

Master's Degree in Informatics Engineering

Dissertation

Smart Monitor Health System: Face Expressions Recognition

September, 2016

Ana Filipa Laranjeira

afolgado@student.dei.uc.pt

Advisors

Prof. Dr. Bernardete Ribeiro

Eng. André Pimentel

Eng. Xavier Frazão



FCTUC DEPARTMENT
OF INFORMATICS ENGINEERING
FACULTY OF SCIENCES AND TECHNOLOGY
UNIVERSITY OF COIMBRA

Smart Monitor Health System: Face Expressions Recognition

Ana Filipa Laranjeira
afolgado@student.dei.uc.pt

MASTER'S DEGREE IN INFORMATICS ENGINEERING
DISSERTATION

University of Coimbra,
Department of Informatics Engineering
September, 2016

DEI Advisors:

Prof. Dr. Bernardete Ribeiro
bribeiro@dei.uc.pt

Eyesee's Advisors:

Eng. André Pimentel
andre.pimentel@eyesee.pt

Eng. Xavier Frazão
xavier.frazao@eyeseesolutions.com

Jury:

Prof.Dr. Carlos Fonseca
cmfonsec@dei.uc.pt

Prof. Dr. António Dourado
dourado@dei.uc.pt

ABSTRACT

The emerging success of digital social media has had an impact on several fields ranging from science to economy and business. This has been particularly relevant for the marketing industry which centers its activity on digital social interactions between branding and the end-consumer, in order to increase their market competitiveness. Therefore, there is an invested interest in emotion detection and recognition technology from facial expressions.

This work is marked by a detailed research about the concepts and the existent methodologies behind the Automatic Facial Expression Recognition (AFER) systems, as well as an evaluation of their effectiveness.

Additionally, the most relevant models were tested, in order to discern the most adequate one for facial expressions recognition. A comparison was made between the traditional methodologies and *Deep Learning*, a recent trend in Pattern Recognition.

Both these domains have challenging inner workings. Traditional methods are strongly dependent on the input, thus any transformation in the dataset will influence the model for more adjustments. The *Deep Learning* methods are more adaptable to variation, however refining their hyperparameters might be an exhaustive work.

Here, we explored the value of Deep Learning by focusing on recent technological breakthroughs, particularly with Convolutional Neural Networks (CNN). Incremental steps were made in order to deploy the better solution to the network architecture. After some preliminary experiments using the more recent and complex networks (such as *GoogLeNet*, *AlexNet*) we ended with the *Lenet-5* as a baseline. We found that *Lenet-5* simplicity was better suited for the system constraints (dataset dimension, faces size and composition).

The Cohn-Kanade Extended dataset was chosen for testing our proposed CNN model. In an attempt to improve the results we also have augmented the dataset with random perturbations from a wide set, including: skew, translation, scale, and horizontal flip.

Our refinements to the model led to a 90% overall accuracy when taking static images as an input. To further validate our results we built and present a real-time framework. Based on the data collected, the deep model has emerged as a promising approach for AFER systems.

Keywords: Facial expression recognition; prototypic-expressions; Machine Learning; Deep Learning; Convolutional Neural Networks

ACKNOWLEDGEMENTS

A friend once quoted,

“ Keep away from people who try to belittle your ambitions. Small people always do that, but the really great make you feel that you, too, can become great.”

- Mark Twain,

This quote has echoed in my mind over the past few years.

At this point, I would like to thank my supervisor Prof. Dr. Bernardete Ribeiro for believing in my capacities and for the significant support during this process.

My thanks to Eyeseer's supervisors (Eng. André Pimentel e Eng. Xavier Frazão) for the long guidance and technical support.

I would like to give a special thank to my family, particularly, to my mother and wiser brother for their support and encouragement.

There are also a group of fellows who never let me give up and made me stick with the plan. Thanks to Guida Rato for the unconditional support and long lasting friendship, Daniel Martins, Diana Gonçalves, Edouard Almeida, Filipa Ferreira, Humberto Alves, Luana Velho, Marcos Góis, Mariana Dias, Raquel Pina and all my friends who have influenced me in a positive way.

Finally, I would like to acknowledge the LARN laboratory for the good work environment.

