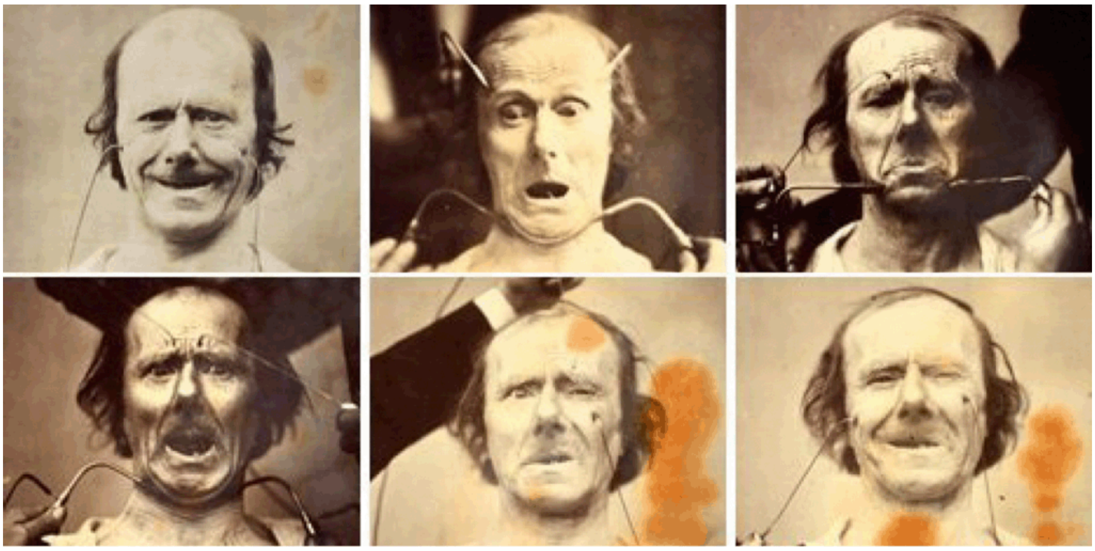
Automatic human behaviour understanding has attracted a great deal of interest over the past two decades, mainly because of its many applications spanning various fields such as psychology, computer technology, medicine and security. It can be regarded as the essence of next-generation computing systems as it plays a crucial role in affective computing technologies (i.e. proactive and affective user interfaces), learner-adaptive tutoring systems, patient-profiled personal well-being technologies, etc. (HIGHFIELD; WISEMAN; JENKINS, 2009).

Facial Expressions (FE) are vital signaling systems of affect, conveying cues on the emotional state of the person. Together with voice, language, hands and posture of the body, they form a fundamental communication system between humans in social contexts. Automatic Facial Expressions Recognition (AFER) is an interdisciplinary domain standing at the crossing of behavioral science, neurology, and artificial intelligence.

Studies of the face were greatly influenced in premodern times by popular theories of physiognomy and creationism. Physiognomy assumed that a person's character or personality could be judged by their outer appearance, especially the face (HIGHFIELD; WISEMAN; JENKINS, 2009). Leonardo Da Vinci was one of the first to refute such claims stating they were without scientific support (VINCI, 2002). In the 17th century in England, John Buwler studied human communication with particular interest in the sign language of persons with hearing impairment. Buwler's book Pathomyotomia or Dissection of the significant Muscles of the Affections of the Mind was the first consistent work in the English language on the muscular mechanism of FE (GREENBLATT et al., 1994). About two centuries later, influenced by creationism, Sir Charles Bell investigated FE as part of his research on sensory and motor control. He believed that FE was endowed by the Creator solely for human communication. Subsequently, Duchenne de Boulogne conducted systematic studies on how FEs are produced (DUCHENNE, 1990). He published beautiful pictures of sometimes strange FEs obtained by electrically stimulating facial muscles (see Figure 1). Approximately in the same historical period, Charles Darwin firmly placed FE in an evolutionary context (DARWIN, 1872). This marked the beginning of modern research of FEs. Important advancements were made through the works of researchers like Carroll Izard and Paul Ekman who inspired by Darwin performed seminal studies of

FEs (IZARD, 1971; EKMAN, 1971; EKMAN; OSTER, 1979).

Figure 1 – In the 19th century, Duchenne de Boulogne conducted experiments on how FEs are produced. In this experiments, the FEs obtained by electrically stimulating facial muscles for analysis of the response.



Source: (DUCHENNE, 1990)

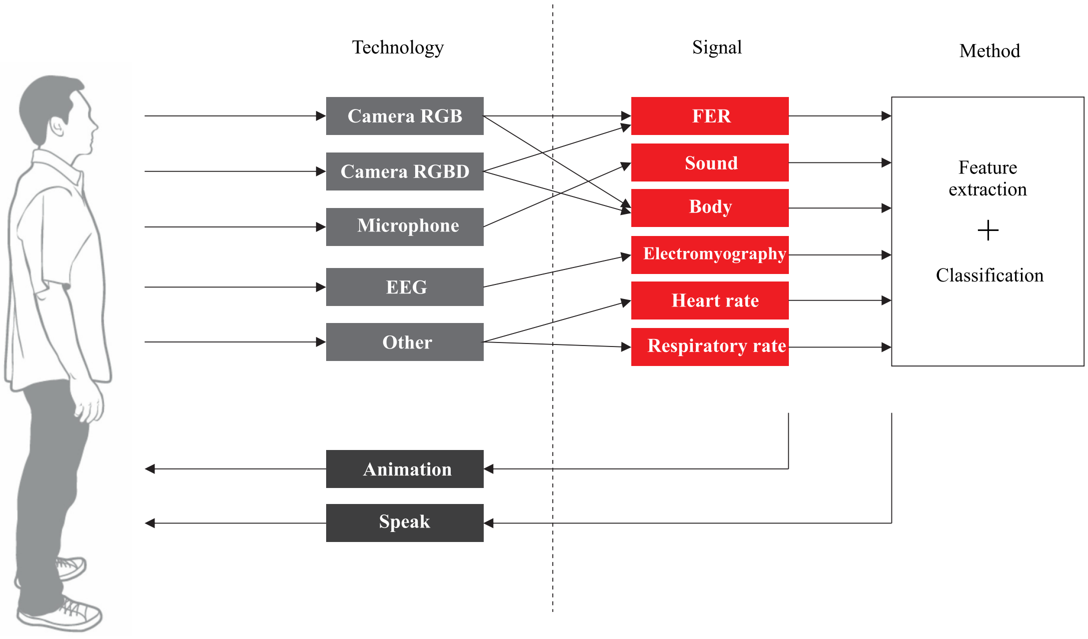
More recently researchers like Maja Pantic have had a great highlight in this area. Pantic published over 100 technical papers in the areas of machine analysis of facial expressions, machine analysis of human body gestures, audiovisual analysis of emotions and social signals, and human centred HCI (MARRERO-FERNÁNDEZ et al., 2014). Finally, in recent years successful commercial applications like Emotient[1] and Affectiva[2], perform large scale internet-based assessments of viewer reactions to ads and related material for predicting buying behaviour(CORNEANU et al., 2016).

Facial expression recognition applications are part of more complex systems known as Multimodal systems. These systems consist of several signals that have to be collected and represented in order to obtain a final estimate of the emotion (see Figure 2). In this direction, the representation of the signal is a fundamental aspect of the creation of more advanced emotion recognition systems. Obtaining facial expressions subspaces with known structures also could contribute to improving individual classification systems by providing more information on the decision boundaries of each class.

[1] www.emotient.com

[2] www.affectiva.com

Figure 2 – Multimodal subsystems for emotion recognition. For different sensors set of various technologies, a set of signals is obtained that will be processed and represented in of features space. The representations are combined to finally generate a classification of emotion.



Source: The author (2019)

## 1.1 PROBLEM

Despite the efforts made by the community of researchers in this area, the proposed systems are still far from the ability of human beings to FER. In (MARRERO-FERNÁNDEZ et al., 2014) we determine some of the current limitations of this area in form of research questions:

RQ1. Is it possible to obtain a efficient semantic representation of the facial expression?

RQ2. How can we easily obtain the fast, efficient and safest recovery from the database?

RQ3. Since the dynamics of known behavioural cues play a crucial role in human behaviour understanding, how continuous representation subspaces of the expression can be learned?

RQ1 addresses the problem of the representation spaces of the emotion. Present studies that address spaces of continuous representations (called dimensional) are limited and most of them impose a prior in the relationships that emotions must have in these spaces. Emotions have a high degree of subjectivity, which requires a trained staff to create such datasets and RQ2 refers to this problem. Finally, RQ3 deals with the need to incorporate temporal dynamics in the characterization of emotions. In this work, we intend to contribute to the answers of these research questions, specifically by providing a mechanism of semantic representation of the expressions that allows: to carry out the fusion of multimodal systems; the recovery and filtering of large volumes of facial expression images; and to be employed by dynamic facial expression recognition systems.

## 1.2 OBJECTIVE

This work contributes to the field of AFER and representation by dealing with these issues (RQ1-RQ3) which helps to brings emotionally aware computing more close to reality. The objective of this research is to develop a more accurate learning architecture for the facial expressions characterization. The specific objectives of the work are:

- Analyze different feature extraction methods and propose a system that combines different facial features.
- Analyze of the main deep neural network architectures applied to FER.
- Develop representation learning method for FE images based on Deep Metric Learning (DML).
- Develop a new deep neural network architecture inspired by the human attention mechanism and the continuous representation of the expression.

## 1.3 DOCUMENT STRUCTURE

Seven chapters are presented in this work. Figure 3 shows the structure of this thesis. Chapters inside the blue box represent articles that were published or submitted during the development of this thesis.

Chapter 2 explain the underlying emotion theories and possible areas of application. This chapter describes the datasets used in the experiments.

Chapter 3 provides a detailed explanation of the feature extraction methods for classification for FER via SR and Multiple Classifier Systems (MCS). We propose a MCS based on Trainable Combination Rules (TCR) for FER.

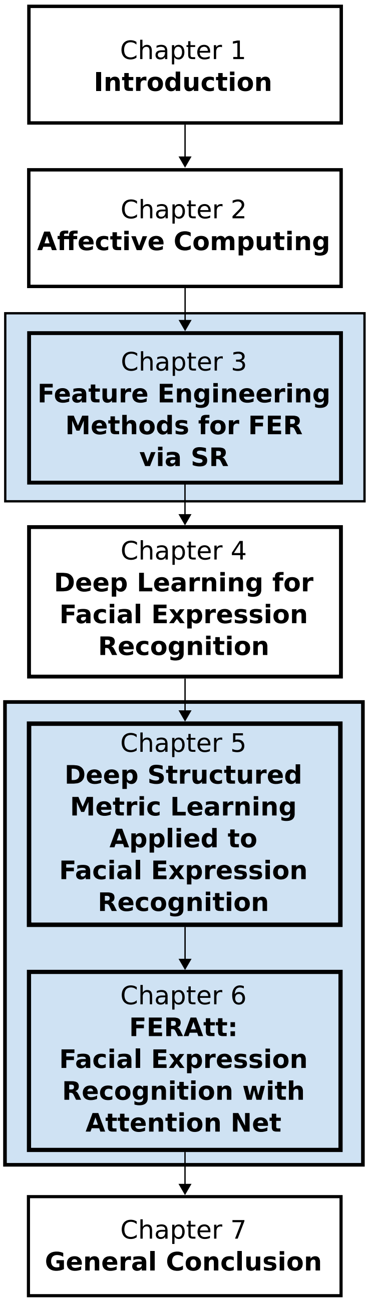
In Chapter 4 a study of the main architectures of Deep Neural Networks applied to FER is presented. A comparative analysis allows us to determine the best deep learning models for the classification of FE.

Chapter 5 proposes two new representation approach based on Deep Metric Learning (DML): 1) Structured Gaussian Manifold Learning and 2) Deep Gaussian Mixture Subspace Learning.

Chapter 6 presents a new end-to-end Deep Learning architecture with an Attention Model for Facial Expression Recognition.

Finally, Chapter 7 provides the conclusion of the thesis and outline the current limitations together with future works.

Figure 3 – Document Structure. Chapters inside the blue box represent articles that were published or submitted during the development of this thesis.


Chapter 1
**Introduction**
Chapter 2
**Affective Computing**
Chapter 3
**Feature Engineering Methods for FER via SR**
Chapter 4
**Deep Learning for Facial Expression Recognition**
Chapter 5
**Deep Structured Metric Learning Applied to Facial Expression Recognition**
Chapter 6
**FERAtt: Facial Expression Recognition with Attention Net**
Chapter 7
**General Conclusion**

Source: The author (2019)

# 2 AFFECTIVE COMPUTING

Affective computing was first popularised by Rosalind Picard's book *"Affective Computing"* which called for research into automatic sensing, detection and interpretation of affect and identified its possible uses in the HCI contexts (PICARD, 2000). Automatic emotion sensing has attracted a lot of interest from various fields and research groups, including psychology, cognitive sciences, linguistics, computer vision, speech analysis, and machine learning. The progress in automatic affect recognition depends on the progress in all of these seemingly disparate fields.

This chapter aims to provide an overview of the field with an emphasis on emotion sensing from FE. We described the datasets used for the experimentation, as well as some of the applications of the area.

## 2.1 THEORIES OF EMOTION

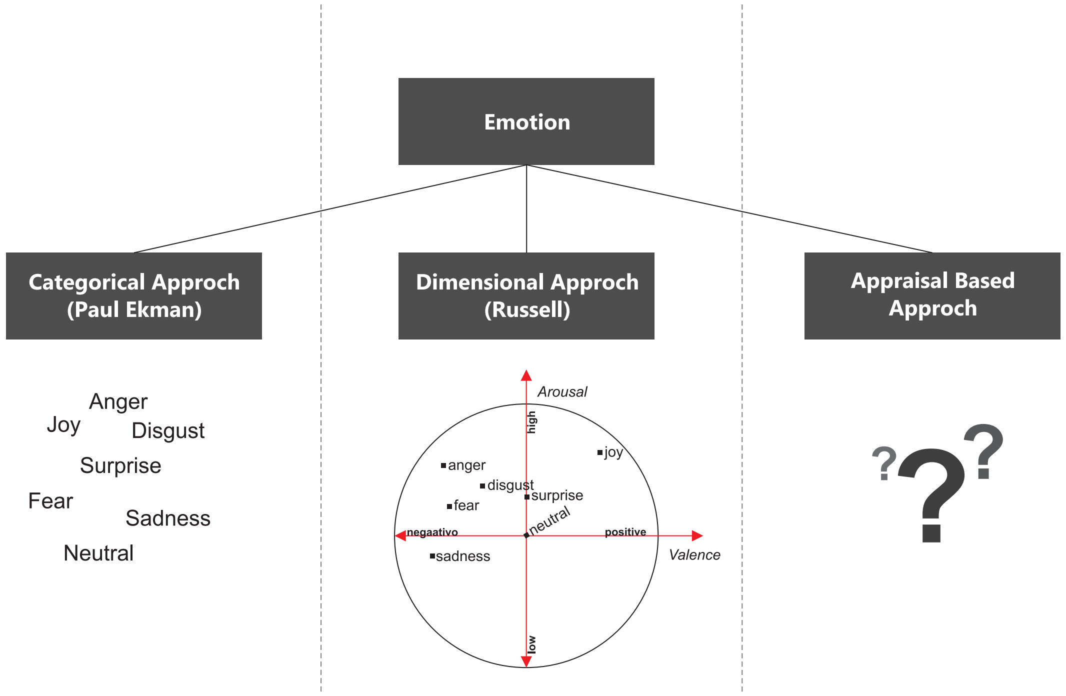
Before discussing on the automatic detection of emotion, we need to understand what emotion is. Unfortunately, psychologists themselves, have not reached a consensus on the definitions of emotion. The three most popular ways that emotion has been conceptualised in psychology research are: *discrete categories*, *dimensional representation*, and *appraisal-based* (see Figure 4). These theories are a good starting point to understanding affect for the purposes of automatic emotion recognition as they provide information about the ways emotion is expressed and interpreted.

### 2.1.1 Categorical

A popular way to describe emotion is in terms of discrete categories using the language from daily life (EKMAN; FRIESEN; ELLSWORTH, 1982). The most popular example of such categorization are the basic emotions proposed by Paul Ekman (EKMAN; FRIESEN; ELLSWORTH, 1982). These are: happiness, sadness, surprise, fear, anger, and disgust. Ekman suggests that they have evolved in the same way for all mankind and their recognition and expression is independent of education. This is supported by a number of cross-cultural studies performed by Ekman et al. (EKMAN; FRIESEN; ELLSWORTH, 1982), suggesting that the FEs of the basic emotions are perceived in the same way, regardless of culture.

The problem of using the basic emotions for automatic affect analysis is that they were never intended as an exhaustive list of possible affective states that a person can exhibit. What makes them basic is their universal expression and recognition, amongst other criteria (EKMAN; FRIESEN; ELLSWORTH, 1982). Finally, they are not the emotions that appear most often in everyday life (ROZIN; COHEN, 2003).

Figure 4 – Different approaches for emotion recognition. In the center, the 2D approach with dimensions Arousal and Valence.



Source: The author (2019)

Despite these shortcomings, basic emotions are very influential in automatic recognition of affect, as the majority of research has focused on detecting specifically these emotions, at least until recently (CORNEANU et al., 2016). However, there is a growing number of evidence that these emotions are not very suitable for the purposes of affective computing, as they do not appear very often in HCI scenarios (D'MELLO; CALVO, 2013).

There exist alternative categorical representations that include complex emotions. An example of such a categorization is the taxonomy developed by Baron-Cohen et al. (BARON-COHEN, 2004). It is a broad taxonomy including 24 groups of 412 different emotions. This taxonomy was created through a linguistic analysis of emotional terms in the English language. In addition to the basic emotions, it includes emotions such as boredom, confusion, interest, frustration, etc. The emotions belonging to some of these categories such as confusion, thinking and interest, seem to be much more common in everyday human-human and human-computer interactions (D'MELLO; CALVO, 2013; ROZIN; COHEN, 2003).

Baron-Cohen's taxonomy has been used by a number of researches in automatic recognition (KALIOUBY; ROBINSON, 2005; SOBOL-SHIKLER; ROBINSON, 2010) and in the description of affect (MAHMOUD et al., 2011). However, it is not nearly as popular as the basic emotion categories. Complex emotions might be a more suitable representation, however, they lack the same level of underlying psychological research when compared to the six basic emotions. Furthermore, little is understood about the universality and cultural specificity of the complex emotions, although there has been some work to suggest the universality of some of them (BARON-COHEN, 1996).

### 2.1.2 Dimensional Representation

Another way of describing affect is by using a dimensional representation (RUSSELL; MEHRABIAN, 1977), in which an affective state is characterised as a point in a multi-dimensional space and the axes represent a small number of affective dimensions. These dimensions attempt to account for similarities and differences in emotional experience (FONTAINE et al., 2007). Examples of such affective dimensions are: valence (pleasant vs. unpleasant); power (sense of control, dominance vs. submission); activation (relaxed vs. aroused); and expectancy (anticipation and appraisals of novelty and unpredictability). Fontaine et al. (FONTAINE et al., 2007) argue that these four dimensions account for most of the distinctions between everyday emotional experiences, and hence form a good set to analyse. Furthermore, there is some evidence of the cross-cultural generality of these dimensions (FONTAINE et al., 2007). FEs which could be associated with certain points in the emotional dimension space.

Dimensional representation allows for more flexibility when analysing emotions when compared to categorical representations. However, problems arise when one tries to use only several dimensions, since some emotions become indistinguishable when projecting high-dimensional emotional states onto lower dimension representations. For example, fear becomes indistinguishable from anger if only valence and activation are used. Furthermore, this representation is not intuitive and requires training in order to label expressive behaviour.

Affective computing researchers have started exploring the dimensional representation of emotion as well. It is often treated as a binary classification problem (GUNES; SCHULLER, 2013; SCHULLER et al., 2011) (active vs. passive, positive vs. negative etc.); or even as a four-class one (classification into quadrants of a 2D space). Treating it as a classification problem loses the added flexibility of this representation, hence there has been some recent work, treating it as a regression one (BALTRUŠAITIS; BANDA; ROBINSON, 2013; IMBRASAITĖ; BALTRUŠAITIS; ROBINSON, 2013; NICOLLE et al., 2012).

### 2.1.3 Appraisal based

The third approach for representing emotion, and very influential amongst psychologists, is the appraisal theory (JUSLIN; SCHERER, 2005). In this representation, an emotion is described through the appraisal of the situation that elicited the emotion, thus accounting for individual differences. Unfortunately this approach does not lend itself well for purposes of automatic affect recognition.

## 2.2 DATASETS

In this section, we analyze the main databases used in FE. The acquisition of tagged data of emotion is a problem, due to the subjectivity of the expressions. There are more than 40 datasets of static images of FE published in the internet.

In most cases, categorical representation for basic emotions is used: Neutral (NE), Happiness (HA), Surprise (SU), Sadness (SA), Anger (AN), Disgust (DI), Fear (FR), Contempt (CO). In this work, we selected two types of facial expression datasets: non-spontaneous and spontaneous. To create non-spontaneous data sets, images of actors that pose a particular expression are taken. These datasets are captured in controlled environments indoor and have few images. For the experiments on the non-spontaneous dataset, three of the most used datasets by the community were selected for this problem: Extended Cohn-Kanade (CK+) (LUCEY et al., 2010), Japanese Female Facial Expression (JAFFE) (LYONS et al., 1998) and Binghamton University 3D Facial Expression (BU-3DFE) (YIN et al., 2006). Spontaneous datasets captured from internet images were labeling for many non-expert peoples using crowdsourcing. There are few datasets of this type available. Two of the most important were selected: Extended Emotion FER (FER+)(BARSOUM et al., 2016a) and Affect from the InterNet (AffectNet)(PICARD, 2000).

Following the recommendation of (LEE et al., 2014), the images in CK+, JAFFE and BU-3DFE were cropped and selected based on eyes locations. The landmarks were obtained using OpenFace[1]. The cropped face image was rescaled to the size of 256×256 pixels.

### 2.2.1 CK+ dataset

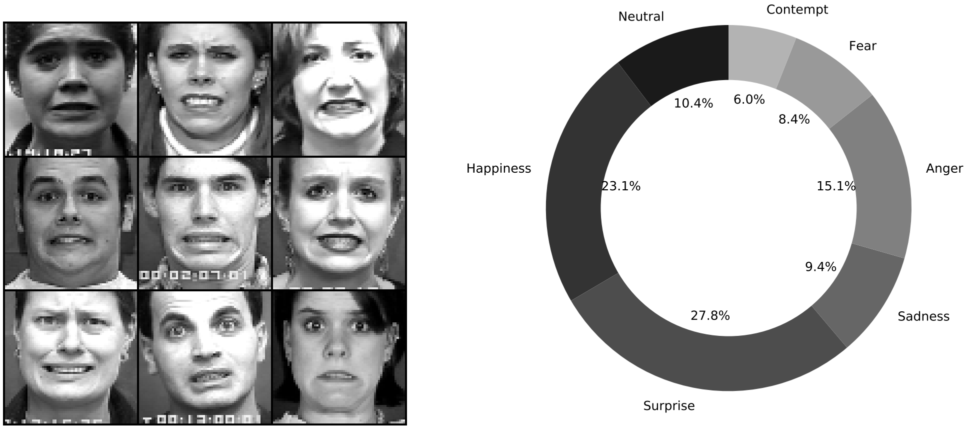
The CK+ dataset includes 593 image sequences from 123 subjects. From the 593 sequences, we selected 325 sequences of 118 subjects, which meet the criteria for one of the seven emotions (LUCEY et al., 2010). The selected 325 sequences consist of 45 AN, 18 CO, 59 DI, 25 FR, 69 HA, 28 SA and 83 SU (LUCEY et al., 2010). In the neutral face case, we selected the first frame of the sequence of 31 random selected subjects. Figure 5 shows examples of this dataset and class distribution.

### 2.2.2 BU-3DFE dataset

The BU-3DFE dataset has been known to be a challenging and difficult mainly due to a variety of ethnic/racial ancestries and expression intensity (YIN et al., 2006). We selected a total of 700 expressive face images (1 intensities × 6 expressions × 100 subjects) and 100 neutral face images (each of which is for one subject) (YIN et al., 2006). Figure 6 shows an example of different face expressions. The final selected 580 sequences consist of expressions of 90 NE, 89 AN, 92 DI, 86 FR, 89 HA, 85 SA and 49 SU.

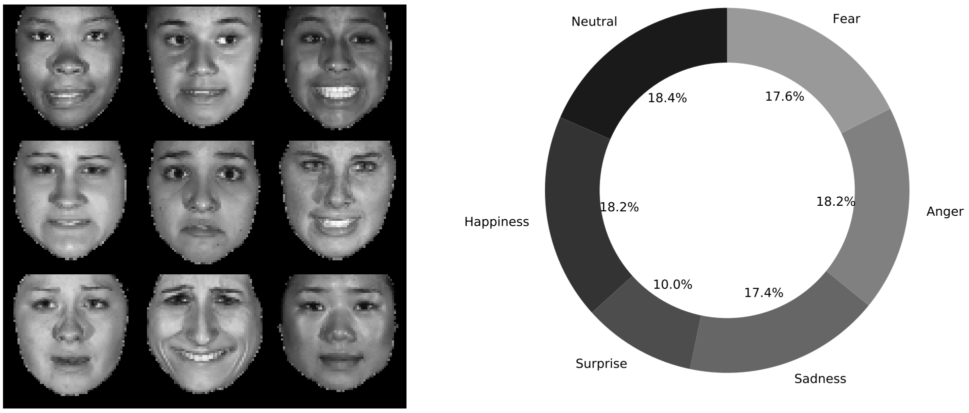
[1] https://github.com/TadasBaltrusaitis/OpenFace

Figure 5 – Examples of FE images from the CK+ database.



Source: The author (2019)

Figure 6 – Examples of FE images from the BU-3DFE database.



Source: The author (2019)

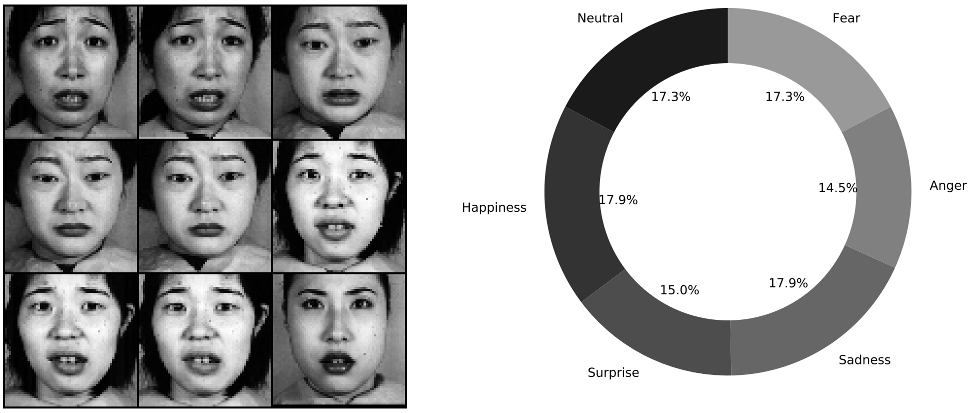
### 2.2.3 JAFFE dataset

The JAFFE dataset (LYONS et al., 1998) contains 10 female subjects and 213 images of FEs. Each image has a resolution of 256 × 256 pixels. The number of images corresponding to each of the 7 categories of expression (neutral, happiness, sadness, surprise, anger, disgust and fear) is almost the same. An example of these categories is presented in Figure 7. Each actor repeats the same expression several times (2,3 or 4 times). We selected 201 sequences consist of 30 NE, 25 AN, 28 DI, 30 FR, 31 HA, 31 SA and 26 SU expressions.

### 2.2.4 FER+ dataset

The FER dataset from the Kaggle Facial Expression Recognition Challenge, comprises 48-by-48-pixel grayscale images of human faces, each labeled with one of 7 emotion cate-

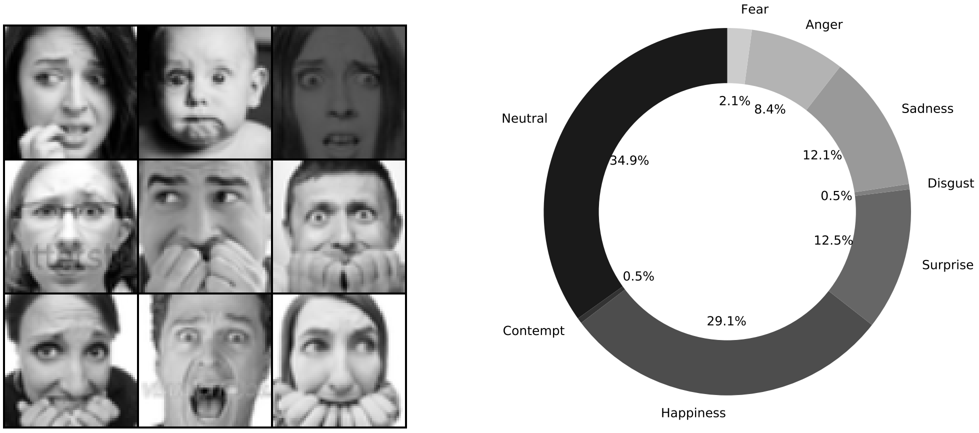
Figure 7 – Examples of FE images from the JAFFE database.



Source: The author (2019)

gories: anger, disgust, fear, happiness, sadness, surprise, and neutral. We used a training set of 28,709 examples, a validation set of 3,589 examples, and a test set of 3,589 examples. Figure 8 show examples of this dataset and class distribution across training data.

Figure 8 – Examples of FE images from the train FER+ database.



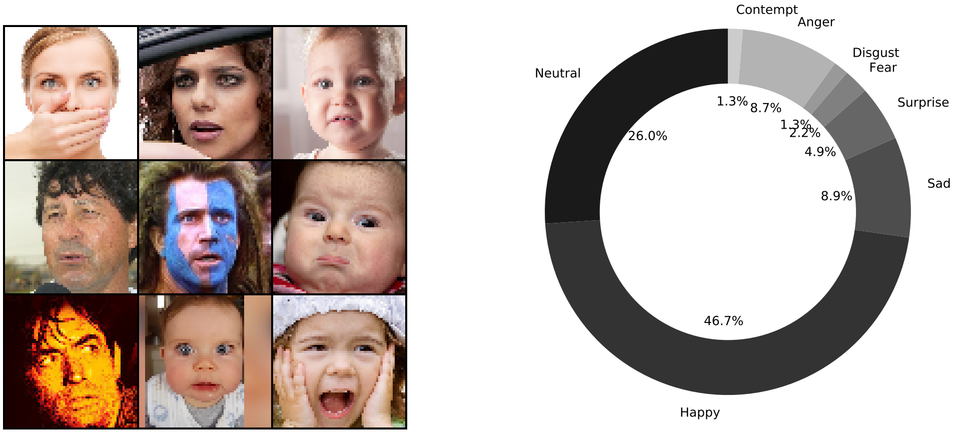
Source: The author (2019)

The FER+ annotations in (BARSOUM et al., 2016a) provide a set of new labels for the standard Emotion FER dataset. In FER+, each image has been labeled by 10 crowd-sourced taggers, which provide better quality ground truth for still image emotion than the original FER labels. Having 10 taggers for each image enables researchers to estimate an emotion probability distribution per face. This allows constructing algorithms that produce statistical distributions or multi-label outputs instead of the conventional single-label output, as described in (BARSOUM et al., 2016a). The distribution in this dataset consist of 8733 NE, 2098 AN, 116 DI, 536 FR, 7284 HA, 3022 SA, 3136 SU and CO 120 expressions.

### 2.2.5 AffectNet dataset

Affect from the InterNet (AffectNet) dataset contains more than one million images from the Internet that were obtained by querying different search engines using emotion-related tags. AffectNet is by far the largest database that provides facial expressions in two different emotion models (categorical model and dimensional model), of which 450000 images have manually annotated labels for eight basic expressions. The distribution in this dataset consist of 74874 NE, 134416 AN, 25459 DI, 14090 FR, 6378 HA, 3803 SA, 24882 SU and CO 3750 expressions.

Figure 9 – Examples of FE images from the train AffectNet database.



Source: The author (2019)

## 2.3 APPLICATIONS

There are a number of areas where the automatic detection and synthesis of affect would be beneficial. Some of the most prominent research areas in commercial and academic research using emotion recognition techniques are described below:

- **Control and Security:** Automatic tracking of attention, boredom and stress is be highly valuable in the safety of critical systems where the attentiveness of the operator is crucial. Examples of such systems are air traffic control, nuclear power plant surveillance, and operating a motor vehicle. An automated tracking tool can make these systems more secure and efficient, because early detection of negative affective states could alert the operator or others around him, thus helping to avoid accidents.

- **Consumer neuroscience and neuromarketing:** Tracking FEs can be leveraged to substantially enrich self-reports with quantified measures of more unconscious emotional responses towards a product or service. Based on FE analysis, products can be optimized, market segments can be assessed, and target audiences and personas can be identified.

- **Media testing & advertisement:** In media research, individual respondents or focus groups can be exposed to TV advertisements, trailers, and full-length pilots while monitoring their FEs. Identifying scenes where emotional responses (particularly smiles) were expected but the audience just did not "get it" is as crucial as to find the key frames that result in the most extreme FEs.

- **Psychological research:** Psychologists analyze FEs to identify how humans respond emotionally towards external and internal stimuli. In systematic studies, researchers can specifically vary stimulus properties (color, shape, duration of presentation) and social expectancies in order to evaluate how personality characteristics and individual learning histories impact FEs.

- **Clinical psychology and psychotherapy:** Clinical populations such as patients suffering from Autism Spectrum Disorder (ASD), depression or Borderline Person-Ality Disorder (BPD) are characterized by strong impairments in modulating, processing, and interpreting their own and others' FEs. Monitoring FEs while patients are exposed to emotionally arousing stimuli or social cues (faces of others, for example) can significantly boost the success of the underlying cognitive-behavioral therapy, both during the diagnostic as well as the intervention phase. An excellent example is the *"Smile Maze"* as developed by the Temporal Dynamics of Learning Center (TDLC) at UC San Diego. Here, autistic children train their FEs by playing a Pacman-like game where smiling steers the game character.

- **Medical applications & plastic surgery:** The effects of facial nerve paralysis can be devastating. The causes of this problem includes Bell's Palsy, tumors, trauma, diseases, and infections. Patients generally struggle with significant changes in their physical appearance, the ability to communicate, and to express emotions. FE analysis can be used to quantify the deterioration and evaluate the success of surgical interventions, occupational and physical therapy targeted towards reactivating the paralyzed muscle groups. Affect sensing systems could also be used to monitor patients in hospitals, or when medical staff are not readily available or overburdened. It could also be used in assisted living scenarios to monitor the patients and inform the medical staff during emergencies. There are some promising developments in medical applications of affective computing. One such development is the automatic detection of pain as proposed by Ashraf et al. (ASHRAF et al., 2009).

- **Software User Interface (UI) & website design:** Ideally, handling software and navigating websites should be a pleasant experience - frustration and confusion levels should certainly be kept as low as possible. Monitoring FEs while testers browse websites or software dialogs can provide insights into the emotional satisfaction of the desired target group. Whenever users encounter road blocks or get lost in complex sub-menus, you might certainly see increased *"negative"* FEs such as brow furrowing or frowning.

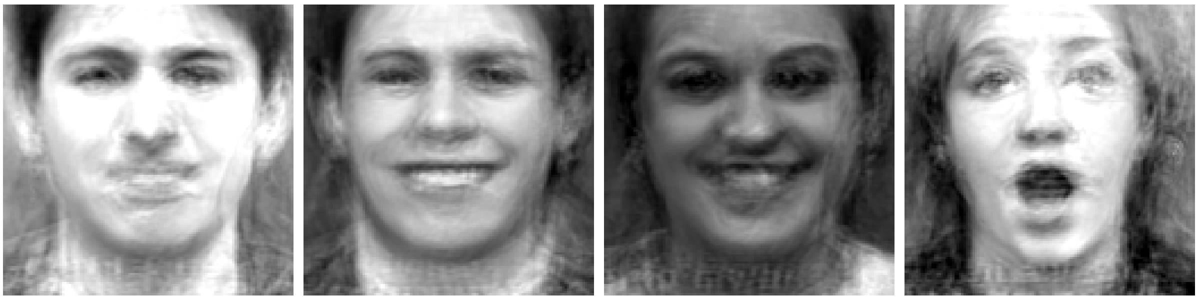
- **Engineering of artificial social agents (avatars):** Until recently, robots and avatars were programmed to respond to user commands based on keyboard and mouse input. Latest breakthroughs in hardware technology, computer vision, and machine learning have laid the foundation for artificial social agents, who are able to reliably detect and flexibly respond to emotional states of the human communication partner.

## 2.4 CONCLUSION

This chapter presented an overview of some essential concepts about the theoretical representation of the emotion. We describe the datasets used in this work and some of the applications that employ this type of technology. The next chapters describe the main results obtained in this work for the representation of the facial expression.

# 3 FEATURE ENGINEERING METHODS FOR FER VIA SPARSE REPRESENTATION[1]

Figure 10 – Examples of the new images generate with SR reconstruction. Signals can be rebuilt with SR, the result depends on the images used for creating the dictionary. These reconstruction signals are using in this work for generated new training samples.



Source: The author (2019)

The objective of this chapter is to propose a new facial expression recognition system for small datasets, that used different representations and combine them through a learned model. For this task we perform the following steps: 1) we obtain several representations of a facial expression applying different methods of representation; 2) we applied a new dataset regeneration method for augmenting the signals; 3) we carry each of the representations obtained to the same domain (the domain of representation errors of Sparse Representation (SR)); 4) and in this new space, we train a combination model to classify the signals. The Extended Cohn-Kanade (CK+) dataset, BU-3DFE dataset, and JAFFE dataset are used to validate the results. We compared 14 combination methods for 247 possible combinations of 8 different features spaces. As a result of this work, we obtained the best combination rule for this type of problem and a process that improves the results compared to other methods from the state of the art.

## 3.1 INTRODUCTION

The notion of Sparse Representation (SR)s, or finding sparse solutions to underdetermined systems, has found applications in a variety of scientific fields. The resulting sparse models are similar in nature to the network of neurons in V1, the first layer of the visual cortex in the human, and more generally, the mammalian brain (OLSHAUSEN;

[1] Pedro D. Marrero Fernandez, Antoni Jaume-i-Capó, Jose M. Buades-Rubio and Tsang Ing Ren; Centro de Informática, Universidade Federal de Pernambuco, Brazil; Computer Graphics, Vision and Artificial Intelligence Group (UGIVIA), Department of Mathematics and Computer Science, University of the Balearic Islands, 07122 Palma de Mallorca, Spain

FIELD, 1997; OLSHAUSEN; others, 1996). Patterns of light are represented by a series of innate or learned basis functions, whereby sparse linear combinations form surrogate input stimuli to the brain. Similarly, for many input signals of interest, such as natural images, a small number of exemplars can form a surrogate representation of a new test image.

In SR systems, new test images are efficiently represented by sparse linear coefficients on a dictionary $D$ of over-complete basis functions. Specifically, SR systems are comprised of an input sample $x \in \mathbb{R}^m$ along with a dictionary $D$ of $n$ samples, $D \in \mathbb{R}^{m \times n}$. SR solves for coefficients $\alpha \in \mathbb{R}^n$ that satisfy the $l_1$ minimization problem $x^\star = D\alpha$.

The advantages of exploiting sparsity in pattern classification have been extensively demonstrated for the FER problems (WRIGHT et al., 2009; WEIFENG; CAIFENG; YANJIANG, 2012; ZHANG; LI; ZHAO, 2012; PTUCHA; SAVAKIS, 2012). The experimental results of (WRIGHT et al., 2009) showed that the magnitude of the representation errors the facial feature vectors obtained for Sparse Representation (SR) is a good metric to classified the facial expressions.

Different techniques of emotion representation have been created and will be created in the next years. Representations can be features designed by experts (feature engineering methods)(WEIFENG; CAIFENG; YANJIANG, 2012; ZHANG; LI; ZHAO, 2012; PTUCHA; SAVAKIS, 2013; PTUCHA; SAVAKIS, 2012) or embedded vectors obtained training a deep network. The main objective of this chapter is to define a new facial expression recognizing system that use representation obtained from different sources and combine them through a learned model.

For this task we perform the following steps: 1) we obtain several representations of a facial expression applying different methods of representation; 2) we carry each of the representations obtained to the same domain, the domain of representation errors of SR; 3) in this new space, we train a combination model to classify the signals. Training models in small data sets is a challenge. We also propose the creation of new training signals using SR to increase the training data and thus increase the performance of the classification models. Figure 10 shows examples of the new images generated from the training images by our method. We used these images to train the FER system.

## 3.2 RELATED WORKS

Ying et. al. (YING; WANG; HUANG, 2010) used Local Binary Patterns (LBP) and Image Raw (RAW) to train two classifiers based on SRC. For each of these schemes (LBP+SRC and RAW+SRC), the approximation error signal is obtained for each class. This error is used as a fuzzy measure for evaluation of a decision rule. The residual ratio is calculated as the ratio between the second smallest residual and the smallest residual for LBP+SRC and RAW+SRC methods. If the results of two classifiers were not the same, the classification with the larger residual ratio is chosen. In Li et al. the classifiers are trained using Local Phase Quantization (LPQ) and GW+Addabost (LI; YING; YANG, 2014). Then, the

Adaboost algorithm is used to select the most effective 100 features from each Gabor filter. As in Ying et al. the classification result with the larger residual ratio is chosen if the classification of two classifiers is not the same (YING; WANG; HUANG, 2010).

Ouyang, Yan (OUYANG; SANG; HUANG, 2015) used Histogram of Oriented Gradients (HOG) and LBP. This approach is based on the idea that the two features are complementary because HOG mainly extracts contour-based shape while LBP primarily extracts the texture information of the images. The output of each classifier is used for evaluating a decision rule and they applied combination rules (KITTLER et al., 1998). The combination rules: Product Rule (RP) and the SR provide the best results. Several studies also employ dynamic features (PTUCHA; SAVAKIS, 2012; JI; IDRISSI, 2012; TSALAKANIDOU; MALASSIOTIS, 2010). In these works, the variability of facial changes is studied using the images or interest points in the face image.

## 3.3 METHODOLOGY

### 3.3.1 Feature Extraction Methods

In this work, we have grouped the feature methods in local, geometric and global. The local features will be defined from the extraction of supervised features on the facial patches. The facial patches are defined as a region in the face that are active during different FEs. It is reported that some facial patches are common during elicitation of all basic expressions, and some are confined to a single expression (ZHONG et al., 2012). The results indicate that these active patches are positioned below the eyes, in between the eyebrows, around the nose and mouth. To extract these patches from the face image, we have first to locate the facial components. In this work, the locations of active patches are defined with respect to the positions of landmarks that can be estimated using OpenFace[1] (BALTRUŠAITIS; ROBINSON; MORENCY, 2016). Happy and Routray (HAPPY; ROUTRAY, 2015) observed that the features from fewer facial patches can replace the high dimensional features without a significant decrease in the recognition accuracy. The Geometric Features (GEO) are defined from the distances and regions between the different landmarks (TSALAKANIDOU; MALASSIOTIS, 2010). The global methods employ unsupervised features obtained by the Deep Learning models pre-trained "VGG" and "VGGFace" over the entire image (PARKHI; VEDALDI; ZISSERMAN, 2015).

**Gabor Wavelets (GW).** Gabor filters have been successfully applied to facial expression recognition (ZHANG; LI; ZHAO, 2012; LI; YING; YANG, 2014). Gabor wavelets were introduced to image analysis because of its importance from the biological point of view since it has been shown to be able to model the properties of the cells in the receptive fields of the visual cortex of animals.

[1] https://github.com/TadasBaltrusaitis/OpenFace

A family of Gabor kernel is the product of a Gaussian envelope and a plane wave, defined as:

$$\Psi_{\mu,\nu}(z) = \frac{\|k_{\mu,\nu}\|^2}{\sigma^2} e^{-\frac{\|k_{\mu,\nu}\|^2\|z\|^2}{2\sigma^2}} \left[e^{ik_{\mu,\nu}z} - e^{-\frac{\sigma^2}{2}}\right] \tag{3.1}$$

where $z = (x, y)$ is the variable in the spatial domain, $k_{\mu,\nu}$ is the frequency vector $k_{\mu,\nu} = k_\nu e^{i\phi_\mu}$, $k_\nu = k_{max}/f^\nu$, $k_{max} = \pi/2$, $f = \sqrt(2)$ and $\phi_\mu = \pi\mu/8$, $\mu$ and $\nu$ are orientation and scale factors, respectively. Different kernels can be obtained by varying $\mu$ and $\nu$.

Given an image $I(z)$, the Gabor transformation at a particular position can be computed by a convolution with the Gabor Kernels using $G_{\mu,\nu} = I(z) \times \Psi_{\mu,\nu}(z)$. The magnitude of the resulting complex image is given by $|G| = \sqrt{Re(G)^2 + Im(G)^2}$. All features are obtained from $|G|$. The feature vector $F$ is defined as:

$$F_{k,l} = \sum_{i=x_l-k}^{x_l+k} \sum_{j=y_l-k}^{y_l+k} |G_{ij}|, \;\; l = 0, 1, \ldots, N, k = 1, 3, 5, 7, 9. \tag{3.2}$$

where $N$ is the number of the landmark point in the face image, $x_l$ and $y_l$ are the coordinates of the landmark point $l$ in $2k + 1$ neighborhood. In this work 68 landmark points were selected. For each point a patch of size $(2k + 1) \times (2k + 1)$ is used to compute the feature vector according to Equation 3.2. Five scale and eight orientation were used to calculate the Gabor Kernels and $k = 7$. This selection generate a vector of 2176 elements.

**Local Binary Patterns (LBP)**. LBP was widely used as a robust illumination invariant feature descriptor. This operator generates a binary number by comparing the neighbouring pixel values with the center pixel value (HUANG; WANG; YING, 2010). The pattern with 8 neighborhoods is given by

$$LBP(x, y) = \sum_{n=0}^{7} s(i_n - i_c) * 2^n \tag{3.3}$$

where $i_c$ is the pixel value at coordinate $(x, y)$ and $i_n$ are the pixel values at coordinates in the neighborhoods of $(x, y)$, and

$$s(x) = \begin{cases} 1 & x \geq 0 \\ 0 & x < 0 \end{cases} \tag{3.4}$$

The histograms of LBP image can be utilized as feature descriptors, given by:

$$H_i = \sum_{x,y}\{LBP(x, y) = i\}, \;\; i = 0, 1, \ldots, n - 1 \tag{3.5}$$

where $n$ is the number of labels produced by the LBP operator.

In addition, uniform LBP and rotation invariant uniform LBP (HUANG; WANG; YING, 2010) are also used in the experiment, and their performances are compared. Uniformity measure $(U)$ corresponds to the number of bitwise transitions from 0 to 1 or vice-versa in a pattern when the bit pattern is traversed circularly. For instance, the pattern $(00000001)_2$ and $(00100110)_2$ have $U$ values 2 and 4 respectively. The pattern is called uniform (LBPu2) when $U <= 2$. This reduces the length of the eight-neighborhood patterns to 59-bin histograms. The effect of rotation can be removed by assigning a unique identifier to each rotation invariant pattern, given by:

$$LBPriu2 = \begin{cases} \sum_{n=0}^{7} s(i_n - i_c) & if\ pattern\ is\ uniform \\ 9 & if\ otherwise \end{cases} \tag{3.6}$$

Thus, the rotational invariant uniform LBP with eight neighborhood produces 10 histogram bins.

**Histograms of Oriented Gradients (HOG)**. The basic idea of HOG features is that the local object appearance and shape can often be well characterized by the distribution of the local intensity gradients or edge directions, even without precise knowledge of the corresponding gradient or edge positions. The orientation analysis is robust to the change in illumination since the histogram gives translational invariance. We formulate a 2D HOG on the $XY$ plane. Given an image $I$, the horizontal and vertical derivatives $I_x$ and $I_y$ are obtained using the convolution operation. More specifically $I_x = I * K^T$ and $I_y = I * K$, where $K = [-1\ 0\ 1]^T$ . For each point of the image, its local gradient direction $\theta$ and gradient magnitude $m$ are computed as follows:

$$\theta = \arg(\nabla I) = \arctan(I_y/I_x), \tag{3.7}$$

$$M = |\nabla I| = \sqrt{I_x^2 + I_y^2}. \tag{3.8}$$

Let the quantization level for $\theta$ be $B$ and $\mathcal{B} = \{1, \cdots, B\}$, note that $\theta \in [-\pi, \pi]$. Thus a quantization function of $\theta$ is a mapping $Q : [-\pi, \pi] \to \mathcal{B}$. As a result, the HOG for a local 2D region (i.e a block) or a sequence of 2D regions (i.e a cuboid) $\mathcal{N}$ is a function $g : \mathcal{B} \to R$. More specifically, it is defined as

$$g(b) = \sum_{x \in \mathcal{N}} \delta(Q(\theta(x)), b) \cdot m(x), \tag{3.9}$$

where $m(x)$ is defined as the magnitude at the point $x$, $b \in \mathcal{B}$ and $\delta(i, j)$ is the Kronecker's delta function:

$$\delta(i, j) = \begin{cases} 1 & if\ i = j \\ 0 & if\ i \neq j \end{cases} \tag{3.10}$$

For HOG, each pixel within the block or cuboid has a weighted vote for a quantized orientation channel $b$ according to the response found in the gradient computation.

**Local Phase Quantization (LQP)**. In digital image processing, a blurred image can be obtained from the original image and the Point Spread Function (PSF), i.e. $G(u) =$

$F(u)H(u)$, where $G(u)$, $F(u)$, $H(u)$, are the Fourier transform of the blurred image, the original images and the PSF respectively. In LPQ, local M-by-M neighborhoods $N_y$ at each pixel position $x$ of the image $I(x)$ can examine the phase:

$$F(u,x) = \sum_{y \in N_x} f(x-y)e^{-j2\pi u^T y} \tag{3.11}$$

where, $u$ is frequency. In LPQ, four complex coefficients are $u_1 = [a,0]^T, u_2 = [0,a]^T, u_3 = [a,a]^T, u_4 = [a,-a]^T$, where $a$ is a small scalar that satisfies $H(u_i) > 0$, so each pixel position $F(x)$ is defined as $F(x) = [F(u_1,x), F(u_2,x), F(u_3,x), F(u_4,x)]$ and $G(x) = [ReF(x), ImF(x)]$. The phase of Fourier coefficient can be achieved by the symbols of each part of the real and imaginary to represent

$$q_j = \begin{cases} 1 & g_j \geq 0 \\ 0 & otherwise \end{cases} \tag{3.12}$$

where $g_j(x)$ is the *j-th* component of $G(x)$. The resulting eight binary coefficients $q_j(x)$ are represented as integer values using binary coding.

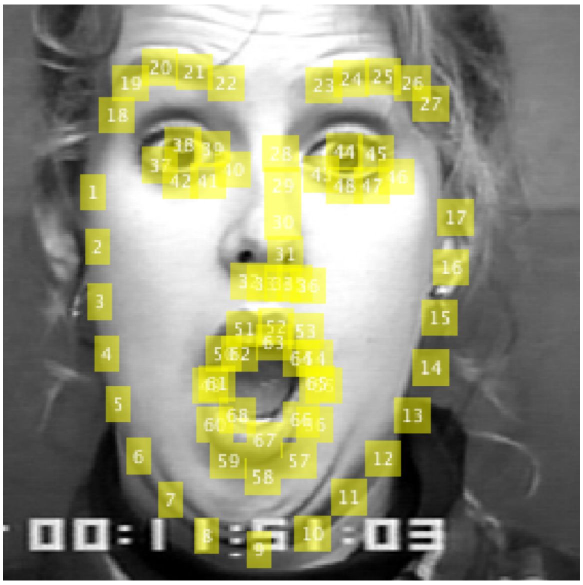
$$f_{LPQ}(x) = \sum_{j=1}^{8} q_j(x) 2^{j-1}. \tag{3.13}$$

**Geometric Features (GEO)**. Geometric measurements are computed using the 68 landmarks (TSALAKANIDOU; MALASSIOTIS, 2010). Figure 11 shows the selected order of the landmarks in this work. The order and the number of landmarks change in some of the works cited. We maintained the appropriate correspondence in all cases. Table 2 shows the 17 geometric distance measures used.

In this work, also used shape measures for the regions of the eyes, nose, and mouth. We define the regions from the landmarks in the region. For example, the region of the left eye is defined by landmarks 37, 38, 38, 40, 41 and 42. The measures are defined for each region as follows: Soliditys ($M_{18}$), returns a scalar specifying the proportion of the pixels in the convex hull that are also in the region, computed as $Area/ConvexArea$; Axes relationship ($M_{19}$), ratio of the distance between minor and major axis length of ellipse that has the same normalized second central moments as the region, computed as $AxisMin/AxisMax$; Circularity factor ($M_{20}$), computed as $4\pi Areas/Perms^2$; Eccentricity ($M_{21}$), scalar that specifies the eccentricity of the ellipse that has the same second-moments as the region; Extent ($M_{22}$), scalar that specifies the ratio of pixels in the region to pixels in the total bounding box. The distances of the centroids from each region to the center of the nose were also calculated.

**Pre-training deep models**. Recently, several deep learning algorithms have been proposed applied to FER (LIU et al., 2014a; LIU et al., 2014b; MOLLAHOSSEINI; CHAN; MAHOOR, 2016; FERNANDEZ et al., 2019). Our interest is to combine different representation

Figure 11 – The 68 landmarks used in this work. The order and the number of landmarks vary in some of the works cited. We maintained the appropriate correspondence in all cases.



Source: The author (2019)

spaces, which is why we select pre-trained models to obtain a representation of the facial expression. Training of deep learning architectures for FER is a problem for data sets such as CK+, JAFFE, BU-3DFE due to the small number of elements in these sets. Models trained in these datasets are overfitting, which could hinder the objective analysis of the contribution of characteristics of this model to our system. Therefore, we selected general classification models (object classification and facial classification models) pre-trained on extensive databases such as the models VGG-face (PARKHI; VEDALDI; ZISSERMAN, 2015) and VGG model (SIMONYAN; ZISSERMAN, 2014a). In VGGFace model case, as a feature, we have selected the output of the Relu layer 33. In the VGG model, we select the output of the Relu layer 34. The dimension of both vectors is 4096.

### 3.3.2 Classification via Sparse Representation

Consider a set of training signals $D = [D_1, D_2, \ldots, D_k] \in \mathbb{R}^{m \times p}$ from $k$ different classes, where the columns of each sub-matrix $D_j$ are signals from the class $w_j$. Ideally, giving sufficient training samples of class $w_j$, where $D_j = [d_1^j, d_2^j, \ldots, d_{n_j}^j] \in \mathbb{R}^{m \times n_j}$, a test signal $x \in \mathbb{R}^m$ that belongs to the same class can be approximated by a linear combination of the training samples from $D_j$, which can be written as:

$$x = \sum_{i=1}^{n_j} \alpha_i^j d_i^j, \tag{3.14}$$

Table 2 – Geometric facial measurements. $d_{ij}$: Euclidean distance between landmarks $i$ y $j$, $\alpha$: angle between two lines, $\xi_{ij}$: line defined by $i$ and $j$. $l_{ijk}$ length of the curve defined by $i$, $j$, $k$, $o_1$ and $o_2$: center of the right and left eye respectively.

| | | |
|---|---|---|
| $M_1$ | Inner eyebrow displacement | $d_{40,22}$, $d_{43,23}$ |
| $M_2$ | Outer eyebrow displacement | $d_{18,o_1}$, $d_{28,o_2}$ |
| $M_3$ | Inner eyebrow corners dist. | $d_{22,23}$ |
| $M_4$ | Eyebrow from nose root dist. | $d_{22,28}$, $d_{23,28}$ |
| $M_5$ | Eye opening | $d_{39,41}$, $d_{44,48}$ |
| $M_6$ | Eye shape | $d_{39,41}/d_{37,40}$, $d_{44,48}/d_{29,46}$ |
| $M_7$ | Nose length | $d_{29,31}$ |
| $M_8$ | Nose width | $d_{32,36}$ |
| $M_9$ | Lower lip boundary length | $l_{55,56,57,58,59,60,49}$ |
| $M_{10}$ | Mouth corners dist. | $d_{49,55}$ |
| $M_{11}$ | Mouth opening | $d_{63,67}$ |
| $M_{12}$ | Mouth shape | $d_{63,67}/d_{49,55}$ |
| $M_{13}$ | Nose–mouth corners angle | $\alpha(\xi_{32,49}, \xi_{36,55})$ |
| $M_{14}$ | Mouth corners to eye dist. | $d_{o_1,49}$, $d_{o_2,55}$ |
| $M_{15}$ | Mouth corners to nose dist. | $d_{49,34}$, $d_{55,34}$ |
| $M_{16}$ | Upper lip to nose dist. | $d_{52,34}$ |
| $M_{17}$ | Lower lip to nose dist. | $d_{67,34}$ |

Source: The author (2019)

that can be rewritten as:

$$x = D\delta_j(\alpha) \in \mathbb{R}^m \tag{3.15}$$

where $\delta_j(\alpha) = [0, \ldots 0, \alpha_1^j, \alpha_2^j, \ldots, \alpha_{n_j}^j, 0 \ldots, 0]^T \in \mathbb{R}^p$ is a vector of coefficient having most of the values equals to zero, except those associated with the class $w_j$. Since a valid test sample $x$ can be sufficiently represented only using the training samples from the same class, and this representation is the sparsest among all others, to find the identity of $x$ is equal to find the sparsest solution of Equation 3.15. This is the same as solving the following optimization problem ($l_0$-minimization):

$$\alpha^{\star} = \arg\min_{\alpha\in\mathbb{R}^p} \|\alpha\|_0 \;\; s.t \; D\alpha = x \tag{3.16}$$

However, solving the $l_0$-minimization of an undetermined system of linear equations is NP-hard. If the sought solution $\alpha^{\star}$ is sparse, the solution of the $l_0$-minimization problem, as defined in Equation 3.16, is equal to the solution of the following $l1$-minimization problem (DONOHO, 2006):

$$\alpha^\star = \arg\min_{\alpha \in \mathbb{R}^p} \|\alpha\|_1 \;\; s.t \; D\alpha = x \tag{3.17}$$

Algorithm 1 below summarizes the complete recognition procedure. The implementation minimizes the $l_1$-norm via a primal-dual algorithm for linear programming based in Sparse toolbox [2].

**Algorithm 1** Sparse Representation-based Classification (SRC). $D = [D_1, \ldots, D_k] \in \mathbb{R}^{m \times p}$ optional error tolerance, for $k$ classes, $x \in \mathbb{R}^m$ a test sample; $\epsilon$ optional error tolerance.

```
1: Normalize the columns of $D$ to have unit $l^2$-norm
2: Solve the $l^1$-minimization problem:
       $\alpha_1^\star \in \arg\min_{\alpha} \|\alpha\|_1 \;\; s.t. \; x = D\alpha$        (3.18)
   (Or alternatively, solve)
       $\alpha_1^\star \in \arg\min_{\alpha} \|\alpha\|_1 \;\; s.t. \; \|D\alpha - x\| \leq \epsilon$        (3.19)
3: Compute the residuals
4: for $i = 1, \ldots, k$ do
5:
       $r_i = \|x - D\delta_i(\alpha_1^\star)\|_2^2$        (3.20)
     where $\delta_i$ selects the entries of $\alpha^\star$ corresponding to the class $i$
6: end for
7: Return:
       $\hat{\imath} \in \arg\max_{i \in \{1,\ldots,k\}} r_i$        (3.21)
   the estimated class $\hat{\imath}$ for the signal $x$.
```

### 3.3.3 Multiclasification Systems

Classifier ensembles are successfully receiving great attention and accolade, not to mention the spawning wealth of research. Theoretical and empirical studies have demonstrated that an ensemble of classifiers is typically more accurate than a single classifier. Research on classifier ensembles permeate many strands machine learning including streaming data, concept drift and incremental learning (ELWELL; POLIKAR, 2011).

The parallel combining of classifiers is computed for different feature sets. This may be especially useful if the objects are represented by different feature sets, when they are described in different physical domains (e.g. sound and vision), or when they are processed by different types of analysis (e.g. moments and frequencies). The original set of features may also be split into subsets in order to reduce the dimensionality and hopefully the

[2] http://spams-devel.gforge.inria.fr/

accuracy of a single classifier. Parallel classifiers are often, but not necessarily, of the same type.

This subsection discusses ten combining methods based on proposals (KUNCHEVA; RODRIGUEZ, 2014; JACOBS, 1995; KITTLER et al., 1998). In (KUNCHEVA; RODRIGUEZ, 2014) be include a common probabilistic framework for the following four combination methods: Majority Vote (MV), Weighted Majority Vote Rule (WMV), Recall Rule (REC) and Naive Bayes Rule (NB). Each combiner is obtained from the previous one when a certain assumption is relaxed or dropped. Proposal (KITTLER et al., 1998) is provided with a theoretical underpinning of many existing classifier combination schemes for fusing the decisions of multiple experts, each employing a distinct pattern representation. It has been demonstrated that under different assumptions, and using different approximations we can derive the commonly used classifier combination schemes such as the Product Rule (RP), Sum Rule (RS), Min Rule (RMI), Max Rule (RMX) and Median Rule (RMD).

The outputs of the input classifiers can be regarded as a mapping to an intermediate space. A combining classifier applied on this space then makes a final decision for the class of a new object. In (JACOBS, 1995) one version of constrained regression for finding the weights that minimize the variance is derived by assuming the expert's errors in approximating the posterior probability.

**Probabilistic set-up**. Consider a set of classes $\Omega = \{w_1, \dots, w_c\}$ and a classifier ensemble of $L$ classifiers. Denote by $s_i$ the class label proposed by classifier $i$ ($s_i \in \Omega$). We are interested in the probability:

$$P(w_k \textit{ is the true class } |s_1, s_2, \dots, s_L), k = 1, \dots, c, \tag{3.22}$$

denoted for short $P(w_k|s)$, where $s = [s_1, s_2, \dots, s_L]^T$ is a label vector. Assume that the classifiers give their decisions independently, conditioned upon the class label which leads to the following decomposition:

$$P(w_k|s) = \frac{P(w_k)}{P(s)} \prod_i P(s_i|w_k). \tag{3.23}$$

Once a set of posterior probabilities $p_{ij}(x),\ i = 1, m;\ j = 1, c$ for $m$ classifiers and $c$ classes is computed for test object $x$, they have to be combined into a new set $\mu_j(x)$ that can be used, by maximum selection, for the final classification. We distinguish two sets of rules, hard combiners and soft combiners.

**The combining rules**. Let $x \in \mathbb{R}^n$ be a feature vector and $\{1, 2, \dots, c\}$ be the label set of $c$ classes. We call a classifier every mapping:

$$D : \mathbb{R}^n \longrightarrow [0, 1]^c - \mathbf{0}, \tag{3.24}$$

where $\mathbf{0} = [0, 0, \dots, 0]^T$ is the origin of $\mathbb{R}^c$. We call the output of $D$ a *"class label"* and denote it by $[\mu_D^1(x), \dots, \mu_D^c(x)]^T$, $\mu_D^i(X) \in [0, 1]$. The components $\mu_D^i(x)$ can be regarded as (estimates of) the posterior probabilities for the classes, give $x$, i.e, $\mu_D^i = P(i|x)$. Alternatively, $\mu_D^i(x)$ can be viewed as typicalness, belief, certainty, possibility, etc. Bezdek et al. (KELLER; KRISNAPURAM; PAL, 2005) define three types of classifiers:

1. Crisp classifier: $\mu_D^i(x) \in \{0, 1\}, \sum_{i=1}^{c} \mu_D^i(x) = 1, \forall x \in \mathbb{R}^n$;
2. Fuzzy classifier: $\mu_D^i(x) \in [0, 1], \sum_{i=1}^{c} \mu_D^i(x) = 1, \forall x$; (Probabilistic interpretation of the outputs fall in this category)
3. Possibilistic classifier: $\mu_D^i(x) \in [0, 1], \sum_{i=1}^{c} \mu_D^i(x) > 0, \forall x$;

The decision of $D$ can be *"hardened"* so that a crisp class label in $\{1, 2, \ldots, c\}$ is assigned to $x$. This is typically done by the maximum membership rule:

$$D(x) = k \Leftrightarrow \mu_D^k = \max_{i=1,\ldots,c} \mu_D^i(x). \tag{3.25}$$

Let $D_1, \ldots, D_L$ be the set of $L$ classifiers. We denote the output of the ith classifier as $D_i(x) = [d_{i,1}(x), \ldots; d_{i,c}(x)]^T$ , where $d_{i,j}(x)$ is the degree of *"support"* given by classifier $D_i$ to the hypothesis that $x$ comes from class $j$. We construct $\widehat{D}$, the fused output of the $L$ first-level classifiers as:

$$\widehat{D} = F(D_1(x), \ldots, D_L(x)), \tag{3.26}$$

where $F$ is called aggregation rule.

**The combining of hard classifiers**. In (KUNCHEVA; RODRIGUEZ, 2014) propose a common probabilistic framework for the following four combination methods: MV, WMV, REC and NB. It is shown a summary of each of the equations.

- Majority Vote:

$$\log(P(w_k|s)) \propto \log(\frac{1-p}{p(c-1)}) \log(P(w_k)) + |I_+^k| \tag{3.27}$$

where $|I_+^k|$ is the number of votes for $w_k$.

- Weighted Majority Vote Rule:

$$\log(P(w_k|s)) \propto \log(P(w_k)) + \sum_{i \in |I_+^k|} \theta_i + |I_+^k| \times \log(c-1) \tag{3.28}$$

where $\theta_i = \log(\frac{p_i}{1-p_i}), 0 < p_i < 1$

- Recall Rule:

$$\log(P(w_k|s)) \propto \log(P(w_k)) + \sum_i \log(1 - p_{ik}) + \sum_{i \in |I_+^k|} v_{ik} + |I_+^k| \times \log(c-1). \quad (3.29)$$

where $v_{ik} = \log(\frac{p_{ik}}{1-p_{ik}}), 0 < p_{ik} < 1$

- Naive Bayes Rule:

$$\log(P(w_k|s)) \propto \log(P(w_k)) + \sum_i \log(p_{i,s_i,k}) \quad (3.30)$$

**The combining of soft classifiers**. In (KITTLER et al., 1998) which provided a theoretical underpinning of many existing classifier combination schemes for fusing the decisions of multiple experts, each employed a distinct pattern representation. It has been demonstrated that under different assumptions and using different approximations we can derive the commonly used classifier combination schemes such as the RP, RS, RMI, RMX, RMD, and MV. It is shown a summary of each of the equations.

- Product Rule:

$$P^{(-(R-1))}(w_j) \prod_i P(w_j|x_i) = \max_k P^{(-(R-1))}(w_k) \prod_i P(w_k|x_i) \quad (3.31)$$

which under the assumption of equal priors, simplifies to the following:

$$\prod_i P(w_j|x_i) = \max_k \prod_i P(w_k|x_i) \quad (3.32)$$

- Sum Rule:

$$(1-R)P(w_j) + \sum_i P(w_j|x_i) = \max_k [(1-R)P(w_k) + \sum_i P(w_k|x_i)] \quad (3.33)$$

which under the assumption of equal priors simplifies to the following:

$$\sum_i P(w_j|x_i) = \max_k \sum_i P(w_k|x_i) \quad (3.34)$$

- Min Rule:

$$(1-R)P(w_j) + R \max_i P(w_j|x_i) = \max_k P^{(-(R-1))}(w_k) \min_i P(w_k|x_i) \quad (3.35)$$

which under the assumption of equal priors simplifies to the following:

$$\max_i P(w_j|x_i) = \max_k \min_i P(w_k|x_i) \quad (3.36)$$

- Median Rule:

$$P^{(-(R-1))}(w_j)\min_i P(w_j|x_i) = \max_k med_i P(w_k|x_i) \tag{3.37}$$

- Majority Vote:

$$\sum_i \Delta_{ji} = \max \sum_i \Delta_{ki} \tag{3.38}$$

where: $\Delta_{ki} = 1$ if $P(w_k|x_i) = \max_j P(w_j|x_i)$ or 0 in otherwise.

**Trainable combining of classifier**. Several trainable combined methods have been proposed in the literature but two fundamental approaches are highlighted, the Weighted Average and Fuzzy Integral (FI) (KUNCHEVA, 2004). The weights are set to express the quality of the classifiers. Accurate and robust classifiers should receive weights with larger value, such weight assignments may come from subjective estimates or theoretical set-ups. Jacobs proposes a version of constrained regression for finding the weights that minimize the variance (JACOBS, 1995). This method is derived by considering the expert's errors in approximating the posterior probability.

One way to set the weights is to fit a linear regression to the posterior probabilities. Take $d_{i,j}(x),\ i = 1,\dots,L$, to be estimates of the posterior probability $P(w_j|x)$.

Consider the largest regression model, where the whole decision profile is involved in approximating each posterior probability as in Equation 3.46. Given a data set $X = x_1,\dots,x_N$ with labels $y_1,\dots,y_N$, $y_j \in [0,1]$, formulate the optimization problem as looking for a weight vector $\theta$ which minimizes:

$$J_i(\theta_i) = \frac{1}{N}\sum_{j=1}^{N}\mathcal{L}(\mu_i(x_j), y_j, w_i, \theta_i) + R(\theta_i) \tag{3.39}$$

where $\mathcal{L}(\mu_i(z_j), y_j, w_i, \theta_i)$ is the loss incurred when labeling object $x_j \in X$, with true label $y_j$, as belonging to class $w_i$. $R(\theta_i)$ is a regularization term which serves to penalize very large weights.

The outputs of the input classifiers can be regarded as a mapping to an intermediate space. A combining classifier applied on this space then makes a final decision for the class of a new object. In (JACOBS, 1995) a version of constrained regression for finding the weights that minimize the variance is derived by assuming the expert's errors in approximating the posterior probability.

Linear Opinion Pools (LOP):

$$P(w_k|s) = \sum_i \theta_{ki} P(w_k|s_i) \tag{3.40}$$

$$J = \sum_i \sum_k \theta_i\theta_k\sigma_{ik} - \lambda(\sum_i \theta_i - 1) \tag{3.41}$$

the solution minimizing $J$ is:

$$\theta = \Sigma^{-1} I (I^T \Sigma^{-1} I)^{-1} \tag{3.42}$$

Fuzzy Integral: The philosophy of the fuzzy integral combiner is to measure the *"strength"* not only for each classifier alone but also for all the subsets of classifiers. Every subset of classifiers has a measure of strength that expresses how good this group of experts is for the given input $x$. The ensemble support for $w_j$, $\mu_j(x)$, is obtained from the support values $d_{i,j}(x)$, $i = 1, \dots,\ L$, by taking into account the competences of the groups of the various subsets of experts. The measure of strength of the subsets is called a fuzzy measure.

**Algorithm 2** Fuzzy Integral for classifier fusion

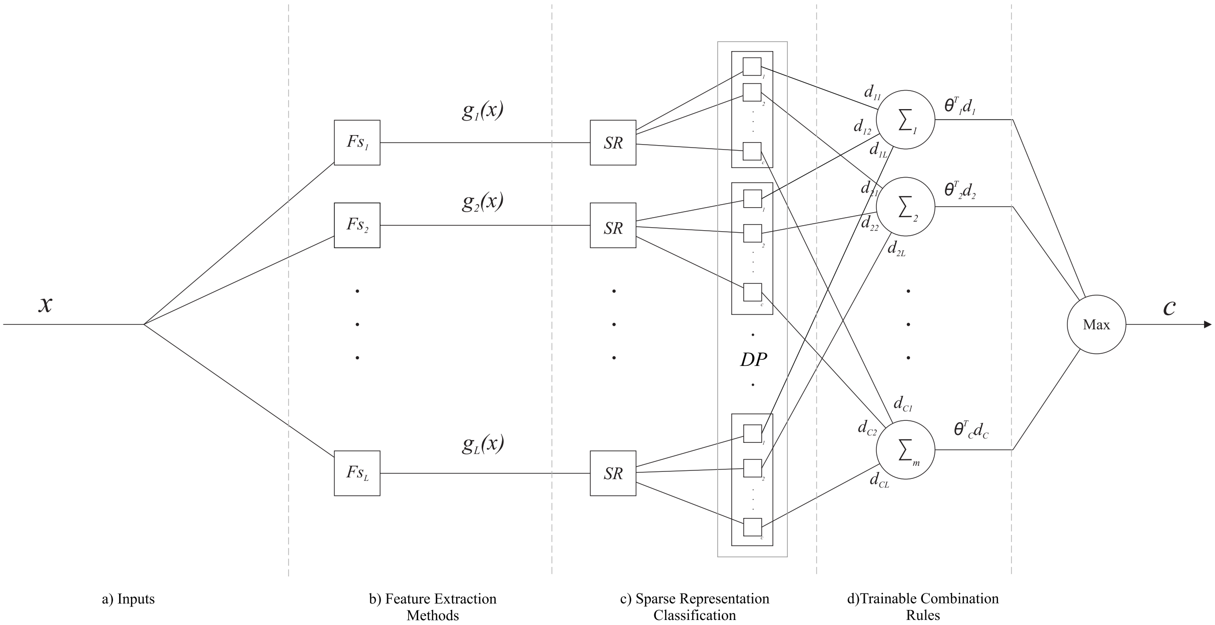
```
Fix the $L$ *fuzzy densities* $g^1, \dots, g^L$.
Calculate $\lambda > -1$ as the only real root greater than $-1$ of the equation
    $$\lambda + 1 = \prod_{i=1}^{L} (1 + \lambda g^i) \tag{3.43}$$
For a give $x$ sort the $k$th columns of $DP(x)$ to obtain $[d_{i_1,k}(x), d_{i_2,k}(x), \dots, d_{i_L,k}(x)]^T$,
    $d_{i_1,k}(x)$ being the highest degree of support, and $d_{i_L,k}(x)$, the lowest.
Arrange the fuzzy densities correspondingly, i.e., $g^{i_1}, \dots, g^{i_1}$ and set $g(1) = g^{i_1}$.
For $t = 2$ to $L$, calculate recursively
    $$g(t) = g^{i_t} + g(t-1) + \lambda g^{i_t} g(t-1) \tag{3.44}$$
Calculate the final degree of support for class $w_k$ by
    $$\mu_k(x) = \max_{t=1,\dots,L} \min(d_{i,k}(x), g(t)) \tag{3.45}$$
```

### 3.3.4 General Framework for Trainable Combination via SR

When different classifiers based in SR are used, the reconstruction error is used to evaluate the combination rules. This supposes that the probability of success of each classifier $D_i$, to each class, $P(w_j|D_i)$ is the same. Combination methods of this type are called *"class-conscious"* (KUNCHEVA; BEZDEK; DUIN, 2001).

Depending on the intrinsic characteristics, each expression can be best characterized in one of the subspaces, or in a particular subset. For example, expressions that involve some kind of facial movement, such as opening the mouth, may be better described by the shape spaces, which are recorded as changes in gradient. On the other hand, changes in the frequency intensity of the image may be characterized by a texture analysis methods. This property suggests that there are classifiers that are specialists (experts) for some classes. These classifiers should have greater weights in the final decision of the classification.

Figure 12 – Architecture proposal for the classification of FE using multiple classifiers. The output of classifiers for each class is merged using an estimated model.


x
Fs1
Fs2
FsL
g1(x)
g2(x)
gL(x)
SR
SR
SR
DP
d11
d12
d1L
d21
d22
d2L
dC1
dC2
dCL
Σ1
Σ2
Σm
θT1d1
θT2d2
θTCdC
Max
c
a) Inputs
b) Feature Extraction Methods
c) Sparse Representation Classification
d)Trainable Combination Rules

Source: The author (2019)

For the calculation of the weights, a new feature space based on the output of each of the classifiers is generated. The variable $d_{i,j}(x)$ denotes the support that classifier $D_i$ gives to the hypothesis that $x$ comes from class $w_j$. The larger the support, the more likely the class label belongs to $w_j$. In this approach, $d_{i,j}(x)$ are features in a new space, defined as the intermediate feature space (KUNCHEVA; BEZDEK; DUIN, 2001). The support for class $w_j$ is calculated as:

$$\mu_j(x) = \sum_{i=1}^{L} \theta_{i,j} d_{i,j}(x). \tag{3.46}$$

Linear regression is the commonly used procedure to derive the weights for this model (ERDOGAN; SEN, 2010). Algorithm 3 describes each step to classify a test pattern. Figure 12 shows the principal component of the system. Note that the output of each classifier $d_{i,j}$ , for the class $w_j$, creates a new feature space. For each output subspace, an estimated model $\theta_j$, weighs the decision of each classifier in the class $w_j$. For the experiments, the FI(KUNCHEVA; BEZDEK; DUIN, 2001), LOP(JACOBS, 1995), SVM and Naive Bayes method are used to adjust the classifiers output (Equation 3.50). Other techniques for combining expert exist, but they need to be trained with a largest number of objects (JORDAN; JACOBS, 1994).

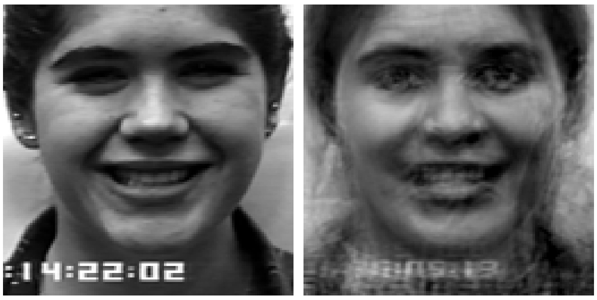
**Algorithm 3** Sparse Representation Fusion Classification SRFC. $D_1, ....D_L$ with $D_i \in \mathbb{R}^{n_i \times m}$; $g_1(x), ..., g_L(x)$ with $g_i(x) \in \mathbb{R}^n_i$; $g_i(x) \neq g_j(x)$

```
Calculate sparse representation:
    For $i = 1, ..., L$
        $\alpha_i^\star \in \arg\min_{\alpha} \|\alpha\|_0 \;\; s.t. \; g_i(x) = D_i\alpha;$    (3.47)
Calculate the vote of each representation to each class:
    For $i = 1, ..., C$ and $j = 1, ..., L$
        $r_{i,j} = \|g_i(x) - D_i\delta_j(\alpha_i^\star)\|_2^2,$    (3.48)
    where $\delta_j$ selects the entries of $\alpha^\star$ corresponding to the class $j$; $r_j$ represents the residual test sample $g_i(x)$ with the linear combination $D_i\delta_j(\alpha^\star)$. Is apply the softmax function about the inverse of the normalization of $r_{i,j}$ for obtain the vote to each class:
        $d_{i,j} = \sigma_{softmax}(1 - \frac{r_{i,j}}{\|r_{\cdot,j}\|_1})$    (3.49)
    where $r_{\cdot,j}$ is referent to the all column value and $d_{i,j}$ represent to the decision profile.
Trained combining rules:
        $\mu_j(x) = \sum_{i=1}^{L} \theta_{i,j} d_{i,j}(x),$    (3.50)
    the weights $\theta_j$ are estimated for decision profile for class $w_j$.
Classification:
        $\hat{J} \in \arg\max_{j \in \{1,...,C\}} \mu_j;$    (3.51)
Return: the estimated class $\hat{J}$ for the signal $x$.
```

### 3.3.5 Regenerate Training Dataset

We propose the creation of new training signals with SR to increase the training data and thus increase the performance of the classification models. Algorithm 4 shows the different steps to create the new dataset. The approximation of the signals obtained by SR contain differences with the original signal but in general, it maintains the high-level features. Figure 13 shows how a reconstruction of a raw signal preserves the facial expression however desirable differences are obtained such as the appearance of the eyes, mouth, etc. (the original image presents almost closed eyes the reconstruction presents open eyes).

Figure 13 – Example of a new training signal generated (right) from an original image (left) using our method. The new images retain high-level features such as facial expression, but some facial features such as the shape of the eyes, mouth, etc, are modified.



Source: The author (2019)

**Algorithm 4** Regenerate Facial Expression Dataset via SR. $S$: Training set; $N$: number of elements generate.

```
$S' \leftarrow \emptyset$
**for** $i = 1, ..., N$ **do**
    Select $x_i$:
    $x, y \sim S$
    Create dictionary $D$:
    Select $k$ random elements for each class in the set $S \cap x$:
    $D = \{x_1^1, x_2^1, \dots, x_k^1, \dots, x_1^c, x_2^c, \dots, x_k^c\}$
    with $x_j \in S \cap x$
    Calculate residuals $r$:
    $\alpha^\star \in \arg\min_\alpha \|\alpha\|_0 \;\; s.t. \; x \leftarrow D\alpha;$
    **for** $j = 1, \dots, C$ **do**
        $r_j \leftarrow \|x_i - D\delta_j(\alpha^\star)\|_2^2,$
    **end for**
    Add $S' \leftarrow r.$
**end for**
**return:** New training set $S'$
```

## 3.4 EXPERIMENTS

In this work, eight feature extraction methods: GW, LBP, HOG, LPQ, RAW, GEO, VGG and VGGF are used, according to the methodology proposed. These features have been used widely in the literature for FER (ZHANG; LI; ZHAO, 2012; LI; YING; YANG, 2014; LI et al., 2015b; YUAN; LIU; YAN, 2012; YING; WANG; HUANG, 2010; WEIFENG; CAIFENG; YANJIANG, 2012; OUYANG; SANG; HUANG, 2013; OUYANG; SANG; HUANG, 2015; DALAL; TRIGGS, 2005). We generated 247 possible scenarios which are all combinations of the selected feature extraction methods. We denote the classification schemes as: G**W**+SR with W, L**B**P+SR with B, **H**oG+SR with H, L**P**Q+SR with P, **R**AW+SR with R, **G**EO+SR with G, **V**GG+SR with V and VGG**F**+SR with F. Then, a possible combination scenarios can be denoted as W/B/H that corresponds to G**W**/L**B**P/**H**OG+SR. We tested 14 combination rules for each scenario: five soft-level rules: Product Rule (RP), Sum Rule (RS), Max Rule (RMX), Min Rule (RMI), Median Rule (RMD) (KITTLER et al., 1998), three hard-level rules: Weighted Majority Vote Rule (WMV), Recall Rule (REC), Naive Bayes Rule (NB)(KUNCHEVA; RODRIGUEZ, 2014) and the trainable methods: Fuzzy Integral (FI)(KUNCHEVA; BEZDEK; DUIN, 2001), Linear Opinion Pools (LOP)(JACOBS, 1995), Bayes Model (MB), SVM Linear Kernel Model (MSVML) and SVM Polynomial Kernel Model (MSVMP). The results are also compared to the individual methods GW+SRC, LBP+SRC, HOG+SRC, LPQ+SRC, RAW+SRC, GEO+SRC, VGG+SRC and VGGF+SRC.

One of the important aspects to consider when evaluating the use of multiple classifier systems is time. The critical component of the system is the features extraction methods ($Fs_i$) and its representation via SRC. As the number of methods increase, time is increased. It is expected that if the method $Fs_1$+SRC has time $t_1$, the method $Fs_2$+SRC has time $t_2, \ldots,$ the method $Fs_n$+SRC has time $t_n$, the system time is $t = \sum t_i$. However, the system components are independent and depend only on the input image (there are no interdependencies) so if there is not a very large number of methods in the pool, each can be assigned to a processing unit. In that case, the mean time could be $t = \max(t_i) + c$, where $c$ is a constant. Nowadays these solutions viable and more accessible.

To validate the proposed method, we performed four experiments. (1) Statistical analysis is employed to determine which combination rule presents the best classification results. We evaluated 247 combinations of feature extraction methods and 14 combination rules in three different datasets. (2) To investigate the generalization performance of the proposed method vs individual classification methods, we performed an inter-database experiment. (3) We analyze the influence of the feature methods for each class. (4) We compare the obtained results with other in the state of the art using the same dataset and experimental protocol. The results obtained from these experiments are described in detail in the next section.

### 3.4.1 Protocol

The parameters for each of the methods were selected from the results obtained in the state of art. The GW representation was obtained by using 5 scales and 8 orientations to construct a set of Gabor filter banks of 25 × 25 and $k = 7$ neighborhoods (ZAVASCHI et al., 2013). The image resolution is 256 × 256 pixels. For extracting LBP features used in Huang et al.(HUANG; WANG; YING, 2010) and Ying et al.(YING; WANG; HUANG, 2010), we adopted a uniform LBP operation with parameters of $P = 8$, $R = 2$ in images of size 256 × 256. The histogram is extracted for each patches of size 25 × 25. For HOG, the bin number is set to 9, the cell size is 16×16 pixels and block size of 2 × 2 in each selected landmark points(OUYANG; SANG; HUANG, 2015). For extracting the LPQ(ZHEN;

ZILU, 2012), we used $M = 5$ and $a = 1/5$. The histograms were extracted from each selected landmark points for patches of size $25 \times 25$. For VGGF case, as a feature we have selected the output of the Relu layer 33 and VGG case, we select the output of the Relu layer 34.

For this analysis different metrics are used. Accuracy is calculated as the average number of successes divided by the total number of observation (in this case each face is considered an observation). The measures precision, recall, F1-score and confusion matrix are also used in the analysis of the effectiveness of the system. Demšar (DEMŠAR, 2006) recommends the Friedman test followed by the pairwise Nemenyi test to compare multiple data. The Friedman test is a nonparametric alternative of the Analysis of Variance (ANOVA) test. The null hypothesis of the test $H_0$ is that all classifier models are equivalent. In this work, the Friedman test is used to identify the best classification scheme between different rules of combination and all features extraction methods. Similar to the methods in (HUANG et al., 2012; LI et al., 2013; LEE et al., 2014; PTUCHA; SAVAKIS, 2013), Leave-One-Subject-Out (LOSO) cross validation was adopted in the evaluation.

### 3.4.2 Statistical Analysis of the Combination Rule

Table 3, 4 and 5 show the accuracies, for 14 combining rules (columns) and the best combinations classifiers in different subspaces (rows) for CK+ dataset, BU-3DFE dataset and JAFFE dataset respectively. Regarding CK+ dataset (Table 3) it has been achieved an accuracy of more than 0.985, in BU-3DFE dataset (Table 4), accuracy of more than 0.821 and for JAFFE (Table 5) 0.776 (in all three cases more than 20 combinations have been selected). The best accuracies for each combination are shown underline. The last row of the tables shows the average accuracies across the 247 combinations. With the large span of classification accuracies, it is unlikely that these accuracies will be commensurable. But even though the average values across the features combining cannot serve as a valid performance gauge, they give a rough reference of the achievements of the combiners. The tables show that the trainable combining rules MSVMP and MSVML present the best results in most cases. The JAFFE database is the dataset with fewer images (213 images). The number of images limits the training phase of the trainable combinators. In this case the best option seems to be the use of non-trainable combinators.

To determine which one is the best for CK+ and BU-3DFE, we calculated the ranks of the combiners. For example, on the combined features R/W/H/P/V/F in the CK+ dataset, the order by rank is as follows: MSVMP (the best), MSVML, MB, RP, RS, RMD, MV, WMV, LOP, REC, NB, RMX, FI, RMI (the worst). In case of a tie, the rank are shared. The average ranks across the combination of features in CK+ (BU-3DFE) dataset were: RP 6.709(5.830), RS 6.749(5.873), RMX 11.699(12.694), RMI 12.956(12.484), RMD 6.749(5.873), MV 7.820(9.982), WMV 7.524(7.401), REC 7.512(6.328), NB 10.757(9.423), LOP 8.124(7.294), FI 11.682(12.611), MB 2.767(3.792), MSVMP 1.984(2.678), MSVML

Table 3 – Accuracy Selection Greater than 0.985 for 14 Rules Combining Methods and 247 Combining of 8 Subspaces in CK+ dataset.

| Features combining | Combining of Soft Classifiers | | | | | Combining of Hard Classifiers | | | | Trainable Classifiers | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | RP | RS | RMX | RMI | RMD | RM | WMV | REC | NB | LOP | FI | MB | MSVMP | MSVML |
| R/B/P/V | 0.986 | 0.986 | 0.950 | 0.913 | 0.986 | 0.975 | 0.972 | 0.969 | 0.958 | 0.978 | 0.950 | 0.978 | 0.975 | 0.975 |
| R/P/V/F | 0.961 | 0.961 | 0.936 | 0.908 | 0.961 | 0.944 | 0.947 | 0.958 | 0.939 | 0.955 | 0.936 | 0.980 | 0.989 | 0.983 |
| R/W/B/P/V | 0.975 | 0.975 | 0.947 | 0.930 | 0.975 | 0.972 | 0.964 | 0.950 | 0.950 | 0.975 | 0.947 | 0.986 | 0.978 | 0.978 |
| R/W/B/V/F | 0.969 | 0.969 | 0.947 | 0.933 | 0.969 | 0.964 | 0.964 | 0.961 | 0.955 | 0.961 | 0.947 | 0.986 | 0.983 | 0.980 |
| R/W/H/P/F | 0.969 | 0.969 | 0.947 | 0.941 | 0.969 | 0.958 | 0.955 | 0.961 | 0.947 | 0.964 | 0.947 | 0.989 | 0.983 | 0.983 |
| R/W/P/G/F | 0.947 | 0.947 | 0.902 | 0.894 | 0.947 | 0.953 | 0.953 | 0.961 | 0.941 | 0.925 | 0.902 | 0.972 | 0.986 | 0.986 |
| R/B/H/V/F | 0.969 | 0.969 | 0.953 | 0.933 | 0.969 | 0.966 | 0.966 | 0.953 | 0.950 | 0.966 | 0.953 | 0.975 | 0.980 | 0.986 |
| R/H/P/V/F | 0.964 | 0.964 | 0.944 | 0.916 | 0.964 | 0.950 | 0.955 | 0.953 | 0.927 | 0.953 | 0.944 | 0.983 | 0.986 | 0.986 |
| R/P/G/V/F | 0.939 | 0.939 | 0.883 | 0.880 | 0.939 | 0.933 | 0.941 | 0.961 | 0.908 | 0.930 | 0.883 | 0.975 | 0.989 | 0.980 |
| W/B/H/P/V | 0.975 | 0.975 | 0.955 | 0.927 | 0.975 | 0.964 | 0.964 | 0.953 | 0.950 | 0.964 | 0.955 | 0.986 | 0.983 | 0.983 |
| W/B/H/V/F | 0.969 | 0.969 | 0.955 | 0.930 | 0.969 | 0.961 | 0.958 | 0.947 | 0.936 | 0.958 | 0.955 | 0.978 | 0.986 | 0.983 |
| R/W/B/H/P/V | 0.978 | 0.978 | 0.947 | 0.933 | 0.978 | 0.975 | 0.972 | 0.955 | 0.955 | 0.972 | 0.947 | 0.989 | 0.986 | 0.983 |
| R/W/B/H/P/F | 0.975 | 0.975 | 0.953 | 0.944 | 0.975 | 0.978 | 0.969 | 0.953 | 0.964 | 0.975 | 0.953 | 0.986 | 0.983 | 0.983 |
| R/W/B/H/V/F | 0.975 | 0.975 | 0.947 | 0.939 | 0.975 | 0.964 | 0.961 | 0.950 | 0.950 | 0.961 | 0.947 | 0.989 | 0.980 | 0.986 |
| R/W/B/P/G/V | 0.958 | 0.958 | 0.908 | 0.888 | 0.958 | 0.972 | 0.966 | 0.950 | 0.941 | 0.941 | 0.908 | 0.986 | 0.983 | 0.983 |
| R/W/B/P/V/F | 0.975 | 0.975 | 0.947 | 0.927 | 0.975 | 0.975 | 0.969 | 0.961 | 0.961 | 0.961 | 0.947 | 0.989 | 0.980 | 0.980 |
| R/W/H/P/G/V | 0.958 | 0.958 | 0.897 | 0.885 | 0.958 | 0.955 | 0.953 | 0.950 | 0.919 | 0.933 | 0.897 | 0.980 | 0.980 | 0.986 |
| R/W/H/P/G/F | 0.955 | 0.955 | 0.905 | 0.894 | 0.955 | 0.958 | 0.955 | 0.953 | 0.941 | 0.933 | 0.905 | 0.978 | 0.986 | 0.986 |
| R/W/H/P/V/F | 0.966 | 0.966 | 0.936 | 0.927 | 0.966 | 0.966 | 0.955 | 0.953 | 0.939 | 0.955 | 0.936 | 0.986 | **0.992** | 0.989 |
| R/B/H/P/V/F | 0.972 | 0.972 | 0.953 | 0.925 | 0.972 | 0.975 | 0.969 | 0.955 | 0.950 | 0.966 | 0.953 | 0.989 | 0.986 | 0.986 |
| R/B/H/G/V/F | 0.961 | 0.961 | 0.899 | 0.888 | 0.961 | 0.964 | 0.964 | 0.953 | 0.947 | 0.933 | 0.899 | 0.975 | 0.986 | 0.986 |
| R/H/P/G/V/F | 0.944 | 0.947 | 0.894 | 0.883 | 0.947 | 0.958 | 0.955 | 0.947 | 0.902 | 0.927 | 0.894 | 0.978 | 0.989 | 0.986 |
| W/B/H/P/V/F | 0.972 | 0.972 | 0.955 | 0.925 | 0.972 | 0.972 | 0.966 | 0.955 | 0.939 | 0.953 | 0.955 | 0.983 | 0.980 | 0.986 |
| R/W/B/H/P/G/F | 0.966 | 0.964 | 0.911 | 0.894 | 0.964 | 0.966 | 0.961 | 0.953 | 0.964 | 0.925 | 0.911 | 0.986 | 0.983 | 0.983 |
| R/W/B/H/P/V/F | 0.972 | 0.972 | 0.947 | 0.933 | 0.972 | 0.966 | 0.969 | 0.955 | 0.955 | 0.958 | 0.947 | 0.989 | 0.980 | 0.983 |
| R/W/B/P/G/V/F | 0.958 | 0.958 | 0.908 | 0.888 | 0.958 | 0.966 | 0.958 | 0.955 | 0.958 | 0.933 | 0.908 | 0.986 | 0.983 | 0.983 |
| R/W/H/P/G/V/F | 0.950 | 0.950 | 0.897 | 0.885 | 0.950 | 0.953 | 0.947 | 0.953 | 0.927 | 0.930 | 0.897 | 0.980 | 0.989 | 0.989 |
| W/B/H/P/G/V/F | 0.955 | 0.955 | 0.891 | 0.880 | 0.955 | 0.961 | 0.961 | 0.953 | 0.939 | 0.930 | 0.891 | 0.983 | 0.986 | 0.986 |
| R/W/B/H/P/G/V/F | 0.964 | 0.961 | 0.908 | 0.891 | 0.961 | 0.966 | 0.966 | 0.955 | 0.955 | 0.930 | 0.958 | 0.986 | 0.989 | 0.989 |
| Average | 0.946 | 0.946 | 0.910 | 0.902 | 0.946 | 0.941 | 0.946 | 0.948 | 0.929 | 0.940 | 0.910 | 0.966 | 0.971 | 0.971 |

Source: The author (2019)

Table 4 – Accuracy Selection Greater than 0.821 for 14 Rules Combining Methods and 247 Combining of 8 Subspaces in BU-3DFE dataset.

| Features combining | Combining of Soft Classifiers | | | | | Combining of Hard Classifiers | | | | Trainable Classifiers | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | RP | RS | RMX | RMI | RMD | RM | WMV | REC | NB | LOP | FI | MB | MSVMP | MSVML |
| R/W/P/G | 0.774 | 0.778 | 0.724 | 0.697 | 0.778 | 0.772 | 0.783 | 0.781 | 0.771 | 0.771 | 0.724 | 0.807 | 0.824 | 0.826 |
| R/W/B/P/G | 0.807 | 0.807 | 0.741 | 0.709 | 0.807 | 0.781 | 0.798 | 0.802 | 0.779 | 0.788 | 0.741 | 0.807 | 0.828 | 0.826 |
| R/W/B/P/V | 0.802 | 0.800 | 0.752 | 0.767 | 0.800 | 0.790 | 0.797 | 0.795 | 0.778 | 0.797 | 0.752 | 0.822 | 0.816 | 0.817 |
| R/W/H/P/G | 0.788 | 0.791 | 0.729 | 0.714 | 0.791 | 0.805 | 0.805 | 0.809 | 0.779 | 0.784 | 0.729 | 0.809 | 0.826 | 0.824 |
| R/H/P/G/V | 0.784 | 0.784 | 0.724 | 0.714 | 0.784 | 0.779 | 0.791 | 0.790 | 0.778 | 0.776 | 0.724 | 0.803 | 0.824 | 0.819 |
| W/H/P/G/V | 0.778 | 0.778 | 0.719 | 0.698 | 0.778 | 0.779 | 0.788 | 0.800 | 0.784 | 0.781 | 0.719 | 0.795 | 0.822 | 0.819 |
| W/H/P/V/F | 0.795 | 0.798 | 0.734 | 0.748 | 0.798 | 0.798 | 0.793 | 0.797 | 0.781 | 0.783 | 0.734 | 0.822 | 0.814 | 0.810 |
| R/W/B/H/P/G | 0.809 | 0.807 | 0.743 | 0.712 | 0.807 | 0.802 | 0.800 | 0.819 | 0.783 | 0.798 | 0.743 | 0.819 | 0.828 | 0.822 |
| R/W/B/H/P/F | 0.817 | 0.816 | 0.745 | 0.752 | 0.816 | 0.810 | 0.807 | 0.817 | 0.795 | 0.809 | 0.745 | 0.821 | 0.822 | 0.824 |
| R/W/B/P/G/V | 0.798 | 0.800 | 0.741 | 0.721 | 0.800 | 0.788 | 0.795 | 0.795 | 0.783 | 0.783 | 0.741 | 0.814 | 0.826 | 0.826 |
| R/W/H/P/G/V | 0.800 | 0.800 | 0.728 | 0.721 | 0.800 | 0.798 | 0.809 | 0.814 | 0.791 | 0.790 | 0.728 | 0.816 | 0.826 | 0.824 |
| R/W/H/P/G/F | 0.803 | 0.803 | 0.728 | 0.722 | 0.803 | 0.798 | 0.807 | 0.828 | 0.784 | 0.793 | 0.728 | 0.816 | 0.812 | 0.807 |
| R/B/H/P/V/F | 0.807 | 0.807 | 0.764 | 0.767 | 0.807 | 0.803 | 0.810 | 0.814 | 0.802 | 0.800 | 0.764 | 0.822 | 0.810 | 0.810 |
| R/H/P/G/V/F | 0.810 | 0.809 | 0.722 | 0.724 | 0.809 | 0.786 | 0.803 | 0.812 | 0.779 | 0.776 | 0.722 | 0.817 | 0.822 | 0.824 |
| W/B/H/P/G/V | 0.800 | 0.798 | 0.741 | 0.707 | 0.798 | 0.795 | 0.795 | 0.807 | 0.790 | 0.784 | 0.741 | 0.809 | 0.822 | 0.821 |
| W/B/H/P/G/F | 0.803 | 0.803 | 0.726 | 0.710 | 0.803 | 0.805 | 0.807 | 0.819 | 0.790 | 0.790 | 0.726 | 0.812 | 0.822 | 0.822 |
| W/B/H/P/V/F | 0.805 | 0.803 | 0.750 | 0.759 | 0.803 | 0.805 | 0.809 | 0.809 | 0.790 | 0.786 | 0.750 | 0.828 | 0.814 | 0.814 |
| R/W/B/H/P/G/V | 0.802 | 0.800 | 0.743 | 0.722 | 0.800 | 0.791 | 0.802 | 0.816 | 0.790 | 0.786 | 0.790 | 0.821 | 0.822 | 0.819 |
| R/W/B/H/P/G/F | 0.821 | 0.821 | 0.741 | 0.722 | 0.821 | 0.816 | 0.819 | 0.812 | 0.795 | 0.798 | 0.791 | **0.828** | 0.826 | 0.822 |
| R/W/B/H/P/V/F | 0.802 | 0.803 | 0.750 | 0.760 | 0.803 | 0.812 | 0.816 | 0.817 | 0.798 | 0.795 | 0.791 | 0.824 | 0.816 | 0.816 |
| R/W/B/H/P/G/V/F | 0.814 | 0.814 | 0.741 | 0.729 | 0.814 | 0.807 | 0.809 | 0.822 | 0.793 | 0.790 | 0.790 | 0.824 | 0.814 | 0.816 |
| Average | 0.779 | 0.779 | 0.727 | 0.727 | 0.779 | 0.762 | 0.774 | 0.777 | 0.766 | 0.7740 | 0.729 | 0.789 | 0.794 | 0.794 |

Source: The author (2019)

Table 5 – Accuracy Selection Greater than 0.776 for 14 Rules Combining Methods and 247 Combining of 8 Subspaces in JAFFE dataset.

| Features combining | Combining of Soft Classifiers | | | | | Combining of Hard Classifiers | | | | Trainable Classifiers | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | RP | RS | RMX | RMI | RMD | RM | WMV | REC | NB | LOP | FI | MB | MSVMP | MSVML |
| R/B/H | 0.746 | 0.746 | 0.726 | 0.706 | 0.746 | 0.701 | 0.711 | 0.706 | 0.692 | 0.761 | 0.726 | 0.751 | 0.726 | 0.741 |
| B/H/P | 0.746 | 0.746 | 0.706 | 0.667 | 0.746 | 0.726 | 0.706 | 0.692 | 0.701 | 0.766 | 0.706 | 0.716 | 0.736 | 0.736 |
| R/W/H/G | 0.751 | 0.746 | 0.647 | 0.622 | 0.746 | 0.721 | 0.731 | 0.736 | 0.597 | 0.731 | 0.647 | 0.761 | 0.736 | 0.736 |
| R/H/P/G | 0.771 | 0.766 | 0.672 | 0.612 | 0.766 | 0.701 | 0.721 | 0.711 | 0.597 | 0.716 | 0.672 | 0.731 | 0.736 | 0.736 |
| R/H/P/F | 0.716 | 0.716 | 0.697 | 0.687 | 0.716 | 0.726 | 0.711 | 0.697 | 0.627 | 0.736 | 0.697 | 0.746 | **0.766** | 0.761 |
| W/B/H/P | 0.721 | 0.721 | 0.682 | 0.682 | 0.721 | 0.687 | 0.706 | 0.751 | 0.647 | 0.751 | 0.682 | 0.721 | 0.761 | 0.756 |
| W/H/P/F | 0.746 | 0.746 | 0.647 | 0.677 | 0.746 | 0.726 | 0.706 | 0.726 | 0.642 | 0.736 | 0.647 | 0.721 | 0.761 | 0.776 |
| R/W/H/P/G | 0.751 | 0.751 | 0.647 | 0.622 | 0.751 | 0.716 | 0.731 | 0.736 | 0.612 | 0.741 | 0.647 | 0.761 | 0.746 | 0.746 |
| R/W/H/G/F | 0.766 | 0.761 | 0.642 | 0.617 | 0.761 | 0.716 | 0.726 | 0.726 | 0.562 | 0.721 | 0.642 | 0.741 | 0.692 | 0.687 |
| R/B/H/G/F | 0.761 | 0.756 | 0.677 | 0.617 | 0.756 | 0.731 | 0.716 | 0.701 | 0.582 | 0.711 | 0.677 | 0.716 | 0.721 | 0.726 |
| R/H/P/G/F | 0.766 | 0.766 | 0.667 | 0.617 | 0.766 | 0.721 | 0.721 | 0.706 | 0.587 | 0.701 | 0.667 | 0.736 | 0.751 | 0.751 |
| W/H/P/V/F | 0.741 | 0.746 | 0.652 | 0.637 | 0.746 | 0.716 | 0.731 | 0.716 | 0.587 | 0.726 | 0.652 | 0.716 | 0.766 | 0.756 |
| R/W/B/H/P/F | 0.741 | 0.741 | 0.692 | 0.682 | 0.741 | 0.726 | 0.721 | 0.726 | 0.637 | 0.731 | 0.692 | 0.731 | 0.746 | 0.761 |
| R/W/B/H/G/V | 0.731 | 0.731 | 0.652 | 0.612 | 0.731 | 0.716 | 0.721 | 0.721 | 0.522 | 0.697 | 0.652 | 0.771 | 0.726 | 0.731 |
| R/W/B/H/G/F | 0.761 | 0.761 | 0.657 | 0.622 | 0.761 | 0.726 | 0.731 | 0.721 | 0.587 | 0.716 | 0.657 | 0.746 | 0.716 | 0.726 |
| R/W/B/H/V/F | 0.761 | 0.756 | 0.692 | 0.637 | 0.756 | 0.716 | 0.731 | 0.731 | 0.572 | 0.721 | 0.692 | 0.741 | 0.697 | 0.692 |
| R/W/H/P/G/F | 0.766 | 0.766 | 0.642 | 0.622 | 0.766 | 0.726 | 0.741 | 0.736 | 0.582 | 0.716 | 0.642 | 0.741 | 0.736 | 0.736 |
| R/B/H/P/G/F | 0.766 | 0.771 | 0.677 | 0.622 | 0.771 | 0.741 | 0.726 | 0.706 | 0.607 | 0.706 | 0.677 | 0.731 | 0.746 | 0.736 |
| R/B/H/P/V/F | 0.761 | 0.761 | 0.687 | 0.647 | 0.761 | 0.746 | 0.751 | 0.731 | 0.602 | 0.746 | 0.687 | 0.736 | 0.721 | 0.711 |
| R/W/B/H/P/G/F | 0.771 | 0.771 | 0.657 | 0.627 | 0.771 | 0.731 | 0.731 | 0.726 | 0.597 | 0.721 | 0.756 | 0.746 | 0.756 | 0.756 |
| R/W/B/H/P/V/F | 0.761 | 0.761 | 0.692 | 0.647 | 0.761 | 0.726 | 0.726 | 0.731 | 0.592 | 0.721 | 0.726 | 0.731 | 0.721 | 0.721 |
| R/B/H/P/G/V/F | 0.761 | 0.761 | 0.657 | 0.597 | 0.761 | 0.746 | 0.751 | 0.736 | 0.562 | 0.711 | 0.741 | 0.726 | 0.736 | 0.731 |
| Average | 0.700 | 0.700 | 0.640 | 0.604 | 0.700 | 0.685 | 0.700 | 0.697 | 0.586 | 0.691 | 0.643 | 0.693 | 0.685 | 0.686 |

Source: The author (2019)

1.970(2.739), showing that MSVMP, MSVML and MB are the best combiner in both cases.

The Friedman nonparametric ANOVA test was executed on the ranks, followed by a multiple comparisons test. The Friedman test is 2326.60 (2249.70), giving a $p$ value of approximately 0 (0) indicating significant differences among the ranks for CK+ (BU-3DFE) dataset. The Nemenyi post-hoc test and Bonferroni-Dunn post-hoc test were applied to obtain the methods that have significant differences.

The result for Nemenyi post-hoc test (two-tailed test), shows that there are significant differences between the MB, MSVMP and MSVML methods and all the others, for a significance level at $\alpha < 0.05$. For MB, MSVMP and MSVML was applied the Bonferroni-Dunn post-hoc test (one-tailed test) to strengthen the power of the test hypotheses. For a significance level of 0.05, the Bonferroni-Dunn post-hoc test did not show significant differences between the MB, MSVMP and MSVML methods. Therefore we can conclude that in general these methods have a similar behavior in this case. For BU-3DFE dataset case, similar results were obtained. The proposed methods (MB, MSVMP and MSVML) are significantly superior to the others to combination rules in FER problems. When there is very little training data, the use of non-trainable combination rules is suggested.

Table 6, 7 and 8 show the Accuracy, Precision, Recall and F1-score mesuremens of the best scheme for each individual and combination rules in CK+ dataset, BU-3DFE dataset and JAFFE dataset respectively. In CK+ dataset case (Table 6), the schema R/W/H/P/G/F+MSVMP shows a 0.992 accuracy. This schema also obtains the best results as Precision, Recall and F1-score with a 0.991, 0.985 and 0.999 respectively. In BU-3DFE dataset case (Table 7), the schema R/W/B/P/G+MSVMP shows a 0.828 accuracy. The best values of Precision, Recall and F1-score are obtained in this scheme for a 0.893, 0.833 and 0.965 respectively. In both cases the MSVMP combination rule gets the best results. In JAFFE dataset case (Table 8), the schema W/H/P/F+MSVML shows a 0.828 accuracy. However, the R/H/P/G+RP scheme shows the best results for Precision, Recall and F1 with 0.871, 0.809 and 0.959 respectively. This is because, as mentioned earlier, the trainable combination rules are not adequate when there is insufficient data in the training dataset.

This result exceeded the results of the state of the art for these types of features. In all cases, the combining rules increase accuracy of the single methods. Figure 14 shows that the classification error of the individual methods is greater than the classification error of combination schemes.

### 3.4.3 Multiple vs Individual Classification Methods

For the experiment, the datasets BU-3DFE and JAFFE were used to construct the training and test sets respectively. The training and test sets contained six expression classes, i.e., Anger, Disgust, Fear, Happiness, Sadness, and Surprise, which were common

Table 6 – The Best Schema for Each Combination Rules ($e * 1000^{-1}$) for CK+ dataset.

| | Combining Schemes | | | | | | | | | | | | | | Individual Schemes | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Measures | RP (R/B/P/V) | RS (R/B/P/V) | RMX (R/B) | RMI (W/B) | RMD (R/B/P/V) | RM (R/W/B/P) | WMV (R/B/H/F) | REC (R/B/P/V) | NB (R/B/H/F) | LOP (R/W/B) | FI (R/B) | MB (R/W/H/P/F) | MSVMP (R/W/H/P/C/F) | MSVML (R/W/H/P/C/F) | R | B | H | W | P | G | V | F |
| Accuracy | 0.986 | 0.986 | 0.964 | 0.975 | 0.986 | 0.980 | 0.975 | 0.969 | 0.966 | 0.978 | 0.964 | 0.989 | 0.992 | 0.989 | 0.953 | 0.947 | 0.930 | 0.947 | 0.897 | 0.718 | 0.754 | 0.852 |
| F1 | 0.992 | 0.992 | 0.968 | 0.982 | 0.992 | 0.987 | 0.981 | 0.975 | 0.973 | 0.982 | 0.968 | 0.989 | 0.991 | 0.989 | 0.961 | 0.966 | 0.951 | 0.964 | 0.940 | 0.774 | 0.798 | 0.898 |
| Precision | 0.985 | 0.985 | 0.945 | 0.968 | 0.985 | 0.977 | 0.967 | 0.957 | 0.953 | 0.968 | 0.945 | 0.980 | 0.985 | 0.980 | 0.936 | 0.942 | 0.918 | 0.938 | 0.900 | 0.674 | 0.699 | 0.835 |
| Recall | 0.998 | 0.998 | 0.994 | 0.996 | 0.998 | 0.997 | 0.996 | 0.995 | 0.995 | 0.997 | 0.994 | 0.998 | 0.999 | 0.998 | 0.993 | 0.992 | 0.990 | 0.992 | 0.985 | 0.936 | 0.948 | 0.976 |

Source: The author (2019)

Table 7 – The Best Schema for Each Combination Rules ($e * 1000^{-1}$) for BU-3DFE dataset.

| | Combining Schemes | | | | | | | | | | | | | | Individual Schemes | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Measures | RP (R/B/H/P/G/F) | RS (R/B/H/P/G/F) | RMX (H/P) | RMI (R/H/V) | RMD (R/B/H/P/G/F) | RM (R/W/B/H/P/G/F) | WMV (R/W/B/H/P/G/F) | REC (R/W/H/P/G/F) | NB (R/B/H/P/F) | LOP (R/W/B/H/P/F) | FI (R/W/B/H/P/G/F) | MB (W/B/H/P/V/F) | MSVMP (R/W/B/P/G) | MSVML (R/W/P/G) | R | B | H | W | P | G | V | F |
| Accuracy | 0.821 | 0.821 | 0.766 | 0.778 | 0.821 | 0.816 | 0.819 | 0.828 | 0.807 | 0.809 | 0.791 | 0.828 | 0.828 | 0.826 | 0.731 | 0.709 | 0.757 | 0.690 | 0.738 | 0.529 | 0.674 | 0.681 |
| F1 | 0.891 | 0.890 | 0.849 | 0.858 | 0.890 | 0.890 | 0.890 | 0.894 | 0.878 | 0.883 | 0.874 | 0.892 | 0.893 | 0.892 | 0.833 | 0.810 | 0.842 | 0.799 | 0.840 | 0.654 | 0.782 | 0.791 |
| Precision | 0.829 | 0.828 | 0.770 | 0.780 | 0.828 | 0.830 | 0.830 | 0.835 | 0.812 | 0.819 | 0.807 | 0.830 | 0.833 | 0.830 | 0.753 | 0.720 | 0.760 | 0.706 | 0.759 | 0.537 | 0.687 | 0.697 |
| Recall | 0.965 | 0.965 | 0.950 | 0.954 | 0.965 | 0.964 | 0.964 | 0.965 | 0.960 | 0.962 | 0.958 | 0.965 | 0.965 | 0.966 | 0.941 | 0.933 | 0.947 | 0.929 | 0.947 | 0.857 | 0.919 | 0.924 |

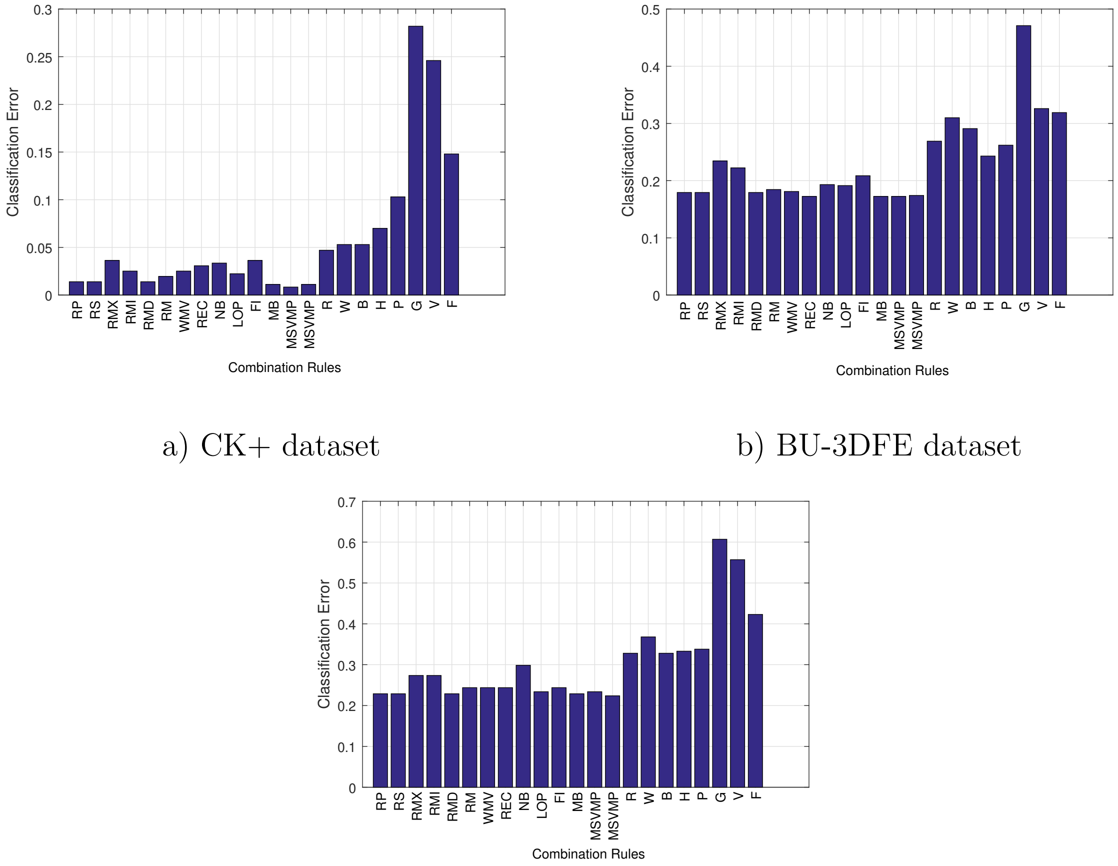
Source: The author (2019)

Table 8 – The Best Schema for Each Combination Rules ($e * 1000^{-1}$) for JAFFE dataset.

| | Combining Schemes | | | | | | | | | | | | | | Individual Schemes | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Measures | RP (R/H/P/G) | RS (R/B/H/P/G/F) | RMX (R/B) | RMI (R/H) | RMD (R/B/H/P/G/F) | RM (R/B/H/P/V) | WMV (R/B/H/G/V) | REC (R/W/H/P/G/V) | NB (B/H/P) | LOP (B/H/P) | FI (R/W/B/H/P/G/F) | MB (R/W/B/H/G/V) | MSVMP (R/H/P/F) | MSVML (W/H/P/F) | R | B | H | W | P | G | V | F |
| Accuracy | 0.771 | 0.771 | 0.726 | 0.726 | 0.771 | 0.756 | 0.756 | 0.756 | 0.701 | 0.766 | 0.756 | 0.771 | 0.766 | 0.776 | 0.672 | 0.672 | 0.667 | 0.632 | 0.662 | 0.393 | 0.443 | 0.577 |
| F1 | 0.871 | 0.870 | 0.838 | 0.830 | 0.870 | 0.867 | 0.850 | 0.848 | 0.803 | 0.859 | 0.860 | 0.856 | 0.844 | 0.859 | 0.787 | 0.785 | 0.787 | 0.733 | 0.793 | 0.541 | 0.578 | 0.701 |
| Precision | 0.809 | 0.807 | 0.764 | 0.750 | 0.807 | 0.804 | 0.776 | 0.768 | 0.707 | 0.788 | 0.794 | 0.779 | 0.762 | 0.784 | 0.689 | 0.694 | 0.695 | 0.629 | 0.699 | 0.416 | 0.462 | 0.583 |
| Recall | 0.959 | 0.958 | 0.944 | 0.942 | 0.958 | 0.957 | 0.950 | 0.951 | 0.934 | 0.955 | 0.953 | 0.954 | 0.950 | 0.956 | 0.927 | 0.924 | 0.926 | 0.898 | 0.932 | 0.804 | 0.809 | 0.889 |

Source: The author (2019)

Figure 14 – Classification error of the best schemes for each combination rule and the individual scheme.



a) CK+ dataset

b) BU-3DFE dataset

c) JAFFE dataset

Source: The author (2019)

in both BU-3DFE and JAFFE databases. Experimental result in Table 9 shows that the fusion method R/W/B/P/G+MSVMP achieves the 0.54 Accuracy and 0.721 F1-score, which is higher than those of the comparison methods. In general, combining the feature methods increases of the system performance. This supports the thesis that the trainable fusion methods are able to find experts for each subspace of features and can efficiently combine. In all cases, the combination rules are better than the individual methods. However, the value of metrics are lower than in the case where face images from a single dataset are used for both training set and test set. This performance degradation is mainly attributed to the fact that the face images are collected under two different controlled conditions. In order to generalize across image acquisition conditions, it is required to collect large training datasets with various image acquisition conditions (LITTLEWORT et al., 2006).

These evidences show the increased of the accuracy of the system regarding the individual systems. The best schemes (R/W/H/P/G/F+MSVMP and R/W/B/P/G+MSVMP)

Table 9 – Generalization Performance on the Two Different Datasets. BU-3DFE dataset and JAFFE dataset were used for training and testing, respectively. ALL, refers to all features (R/B/H/W/P/G/V/F).

| Method | Accuracy | F1-measure |
|---|---|---|
| ALL | 0.523 | 0.726 |
| R/H/P/F | 0.517 | 0.703 |
| R/W/B/H/P/G/F | 0.527 | 0.722 |
| R/W/B/P/G | **0.547** | **0.721** |
| R | 0.433 | 0.635 |
| B | 0.368 | 0.494 |
| H | 0.488 | 0.626 |
| W | 0.353 | 0.594 |
| P | 0.463 | 0.592 |
| G | 0.378 | 0.582 |
| V | 0.189 | 0.332 |
| F | 0.468 | 0.720 |

Source: The author (2019)

combine the best and the worst individual schemes. This increases the diversity of the system, giving the opportunity to select in each case the best subspaces of representation of each class.

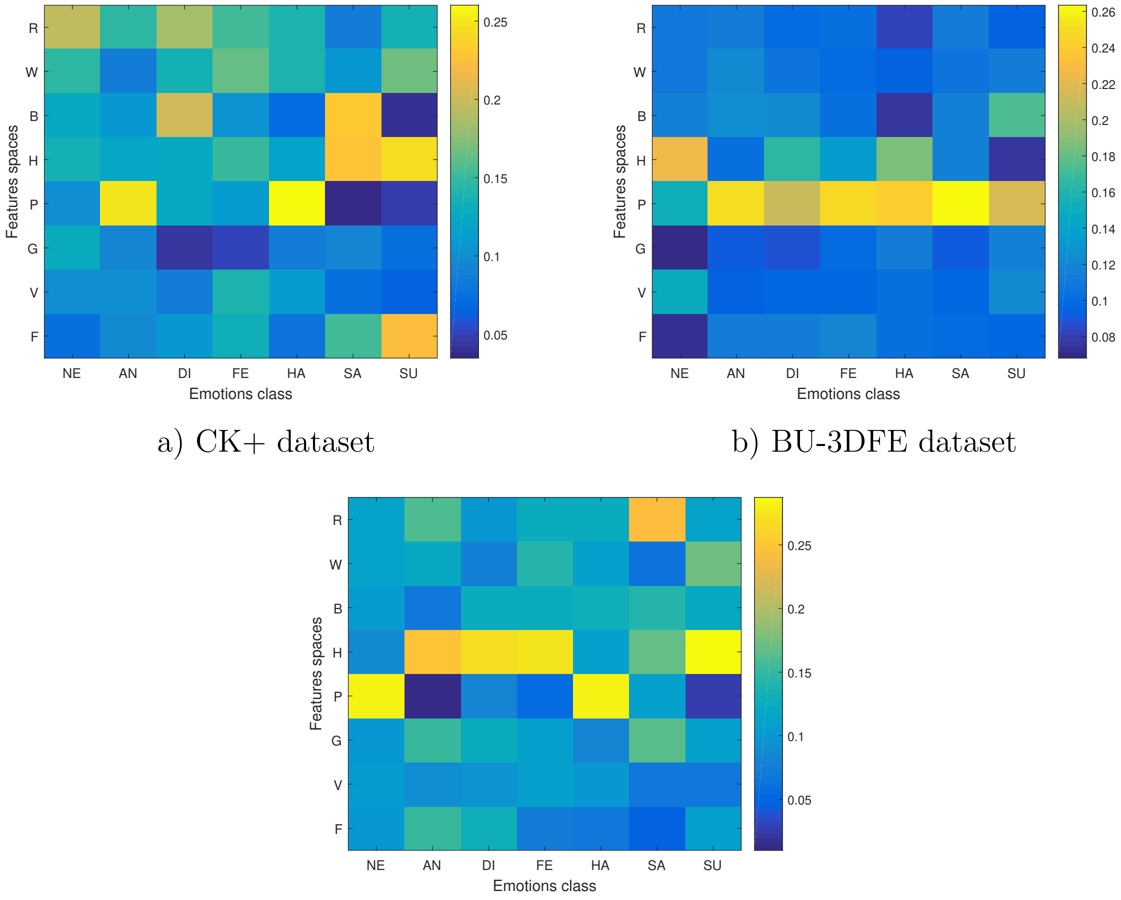
### 3.4.4 Analyze the Influence of the Feature Methods

The visualization of the learning models (subspaces generated), using classifiers like SVM and Naive Bayes method, is not trivial. To analyze the contribution of each feature to each class, a logistic regression model was trained on the decision profile. From the weights obtained for each class a weight map is created. The weight maps obtained for each datasets are shown in Figure 15. In the horizontal axis we can see the weight of a features space on each of the classes. In the vertical axis we can see how important each subspace of features is over a class.

As can be seen, the LPQ, HOG and RAW methods present high specialization values for some of the classes. For example, for CK+ dataset, LPQ has great decision power over the AN and HA classes. The feature space RAW specializes in the NE class and HOG specializes in SU and SA classes. In general we can observe that the spaces of characteristics do not always specialize on the same classes. This is because there are significant differences between the images in the different datasets.

We calculated the frequency of the feature methods from tables 3, 4 and 5. These frequencies are shown in Figure 16. It can be observed that LPQ, HOG and RAW methods are present in more than 80 percent of the selected combinations. This demostrates that

Figure 15 – Weight map learning for the combining of all features. NE: Neutral; AN: Anger; DI: Disgust; FE: Fear; HA: Happy; SA: Sadness; SU: Surprise



a) CK+ dataset

b) BU-3DFE dataset

c) JAFFE dataset

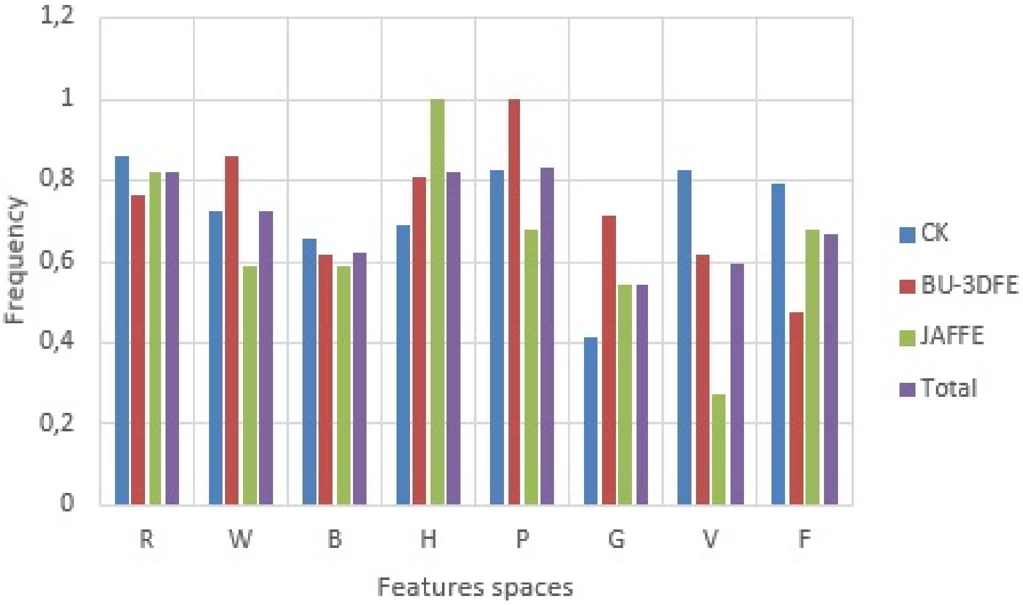
Source: The author (2019)

they have a significant weight in the classification of some emotions.

### 3.4.5 Our vs State of the Art

Table 10 shows the resulting comparisons between the different FER methods that use deep learning (BURKERT et al., 2015; LI et al., 2015a; JUNG et al., 2015; MENG et al., 2017; ZENG et al., 2018, 2018) on CK+ dataset. Although some results in Tables 10 cannot be directly compared due to different experimental setups, different expression classes and different preprocessing methods (e.g. face alignment), it is demonstrated that the proposed method can yield a feasible and promising recognition rate (around 99.2 percent) with static facial images under person-independent recognition scenario.

Figure 16 – Frequency of the feature methods select in the Table 3, 4 and 5 for CK+ dataset, BU-3DFE dataset and JAFFE dataset respectively. RAW: (R), GW: (W), LBP: (B), LPQ: (P), GEO: (G), VGGF: (F)



Source: The author (2019)

Table 10 – Comparisons with Deep Learning Technique in Expression Recognition on CK+ dataset. ACC: Accuracy, NE: class expression number. †: Six basic expressions + neutral class and contempt class. ‡: Six basic expressions (contempt is excluded). ∗: Six basic expressions + contempt class. LOSO: Leave-one-subject-out cross validation, L-X-SO: Leave-X-subjects-out cross validation.

| Methods | ACC(%) | NE | Validation |
|---|---|---|---|
| (KHORRAMI; PAINE; HUANG, 2015) | 98.30 | $6^{‡}$ | L-12-SO |
| (BURKERT et al., 2015) | 99.60 | $7^{*}$ | 10-fold |
| (LI et al., 2015a) | 83.00 | $7^{*}$ | LOSO |
| (JUNG et al., 2015) | 97.30 | $7^{*}$ | L-12-SO |
| (MENG et al., 2017) | 95.37 | $7^{*}$ | L-10-SO |
| (ZENG et al., 2018) | 95.79 | $7^{*}$ | L-10-SO |
| (YANG; CIFTCI; YIN, 2018) | 97.30 | $7^{*}$ | L-10-SO |
| (ZENG et al., 2018) | 89.84 | $8^{†}$ | L-10-SO |
| Our | **99.20** | $8^{†}$ | LOSO |

Source: The author (2019)

## 3.5 CONCLUSIONS

We show that the combination of classifiers improves the performance of individual classifiers. The proposed methods (MB, MSVMP, and MSVML) are significantly superior to the others combination rules in facial expression recognition problems. With the regeneration method proposed in this work, we obtained new training data. The new data retain high-level characteristics of the signal and allows training of the multi-classifier system.

# 4 DEEP LEARNING FOR FACIAL EXPRESSION RECOGNITION

Deep Convolutional Neural Network (DCNN) has recently yielded excellent performance in a wide variety of image classification tasks (KRIZHEVSKY; SUTSKEVER; HINTON, 2012; RUSSAKOVSKY et al., 2015; SZEGEDY et al., 2015; SIMONYAN; ZISSERMAN, 2014b). The careful design of local to global feature learning with convolution, pooling and layered architecture renders very strong visual representation ability, making it a powerful tool for FER. Research challenges such as Emotion Recognition in the Wild (EmotiW) and Kaggle's Facial Expression Recognition Challenge show the growing interest of the community in the use of this technique for the solution of this problem.

This chapter aims to explore the different solutions based on deep learning for FER. A description of the main problems found based on the experimental results is made. This chapter addressed the following research questions: *RQ1) What DCNN offers the best results for FER problem?*; *RQ2) What is the generalization capacity of the models in other domains?*; and *RQ3) What are the main problems presented by these DCNN architectures?*

In RQ1, we evaluate some architectures designed specifically for FER and some of the most powerful architectures on ImageNet[1]. In RQ2, we evaluate the results obtained on different datasets with different capture conditions. In RQ3, we analyze some problems that these architectures present.

## 4.1 RELATED WORKS

The deep learning models for FER and ER were reported in (KAHOU et al., 2013; TANG, 2013; LIU et al., 2014c; LIU et al., 2014a; LIU et al., 2014b; MOLLAHOSSEINI; CHAN; MAHOOR, 2016).

Tang (TANG, 2013) proposed a method for jointly learning a deep CNN with a linear Support Vector Machine (SVM) output which achieved the first place on both public (validation) and private data on the FER-2013 Challenge (GOODFELLOW et al., 2013).

Liu et al. (LIU et al., 2014c) introduced a facial expression recognition framework using 3DCNN together with deformable action parts constraints to jointly localize facial action parts and learn part-based representations for expression recognition. (LIU et al., 2014b) followed by including the pre-trained Caffe CNN models to extract image-level features.

In 2015, Yu and Zhang (YU; ZHANG, 2015a) achieved state-of-the-art results in the EmotiW challenge using CNNs. They used an ensemble of CNNs each with five convolutional layers and showed that randomly perturbing the input images yielded a 2-3% boost in accuracy. Specifically, they applied transformations to the input images at training time.

[1] http://www.image-net.org/

At testing time, their model generated predictions for multiple perturbations of each test example and voted on the class label to produce a final answer. They used stochastic pooling (GRAHAM, 2014) rather than max pooling due to its good performance on limited training data. (MOLLAHOSSEINI; CHAN; MAHOOR, 2016) have also obtained state of the art results with a network consisting of two convolutional layers, max-pooling, and four inception layers, the latter introduced by GoogLeNet.

## 4.2 DATA AUGMENTATION

While the FER+ dataset contains more than 35000 labeled samples, the classification performance can be further improved if we randomly perturb the input faces with additional transforms. The random perturbation essentially generates additional unseen training samples and therefore makes the network even more robust to deviated and rotated faces.

A similar method is reported in (KAHOU et al., 2013) where the authors generate perturbed training data by feeding their network with randomly cropped and flipped $40 \times 40$ face images from the original ones. As in (YU; ZHANG, 2015b), we consider a much more comprehensive set of perturbations through the following randomized affine image warping:

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} c & 0 \\ 0 & c \end{bmatrix} \begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix} \begin{bmatrix} 1 & s_1 \\ s_2 & 1 \end{bmatrix} \begin{bmatrix} x - t_1 \\ y - t_2 \end{bmatrix} \tag{4.1}$$

where $s_1$ and $s_2$ are the skew parameters along $x$ and $y$ directions and are both randomly sampled from $\{-0.1, 0, 0.1\}$. $\theta$ is the rotation angle randomly sampled from three different values: $\{-\frac{\pi}{18}, 0, \frac{\pi}{18}\}$. $c$ is a random scale parameter defined as $c = \frac{47}{(47-\delta)}$, where $\delta$ is a randomly sampled integer on $[0, 4]$. $t_1$ and $t_2$ are two translation parameters whose values are sampled from $0, \delta$ and are coupled with $c$. In reality one generates the warped image with the following inverse mapping:

$$\begin{bmatrix} x \\ y \end{bmatrix} = \mathbf{A}^{-1} \begin{bmatrix} x' \\ y' \end{bmatrix} + \begin{bmatrix} t_1 \\ t_2 \end{bmatrix}, \tag{4.2}$$

where $A$ is the composition of the skew, rotation and scale matrices. The input ($x' \in [0, 47]$, $y' \in [0, 47]$) are the pixel coordinates of the warped image. Equation 4.2 simply computes an inverse mapping to find the corresponding $(x, y)$. As the computed mappings mostly contain non-integer coordinates, bilinear interpolation is used to obtain the perturbed image pixel values. For pixels mapped outside the original image, we take pixel value of its mirrored position. The input training faces are also randomly flipped. Finally, we applied a random color transformation to change the brightness, contrast, gamma, blur, and noise to further introduce additional robustness.

Table 11 – Convolutional Neural Networks Selected for this Experiments. These five architectures are the most used in the problems of facial expression recognition.

| Arch. | Input size | Fine-tuning | Reference |
|---|---|---|---|
| FMPNet | 48 × 48 | - | (YU; ZHANG, 2015b) |
| CVGG13 | 64 × 64 | - | (BARSOUM et al., 2016b) |
| AlexNet | 227 × 227 | × | (KRIZHEVSKY, 2014) |
| ResNet18 | 224 × 224 | × | (HE et al., 2015) |
| PreActResNet18 | 32 × 32 | - | (HE et al., 2016) |

Source: The author (2019)

## 4.3 EXPERIMENTS

### 4.3.1 Protocol

Five databases were used to carry out the experiments: FER+(BARSOUM et al., 2016a), CK+ dataset(LUCEY et al., 2010), BU-3DFE (YIN et al., 2006), JAFFE (LYONS et al., 1998) and AffectNet (MOLLAHOSSEINI; HASANI; MAHOOR, 2017). We train the neural networks on the FER+ dataset and AffectNet dataset employing the same split between training, validation and testing data provided in the original FER dataset and training and validation provided in the original AffectNet dataset. CK+, BU-3DFE and JAFFE datasets are used exclusively in the test stage. PyTorch[2] was used as a deep learning framework in all cases. The neural net hyper-parameters were set to generate the same conditions in the experimentation: 150 epoch; learning rate 0.0001; loss function cross entropy and Adam optimizer. We used the weighted random sampler in all experiment for imbalance problems.

Table 11 show the select architectures in this study. We selected the pre-trained models: AlexNet(KRIZHEVSKY, 2014) and ResNet18(HE et al., 2015) on ImageNet to be re-trained on FER+. The models were obtained from the torchvision[3] repository for PyTorch.

### 4.3.2 Results

Accuracy, Precision, Recall, and F1 score were calculated to evaluate the results on FER+ dataset and AffectNet dataset (see Table 12). ResNet18 and AlexNet obtained the best results in both cases. In particular, ResNet obtained an F1 score of 0.769 as the best result on FER+ and an F1 score of 0.591 on AffectNet. The obtained models improvement the results of the (BARSOUM et al., 2016a) and (MOLLAHOSSEINI; HASANI; MAHOOR, 2017) for FER+ and AffectNet datasets respectively. As a result, we have a set

[2] http://pytorch.org/

[3] https://download.pytorch.org/models/

Table 12 – Classification results for the FER+ database and AffectNet database. Acc.: Accuracy, Prec.: Precision, Rec.: Recall, F1: F1 measurement, Arch: Architecture.

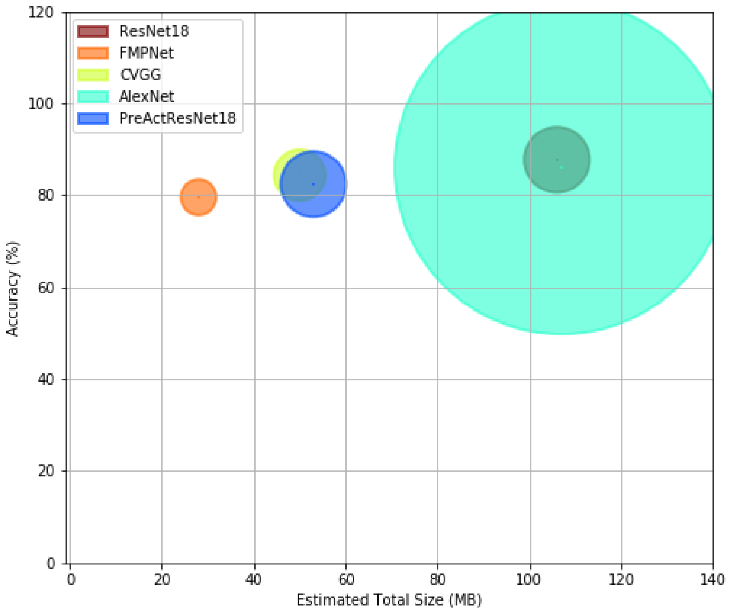
| | FER+ | | | | AffectNet | | | |
|---|---|---|---|---|---|---|---|---|
| Arch. | Acc. | Prec. | Rec. | F1 | Acc. | Prec. | Rec. | F1 |
| FMPNet | 79.535 | 66.697 | 68.582 | 67.627 | 54.000 | 54.629 | 54.000 | 54.313 |
| CVGG13 | 84.316 | 75.151 | 67.425 | 71.079 | 56.675 | 58.139 | 56.675 | 57.398 |
| AlexNet | 86.038 | 77.658 | 68.657 | 72.881 | 57.875 | 59.737 | 57.875 | 58.791 |
| ResNet18 | **87.695** | **85.956** | **69.659** | **76.954** | **58.550** | **59.847** | **58.550** | **59.191** |
| PreActResNet18 | 82.372 | 76.915 | 65.238 | 70.597 | 51.425 | 52.692 | 51.425 | 52.051 |

Source: The author (2019)

of baseline models that can be compared with the different approaches proposed in the next chapters.

CVGG13 is a particular adaptation of VGG architecture for FER problems. This network does not obtain the best results but it is considerably smaller compared to ResNet and AlexNet. Figure 17 shows the relationship between accuracy (%), the estimated total size (MB) and the number of parameters to be optimized. CVGG13 obtained a good ratio between memory size and accuracy which makes this architecture one of the most suitable for low-cost applications, such as mobile applications for FER.

Figure 17 – Accuracy (%) vs. estimated total size (MB), size $\propto$ parameters. The accuracy is calculated with models trained on FER+.



Source: The author (2019)

The models were also tested on the JAFFE, CK+ and BU-3DFE datasets. This ex-

Table 13 – Accuracy (%) of the classification on all datasets for models training on FER+. Arch: Architecture, FER+†: FER+ test dataset, AffectNet‡ validation dataset.

| Arch. | FER+† | AffectNet‡ | JAFFE | CK+ | BU-3DFE |
|---|---|---|---|---|---|
| FMPNet | 79.790 | 29.200 | 46.766 | 65.363 | 41.379 |
| CVGG13 | 84.316 | 31.150 | 46.269 | 66.201 | 42.069 |
| AlexNet | 86.038 | **35.075** | 46.269 | 70.670 | **46.379** |
| ResNet18 | **87.695** | 34.400 | **50.746** | **71.508** | 45.345 |
| PreActResNet18 | 82.372 | 26.100 | 36.318 | 55.307 | 39.828 |

Source: The author (2019)

Table 14 – Accuracy (%) of the classification on all datasets for models training on AffectNet. Arch: Architecture, FER+†: FER+ test dataset, AffectNet‡ validation dataset.

| Arch. | AffectNet‡ | FER+† | JAFFE | CK+ | BU-3DFE |
|---|---|---|---|---|---|
| FMPNet | 54.000 | 56.073 | 53.234 | **87.989** | 62.586 |
| CVGG13 | 56.675 | 60.822 | 55.224 | 83.520 | **66.034** |
| AlexNet | 57.875 | 59.770 | 57.214 | 87.151 | 63.793 |
| ResNet18 | **58.550** | **62.894** | **59.204** | 87.430 | 64.310 |
| PreActResNet18 | 51.425 | 58.368 | 47.761 | 84.078 | 61.379 |

Source: The author (2019)

periment allows us to know the degree of generalization of the networks with respect to emotions captured in different scenarios (controlled environments). Table 13 and Table 14 shows the results of the models trained on FER+ and AffectNet respectively. As can be seen, there is a degradation of the results, especially in the JAFFE and BU-3DFE databases. This is because these databases have marked ethnicity differences (JAFFE is a base for Asian women for example) and capture conditions very different from those of the FER+ dataset and AffectNet. In general, architectures designed specifically for FERs such as CVGG and FMPNet improved their performance. However, the best architectures in terms of accuracy were ResNet18 and AlexNet.

### 4.3.3 Discussion

What neural net architecture (of the all used in this works) offers the best results for this problem?

ResNet18 and AlexNet obtained the best results for FER+ and AffectNet respectively under the conditions proposed. It is important to note that in our experiments, we set the hyper-parameters to be able to make general assessments in the same scenario. Although, the results exceeded (BARSOUM et al., 2016b) and (MOLLAHOSSEINI; HASANI; MAHOOR, 2017), good results were not achieved in terms of the generalization on classification of the JAFFE, CK+ and BU-3DFE databases. VGG13 shows a good relationship between the performance level and the memory capacity (see Figure 17), which makes it a good choice for mobile applications and embedded systems.

A critical problem observed was the imbalance of the classes. There is a relationship between the error obtained and the frequency of objects per class. Disgust and contempt present the smallest number of objects in the database (the sum of them does not exceed 1% of the data of training FER+ dataset), which implies that the samples of these classes may not represent the data of their respective populations. Despite having used a weighted random sampler and data augmentation, apparently, it was not possible to obtain a good representation of the high-level features for these classes.

What is the generalization capacity of the models in other domains?

The degradation of the results in the data sets of JAFFE, CK+, BU-3DFE, shows that the methods, in general, do not get a good generalization in these domains. As in all machine learning problems, the data are essential to achieve good results. This is particularly dramatic in the case of facial expression recognition. The experiments reveal that exist two factors that affect the generalization of models to other domains.

**Dataset biases**: Datasets such as CK+, JAFFE, and BU-3DFE present specific capture conditions such as lighting, exposure, type of actors, position, etc. For example, JAFFE is a dataset of japanese females who repeat the same expression several times to capture different states of the same emotion. If care is not taken, the field can end up putting efforts in attaining an “algorithmic local maximum”.

**Label subjectivity**: The subjective criterion of the manual annotator often plays an important role on which labels are assigned to a specific data point. When these subjective effects are large, then the number of manual annotators required to obtain a consistent labelling grows, and with that the resources required to perform the manual annotation grow accordingly. Important factors affecting the inter-rater reliability include both the expertise of the manual annotator for that specific annotation task, and the nature of the annotation task.

In this particular case, there are some relevant differences between the datasets. It should be noted that datasets CK+, JAFFE, and BU-3DFE were labeled by experts and that the expressions are not spontaneous. In the case of FER+ and AffectNet, natural expressions and crowdsourcing were used.

What are the main problems presented?

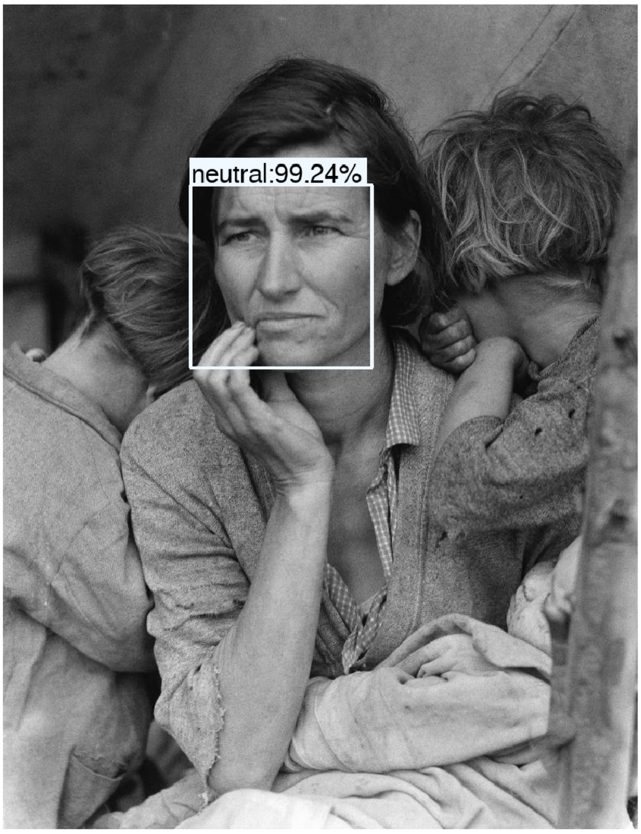
Recent developments for the facial expression recognition problem consider processing the entire image regardless of the face crop location within the image (YU; ZHANG, 2015b). Such developments bring in extraneous artifacts, including noise, which might be harmful for classification. This is problematic as the *minutiae* that characterizes facial expressions can be affected by elements such as hair, jewelry, and other environmental objects not defining the actual face and as part of the image background. Some methods use heuristics to decrease the searching size of the facial regions to avoid considering objects beyond the face itself.

We tested ResNet18 for a set of natural images. These images were visually assessed by the team and we observed a good performance. The results obtained were compared with the Microsoft Tool[4] for FER and coincided in all cases.

Carey Dunne in (DUNNE, 2015) makes a small criticism of the Microsoft Tool in which he evaluates the results obtained in some emotions of famous portrait subjects. Figure 18 and Figure 19 shows the results of our system versus two of the portrait used for the analysis in (DUNNE, 2015). Like the results obtained by the Microsoft Tool, our system gives an insignificant score (0.66%) to the deep melancholy of Florence Thompson (see Figure 18) and qualifies with 99.9% happiness the super creepy, painted-on grin the of Cindy Sherman in *"Untitled # 414"* (see Figure 19).

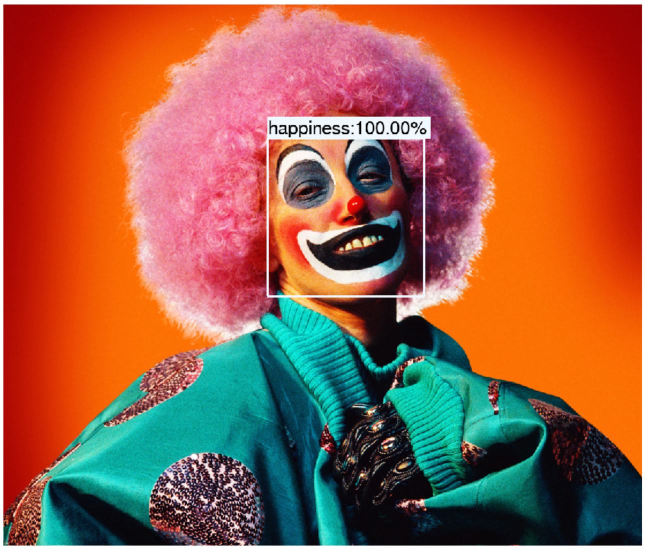
[4] https://azure.microsoft.com/en-us/services/cognitive-services/emotion/

Figure 18 – Dorothea Lange, *"Madre migrante"* (1936). The face of Florence Thompson, a "destitute pea-picker" and 32-year-old mother of seven in Depression-era California, shot by Dorothea Lange in 1936.



Source: The author (2019)

Figure 19 – Cindy Sherman, *"Untitled # 414"* (2003). The super creepy, painted-on grin the of Cindy Sherman.



Source: The author (2019)

## 4.4 CONCLUSION

In this chapter, we present an analysis of the main problems of the traditional techniques of facial expression recognition based on deep learning. Some of the most used architectures for this problem were selected to answer three questions about this type of technique. The answers to these questions allow identifying the current limitations of these techniques.

# 5 DEEP STRUCTURED METRIC LEARNING APPLIED TO FACIAL EXPRESSION RECOGNITION[2]

We propose a deep metric learning model to create embedded sub–spaces with a well defined structure. A new loss function that imposes Gaussian structures on the output space is introduced to create these sub–spaces thus shaping the distribution of the data. Having a mixture of Gaussians solution space is advantageous given its simplified and well established structure. It allows fast discovering of classes within classes and the identification of mean representatives at the centroids of individual classes. We also propose a new semi–supervised method to create sub–classes. We illustrate our methods on the facial expression recognition problem and validate results on the FER+, AffectNet, Extended Cohn-Kanade (CK+), BU-3DFE, and JAFFE datasets. We experimentally demonstrate that the learned embedding can be successfully used for various applications including expression retrieval and emotion recognition.

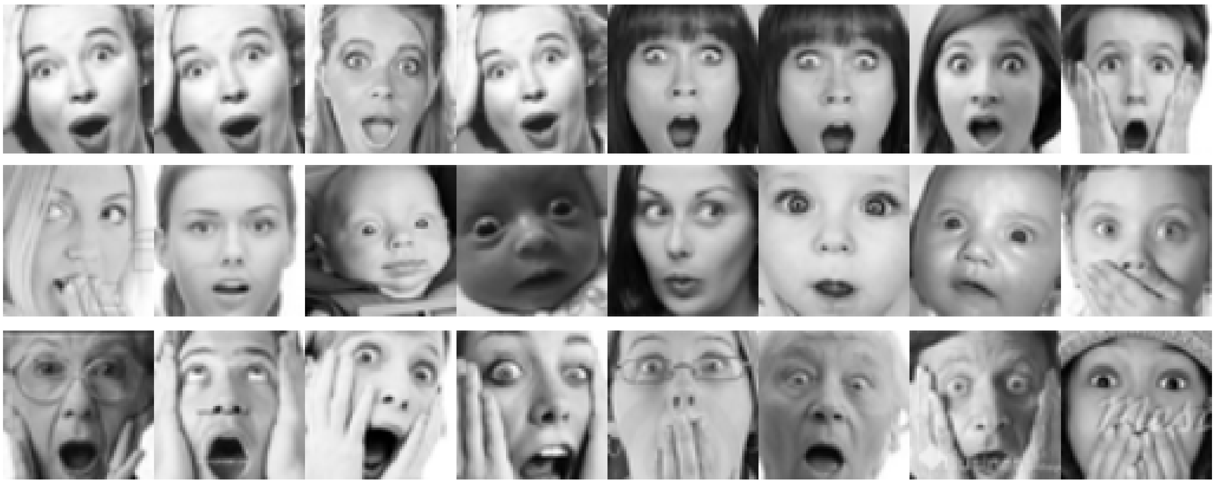
## 5.1 INTRODUCTION

Classical distance metrics like Euclidean distance and cosine similarity are limited and do not always perform well when computing distances between images or their parts. Recently, end–to–end methods (SCHROFF; KALENICHENKO; PHILBIN, 2015; BALNTAS et al., 2016; SONG et al., 2016a; WANG et al., 2014) have shown much progress in learning an intrinsic distance metric. They train a network to discriminatively learn embeddings so that similar images are close to each other and images from different classes are far away in the feature space. These methods are shown to outperform others adopting manually crafted features such as SIFT and binary descriptors (DOSOVITSKIY et al., 2016; SIMO-SERRA et al., 2015). Feedforward networks trained by supervised learning can be seen as performing representation learning, where the last layer of the network is typically a linear classifier, *e.g.* a softmax regression classifier.

Representation learning is of great interest as a tool to enable semi-supervised and unsupervised learning. It is often the case that datasets are comprised of vast training data but with relatively little labeled training data. Training with supervised learning techniques on a reduced labeled subset generally results in severe overfitting. Semi-supervised learning is an alternative to resolve the overfitting problem by learning from the vast unlabeled data. Specifically, it is possible to learn good representations for the unlabeled data and use them to solve the supervised learning task.

[2] Pedro D. Marrero Fernandez, Fidel A. Guerrero Peña, Tsang Ing Ren, Tsang Ing Jyh and Alexandre Cunha; Centro de Informática, Universidade Federal de Pernambuco, Brazil; University of Antwerp - IMEC, IDLab research group, Sint-Pietersvliet 7, 2000 Antwerp, Belgium; Center for Advanced Methods in Biological Image Analysis, California Institute of Technology, USA

Figure 20 – **Classes within classes**. The figure depicts some faces in the FER+ dataset classified by our method as having a surprise expression. Our method further separates these faces into other sub–classes, as shown in the three examples above. Each row contains the top eight images identified to be the closest ones to the centroid of their respective sub–class, and each represented by its own Gaussian. One could tentatively visually describe the top row as faces with strong eye and mouth expressions of surprise, the middle row with mostly mildly surprised eyes, and the bottom row faces with strong surprise expressed with wide open eyes and mouth, and hands on face. Observe the face similarities in each sub–class.



Source: The author (2019)

The adoption of a particular cost function in learning methods imposes constraints on the solution space, whose shape can take any form satisfying the underlying properties induced by the loss function. For example, in the case of triplet loss (SCHROFF; KALENICHENKO; PHILBIN, 2015), the optimization of the cost function leads to the creation of a solution space where every object has the nearest neighbors within the same class. Unfortunately, it does not generate a much desired probability distribution function, which is achieved by our formulation.

In theory, we would like to have the solution manifold to be a continuous function representing the true original information, because, as in the case of the facial expression recognition problem, face expressions are points in the continuous facial action space resulting from the smooth activation of facial muscles (EKMAN; FRIESEN; J, 2002). The transition from one expression to another is represented as the trajectory between the embedded vectors on the manifold surface.

The objective of this work is to offer a formulation for the creation of separable sub–spaces each with a defined structure and with a fixed data distribution. We propose a new loss function that imposes Gaussian structures in the creation of these sub-spaces. In addition, we also propose a new semi-supervised method to create sub–classes within each facial expression class, as exemplified in Figure 20.

## 5.2 RELATED WORKS

Siamese networks applied to signature verification showed the ability of neural networks to learn compact embedding (BROMLEY et al., 1994). OASIS (CHECHIK et al., 2010) and local distance learning (FROME; SINGER; MALIK, 2007) learn fine-grained image similarity ranking models using hand-crafted features that are not based on deep-learning. Recent methods such as (SCHROFF; KALENICHENKO; PHILBIN, 2015; BALNTAS et al., 2016; SONG et al., 2016a; WANG et al., 2014) approaches the problem of learning a distance metric by discriminatively training a neural network. Features generated by those approaches are shown to outperform manually crafted features (BALNTAS et al., 2016), such as SIFT and various binary descriptors (DOSOVITSKIY et al., 2016; SIMO-SERRA et al., 2015).

Deep Metric Learning (DML) can be broadly divided into contrastive loss based methods, triplet networks, and approaches that go beyond triplets such as quadruplets, or even batch-wise loss. Contrastive embedding is trained on paired data, and it tries to minimize the distance between pairs of examples with the same class label while penalizing examples with different class labels that are closer than a margin $\alpha$ (HADSELL; CHOPRA; LECUN, 2006). Triplet embedding is trained on triplets of data with anchor points, a positive that belongs to the same class, and a negative that belongs to a different class (WEINBERGER; SAUL, 2009; HOFFER; AILON, 2015). Triplet networks use a loss over triplets to push the anchor and positive closer, while penalizing triplets where the distance between the anchor and negative is less than the distance between the anchor and positive, plus a margin $\alpha$. Contrastive embedding has been used for learning visual similarity for products (BELL; BALA, 2015), while triplet networks have been used for face verification, person reidentification, patch matching, for learning similarity between images and for fine-grained visual categorization (SCHROFF; KALENICHENKO; PHILBIN, 2015; SHI et al., 2016; WANG et al., 2014; CUI et al., 2016; BALNTAS et al., 2016).

Several works are based on triplet-based loss functions for learning image representations. However, the majority of them use category label-based triplets (ZHUANG et al., 2016; WANG et al., 2017; SONG et al., 2016b). Some existing works such as (CHECHIK et al., 2010; WANG et al., 2014) have focused on learning fine-grained representations. In addition, (ZHUANG et al., 2016) used a similarity measure computing several existing feature representations to generate ground truth annotations for the triplets, while (WANG et al., 2014) used text image relevance, based on Google image search to annotate the triplets. Unlike those approaches, we use human raters to annotate the triplets. None of those works focus on facial expressions, only recently (VEMULAPALLI; AGARWALA, 2019) proposed a system of facial expression recognition based on triplet loss.

## 5.3 CONTRASTIVE EMBEDDING

Contrastive embedding (HADSELL; CHOPRA; LECUN, 2006) is trained on the paired data $\{(x_i, x_j, y_{ij})\}$. Intuitively, the contrastive training minimizes the distance between a pair of examples with the same class label and penalizes the negative pair distances for being smaller than the margin parameter $\alpha$. Concretely, the cost function is defined as,

$$\mathcal{L} = \frac{1}{m} \sum_{i,j}^{m/2} y_{i,j} \Delta_{i,j}^2 + (1 - y_{i,j})[\alpha - \Delta_{i,j}]_+^2 \tag{5.1}$$

where $m$ stands for the number of images in the batch, $f(\cdot)$ is the feature embedding output from the network, $\Delta_{i,j} = ||f(x_i) - f(x_j)||_2$, and the label $y_{i,j} \in 0, 1$ indicates whether a pair $(x_i, x_j)$ is from the same class or not. The $[\cdot]_+$ operation indicates the hinge function $max(0, \cdot)$. For more details we refer to the works of (HADSELL; CHOPRA; LECUN, 2006; BELL; BALA, 2015).

## 5.4 TRIPLET LOSS

Triplet Loss is trained on the triplet data $x_a^{(i)}, x_p^{(i)}, x_n^{(i)}$ where $x_a^{(i)}$, $x(i)_p$ have the same class labels and $x_a^{(i)}$, $x_n^{(i)}$ have different class labels. The $x^{(i)}$ term is referred to as an anchor of a triplet. Intuitively, the training process encourages the network to find an embedding where the distance between $x_a^{(i)}$ and $x_n^{(i)}$ is larger than the distance between $x_a^{(i)}$ and $x_p^{(i)}$ plus the margin parameter $\alpha$. Let $\Delta_{ia,ip}$ denote the distance between normalized anchor and positive features and $\Delta_{ia,in}$ denote the distance between normalized anchor and negative features, computed using $L_2$ distance, that is $\Delta_{ia,ip} = ||f(x_i^a) - f(x_i^p)||$ and $\Delta_{ia,in} = ||f(x_i^a) - f(x_i^n)||$. There are various ways to compute triplet loss. The most commonly used is the Hinge Loss function, with a hyper-parameter, margin, $\alpha$(SCHROFF; KALENICHENKO; PHILBIN, 2015; WEINBERGER; SAUL, 2009; HOFFER; AILON, 2015). The loss function is expressed as:

$$\mathcal{L} = \frac{3}{2m} \sum_{i,j}^{m/3} [\Delta_{ia,ip}^2 - \Delta_{ia,in}^2 + \alpha]_+ \tag{5.2}$$

## 5.5 DEEP METRIC LEARNING VIA LIFTED STRUCTURED FEATURE EMBEDDING

The structured loss function is define based on all positive and negative pairs of samples in the training set:

$$\begin{aligned}\mathcal{L} &= \frac{1}{2|\hat{P}|}\sum_{i,j\in\hat{P}}\max(0,\mathcal{L}_{i,j}),\\ \mathcal{L}_{i,j} &= \max\left(\max_{i,k\in\hat{N}}\alpha-\Delta_{i,k},\max_{j,l\in\hat{N}}\alpha-\Delta_{j,l}\right)+\Delta_{i,j}\end{aligned} \tag{5.3}$$

where $\hat{P}$ is the set of positive pairs and $\hat{N}$ is the set of negative pairs in the training set. This function poses two computational challenges: (1) it is non-smooth, and (2) both evaluating it and computing the subgradient requires mining all pairs of examples several times (SONG et al., 2016a).

## 5.6 METHODOLOGY

### 5.6.1 Structured Gaussian Manifold Loss

Let $S = \{x_i|x_i \in \Re D\}$ be a collection of *i.i.d.* samples $x_i$ to be classified into $c$ classes, and let $w_j$ represent the $j$–th class, for $j = 1,\dots,c$. The computed class function $l(x) = \arg\max p(w|f_\Theta(x))$ returns the class $w_j$ of sample $x$ – maximum *a posteriori* probability estimate – for the neural net function $f_\Theta : \Re D \rightarrow \Re d$ drawn independently according to probability $p(x|w_j)$ for input $x$. Suppose we separate $S$ in an embedded space such that each set $C_j = \{x|x \in S, l(x) = w_j\}$ contains the samples belonging to class $w_j$. Our goal is to find a Gaussian representation for each $C_j$ which would allow a clear separation of $S$ in a reduced space, $d \ll D$.

We assume that $p(f_\Theta(x)|w_j)$ has a known parametric form, and it is therefore determined uniquely by the value of a parameter vector $\theta_j$. For example, we might have $p(f_\Theta(x)|w_j) \sim N(\mu_j, \Sigma_j)$, where $\theta_j = (\mu_j, \Sigma_j)$, for $N(.,.)$ the normal distribution with mean $\mu_j$ and variance $\Sigma_j$. To show the dependence of $p(f_\Theta(x)|w_j)$ on $\theta_j$ explicitly, we write $p(f_\Theta(x)|w_j)$ as $p(f_\Theta(x)|w_j, \theta_j)$. Our problem is to use the information provided by the training samples to obtain a good transformation function $f_\Theta(x_j)$ that generates embedded spaces with a known distribution associated with each category. Then the *a posteriori* probability $P(w_j|f_\Theta(x))$ can be computed from $p(f_\Theta(x)|w_j)$ by the Bayes' formula:

$$P(w_j|f_\Theta(x)) = \frac{p(w_j)p(f_\Theta(x)|w_j,\theta_i)}{\sum_i^c p(w_i)p(f_\Theta(x)|w_i,\theta_i)} \tag{5.4}$$

We use the normal density function for $p(x|w_j, \theta_j)$. The objective is to generate embedded sub-spaces with defined structure. Thus, using the Gaussian structures:

$$p(f_\Theta(x)|w_j, \mu_j, \Sigma_j) = \frac{1}{(2\pi)^{n/2}|\Sigma_j|^{1/2}} \exp(-\frac{1}{2}X^T\Sigma_j^{-1}X) \tag{5.5}$$

where $X = (f_\Theta(x) - \mu_j)$. For the case $\Sigma_j = \sigma^2 I$, where $I$ is the identity matrix:

$$p(x|w_j, \mu_j, \sigma_j) = \frac{1}{\sqrt{(2\pi)^n}\sigma_j} \exp(-\frac{||f_\Theta(x) - \mu_j||^2}{2\sigma_j^2}) \tag{5.6}$$

In a supervised problem, we know the *a posteriori* probability $P(w_j|x)$ for the input set. From this, we can define our structured loss function as the mean square error between the *a posteriori* probability of the input set and the *a posteriori* probability estimated for the embedded space:

$$\mathcal{L}_{rep} = \mathbb{E}\left\{||P(w_j|f_\Theta(x_i)) - P(w_j|x_i)||_2^2\right\} \tag{5.7}$$

We applied the steps described in Algorithm 5 to train the system. The batch size is given by $n \times c$ where $c$ is the number of classes, and $n$ is the sample size. In this work, we use $n = 30$, thus for eight classes the batch size is 240, which was used for the estimation of the parameters in Equation 5.7.

**Algorithm 5** Structured Gaussian Manifold Learning. $f_\Theta$: Neural Network; $S$: dataset; $C_j$ are the subset of the elements of class $w_j$; $N$: number of updates;

```
k ← 0
while  $k < N$  do
    $\{\text{Sample}(x_i, w_i)\} \sim S$, get current batch.
    $z \leftarrow f_\Theta(x_i)$, representation.
    $\theta_j \leftarrow \{\mu_j, \sigma j\}$
    where $\sigma$ is a parameters ($\sigma = 0.5$ in this work) and $\mu_j$ is the mean of the elements
    of the class $w_j$:
        $$\mu_j = \frac{1}{|C_j|}\sum_{k \in C} z_k$$
    where $|.|$ denotes set cardinality.
    Evaluation of the Loss function. For the explanation of the loss representation see
    equation 5.7:
        $$\mathcal{L} \leftarrow \mathcal{L}_{rep}(z_i, w_i, \mu, \sigma) + \frac{\lambda}{|\Omega|}\sum_k ||f_\Theta(x_k)||_2$$
    $\Theta^{t+1} = \Theta^t - \nabla\mathcal{L}$, backward and optimization steps.
end while
```

We define the accuracy of the model as the ability of the parameter vector $\theta$ to represent the test dataset in the embedded space. The prediction of a class can be calculated as:

$$\hat{j} = \max_{j} P(w_j, f_{\Theta}(x_k)) \tag{5.8}$$

### 5.6.2 Deep Gaussian Mixture Subspace

The same facial expression may possess a different set of global features. For example, ethnicity can determine specific color and shape, while age provides physiological differences of facial characteristics; moreover, gender, weight, and other features can determine different facial characteristics, while having the same expression. Our proposal can group and extract these characteristics automatically. We propose to represent each facial expression class as a Gaussians Mixture. These Gaussian parameters are obtained in an unsupervised way as part of the learning processes. We start from a representation space given by Algorithm 5. Subsequently, a clustering algorithm is applied to separate each class into a new class subset. This process is repeated until reaching the desired granularity level. Algorithm 6 shows the set of steps to obtain the new sub-classes.

**Algorithm 6** Deep Gaussian Mixture Sub-space Learning. $L$: Maximum level of subdivisions for the class; $f_{\Theta}$: Neural Network; $StructureGaussianManifold$: Structure Gaussian Manifold Algorithm 5; $EM$: Expectation Maximization Algorithm; $S$: dataset; $C_j$ are the subset of the elements of class $w_j$; $N$: number of updates;

```
$l \leftarrow 1$
$X, Y \sim S$
$\hat{Y} \leftarrow Y$
while $l < L$ do
    $\Theta \leftarrow$StructureGaussianManifold$(\{X, \hat{Y}\}, N)$
    $Z = f_{\Theta}(X)$
    $\hat{Y} \leftarrow \{\emptyset\}$
    $k \leftarrow 0$
    for all class $w_j$ do
        $Z_c = \{z \mid \forall z \in C_j\}$
        $\hat{l} \leftarrow \min(l, |Z_c|/Mc)$
        $g \leftarrow EM(\hat{l}, Z_c)$
        $\hat{Y} = \{\hat{Y}, g + k\}$
        $k + = \hat{l}$
    end for
end while
```

## 5.7 EXPERIMENTS

### 5.7.1 Protocol

For the evaluation of the clustering task, we use the F1-measure and Normalized Mutual Information (NMI) measures. The F1-measure computes the harmonic mean of the precision and recall, $F1 = \frac{2PR}{P+R}$. The NMI measure take as input a set of clusters $\Omega = \{o_1, \ldots, o_k\}$ and a set of ground truth classes $\mathcal{G} = \{g_1, \ldots, g_k\}$, $o_i$ indicates the set of examples with cluster assignment $i$ and $g_j$ indicates the set of examples with the ground truth class label $j$. Normalized mutual information is defined by the ratio of mutual information and the average entropy of the clusters and the entropy of the labels, $NMI(\Omega, \mathcal{G}) = \frac{I(\Omega;\mathcal{G})}{2(H(\Omega)+H(\mathcal{G}))}$, for complete details see (MANNING et al., 2008). For the retrieval task, we use the Recall@K (JEGOU; DOUZE; SCHMID, 2011) measure. Each test image (query) first retrieves K Nearest Neighbour (KNN) from the test set and receives score 1 if an image of the same class is retrieved among the KNN, and 0 otherwise. Recall@K averages those score over all the images. Moreover, we also evaluate accuracy, i.e. the fraction of results that are the same class as queried image, averaged over all queries. While the classification task is evaluated using KNN on the training set.

For the training process, we use the Adam method (KINGMA; BA, 2014) with a learning rate of 0.0001 and batch size of 256 (samples of size 32 to estimate the parameters in each iteration). In the TripletLoss case, we used 128 triplets in each batch. The neural networks were initialized with the same weights in all cases.

### 5.7.2 Representation and Recover

The groups used for the evaluation of the measures are obtained using K-means, whereas K equals the number of classes (8 in the case of the FER+, AffectNet, CK+ datasets, and 7 for JAFFE and BU-3DFE datasets).

The results obtained for the clustering task show that the proposed method presents good group quality (see table 15) in similar domains. As can be observed, the results are degraded for different domains. In general, we observe that the TripletLoss is most robust to the change of domains on all models. However, the best result is achieved using the proposed method for the RestNet18 model in FER+, CK+ and BU-3DFE.

Figure 21 shows a 2D t-SNE (MAATEN, 2014) visualization of the learned TripletLoss (left) and SGMLoss (right) embedding space using the FER+ training set. The amount of overlap between two categories in this figure roughly tells us about the extent of visual similarity between them. For example, in the SGMLoss case, happy and neutral have some objects overlap indicating that these cases could be confused easily, and both of them have a very low overlap with fear indicating that they are visually very distinct from fear. Also, the spread of a category in this figure tells us about the visual diversity

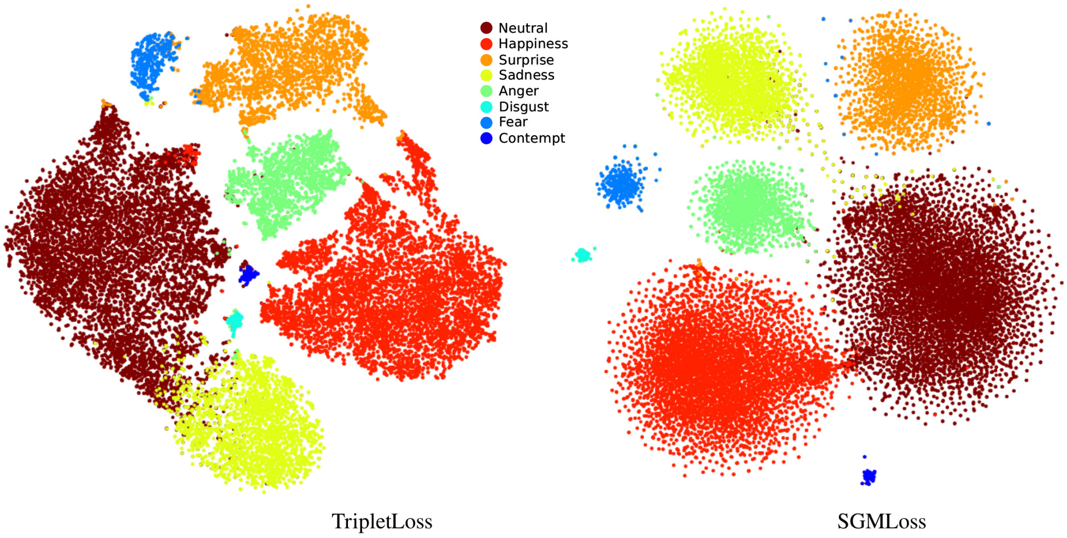
Table 15 – NMI (%) of the clustering task on all datasets of the TripletLoss and SGMLoss models trained on FER+. SGMLoss: Structured Gaussian Manifold Loss, Arch: Architecture, FER+†: FER+ test dataset, AffectNet‡ validation dataset.

| Method | Arch. | FER+† | AffectNet‡ | JAFFE | CK+ | BU-3DFE |
|---|---|---|---|---|---|---|
| TripletLoss | FMPNet | 55.257 | 10.627 | 19.528 | 71.129 | 34.901 |
| | CVGG13 | 67.384 | 9.103 | 28.295 | 68.303 | 27.275 |
| | AlexNet | 67.035 | 12.945 | 30.241 | 68.800 | 27.039 |
| | ResNet18 | 64.457 | **15.588** | **31.046** | 74.028 | 36.708 |
| | PreActResNet18 | 57.904 | 8.452 | 20.699 | 70.079 | 27.580 |
| SGMLoss | FMPNet | 57.880 | 10.469 | 26.196 | **77.839** | 36.559 |
| | CVGG13 | 65.139 | 10.355 | 24.293 | 66.062 | 27.233 |
| | AlexNet | 62.091 | 10.582 | 24.560 | 65.230 | 28.115 |
| | ResNet18 | **68.840** | 12.333 | 30.382 | **77.902** | **37.545** |
| | PreActResNet18 | 51.425 | 6.886 | 23.216 | 61.413 | 26.104 |

Source: The author (2019)

within that category. For example, happiness category maps to some distinct regions indicating that there are some visually distinct modes within this category.

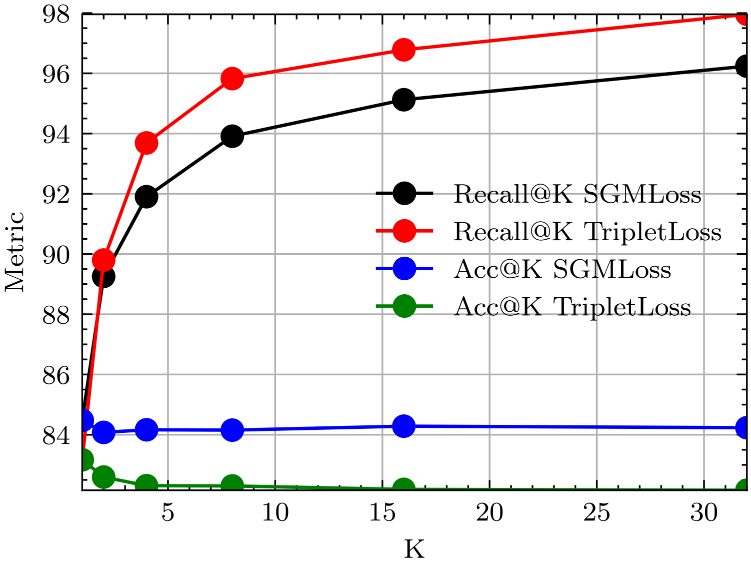
Figure 21 – Barnes-Hut t-SNE visualization (MAATEN, 2014) of the TripletLoss (left) and SGMLoss (right) for the FER+ database. Each color represents one of the eight emotions including neutral.



Source: The author (2019)

Figure 22 shows the results obtained in the recovery task (Recall@K and Acc@K measures) for $K = \{1, 2, 4, 8, 16, 32\}$. TripletLoss obtains better recovery results for all K but to the detriment of accuracy. Our method manages to increase its recovery value while preserving quality. It means that most neighbors are of the same class. Figure 23 shows the top-5 retrieved images for some of the queries on CelebA dataset (LIU et al., 2015). The overall results of the proposed SGMLoss embedding are clearly better than the results of TripletLoss embedding.

Figure 22 – Recall@K and Acc@K measures for the test split FER+ dataset. The applied model was the ResNet18 having $K = \{1, 2, 4, 8, 16, 32\}$.



Source: The author (2019)

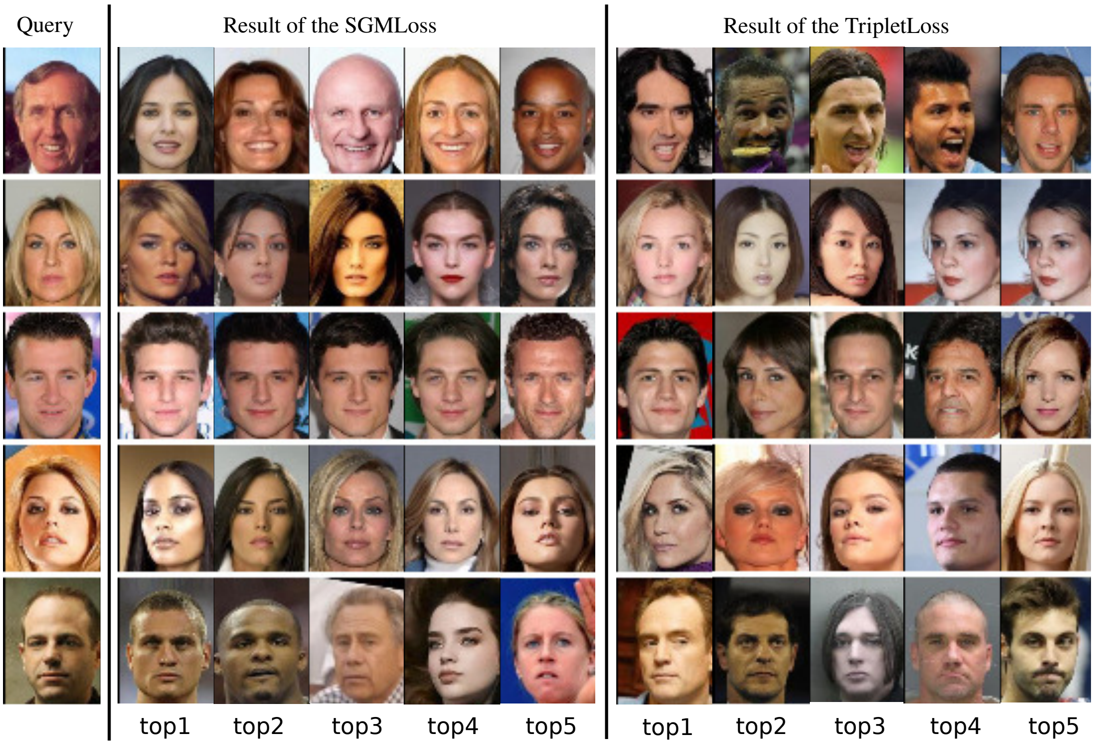
### 5.7.3 Classification

The proposed SGMLoss method can be used for FER by combining it with the KNN classifier. Figure 24 shows the average F1-score of the SGMLoss and TripletLoss on the FER+ validation set as a function of the number of neighbors used. F1-score is maximized for K=11.

Table 16 compares the classification performance of the SGMLoss embedding (using 11 neighbors) with TripletLoss and CNN models. In general, our method obtains the best classification results for all architectures. ResNet18 CNN model does not obtains a significant higher accuracy. Moreover, our results surpass the accuracy 84.99 presented in (BARSOUM et al., 2016a).

The Facial Expression dataset constitute a great challenge due to the subjectivity of the emotions (MARRERO-FERNÁNDEZ et al., 2014). The labeling process requires the effort of a group of specialists to make the annotations. FER+ and AffectNet datasets contains many problems in the labels. In (BARSOUM et al., 2016a) an effort was made to improve the quality of the labels of the FER+ (dataset used in our experiments) by re-tagging the dataset using crowd sourcing. Figure 25 shows some mislabeled images retrieved by

Figure 23 – Top-5 images retrieved using SGMLoss (left) and TripletLoss (right) embeddings. The overall results of the SGMLoss match the query set apparently better when compared to TripletLoss.



Source: The author (2019)

our method. The scale, position, and context could influence the decision of a non-expert tagger such as those in crowd sourcing.

Experimental results show the quality of the embedded representation obtained by SGMLoss in the classification problems. Our representation improves the representation obtained by TripletLoss, which is the method most used in the identification and representation problems.

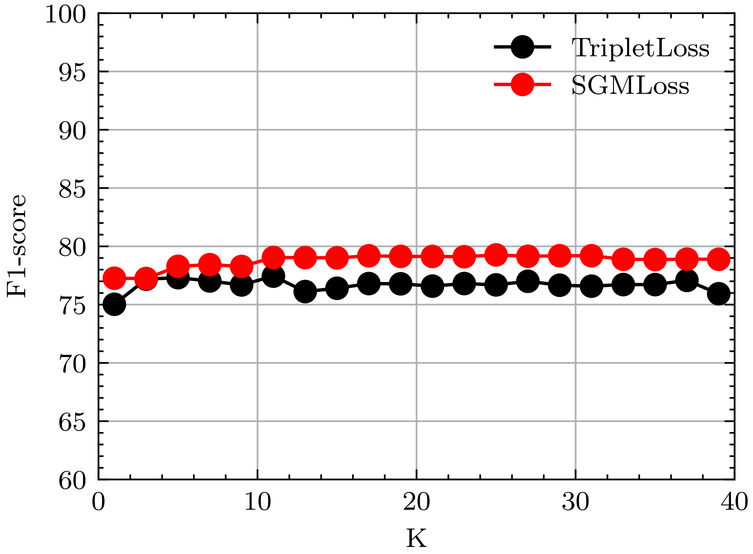
### 5.7.4 Clustering

For the training process, we use the Adam method (KINGMA; BA, 2014) with a learning rate of 0.0001, a batch size of 640 and 500 epoch. The maximum level of subdivision used is L=5 (this value guarantee that the batch for a subclass in this level to be 128). The ResNet18 architecture is selected to train the FER+ dataset. The objective of this experiment is to visually analyse the clustering obtained by this approach.

The results shown in Figure 26 present 64-dimensional embedded space using the Barnes-Hut t-SNE visualization scheme (MAATEN, 2014) using the Deep Gaussian Mixture Sub-space model for the FER+ dataset. The method created five Gaussian sub-spaces for the unsupervised case for each class.

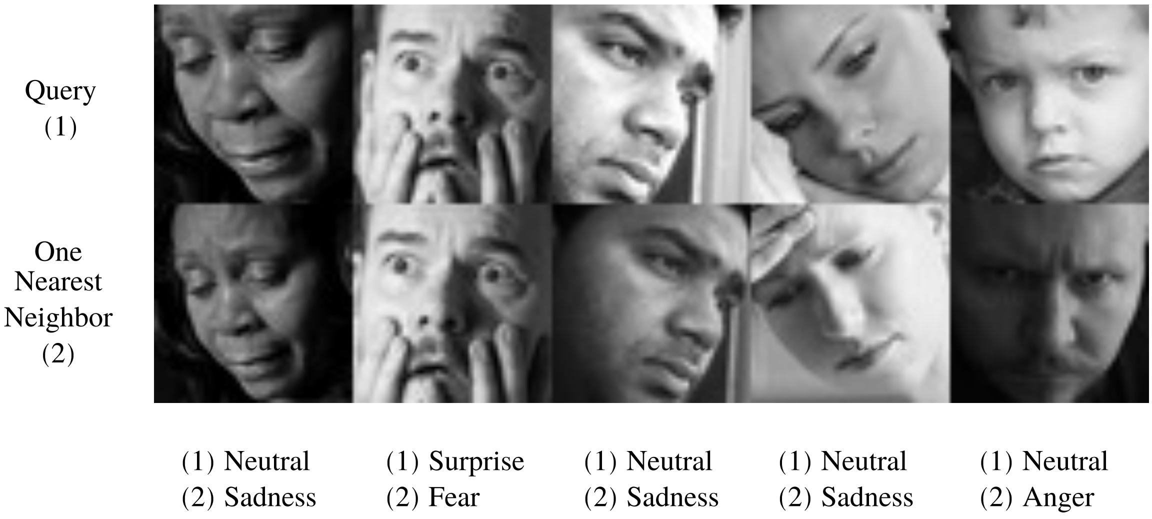
For the clustering task, all embedded vectors are calculated and EM method is applied

Figure 24 – Classification performance of the SGMLoss and TripletLoss on the FER+ validation set when combined with KNN classifier.



Source: The author (2019)

Figure 25 – Examples of mislabeled images on the FER+ dataset that were recovery using SGMLoss. The first row show the result of the query (1) and the second row the result (2). We can clearly observer that two very similar images have different labels in the dataset.



Source: The author (2019)

creating 40 groups. For each group, the medoid is calculated. The medoid is the object in the group closest to the centroid (mean to the sample). The Top-k of a group contains the *k-objects* nearer to the medoid of the group.

Figure 27 shows the Top-16 images obtained for the happiness category. The first group (Figure 27 (a) ) shows an expression of happiness closer to surprise (raised eyebrows and open mouth) with the shape of the eyes similar to each other. The second group (Figure 27

Table 16 – Classification results of the CNN, TripletLoss and SGMLoss models trained on FER+. SGMLoss: Structured Gaussian Manifold Loss, Arch: Architecture, FER+†: FER+ test dataset, AffectNet‡ validation dataset.

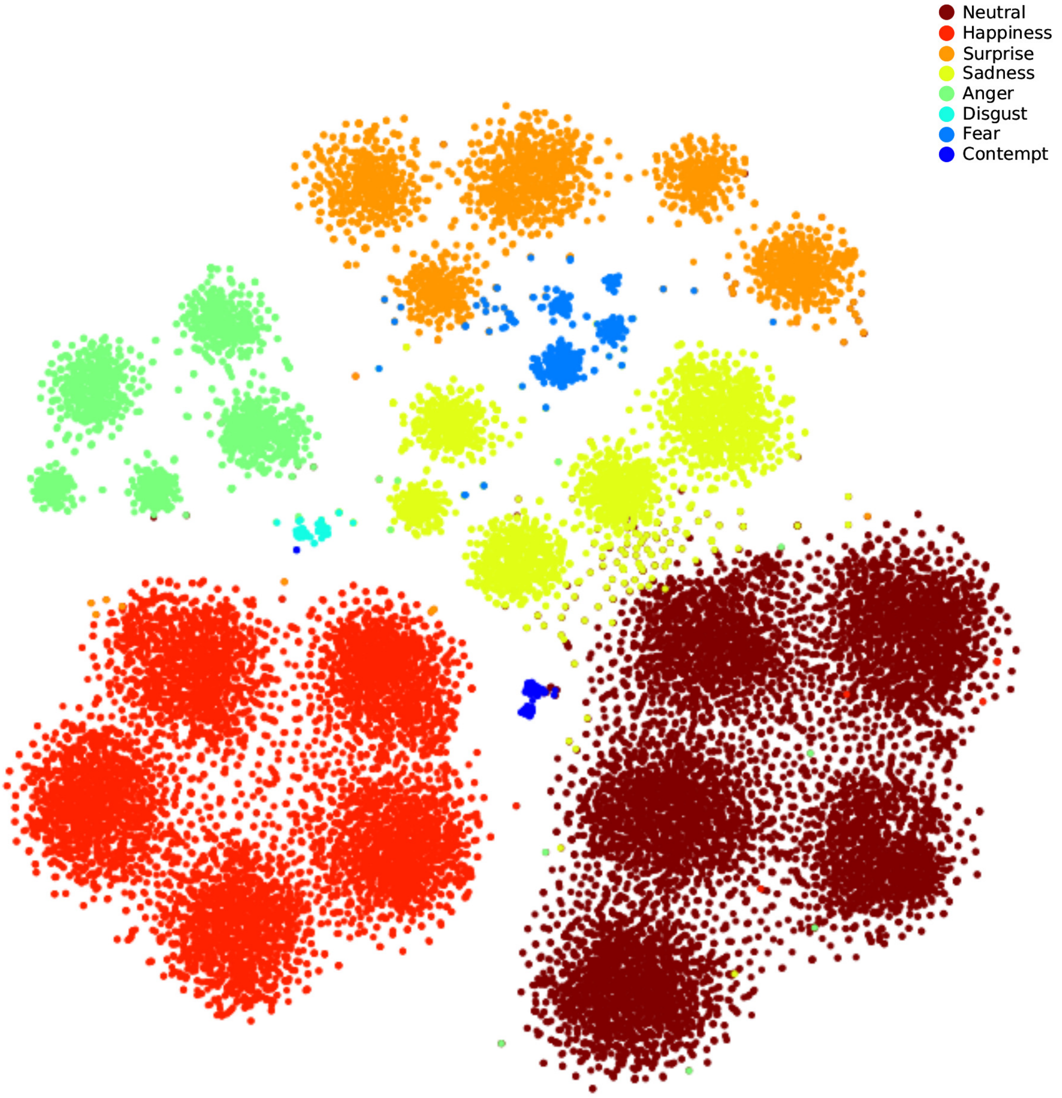
| Method | Arch. | Acc. | Prec. | Rec. | F1 |
|---|---|---|---|---|---|
| CNN | FMPNet | 79.535 | 66.697 | **68.582** | 67.627 |
| | CVGG13 | 84.316 | 75.151 | 67.425 | 71.079 |
| | AlexNet | 86.038 | 77.658 | **68.657** | 72.881 |
| | ResNet18 | **87.695** | 85.956 | **69.659** | 76.954 |
| | PreActResNet18 | 82.372 | 76.915 | 65.238 | 70.597 |
| TripletLoss | FMPNet | 82.563 | **79.554** | 62.406 | 69.944 |
| | CVGG13 | 85.974 | 82.034 | **68.112** | 74.428 |
| | AlexNet | 86.038 | 80.598 | 67.895 | 73.703 |
| | ResNet18 | 87.121 | 78.543 | 68.378 | 73.109 |
| | PreActResNet18 | 83.519 | 74.081 | 64.856 | 69.162 |
| SGMLoss | FMPNet | **83.360** | 78.806 | 66.520 | **72.143** |
| | CVGG13 | **86.261** | **86.321** | 67.341 | **75.659** |
| | AlexNet | **86.643** | **86.182** | 67.673 | **75.814** |
| | ResNet18 | 87.631 | **88.614** | 68.724 | **77.412** |
| | PreActResNet18 | **84.316** | **89.008** | **66.519** | **76.138** |

Source: The author (2019)

(b)) represents an expression closer to contempt. The third group (Figure 27 (c)) shows an expression of more intense happiness (the teeth are shown in all cases) with the shape of the mouth very similar to each other. In the fourth case (Figure 27 (d) ) shows a subcategory that is present in all facial expressions. Babies are a typically expected subset due to the intensity of expression and the physiognomical formation. Generally babies and children from 1 to 4 years old present facial expressions of greater intensity. The last group (Figure 27 (f) ) represents people with glasses and large eyes.

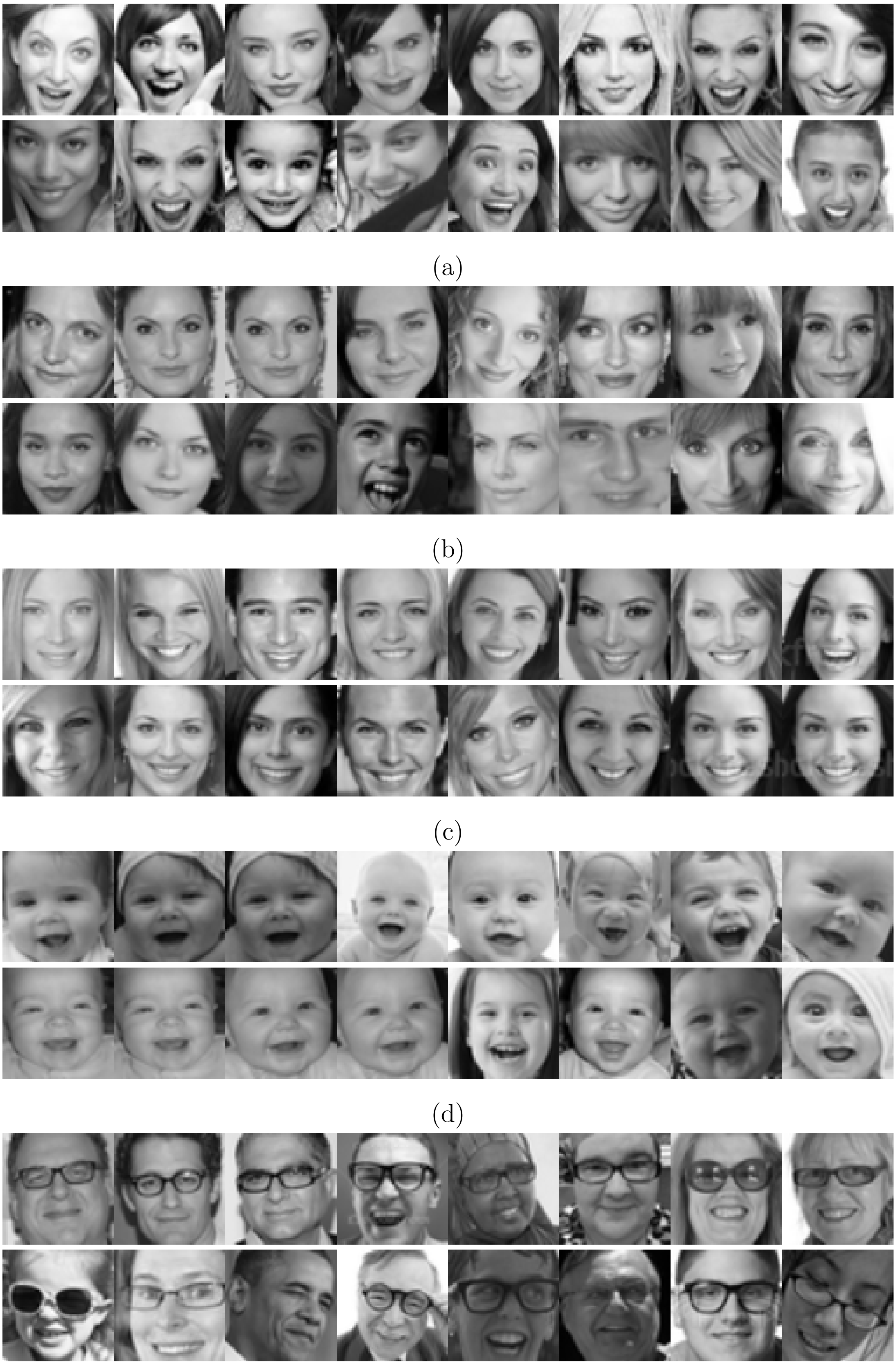
This method is a powerful tool for tasks such as photo album summarization. In this task, we are interested in summarizing the diverse expression content present in a given photo album using a fixed number of images. Figure 28 shows 5 of the 40 groups obtained on AffectNet dataset. The obtained groups show great similarity in terms of FER. These results demonstrate the generalization capacity of the proposed method and its applicability to problems of FER clustering.

Figure 26 – Barnes-Hut t-SNE visualization (MAATEN, 2014) of the DMGSubspace on the FER+ database. Each color represents one of the eight emotions including neutral.



Source: The author (2019)

Figure 27 – Top-16 of clustering obtained of the happiness class on FER+ dataset.



(a)

(b)

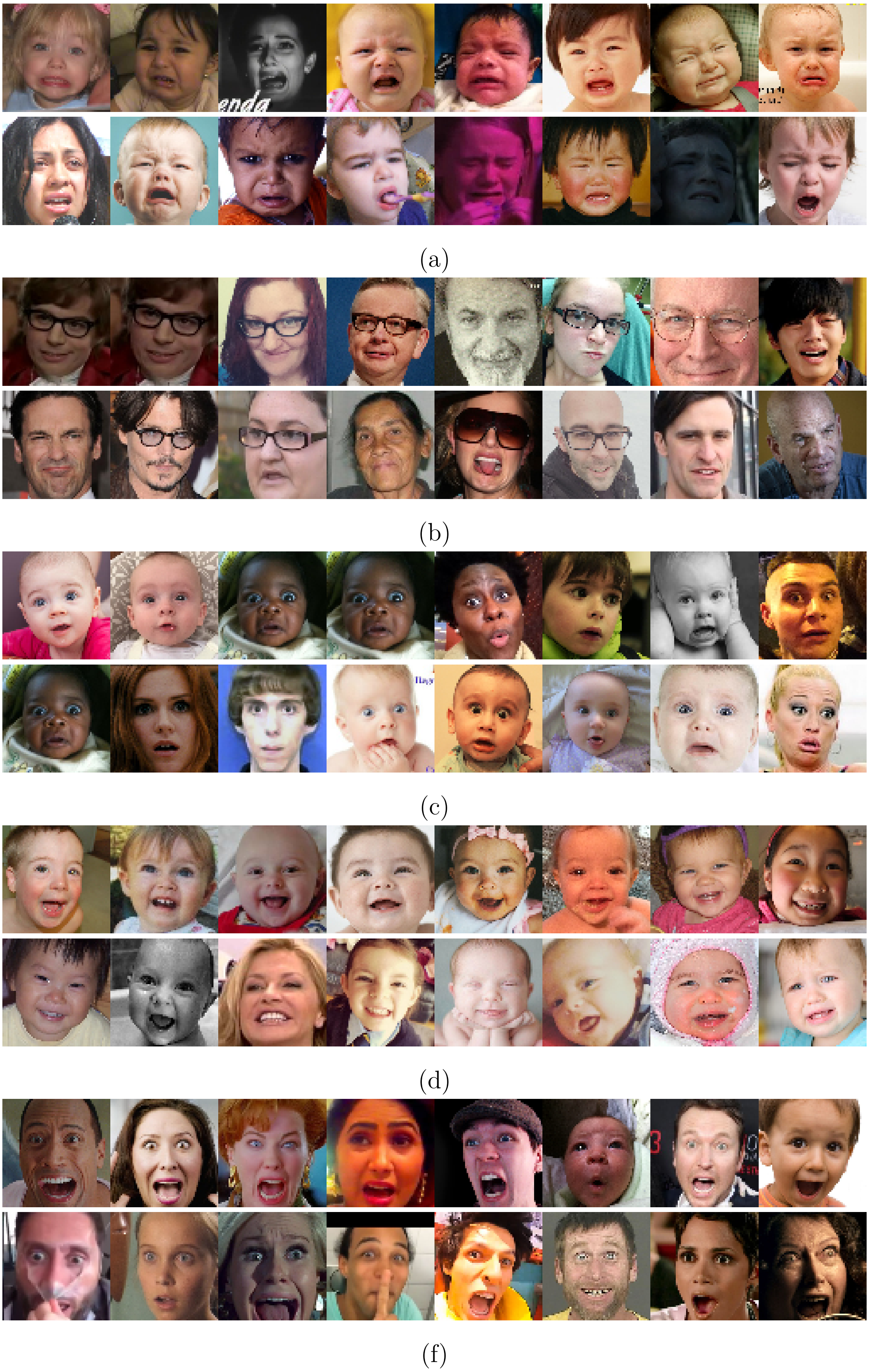
(c)

(d)

(f)

Source: The author (2019)

Figure 28 – Top-16 of fives clustering obtained on AffectNet dataset.



(a)

(b)

(c)

(d)

(f)

Source: The author (2019)

## 5.8 CONCLUSIONS

We introduced two new metric learning representation models in this work, namely Deep Gaussian Mixture Subspace Learning and Structured Gaussian Manifold Learning. In the first model, we build a Gaussian representation of expressions leading to a robust classification and grouping of facial expressions. We illustrate through many examples, the high quality of the vectors obtained in recovery tasks, thus demonstrating the effectiveness of the proposed representation. In the second case, we provide a semi-supervised method for grouping facial expressions. We were able to obtain embedded subgroups sharing the same facial expression group. These subgroups emerged due to shared specific characteristics other than the general appearance. For example, individuals with glasses expressing a happy appearance.

# 6 FERATT: FACIAL EXPRESSION RECOGNITION WITH ATTENTION NET[3]

We present a new end-to-end network architecture for facial expression recognition with an attention model. It focuses attention in the human face and uses a Gaussian space representation for expression recognition. We devise this architecture based on two fundamental complementary components: (1) facial image correction and attention and (2) facial expression representation and classification. The first component uses an encoder-decoder style network and a convolutional feature extractor that are pixel-wise multiplied to obtain a feature attention map. The second component is responsible for obtaining an embedded representation and classification of the facial expression. We propose a loss function that creates a Gaussian structure on the representation space. To demonstrate the proposed method, we create two larger and more comprehensive synthetic datasets using the traditional BU3DFE and CK+ facial datasets. We compared results with the PreActResNet18 baseline. Our experiments on these datasets have shown the superiority of our approach in recognizing facial expressions.

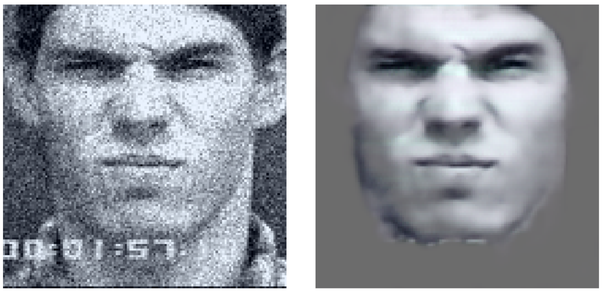
## 6.1 INTRODUCTION

Recent developments for the facial expression recognition problem consider processing the entire image regardless of the face crop location within the image (YU; ZHANG, 2015b). Such developments bring in extraneous artifacts, including noise, which might be harmful for classification as well as incur in unnecessary additional computational cost. This is problematic as the *minutiae* that characterizes facial expressions can be affected by elements such as hair, jewelry, and other environmental objects not defining the actual face and as part of the image background. Some methods use heuristics to decrease the searching size of the facial regions to avoid considering objects beyond the face itself. Such approaches contrast to our understanding of the human visual perception, which quickly parses the field of view, discards irrelevant information, and then focus the main processing on a specific target region of interest – the so called *visual attention* mechanism (ITTI; KOCH, 2001; WANG et al., 2017). Our approach tries to mimic this behavior as it aims to suppress the contribution of surrounding deterrent elements by segmenting the face in the image and thus concentrating recognition solely on facial regions. Figure 29 illustrates how the attention mechanism works in a typical scene.

Attention mechanisms have recently been explored in a wide variety of contexts (VINYALS et al., 2015; JADERBERG et al., 2015), often providing new capabilities for known

[3] Pedro D. Marrero Fernandez, Fidel A. Guerrero Peña, Tsang Ing Ren and Alexandre Cunha; Centro de Informática, Universidade Federal de Pernambuco, Brazil; Center for Advanced Methods in Biological Image Analysis, California Institute of Technology, USA; Published in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2019.

Figure 29 – Example of attention in an image. Facial expression is recognized on the front face which is separated from the less prominent components of the image by our approach. The goal is to jointly train for attention and classification where a probability map of the faces are created and their expressions learned by a dual–branch network. By focusing attention on the face features, we try to eliminate the detrimental influence possibly present on the other elements in the image during the facial expression classification. In this formulation, we explicitly target learning expressions solely on learned faces and not on other irrelevant parts of the image (background).



Source: (FERNANDEZ et al., 2019)

neural networks models (GRAVES et al., 2016; GREGOR et al., 2015; ESLAMI et al., 2016). While they improve efficiency (MNIH et al., 2014) and performance on state-of-the-art machine learning benchmarks (VINYALS et al., 2015), their computational architecture is much simpler than those comprising the mechanisms in the human visual cortex (DAYAN; ABBOTT et al., 2003). Attention has also been long studied by neuroscientists (UNGERLEIDER; G, 2000), who believe it is crucial for visual perception and cognition (CHEUNG; WEISS; OLSHAUSEN, 2016) as it is inherently tied to the architecture of the visual cortex and can affect its information.

Our contributions are summarized as follows: (1) We propose a CNN-based method using attention to jointly solve for representation and classification in FER problems; (2) We introduce a new dual-branch network to extract an attention map which in turn improves the learning of kernels specific to facial expression; and (3) We offer a new synthetic generator to render face expressions which significantly augments training data and consequently improves the overall classification.

## 6.2 RELATED WORKS

Recent method the De-expression Residue Learning (DeRL) (YANG; CIFTCI; YIN, 2018), trains a generative model to create a corresponding neutral face image for any input face. Then, another model is trained to learn the deposition (or residue) that remains in the intermediate layers of the generative model for the classification of facial expression.

Zhang *et al.* (ZHANG et al., 2018) proposed an end-to-end learning model based on Generative Adversarial Network (GAN). The architecture incorporates a generator, two discriminators, and a classifier. The GAN is used for generating multiples variation of one image, which is used to train a convolutional neural network.

## 6.3 METHODOLOGY

In this section, we describe our contributions in designing a new network architecture, in the formulation of the loss function used for training, and in the method to generate synthetic data.

### 6.3.1 Network architecture

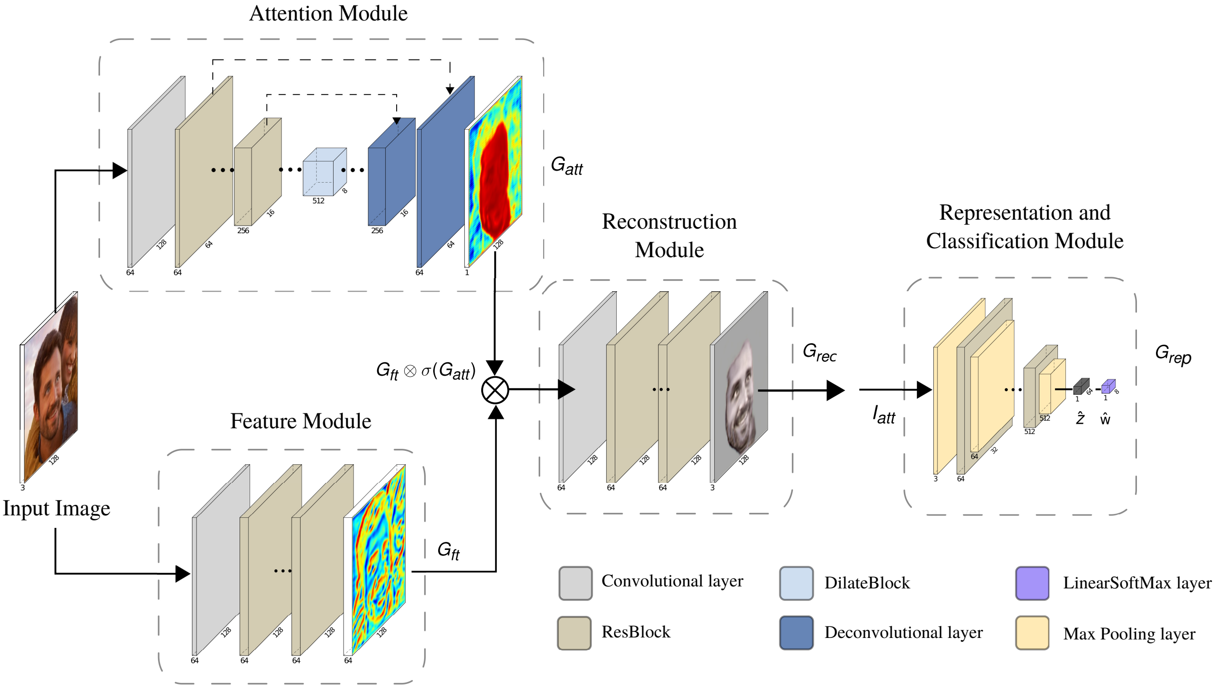
Given a facial expression image $I$, our objective is to obtain a good representation and classification of $I$. The proposed model, Facial Expression Recognition with Attention Net (FERAtt), is based on the dual-branch architecture (HE et al., 2017; LI et al., 2016; PAN et al., 2018; ZHU et al., 2016) and consists of four major modules: (i) an attention module $G_{att}$ to extract the attention feature map, (ii) a feature extraction module $G_{ft}$ to obtain essential features from the input image $I$, (iii) a reconstruction module $G_{rec}$ to estimate a proper attention image $I_{att}$, and (iv) a representation module $G_{rep}$ that is responsible for the representation and classification of the facial expression image. An illustration of the proposed model is shown in Figure 30.

**Attention module.** We use an encoder-decoder style network, which has been shown to produce good results for many generative (SHOCHER; COHEN; IRANI, 2018; ZHU et al., 2016) and segmentation tasks (RONNEBERGER; P.FISCHER; BROX, 2015). In particular, we choose a variation of the fully convolutional model proposed in (RONNEBERGER; P.FISCHER; BROX, 2015) for semantic segmentation. Also, we applied four layers in the coder with skip connections and dilation of 2x. The decoder layer is initialized with pre-trained ResNet34 (HE et al., 2015) layers. This significantly accelerates the convergence. The output features of the decoder are denoted by $G_{att}$, which is used to determine the attention feature map. This is a probability map that is not the same as a simple segmentation procedure.

**Feature extraction module.** Four ResBlocks (LIM et al., 2017) were used to extract high-dimensional features for image attention and to maintain spatial information; no pooling or strided convolutional layers were used. We denote the extracted features as $G_{ft}$ – see Figure 31b.

**Reconstruction module.** The reconstruction layer adjusts the attention map to create an enhanced input to the representation module. This module has two convolutional layers, a Relu layer, and an Average Pooling layer which, by design choice, resizes the input image of $128 \times 128$ to $32 \times 32$. This reduced size was chosen for the input of the representation and classification module (PreActivationResNet (HE et al., 2016)), the image size number we borrowed from the literature to facilitate comparisons. We plan to

Figure 30 – **Architecture of FERAtt**. Our model consists of four major modules: attention module $G_{att}$, feature extraction module $G_{ft}$, reconstruction module $G_{rec}$, and classification and representation module $G_{rep}$. The features extracted by $G_{att}$, $G_{ft}$ and $G_{rec}$ are used to create the attention map $I_{att}$ which in turn is fed into $G_{rep}$ to create a representation of the image. Input images $I$ have $128 \times 128$ pixels and are reduced to $32 \times 32$ by an Averaging Pooling layer on the reconstruction module. Classification is thus done on these smaller but richer representations of the original image.



Source: (FERNANDEZ et al., 2019)

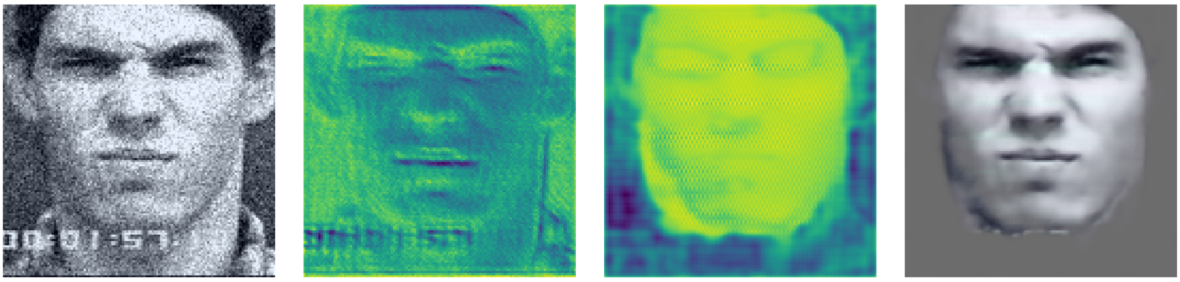
experiment with other sizes in the future. We denote the feature attention map as $I_{att}$ – see Figure 31d.

**Representation and classification module.** For the representation and classification of facial expressions, we have chosen a Fully Convolutional Network (FCN) of PreActivateResNet (HE et al., 2016). This architecture has shown excellent results when applied on classification tasks. The output of the FCN, vector $z$, is evaluated in a linear layer to obtain a vector $\hat{z} \in \mathbb{R}^d$ with the desired dimensions. $f_\Theta : \mathbb{R}^D \to \mathbb{R}^d$, the network function, builds a representation for a sample image $x \in \mathbb{R}^D$, (e.g. $D = 128 \times 128$ pixels) in an embedded space of reduced dimension $\mathbb{R}^d$ (we use $d = 64$ in our experiments). Vector $\hat{z}$ is then evaluated in a regression layer to estimate the probability $p(w|\hat{z})$ for each class $w_j$, $w = [w_1, w_2, \ldots, w_c]$.

### 6.3.2 Loss functions

The FERAtt network generates three outputs: a feature attention map $\hat{I}_{att}$, a representation vector $\hat{z}$, and a classification vector $\hat{w}$. In our training data, each image $I$

Figure 31 – Generation of attention map $I_{att}$. A 128 × 128 noisy input image (a) is processed by the feature extraction $G_{ft}$ and attention $G_{att}$ modules whose results, shown, respectively, in panels (b) and (c), are combined and then fed into the reconstruction module $G_{rec}$. This in turn produces a clean and focused attention map $I_{att}$, shown on panel (d), that is classified by the last module $G_{rep}$ of FERAtt. The $I_{att}$ image shown here is before reduction to 32 × 32 size.



Source: (FERNANDEZ et al., 2019)

has an associated binary ground truth mask $I_{mask}$ corresponding to a face in the image and its expression class vector $w$. We train the network by jointly optimizing the sum of attention, representation, and classification losses:

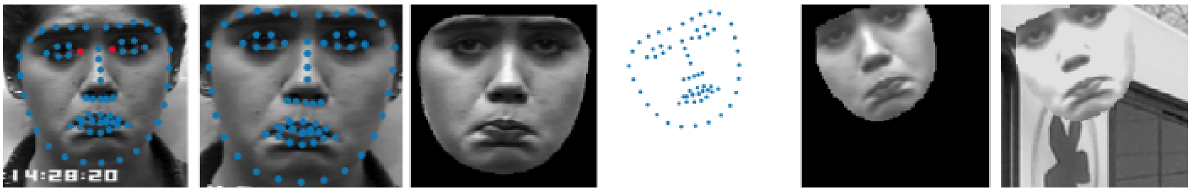
$$\hat{\theta} = \arg\min_{\theta \in \Theta} \{\mathcal{L}_{att}(I_{att}, I \otimes I_{mask}) + \mathcal{L}_{rep}(\hat{z}, w) + \mathcal{L}_{cls}(\hat{w}, w)\} \tag{6.1}$$

where $\Theta$ represents the collective parameters that need be optimized. We use the pixel-wise MSE (Mean Square Error) loss function for $\mathcal{L}_{att}$, and for $\mathcal{L}_{cls}$ we use the CE (Cross Entropy) loss function. The Structured Gaussian Manifold Loss function $\mathcal{L}_{rep}$ were proposed in this work (Equation 5.7).

### 6.3.3 Synthetic image generator

A limiting problem of currently available face expression datasets for supervised learning is the reduced number of correctly labeled data. We propose a data augmentation strategy to mitigate this problem in the lines of what has been introduced in (FERNÁNDEZ et al., 2019). Our image renderer $R$ creates a synthetic larger dataset using real face datasets by making background changes and geometric transformations of face images. The example in Figure 32 shows a synthetic image generated pipeline by combining an example face of the CK+ dataset and a background image.

Figure 32 – The pipeline of the synthetic image generation. The horizontal alignment of the image (b) is based on the inner points of the eyes (red points in (a)). The face is obtained as the convex hull of the landmarks set (c) and a random transform matrix is generated (d). The face image is projected on the background image (e). A face image and a general cropped background image are combined to generate a composite image (f). By using distinct background images for every face image, we are able to generate a much larger training data set. We create a large quantity of synthetic new images for every face of a database: approximately 9,231 synthetic images are generated for each face in the CK+ database, and 5,000 for the BU-3DFE database. This covers a great variety of possible tones and different backgrounds.



(a) (b) (c) (d) (e) (f)

Source: (FERNANDEZ et al., 2019)

## 6.4 EXPERIMENTS

We describe here the creation of the dataset used for training our network and its implementation details. We discuss two groups of experimental results: (1) Expression recognition result, to measure the performance of the method regarding the relevance of the attention module and the proposed loss function for attention, and (2) Correction result, to analyze the robustness to noise.

### 6.4.1 Datasets

We employ two public facial expression datasets, namely Extended Cohn-Kanade (CK+) (LUCEY et al., 2010) and Binghamton University 3D Facial Expression (BU-3DFE) (YIN et al., 2006) to evaluate our method. We apply in all experiments person-independent FER scenarios (ZENG et al., 2009). Subjects in the training set are completely different from the subjects in the test set, i.e., the subjects used for training are not used for testing. The CK+ dataset includes 593 image sequences from 123 subjects. We selected 325 sequences of 118 subjects from this set, which meet the criteria for one of the seven emotions (LUCEY et al., 2010). The selected 325 sequences consist of 45 Angry, 18 Contempt, 58 Disgust, 25 Fear, 69 Happy, 28 Sadness and 82 Surprise (LUCEY et al., 2010) facial expressions. In the neutral face expression case, we selected the first frame of the sequence of 33 random selected subjects. The BU-3DFE dataset is known to be challenging mainly due to a variety of ethnic/racial ancestries and expression intensity (YIN et al., 2006). A total of 600 expressive face images (1 intensity x 6 expressions x 100 subjects) and 100

Figure 33 – Examples from the synthetic BU-3DFE dataset. Different faces are transformed and combined with randomly selected background images from the COCO dataset. We then augment images after transformation by changing brightness and contrast and applying Gaussian blur and noise.

Source: (FERNANDEZ et al., 2019)

neutral face expression images, one for each subject, were used (YIN et al., 2006).

We employed our renderer $R$ to augment training data for the neural network. $R$ uses a facial expression dataset (we use BU-3DFE and CK+, which were segmented to obtain face masks) and a dataset of background images chosen from the COCO dataset. Figure 33 shows some examples of images generated by the renderer on the BU-3DFE dataset.

### 6.4.2 Implementation and training details

In all experiments, we considered the neural network architecture PreActResNet18 for the classification and representation processes. We adopted two approaches: (1) a model with attention and classification, FERAtt+Cls, and (2) a model with attention, classification, and representation, FERAtt+Rep+Cls. These models were compared with the classification results. For representation, the last convolutional layer of PreActResNet is evaluated by a linear layer to generate a vector of selected size. We have opted for 64 dimensions for the representation vector $\hat{z}$.

All models were trained on Nvidia GPUs (P100, K80, Titan XP) using PyTorch[1] for 60 epochs for the training set with 200 examples per mini batch and employing Adam optimizer. Face images were rescaled to 32×32 pixels. The code for the FERAtt is available in a public repository[2].

### 6.4.3 Expression recognition results

This set of experiments makes comparisons between a baseline architecture and the different variants of the proposed architecture. The objective is to evaluate the relevance of the attention module and the proposed loss function for attention.

[1] http://pytorch.org/

[2] https://github.com/pedrodiamel/ferattention

Table 17 – Classification results for the Synthetic/Real BU-3DFE database (6 expression + neutral) and CK+ database (7 expression classes + neutral). Baseline: PreActResNet18(HE et al., 2016), Acc.: Accuracy, Prec.: Precision, Rec.: Recall, F1: F1 measurement. Leave-10-subjects-out cross-validation is used for all experiments.

| | | Synthetic | | | | Real | | | |
|---|---|---|---|---|---|---|---|---|---|
| Database | Method | Acc. | Prec. | Rec. | F1 | Acc. | Prec. | Rec. | F1 |
| BU-3DFE | Baseline | 69.37<br>±2.84 | 71.48<br>±1.46 | 69.56<br>±2.76 | 70.50<br>±2.05 | 75.22<br>±4.60 | 77.58<br>±3.72 | 75.49<br>±4.68 | 76.52<br>±4.19 |
| | FERAtt+Cls | 75.15<br>±3.13 | 77.34<br>±1.40 | 75.45<br>±2.57 | 76.38<br>±1.98 | 80.41<br>±4.33 | 82.30<br>±2.99 | 80.79<br>±3.75 | 81.54<br>±3.38 |
| | FERAtt+Rep+Cls | **77.90**<br>±2.59 | **79.58**<br>±1.77 | **78.05**<br>±2.34 | **78.81**<br>±2.01 | **82.11**<br>±4.39 | **83.72**<br>±3.09 | **82.42**<br>±4.08 | **83.06**<br>±3.59 |
| CK+ | Baseline | 77.63<br>±2.11 | 68.42<br>±2.97 | 68.56<br>±1.91 | 68.49<br>±2.43 | 86.67<br>±3.15 | 81.62<br>±7.76 | 80.15<br>±9.50 | 80.87<br>±8.63 |
| | FERAtt+Cls | 84.60<br>±0.93 | 74.94<br>±0.38 | 76.30<br>±1.19 | 75.61<br>±0.76 | 85.42<br>±2.89 | 75.65<br>±2.77 | 78.79<br>±2.30 | 77.18<br>±2.55 |
| | FERAtt+Rep+Cls | **85.15**<br>±1.07 | **74.68**<br>±1.37 | **77.45**<br>±0.55 | **76.04**<br>±0.97 | **90.30**<br>±1.36 | **83.64**<br>±5.28 | **84.90**<br>±8.52 | **84.25**<br>±6.85 |

Source: (FERNANDEZ et al., 2019)

**Protocol**. We used different metrics to evaluate the proposed methods. Accuracy is calculated as the average number of successes divided by the total number of observations (in this case each face is considered an observation). Precision, recall, F1 score, and confusion matrix are also used in the analysis of the effectiveness of the system. Demšar (DEMŠAR, 2006) recommends the Friedman test followed by the pairwise Nemenyi test to compare multiple data. The Friedman test is a nonparametric alternative of the analysis of variance (ANOVA) test. The null hypothesis of the test $H_0$ stipulates that models are equivalent. Similar to the methods in (PTUCHA; SAVAKIS, 2013), Leave-10-subject-out (L-10-SO) cross-validation was adopted in the evaluation.

**Results**. Tables 17 shows the mean and standard deviation for the results obtained on the real and synthetic datasets. For the BU-3DFE database the Friedman nonparametric ANOVA test reveals significant differences ($p = 0.0498$) between the methods. The Nemenyi post-hoc test was applied to determine which method present significant differences. The result for the Nemenyi post-hoc test (two-tailed test) shows that there are significant differences between the FERAtt+Cls+Rep and all the others, for a significance level at $\alpha < 0.05$.

In the CK+ database case, the Friedman test found significant differences between the methods with a level of significance of $p = 0.0388$ for the Synthetic CK+ dataset and $p = 0.0381$ for Real CK+ dataset. In this case, we applied the Bonferroni-Dunn post-hoc test (one-tailed test) to strengthen the power of the hypotheses test. For a significance level

Table 18 – Comparison of the average recognition accuracy with state-of-the-art FER methods for the BU-3DFE database. NE: number of expressions, †: six basic expressions + neutral class. Leave-10-subjects-out cross-validation is used for all methods.

| Methods | Accuracy | NE |
|---|---|---|
| (LOPES et al., 2017) | 72.89 | 7† |
| (JAMPOUR; MAUTHNER; BISCHOF, 2015) | 78.64 | 7† |
| (ZHANG et al., 2016) | 80.10 | 7† |
| (ZHANG et al., 2018) | 80.95 | 7† |
| Our | **82.11** | 7† |

Source: (FERNANDEZ et al., 2019)

Table 19 – Comparison of the average recognition accuracy with state-of-the-art FER methods for the CK+ database. NE: number of expressions, †: six basic expressions + neutral class and contempt class, ‡: six basic expressions + contempt class (neutral is excluded). ∗: the value in parentheses is the mean accuracy, which is calculated with the confusion matrix given by the authors. Leave-10-subjects-out cross-validation is used for all methods.

| Methods | Accuracy* | NE |
|---|---|---|
| (MENG et al., 2017) | 95.37 | 7‡ |
| (ZENG et al., 2018) | 95.79 (93.78) | 7‡ |
| (YANG; CIFTCI; YIN, 2018) | 97.30 (96.57) | 7‡ |
| Our | **97.50** | 7‡ |
| (ZENG et al., 2018) | 89.84 (86.82) | 8† |
| Our | **90.30** | 8† |

Source: (FERNANDEZ et al., 2019)

of 0.05, the Bonferroni-Dunn post-hoc test did not show significant differences between the FERAtt+Cls and the Baseline for Synthetic CK+ with $p = 0.0216$. When considering FERAtt+Rep+Cls and Baseline methods, it shows significant differences for the Real CK+ dataset with $p = 0.0133$.

Table 18 and 19 show the comparisons results between the different FER methods for the BU-3DFE database (YANG; CIFTCI; YIN, 2018; LOPES et al., 2017; JAMPOUR; MAUTHNER; BISCHOF, 2015; ZHANG et al., 2016; ZHANG et al., 2018) and for the CK+ database (MENG et al., 2017; ZENG et al., 2018; YANG; CIFTCI; YIN, 2018). Although some results cannot be directly compared due to different experimental setups, different expression classes and different preprocessing methods (e.g. face alignment), it is demonstrated that the proposed method can yield a feasible and promising recognition rate (around 82.11 percent for the BU-3DFE database and 90.30 for the CK+ database) with static facial images under person-independent recognition scenario.

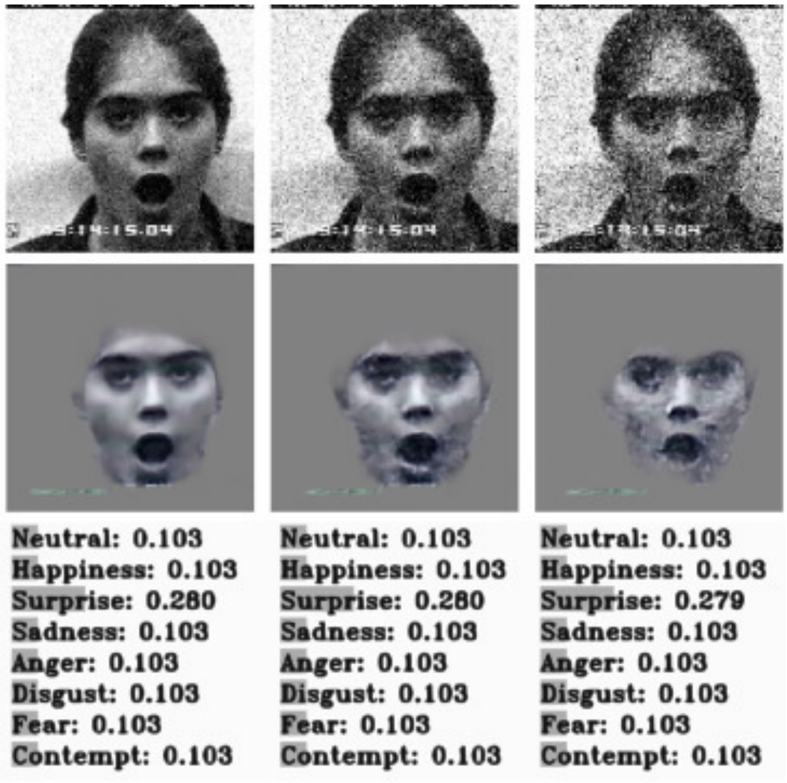
### 6.4.4 Robustness to noise

The objective of this set of experiments is to demonstrate the robustness of our method to the presence of image noise when compared to the baseline architecture PreActResNet18.

**Protocol**. To carry out this experiment, the Baseline, FERAtt+Class, and FERAtt+Rep+Class models were trained on the Synthetic CK+ dataset. Each of these models was readjusted with increasing noise in the training set ($\sigma \in [0.05, 0.30]$). We maintained the parameters in the training for fine-tuning and used the real database CK+, so that 2000 images were generated for the synthetic dataset for test.

**Results**. One of the advantages of the proposed approach is that we can evaluate the robustness of the method under different noise levels by visually assessing the changes in the attention map $I_{att}$. Figure 34 shows the attention maps for an image for white zero mean Gaussian noise levels $\sigma = [0.01, 0.05, 0.07, 0.09, 0.1, 0.2, 0.3]$. We observe that our network is quite robust to noise for the range of 0.01 to 0.1 and maintains a distribution of homogeneous intensity values. This aspect is beneficial to the subsequent performance of the classification module. Figures 35 and 36 present classification accuracy results of the evaluated models in the Real CK+ dataset and for 2000 synthetic images. The proposed method FERAtt+CLs+Rep provides the best classification in both cases.

Figure 34 – Attention maps $I_{att}$ under increasing noise levels. We progressively added higher levels (increasing variance $\sigma$) of zero mean white Gaussian noise to the same image and tested them using our model. The classification numbers above show the robustness of the proposed approach under different noise levels, $\sigma = 0.10, 0.20, 0.30$, where the Surprise and all other scores are mostly maintained throughout all levels, with only a minor change of the Surprise score, from 0.280 to 0.279, occurring for the highest noise contamination of $\sigma = 0.30$.



**Neutral: 0.103**
**Happiness: 0.103**
**Surprise: 0.280**
**Sadness: 0.103**
**Anger: 0.103**
**Disgust: 0.103**
**Fear: 0.103**
**Contempt: 0.103**

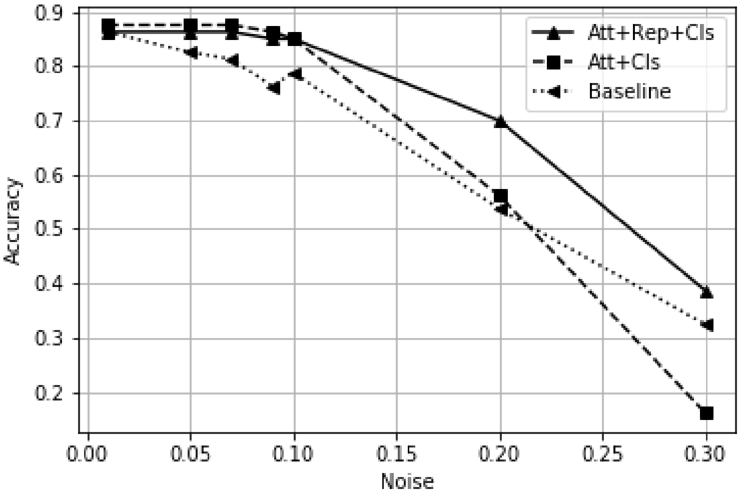
**Neutral: 0.103**
**Happiness: 0.103**
**Surprise: 0.280**
**Sadness: 0.103**
**Anger: 0.103**
**Disgust: 0.103**
**Fear: 0.103**
**Contempt: 0.103**

**Neutral: 0.103**
**Happiness: 0.103**
**Surprise: 0.279**
**Sadness: 0.103**
**Anger: 0.103**
**Disgust: 0.103**
**Fear: 0.103**
**Contempt: 0.103**

(a) $\sigma = 0.10$ (b) $\sigma = 0.20$ (c) $\sigma = 0.30$

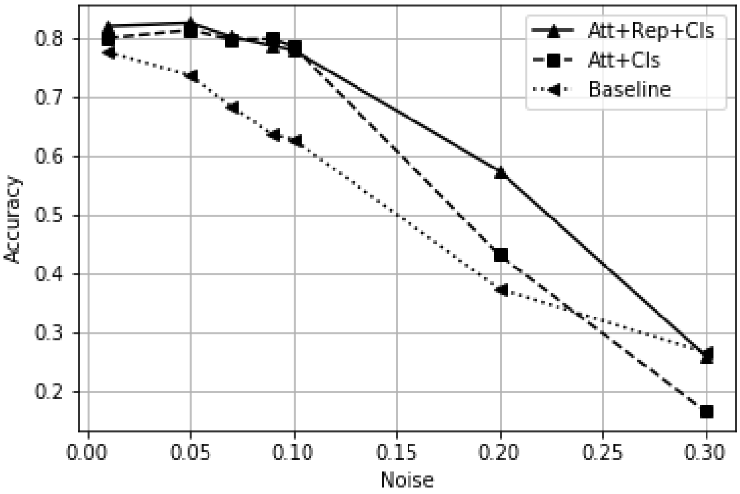
Source: (FERNANDEZ et al., 2019)

Figure 35 – Classification accuracy after adding incremental noise on the Real CK+ dataset. Our approach results in higher accuracy when compared to the baseline, specially for stronger noise levels. Our representation model clearly leverages results showing its importance for classification. Plotted values are the average results for all 325 images in the database.



Source: (FERNANDEZ et al., 2019)

Figure 36 – Average classification accuracy after adding incremental noise on the Synthetic CK+ dataset. The behavior of our method in the synthetic data replicates what we have found for the original Real CK+ database, *i.e.*, our method is superior to the baseline for all levels of noise. Plotted average values are for 2,000 synthetic images.



Source: (FERNANDEZ et al., 2019)

## 6.5 CONCLUSIONS

In this chapter, we present a new end-to-end neural network architecture with an attention model for facial expression recognition. We create a generator of synthetic images which is used for training our models. The results show that, for these experimental conditions, the attention module improves the system classification performance in comparison to other methods from the state-of-the-art. The loss function presented works as a regularization method on the embedded space.

# 7 GENERAL CONCLUSION

The main goal of this thesis was to develop a more accurate learning architecture for the facial expressions characterization. In this work, we propose FERAtt, a new deep architecture with attention net for facial expression recognition. This architecture provides a robust representation of the facial expression obtained through our Structured Gaussian Manifold Loss algorithm.

## 7.1 CONTRIBUTIONS

The main contributions of this dissertation are as follows. First, we present an analysis of the combination of feature engineering and classification methods based on SR applied to the problems of FER. Second, a study is made on the main neural network architectures applied to this problem. Third, we propose a new supervised and semi-supervised representation approach based on DML. Fourth, we offer a new synthetic generator to render face expressions. Last, we propose a new end-to-end network architecture for facial expression recognition with an attention model. A brief description of these contributions follows.

### 7.1.1 Combination of feature engineering methods based on SR

We proposed a FER system based on SR. We show that the combination of classifiers improves the performance of individual classifiers. The proposed methods (MB, MSVMP, and MSVML) are significantly superior to the others combination rules in FER problems. The regeneration method proposed in this work allowed the use of a trainable multiclassifier system.

### 7.1.2 Study on the main neural network architectures for FER

A comparative analysis allowed us to determine that there are two important reasons for the problem of the generalization of models: 1) Biases of the dataset, the datasets are captured with particular characteristics of illuminations and different types of sources of spontaneous emotion vs non-spontaneous; 2) Label subjectivity, where the intrinsic subjectivity of the emotion in the images that contain different emotional components and complex contexts causes the labelers to commit mistakes. We also show the difficulty of current systems to deal with sarcasm and irony. We suppose that some of the possible causes of the errors due to the artifacts in the images given to the network such as jewelry, hair,etc.

### 7.1.3 New representation approach based on DML

Two new methods of representation were presented in this work: 1) Deep Gaussian Mixture Subspace and 2) Structured Gaussian Manifold Loss. The first one obtains Gaussian representation of the expressions that allow classifying and grouping facial expressions. We show that the high quality of the vectors obtained in recovery tasks, which explains the quality of the obtained representation. In the second case, we provide a semi-supervised method for grouping the facial expressions. New subgroups of the same facial expression were obtained. In these subgroups, different types of characteristics that have to do with the appearance and, in some cases, with different artifacts such as glasses are observed.

### 7.1.4 New synthetic generator to render FE

We proposed a new image renderer $R$, for creating a larger synthetic dataset using real face datasets by making background changes and geometric transformations of the face images. In this way, it is possible to train classification systems under extreme global conditions.

### 7.1.5 New end-to-end attention network architecture for FER

We present a new end-to-end neural network architecture with an attention model for facial expression recognition. The results show that, for these experimental conditions, the attention module improves the system classification performance in comparison to other methods from the state-of-the-art. The Structured Gaussian Manifold loss function presented works as a regularization method on the embedded space.

## 7.2 FUTURE WORKS

This work addressed a number of existing issues in the field of facial expression recognition. It also points to certain areas which could benefit from further research.

Obtaining a high-quality representation allows us to use a recurrence approach for temporal analysis in videos. It would also allow the multimodal combination from different data sources with our classifier combination algorithm.

One of the challenges we have is to eliminate any method of preprocessing the image. The attention module could be the solution to this problem. We are adapting the attention module so that it is able to adapt to extreme conditions of lighting changes, misalignment of the facial image and occlusion.

We will evaluate the trained representation models in the proposed multiclassifier system. The differences between the representation spaces obtained could contribute to the diversity of the system and increase its performance.

We proposed two methods of data generating. The algorithm for regenerate datasets obtained new reconstructed data with low-level features. Our image renderer $R$ creates a larger synthetic dataset using real face datasets by making background changes and geometric transformations of face images. With the combination of these approaches, more robust and diverse datasets could be created.

## 7.3 SUMMARY OF PUBLICATIONS

Articles published during this thesis:

- Fernández, P. D. M., Peña, F. A. G., Ren, T. I., and Leandro, J. J. (2019). Fast and robust multiple ColorChecker detection using deep convolutional neural networks. *Image and Vision Computing*, 81, 15-24.
- Peña, F. A. G., Fernández, P. D. M., Ren, T. I., Leandro, J. J., and Nishihara, R. (2019). Burst ranking for blind multi-image deblurring. *In Transactions on Image Processing.*
- Marrero Fernandez, P. D., Guerrero Pena, F. A., Ren, T., and Cunha, A. (2019). FERAtt: Facial Expression Recognition With Attention Net. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops.*
- Guerrero-Pena, F. A., Fernandez, P. D. M., Ren, T. I., Yui, M., Rothenberg, E., and Cunha, A. (2018). Multiclass Weighted Loss for Instance Segmentation of Cluttered Cells, *2018 25th IEEE International Conference on Image Processing (ICIP)*, Athens, Greece, pp. 2451-2455.
- Oliveros, E. R., Coello, G., Fernandez, P. D., Buades, J. M., Jaume-i-Capó A,. (2016). Evaluation of K-SVD method in facial expression recognition based on sparse representation problems. *In International Conference on Articulated Motion and Deformable Objects*, (pp. 135-146). Springer, Cham.

Article in preparation:

- Fernández, P. D. M., Jaume-i-Capó A., Buades-Rubio J. M., Ren, T. I.,Facial Expression Recognition with Regenerate Datasets. *Journal of Visual Communication and Image Representation.* (status: under review)
- Fernandez, P. D. M., Guerrero Pena, F. A., Ren, T., and Cunha, A. Deep Metric Structured Learning For Facial Expression. *In Transactions on Image Processing.* (status: on submission)

# REFERENCES

ASHRAF, A. B.; LUCEY, S.; COHN, J. F.; CHEN, T.; AMBADAR, Z.; PRKACHIN, K. M.; SOLOMON, P. E. The painful face–pain expression recognition using active appearance models. *Image and vision computing*, Elsevier, v. 27, n. 12, p. 1788–1796, 2009.

BALNTAS, V.; RIBA, E.; PONSA, D.; MIKOLAJCZYK, K. Learning local feature descriptors with triplets and shallow convolutional neural networks. In: *BMVC*. [S.l.: s.n.], 2016. v. 1, n. 2, p. 3.

BALTRUŠAITIS, T.; BANDA, N.; ROBINSON, P. Dimensional affect recognition using continuous conditional random fields. In: IEEE. *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*. [S.l.], 2013. p. 1–8.

BALTRUŠAITIS, T.; ROBINSON, P.; MORENCY, L.-P. OpenFace: an open source facial behavior analysis toolkit. In: *IEEE Winter Conference on Applications of Computer Vision*. [S.l.: s.n.], 2016.

BARON-COHEN, S. Reading the mind in the face: A cross-cultural and developmental study. *Visual Cognition*, Taylor & Francis, v. 3, n. 1, p. 39–60, 1996.

BARON-COHEN, S. *Mind Reading: The Interactive Guide to Emotions*. Jessica Kingsley Publishers, 2004. ISBN 9781843102151. Disponível em: <https://books.google.com.br/books?id=aWyTAAAACAAJ>.

BARSOUM, E.; ZHANG, C.; FERRER, C. C.; ZHANG, Z. Training deep networks for facial expression recognition with crowd-sourced label distribution. In: *ACM International Conference on Multimodal Interaction (ICMI)*. [S.l.: s.n.], 2016.

BARSOUM, E.; ZHANG, C.; FERRER, C. C.; ZHANG, Z. Training deep networks for facial expression recognition with crowd-sourced label distribution. In: *ACM International Conference on Multimodal Interaction (ICMI)*. [S.l.: s.n.], 2016.

BELL, S.; BALA, K. Learning visual similarity for product design with convolutional neural networks. *ACM Transactions on Graphics (TOG)*, ACM, v. 34, n. 4, p. 98, 2015.

BROMLEY, J.; GUYON, I.; LECUN, Y.; SÄCKINGER, E.; SHAH, R. Signature verification using a" siamese" time delay neural network. In: *Advances in Neural Information Processing Systems*. [S.l.: s.n.], 1994. p. 737–744.

BURKERT, P.; TRIER, F.; AFZAL, M. Z.; DENGEL, A.; LIWICKI, M. Dexpression: Deep convolutional neural network for expression recognition. *arXiv preprint arXiv:1509.05371*, 2015.

CHECHIK, G.; SHARMA, V.; SHALIT, U.; BENGIO, S. Large scale online learning of image similarity through ranking. *Journal of Machine Learning Research*, v. 11, n. Mar, p. 1109–1135, 2010.

CHEUNG, B.; WEISS, E.; OLSHAUSEN, B. Emergence of foveal image sampling from learning to attend in visual scenes. *arXiv preprint arXiv:1611.09430*, 2016.

CORNEANU, C. A.; SIMÓN, M. O.; COHN, J. F.; GUERRERO, S. E. 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*, IEEE, v. 38, n. 8, p. 1548–1568, 2016.

CUI, Y.; ZHOU, F.; LIN, Y.; BELONGIE, S. Fine-grained categorization and dataset bootstrapping using deep metric learning with humans in the loop. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. [S.l.: s.n.], 2016. p. 1153–1162.

DALAL, N.; TRIGGS, B. Histograms of oriented gradients for human detection. In: IEEE. *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*. [S.l.], 2005. v. 1, p. 886–893.

DARWIN, C. *The expression of emotion in man and animals*. [S.l.]: Oxford University Press, USA, 1872.

DAYAN, P.; ABBOTT, L. et al. Theoretical neuroscience: computational and mathematical modeling of neural systems. *Journal of Cognitive Neuroscience*, v. 15, n. 1, p. 154–155, 2003.

DEMŠAR, J. Statistical comparisons of classifiers over multiple data sets. *The Journal of Machine Learning Research*, JMLR. org, v. 7, p. 1–30, 2006.

D'MELLO, S.; CALVO, R. A. Beyond the basic emotions: what should affective computing compute? In: ACM. *CHI'13 Extended Abstracts on Human Factors in Computing Systems*. [S.l.], 2013. p. 2287–2294.

DONOHO, D. L. For most large underdetermined systems of linear equations the minimal ?1-norm solution is also the sparsest solution. *Communications on pure and applied mathematics*, Wiley Online Library, v. 59, n. 6, p. 797–829, 2006.

DOSOVITSKIY, A.; FISCHER, P.; SPRINGENBERG, J. T.; RIEDMILLER, M.; BROX, T. Discriminative unsupervised feature learning with exemplar convolutional neural networks. *IEEE transactions on pattern analysis and machine intelligence*, IEEE, v. 38, n. 9, p. 1734–1747, 2016.

DUCHENNE, G.-B. *The mechanism of human facial expression*. [S.l.]: Cambridge university press, 1990.

DUNNE, C. *Microsoft's New Emotion-Detecting App Deems the Mona Lisa 43% Happy*. 2015. Disponível em: <https://hyperallergic.com/261508/microsofts-new-emotion-detecting-app-deems-the-mona-lisa-43-happy/>.

EKMAN, P. Universals and cultural differences in facial expressions of emotion. In: UNIVERSITY OF NEBRASKA PRESS. *Nebraska symposium on motivation*. [S.l.], 1971.

EKMAN, P.; FRIESEN, W. V.; ELLSWORTH, P. *Emotion in the Human Face*. [S.l.]: Cambridge University Press, 1982.

EKMAN, P.; FRIESEN, W. V.; J, H. *Facial action coding system*. [S.l.]: A Human Face, 2002.

EKMAN, P.; OSTER, H. Facial expressions of emotion. *Annual review of psychology*, Annual Reviews 4139 El Camino Way, PO Box 10139, Palo Alto, CA 94303-0139, USA, v. 30, n. 1, p. 527–554, 1979.

ELWELL, R.; POLIKAR, R. Incremental learning of concept drift in nonstationary environments. *IEEE Transactions on Neural Networks*, IEEE, v. 22, n. 10, p. 1517–1531, 2011.

ERDOGAN, H.; SEN, M. U. A unifying framework for learning the linear combiners for classifier ensembles. In: IEEE. *Pattern Recognition (ICPR), 2010 20th International Conference on*. [S.l.], 2010. p. 2985–2988.

ESLAMI, S. M. A.; HEESS, N.; WEBER, T.; TASSA, Y.; SZEPESVARI, D.; KAVUKCUOGLU, k.; HINTON, G. E. Attend, infer, repeat: Fast scene understanding with generative models. In: LEE, D. D.; SUGIYAMA, M.; LUXBURG, U. V.; GUYON, I.; GARNETT, R. (Ed.). *Advances in Neural Information Processing Systems 29*. Curran Associates, Inc., 2016. p. 3225–3233. Disponível em: <http://papers.nips.cc/paper/6230-attend-infer-repeat-fast-scene-understanding-with-generative-models.pdf>.

FERNÁNDEZ, P. D. M.; PEÑA, F. A. G.; REN, T. I.; LEANDRO, J. J. Fast and robust multiple colorchecker detection using deep convolutional neural networks. *Image and Vision Computing*, Elsevier, v. 81, p. 15–24, 2019.

FERNANDEZ, P. D. M.; PENA, F. A. G.; REN, T. I.; CUNHA, A. Feratt: Facial expression recognition with attention net. In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. [S.l.: s.n.], 2019.

FONTAINE, J. R.; SCHERER, K. R.; ROESCH, E. B.; ELLSWORTH, P. C. The world of emotions is not two-dimensional. *Psychological science*, Sage Publications Sage CA: Los Angeles, CA, v. 18, n. 12, p. 1050–1057, 2007.

FROME, A.; SINGER, Y.; MALIK, J. Image retrieval and classification using local distance functions. In: *Advances in neural information processing systems*. [S.l.: s.n.], 2007. p. 417–424.

GOODFELLOW, I. J.; ERHAN, D.; CARRIER, P. L.; COURVILLE, A.; MIRZA, M.; HAMNER, B.; CUKIERSKI, W.; TANG, Y.; THALER, D.; LEE, D.-H. et al. Challenges in representation learning: A report on three machine learning contests. In: SPRINGER. *International Conference on Neural Information Processing*. [S.l.], 2013. p. 117–124.

GRAHAM, B. Fractional max-pooling. *arXiv preprint arXiv:1412.6071*, 2014.

GRAVES, A.; WAYNE, G.; REYNOLDS, M.; HARLEY, T.; DANIHELKA, I.; GRABSKA-BARWIŃSKA, A.; COLMENAREJO, S. G.; GREFENSTETTE, E.; RAMALHO, T.; AGAPIOU, J. et al. Hybrid computing using a neural network with dynamic external memory. *Nature*, Nature Publishing Group, v. 538, n. 7626, p. 471, 2016.

GREENBLATT, S. et al. Toward a universal language of motion: Reflections on a seventeenth century muscle man. *LiNQ*, James Cook University, Department of Humanities, School of Arts and Social Sciences, v. 21, n. 2, p. 56, 1994.

GREGOR, K.; DANIHELKA, I.; GRAVES, A.; REZENDE, D. J.; WIERSTRA, D. Draw: A recurrent neural network for image generation. *arXiv preprint arXiv:1502.04623*, 2015.

GUNES, H.; SCHULLER, B. Categorical and dimensional affect analysis in continuous input: Current trends and future directions. *Image and Vision Computing*, Elsevier, v. 31, n. 2, p. 120–136, 2013.

HADSELL, R.; CHOPRA, S.; LECUN, Y. Dimensionality reduction by learning an invariant mapping. In: IEEE. *Computer vision and pattern recognition, 2006 IEEE computer society conference on*. [S.l.], 2006. v. 2, p. 1735–1742.

HAPPY, S. L.; ROUTRAY, A. Automatic Facial Expression Recognition Using Features of Salient Facial Patches. *Affective Computing, IEEE Transactions on*, IEEE, v. 6, n. 1, p. 1–12, 2015.

HE, K.; GKIOXARI, G.; DOLLÁR, P.; GIRSHICK, R. Mask r-cnn. In: IEEE. *Computer Vision (ICCV), 2017 IEEE International Conference on*. [S.l.], 2017. p. 2980–2988.

HE, K.; ZHANG, X.; REN, S.; SUN, J. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015. Disponível em: <http://arxiv.org/abs/1512.03385>.

HE, K.; ZHANG, X.; REN, S.; SUN, J. Identity mappings in deep residual networks. In: SPRINGER. *European conference on computer vision*. [S.l.], 2016. p. 630–645.

HIGHFIELD, R.; WISEMAN, R.; JENKINS, R. How your looks betray your personality. *New Scientist*, v. 201, n. 2695, p. 28–32, 2009.

HOFFER, E.; AILON, N. Deep metric learning using triplet network. In: SPRINGER. *International Workshop on Similarity-Based Pattern Recognition*. [S.l.], 2015. p. 84–92.

HUANG, M.-W.; WANG, Z.-w.; YING, Z.-L. A new method for facial expression recognition based on sparse representation plus LBP. In: IEEE. *Image and Signal Processing (CISP), 2010 3rd International Congress on*. [S.l.], 2010. v. 4, p. 1750–1754.

HUANG, X.; ZHAO, G.; ZHENG, W.; PIETIKÄINEN, M. Spatiotemporal local monogenic binary patterns for facial expression recognition. *Signal Processing Letters, IEEE*, IEEE, v. 19, n. 5, p. 243–246, 2012.

IMBRASAITĖ, V.; BALTRUŠAITIS, T.; ROBINSON, P. Emotion tracking in music using continuous conditional random fields and relative feature representation. In: IEEE. *Multimedia and Expo Workshops (ICMEW), 2013 IEEE International Conference on*. [S.l.], 2013. p. 1–6.

ITTI, L.; KOCH, C. Computational modelling of visual attention. *Nature Reviews Neuroscience*, Nature Publishing Group, v. 2, n. 3, p. 194, 2001.

IZARD, C. *The face of emotion*. Appleton-Century-Crofts, 1971. (Century psychology series). Disponível em: <https://books.google.com.br/books?id=7DQNAQAAMAAJ>.

JACOBS, R. A. Methods for combining experts' probability assessments. *Neural computation*, MIT Press, v. 7, n. 5, p. 867–888, 1995.

JADERBERG, M.; SIMONYAN, K.; ZISSERMAN, A.; KAVUKCUOGLU, k. Spatial transformer networks. In: CORTES, C.; LAWRENCE, N. D.; LEE, D. D.; SUGIYAMA, M.; GARNETT, R. (Ed.). *Advances in Neural Information Processing Systems 28*. Curran Associates, Inc., 2015. p. 2017–2025. Disponível em: <http://papers.nips.cc/paper/5854-spatial-transformer-networks.pdf>.

JAMPOUR, M.; MAUTHNER, T.; BISCHOF, H. Multi-view facial expressions recognition using local linear regression of sparse codes. In: *Proceedings of the 20th Computer Vision Winter Workshop Paul Wohlhart*. [S.l.: s.n.], 2015.

JEGOU, H.; DOUZE, M.; SCHMID, C. Product quantization for nearest neighbor search. *IEEE transactions on pattern analysis and machine intelligence*, IEEE, v. 33, n. 1, p. 117–128, 2011.

JI, Y.; IDRISSI, K. Automatic facial expression recognition based on spatiotemporal descriptors. *Pattern Recognition Letters*, Elsevier, v. 33, n. 10, p. 1373–1380, 2012.

JORDAN, M. I.; JACOBS, R. A. Hierarchical mixtures of experts and the EM algorithm. *Neural computation*, MIT Press, v. 6, n. 2, p. 181–214, 1994.

JUNG, H.; LEE, S.; YIM, J.; PARK, S.; KIM, J. Joint fine-tuning in deep neural networks for facial expression recognition. In: *Proceedings of the IEEE International Conference on Computer Vision*. [S.l.: s.n.], 2015. p. 2983–2991.

JUSLIN, P. N.; SCHERER, K. R. Vocal expression of affect. *The new handbook of methods in nonverbal behavior research*, New York: Oxford University, p. 65–135, 2005.

KAHOU, S. E.; PAL, C.; BOUTHILLIER, X.; FROUMENTY, P.; GULCEHRE, C.; MEMISEVIC, R.; VINCENT, P.; COURVILLE, A.; BENGIO, Y.; FERRARI, R. C.; MIRZA, M.; JEAN, S.; CARRIER, P.-L.; DAUPHIN, Y.; BOULANGER-LEWANDOWSKI, N.; AGGARWAL, A.; ZUMER, J.; LAMBLIN, P.; RAYMOND, J.-P.; DESJARDINS, G.; PASCANU, R.; WARDE-FARLEY, D.; TORABI, A.; SHARMA, A.; BENGIO, E.; COTE, M.; KONDA, K. R.; WU, Z. Combining modality specific deep neural networks for emotion recognition in video. In: *Proceedings of the 15th ACM on International Conference on Multimodal Interaction*. New York, NY, USA: ACM, 2013. (ICMI '13), p. 543–550. ISBN 978-1-4503-2129-7. Disponível em: <http://doi.acm.org/10.1145/2522848.2531745>.

KALIOUBY, R. E.; ROBINSON, P. Real-time inference of complex mental states from facial expressions and head gestures. In: *Real-time vision for human-computer interaction*. [S.l.]: Springer, 2005. p. 181–200.

KELLER, J.; KRISNAPURAM, R.; PAL, N. R. *Fuzzy models and algorithms for pattern recognition and image processing*. [S.l.]: Springer Science & Business Media, 2005. v. 4.

KHORRAMI, P.; PAINE, T.; HUANG, T. Do deep neural networks learn facial action units when doing expression recognition? In: *Proceedings of the IEEE International Conference on Computer Vision Workshops*. [S.l.: s.n.], 2015. p. 19–27.

KINGMA, D. P.; BA, J. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014. Disponível em: <http://arxiv.org/abs/1412.6980>.

KITTLER, J.; HATEF, M.; DUIN, R. P. W.; MATAS, J. On combining classifiers. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, IEEE, v. 20, n. 3, p. 226–239, 1998.

KRIZHEVSKY, A. One weird trick for parallelizing convolutional neural networks. *CoRR*, abs/1404.5997, 2014. Disponível em: <http://arxiv.org/abs/1404.5997>.

KRIZHEVSKY, A.; SUTSKEVER, I.; HINTON, G. E. Imagenet classification with deep convolutional neural networks. In: PEREIRA, F.; BURGES, C. J. C.; BOTTOU, L.; WEINBERGER, K. Q. (Ed.). *Advances in Neural Information Processing Systems 25*. Curran Associates, Inc., 2012. p. 1097–1105. Disponível em: <http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf>.

KUNCHEVA, L. I. *Combining pattern classifiers: methods and algorithms*. [S.l.]: John Wiley & Sons, 2004.

KUNCHEVA, L. I.; BEZDEK, J. C.; DUIN, R. P. W. Decision templates for multiple classifier fusion: an experimental comparison. *Pattern recognition*, Elsevier, v. 34, n. 2, p. 299–314, 2001.

KUNCHEVA, L. I.; RODRIGUEZ, J. J. A weighted voting framework for classifiers ensembles. *Knowledge and Information Systems*, Springer, v. 38, n. 2, p. 259–275, 2014.

LEE, S. H.; PLATANIOTIS, K.; KONSTANTINOS, N.; RO, Y. M. Intra-Class Variation Reduction Using Training Expression Images for Sparse Representation Based Facial Expression Recognition. *Affective Computing, IEEE Transactions on*, IEEE, v. 5, n. 3, p. 340–351, 2014.

LI, L.; YING, Z.; YANG, T. Facial expression recognition by fusion of gabor texture features and local phase quantization. In: IEEE. *Signal Processing (ICSP), 2014 12th International Conference on*. [S.l.], 2014. p. 1781–1784.

LI, W.; LI, M.; SU, Z.; ZHU, Z. A deep-learning approach to facial expression recognition with candid images. In: IEEE. *Machine Vision Applications (MVA), 2015 14th IAPR International Conference on*. [S.l.], 2015. p. 279–282.

LI, X.; HONG, X.; MOILANEN, A.; HUANG, X.; PFISTER, T.; ZHAO, G.; PIETIKÄINEN, M. Reading Hidden Emotions: Spontaneous Micro-expression Spotting and Recognition. *arXiv preprint arXiv:1511.00423*, 2015.

LI, Y.; HUANG, J.-B.; AHUJA, N.; YANG, M.-H. Deep joint image filtering. In: SPRINGER. *European Conference on Computer Vision*. [S.l.], 2016. p. 154–169.

LI, Y.; WANG, S.; ZHAO, Y.; JI, Q. Simultaneous facial feature tracking and facial expression recognition. *Image Processing, IEEE Transactions on*, IEEE, v. 22, n. 7, p. 2559–2573, 2013.

LIM, B.; SON, S.; KIM, H.; NAH, S.; LEE, K. M. Enhanced deep residual networks for single image super-resolution. In: *The IEEE conference on computer vision and pattern recognition (CVPR) workshops*. [S.l.: s.n.], 2017. v. 1, n. 2, p. 4.

LITTLEWORT, G.; BARTLETT, M. S.; FASEL, I.; SUSSKIND, J.; MOVELLAN, J. Dynamics of facial expression extracted automatically from video. *Image and Vision Computing*, Elsevier, v. 24, n. 6, p. 615–625, 2006.

LIU, M.; LI, S.; SHAN, S.; WANG, R.; CHEN, X. Deeply learning deformable facial action parts model for dynamic expression analysis. In: SPRINGER. *Asian conference on computer vision*. [S.l.], 2014. p. 143–157.

LIU, M.; WANG, R.; LI, S.; SHAN, S.; HUANG, Z.; CHEN, X. Combining multiple kernel methods on riemannian manifold for emotion recognition in the wild. In: ACM. *Proceedings of the 16th International Conference on Multimodal Interaction*. [S.l.], 2014. p. 494–501.

LIU, P.; HAN, S.; MENG, Z.; TONG, Y. Facial expression recognition via a boosted deep belief network. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. [S.l.: s.n.], 2014. p. 1805–1812.

LIU, Z.; LUO, P.; WANG, X.; TANG, X. Deep learning face attributes in the wild. In: *Proceedings of the IEEE international conference on computer vision*. [S.l.: s.n.], 2015. p. 3730–3738.

LOPES, A. T.; AGUIAR, E. de; SOUZA, A. F. D.; OLIVEIRA-SANTOS, T. Facial expression recognition with convolutional neural networks: coping with few data and the training sample order. *Pattern Recognition*, Elsevier, v. 61, p. 610–628, 2017.

LUCEY, P.; COHN, J. F.; KANADE, T.; SARAGIH, J.; AMBADAR, Z.; MATTHEWS, I. The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression. In: *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*. [S.l.]: IEEE, 2010. p. 94–101.

LYONS, M.; AKAMATSU, S.; KAMACHI, M.; GYOBA, J. Coding facial expressions with gabor wavelets. In: IEEE. *Automatic Face and Gesture Recognition, 1998. Proceedings. Third IEEE International Conference on*. [S.l.], 1998. p. 200–205.

MAATEN, L. V. D. Accelerating t-sne using tree-based algorithms. *Journal of machine learning research*, v. 15, n. 1, p. 3221–3245, 2014.

MAHMOUD, M.; BALTRUŠAITIS, T.; ROBINSON, P.; RIEK, L. D. 3d corpus of spontaneous complex mental states. In: SPRINGER. *International Conference on Affective Computing and Intelligent Interaction*. [S.l.], 2011. p. 205–214.

MANNING, C. D.; RAGHAVAN, P.; SCHÜTZE, H. et al. *Introduction to information retrieval*. [S.l.]: Cambridge university press Cambridge, 2008. v. 1.

MARRERO-FERNÁNDEZ, P.; MONTOYA-PADRÓN, A.; CAPÓ, A. Jaume-i; RUBIO, J. M. B. Evaluating the research in automatic emotion recognition. *IETE Technical Review*, Taylor & Francis, v. 31, n. 3, p. 220–232, 2014.

MARRERO-FERNÁNDEZ, P.; MONTOYA-PADRÓN, A.; JAUME-I-CAPÓ, A.; RUBIO, J. M. B. Evaluating the research in automatic emotion recognition. *IETE Technical Review (Institution of Electronics and Telecommunication Engineers, India)*, Medknow Publications, v. 31, n. 3, p. 220–232, 2014. ISSN 09745971.

MENG, Z.; LIU, P.; CAI, J.; HAN, S.; TONG, Y. Identity-aware convolutional neural network for facial expression recognition. In: IEEE. *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*. [S.l.], 2017. p. 558–565.

MNIH, V.; HEESS, N.; GRAVES, A.; KAVUKCUOGLU, k. Recurrent models of visual attention. In: GHAHRAMANI, Z.; WELLING, M.; CORTES, C.; LAWRENCE, N. D.; WEINBERGER, K. Q. (Ed.). *Advances in Neural Information Processing Systems 27*. Curran Associates, Inc., 2014. p. 2204–2212. Disponível em: <http://papers.nips.cc/paper/5542-recurrent-models-of-visual-attention.pdf>.

MOLLAHOSSEINI, A.; CHAN, D.; MAHOOR, M. H. Going deeper in facial expression recognition using deep neural networks. In: IEEE. *Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on*. [S.l.], 2016. p. 1–10.

MOLLAHOSSEINI, A.; HASANI, B.; MAHOOR, M. H. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, IEEE, 2017.

NICOLLE, J.; RAPP, V.; BAILLY, K.; PREVOST, L.; CHETOUANI, M. Robust continuous prediction of human emotions using multiscale dynamic cues. In: ACM. *Proceedings of the 14th ACM international conference on Multimodal interaction*. [S.l.], 2012. p. 501–508.

OLSHAUSEN, B. A.; FIELD, D. J. Sparse coding with an overcomplete basis set: A strategy employed by V1? *Vision Research*, v. 37, n. 23, p. 3311–3325, 1997. ISSN 0042-6989. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0042698997001697>.

OLSHAUSEN, B. A.; others. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, v. 381, n. 6583, p. 607–609, 1996.

OUYANG, Y.; SANG, N.; HUANG, R. Robust automatic facial expression detection method based on sparse representation plus LBP map. *Optik-International Journal for Light and Electron Optics*, Elsevier, v. 124, n. 24, p. 6827–6833, 2013.

OUYANG, Y.; SANG, N.; HUANG, R. Accurate and robust facial expressions recognition by fusing multiple sparse representation based classifiers. *Neurocomputing*, Elsevier, v. 149, p. 71–78, 2015.

PAN, J.; LIU, S.; SUN, D.; ZHANG, J.; LIU, Y.; REN, J.; LI, Z.; TANG, J.; LU, H.; TAI, Y.-W. et al. Learning dual convolutional neural networks for low-level vision. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. [S.l.: s.n.], 2018. p. 3070–3079.

PARKHI, O. M.; VEDALDI, A.; ZISSERMAN, A. Deep Face Recognition. In: *BMVC*. [S.l.: s.n.], 2015. v. 1, n. 3, p. 6.

PICARD, R. *Affective Computing*. MIT Press, 2000. ISBN 9780262661157. Disponível em: <https://books.google.com.br/books?id=GaVncRTcb1gC>.

PTUCHA, R.; SAVAKIS, A. Fusion of static and temporal predictors for unconstrained facial expression recognition. In: IEEE. *Image Processing (ICIP), 2012 19th IEEE International Conference on*. [S.l.], 2012. p. 2597–2600.

PTUCHA, R.; SAVAKIS, A. Manifold based sparse representation for facial understanding in natural images. *Image and Vision Computing*, Elsevier, v. 31, n. 5, p. 365–378, 2013.

RONNEBERGER, O.; P.FISCHER; BROX, T. U-net: Convolutional networks for biomedical image segmentation. In: *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. Springer, 2015. (LNCS, v. 9351), p. 234–241. (available on arXiv:1505.04597 [cs.CV]). Disponível em: <http://lmb.informatik.uni-freiburg.de/Publications/2015/RFB15a>.

ROZIN, P.; COHEN, A. B. High frequency of facial expressions corresponding to confusion, concentration, and worry in an analysis of naturally occurring facial expressions of americans. *Emotion*, American Psychological Association, v. 3, n. 1, p. 68, 2003.

RUSSAKOVSKY, O.; DENG, J.; SU, H.; KRAUSE, J.; SATHEESH, S.; MA, S.; HUANG, Z.; KARPATHY, A.; KHOSLA, A.; BERNSTEIN, M.; BERG, A. C.; FEI-FEI, L. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, v. 115, n. 3, p. 211–252, Dec 2015. ISSN 1573-1405. Disponível em: <https://doi.org/10.1007/s11263-015-0816-y>.

RUSSELL, J. A.; MEHRABIAN, A. Evidence for a three-factor theory of emotions. *Journal of research in Personality*, Elsevier, v. 11, n. 3, p. 273–294, 1977.

SCHROFF, F.; KALENICHENKO, D.; PHILBIN, J. Facenet: A unified embedding for face recognition and clustering. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. [S.l.: s.n.], 2015. p. 815–823.

SCHULLER, B.; VALSTAR, M.; EYBEN, F.; MCKEOWN, G.; COWIE, R.; PANTIC, M. Avec 2011–the first international audio/visual emotion challenge. In: *Affective Computing and Intelligent Interaction*. [S.l.]: Springer, 2011. p. 415–424.

SHI, H.; YANG, Y.; ZHU, X.; LIAO, S.; LEI, Z.; ZHENG, W.; LI, S. Z. Embedding deep metric for person re-identification: A study against large variations. In: SPRINGER. *European Conference on Computer Vision*. [S.l.], 2016. p. 732–748.

SHOCHER, A.; COHEN, N.; IRANI, M. Zero-shot” super-resolution using deep internal learning. In: *Conference on computer vision and pattern recognition (CVPR)*. [S.l.: s.n.], 2018.

SIMO-SERRA, E.; TRULLS, E.; FERRAZ, L.; KOKKINOS, I.; FUA, P.; MORENO-NOGUER, F. Discriminative learning of deep convolutional feature point descriptors. In: IEEE. *Computer Vision (ICCV), 2015 IEEE International Conference on*. [S.l.], 2015. p. 118–126.

SIMONYAN, K.; ZISSERMAN, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *CoRR*, abs/1409.1, 2014.

SIMONYAN, K.; ZISSERMAN, A. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014. Disponível em: <http://arxiv.org/abs/1409.1556>.

SOBOL-SHIKLER, T.; ROBINSON, P. Classification of complex information: Inference of co-occurring affective states from their expressions in speech. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, IEEE, v. 32, n. 7, p. 1284–1297, 2010.

SONG, H. O.; XIANG, Y.; JEGELKA, S.; SAVARESE, S. Deep metric learning via lifted structured feature embedding. In: IEEE. *Computer Vision and Pattern Recognition (CVPR), 2016 IEEE Conference on*. [S.l.], 2016. p. 4004–4012.

SONG, H. O.; XIANG, Y.; JEGELKA, S.; SAVARESE, S. Deep metric learning via lifted structured feature embedding. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. [S.l.: s.n.], 2016. p. 4004–4012.

SZEGEDY, C.; LIU, W.; JIA, Y.; SERMANET, P.; REED, S.; ANGUELOV, D.; ERHAN, D.; VANHOUCKE, V.; RABINOVICH, A. Going deeper with convolutions. In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. [S.l.: s.n.], 2015.

TANG, Y. Deep learning using linear support vector machines. *arXiv preprint arXiv:1306.0239*, 2013.

TSALAKANIDOU, F.; MALASSIOTIS, S. Real-time 2D+ 3D facial action and expression recognition. *Pattern Recognition*, Elsevier, v. 43, n. 5, p. 1763–1775, 2010.

UNGERLEIDER, S. K.; G, L. Mechanisms of visual attention in the human cortex. *Annual review of neuroscience*, Annual Reviews 4139 El Camino Way, PO Box 10139, Palo Alto, CA 94303-0139, USA, v. 23, n. 1, p. 315–341, 2000.

VEMULAPALLI, R.; AGARWALA, A. A compact embedding for facial expression similarity. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. [S.l.: s.n.], 2019. p. 5683–5692.

VINCI, L. D. *Leonardo on Art and the Artist*. [S.l.]: Courier Corporation, 2002.

VINYALS, O.; KAISER, L. u.; KOO, T.; PETROV, S.; SUTSKEVER, I.; HINTON, G. Grammar as a foreign language. In: CORTES, C.; LAWRENCE, N. D.; LEE, D. D.; SUGIYAMA, M.; GARNETT, R. (Ed.). *Advances in Neural Information Processing Systems 28*. Curran Associates, Inc., 2015. p. 2773–2781. Disponível em: <http://papers.nips.cc/paper/5635-grammar-as-a-foreign-language.pdf>.

WANG, F.; JIANG, M.; QIAN, C.; YANG, S.; LI, C.; ZHANG, H.; WANG, X.; TANG, X. Residual attention network for image classification. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. [S.l.: s.n.], 2017. p. 3156–3164.

WANG, J.; LEUNG, T.; ROSENBERG, C.; WANG, J.; PHILBIN, J.; CHEN, B.; WU, Y. et al. Learning fine-grained image similarity with deep ranking. *arXiv preprint arXiv:1404.4661*, 2014.

WANG, J.; ZHOU, F.; WEN, S.; LIU, X.; LIN, Y. Deep metric learning with angular loss. In: *Proceedings of the IEEE International Conference on Computer Vision*. [S.l.: s.n.], 2017. p. 2593–2601.

WEIFENG, L.; CAIFENG, S.; YANJIANG, W. Facial expression analysis using a sparse representation based space model. In: IEEE. *Signal Processing (ICSP), 2012 IEEE 11th International Conference on*. [S.l.], 2012. v. 3, p. 1659–1662.

WEINBERGER, K. Q.; SAUL, L. K. Distance metric learning for large margin nearest neighbor classification. *Journal of Machine Learning Research*, v. 10, n. Feb, p. 207–244, 2009.

WRIGHT, J.; YANG, A. Y.; GANESH, A.; SASTRY, S. S.; MA, Y. Robust face recognition via sparse representation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, IEEE, v. 31, n. 2, p. 210–227, 2009.

YANG, H.; CIFTCI, U.; YIN, L. Facial expression recognition by de-expression residue learning. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. [S.l.: s.n.], 2018. p. 2168–2177.

YIN, L.; WEI, X.; SUN, Y.; WANG, J.; ROSATO, M. J. A 3D facial expression database for facial behavior research. In: IEEE. *Automatic face and gesture recognition, 2006. FGR 2006. 7th international conference on*. [S.l.], 2006. p. 211–216.

YING, Z.-L.; WANG, Z.-W.; HUANG, M.-W. Facial expression recognition based on fusion of sparse representation. In: *Advanced Intelligent Computing Theories and Applications. With Aspects of Artificial Intelligence*. [S.l.]: Springer, 2010. p. 457–464.

YU, Z.; ZHANG, C. Image based static facial expression recognition with multiple deep network learning. In: ACM. *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*. [S.l.], 2015. p. 435–442.

YU, Z.; ZHANG, C. Image based static facial expression recognition with multiple deep network learning. In: *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*. New York, NY, USA: ACM, 2015. (ICMI '15), p. 435–442. ISBN 978-1-4503-3912-4. Disponível em: <http://doi.acm.org/10.1145/2818346.2830595>.

YUAN, X.-T.; LIU, X.; YAN, S. Visual classification with multitask joint sparse representation. *Image Processing, IEEE Transactions on*, IEEE, v. 21, n. 10, p. 4349–4360, 2012.

ZAVASCHI, T. H. H.; BRITTO, A. S.; OLIVEIRA, L. E. S.; KOERICH, A. L. Fusion of feature sets and classifiers for facial expression recognition. *Expert Systems with Applications*, Elsevier, v. 40, n. 2, p. 646–655, 2013.

ZENG, N.; ZHANG, H.; SONG, B.; LIU, W.; LI, Y.; DOBAIE, A. M. Facial expression recognition via learning deep sparse autoencoders. *Neurocomputing*, Elsevier, v. 273, p. 643–649, 2018.

ZENG, Z.; PANTIC, M.; ROISMAN, G.; HUANG, T. S.; others. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, IEEE, v. 31, n. 1, p. 39–58, 2009.

ZHANG, F.; ZHANG, T.; MAO, Q.; XU, C. Joint pose and expression modeling for facial expression recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. [S.l.: s.n.], 2018. p. 3359–3368.

ZHANG, S.; LI, L.; ZHAO, Z. Facial expression recognition based on Gabor wavelets and sparse representation. In: IEEE. *Signal Processing (ICSP), 2012 IEEE 11th International Conference on*. [S.l.], 2012. v. 2, p. 816–819.

ZHANG, T.; ZHENG, W.; CUI, Z.; ZONG, Y.; YAN, J.; YAN, K. A deep neural network-driven feature learning method for multi-view facial expression recognition. *IEEE Transactions on Multimedia*, IEEE, v. 18, n. 12, p. 2528–2536, 2016.

ZHEN, W.; ZILU, Y. Facial expression recognition based on local phase quantization and sparse representation. In: IEEE. *Natural Computation (ICNC), 2012 Eighth International Conference on.* [S.l.], 2012. p. 222–225.

ZHONG, L.; LIU, Q.; YANG, P.; LIU, B.; HUANG, J.; METAXAS, D. N. Learning active facial patches for expression analysis. In: IEEE. *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on.* [S.l.], 2012. p. 2562–2569.

ZHU, S.; LIU, S.; LOY, C. C.; TANG, X. Deep cascaded bi-network for face hallucination. In: SPRINGER. *European Conference on Computer Vision.* [S.l.], 2016. p. 614–630.

ZHUANG, B.; LIN, G.; SHEN, C.; REID, I. Fast training of triplet-based deep binary embedding networks. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* [S.l.: s.n.], 2016. p. 5955–5964.