CONTENTS

1	INTRODUCTION	19
1.1	Context	20
1.2	Objectives	20
1.3	Outline	21
2	FACIAL EXPRESSIONS ANALYSIS-REVIEW	23
2.1	Emotion and Expression	24
2.1.1	Sadness	25
2.1.2	Anger	26
2.1.3	Surprise	26
2.1.4	Fear	27
2.1.5	Disgust and Contempt	28
2.1.6	Happiness	29
2.2	Face Acquisition	30
2.2.1	Perception & Recognition	31
2.2.2	Tracker methods	34
2.3	Facial Expression and Representation	37
2.3.1	Facial Expression Recognition Methods	41
2.4	Deep Learning	44
2.5	Challenges/Application	45
3	DEEP LEARNING	47
3.1	Convolution Neural Networks	48
3.1.1	Lenet-5	49
4	EXPERIMENTAL SETUP	51
4.1	Datasets	51
4.1.1	Cohen-Kanade Extended Dataset	53
4.2	Methods and Algorithms	54
4.2.1	Facial Expression Extraction - Background Model	54
4.2.2	Facial Expression Extraction - Background Branch Model	55
4.2.3	Facial Expression Extraction and Recognition - CNNs	56
5	EXPERIMENTAL RESULTS AND DISCUSSION	61
6	CONCLUSION AND FUTURE WORK	65
6.1	Face Anatomy	67
6.2	FACS Action Units	68
6.3	HMM	69
6.4	LeNet-5	69

LIST OF FIGURES

Figure 1	Basic Structure of Facial Expression Analysis Systems based on fig.11.1 at [54]	24
Figure 2	Expression of Sadness from [20]	25
Figure 3	Expression of Anger from [20]	26
Figure 4	Expression of Surprise from [20]	27
Figure 5	Expression of Fear from [20]	27
Figure 6	Expression of Disgust from [20]	28
Figure 7	Disgust (partial face representation) and Contempt from [20]	29
Figure 8	<i>Duchenne Smile</i> from [20]	30
Figure 9	Distal stimulus, proximal stimulus, and percepts adapted from fig.3.1 at [38]	33
Figure 10	The integral image reprinted from [30]	34
Figure 11	Sum of calculations reprinted from [30]	35
Figure 12	The different types of features reprinted from [30]	35
Figure 13	A detection cascade depiction inspired from [56]	37
Figure 14	Feature points reprinted from fig.3 at [5]	39
Figure 15	Bank of Gabor filters with 5 frequencies and 8 orientations reprinted from [30]	41
Figure 16	Max Pooling Layer (2×2 filter and stride 2) detailed with one depth slice	49
Figure 17	CKP static images vs our frame sequence, both cropped to 224×224 pixels	54
Figure 18	Flowchart of the different stages of our CNN model, adapted from the classic LeNet-5 - Lenet Ov (Our version)	58
Figure 19	Normalized Confusion Matrices - (a) Alexnet/Ada; (b) Alexnet/SGD; (c) Googlenet/SGD; (d) Lenet changed/SGD.	60
Figure 20	Confusion Matrix for the classification of the Framework in the test set	63
Figure 21	http://medicalart-work.co.uk/wordpress/?page_id=8 (accessed gallery: Sept 1, 2016)	67
Figure 22	FACS Action Units Table.11.1 at [54]	68
Figure 23	Multilevel HMM architecture for automatic segmentation and recognition of emotion reprinted from [8]	69
Figure 24	Architecture of LeNet-5 representing Figure 2 from [34]	69

LIST OF TABLES

Table 1	Face Detection performance completed from Dissertation References	31
Table 2	Facial Feature Extraction based on hybrid algorithms	38
Table 3	Facial Recognition Methods using labeled data (L) and Unlabeled data (UL) or both (LUL)	42
Table 4	Train loss results used to select the shallow network	57
Table 5	Test Accuracy with three versions of the CNN model (a), (b) and (c).	60
Table 6	Methods for comparison	61
Table 7	Precision, Recall and F1 from three above versions (a)(b)(c).	62

NOTATIONS

Adaboost

$h(x, f, p, \theta)$	weak classifier
$w_{1,i}$	initial weights
$w_{t,i}$	normalized weights
$w_{t+1,i}$	updated weights
ϵ_t	weight error
β_t	binary classification (correctiveness) $\frac{\epsilon_t}{1-\epsilon_t}$
$C(x)$	stronger classifier
α_t	$\log \frac{1}{\beta_t}$

Gabor Filter

(x, y)	pixel position in the spatial domain
γ	wavelength (a reciprocal of frequency) in pixels
θ	orientation of a Gabor filter
S_x	standard deviation along the x direction
S_y	standard deviation along the y direction
ϵ	error rate

Deep Learning

$f_j(z)$	softmax function
L_i	multinomial logistic loss

ACRONYMS

AFER Automatic Facial Expression Recognition

AFEW Acted Facial Expressions in the Wild

AR Aleix Martinez and Robert Benavente (database)

AU Action Units

CKP Cohen-Kanade Extended dataset

AFER Automatic Facial Expression Recognition

CNN Convolutional Neural Networks

CVC Computer Vision Center

DARPA Defense Advanced Research Products Agency

DCNN Deep Convolutional Neural Networks

DDFD Deep Dense Face Detector

DWT Discrete Wavelet Transform

EmotiW Emotion Recognition in the Wild

FACS Facial Action Coding System

FAP Facial Animation Parameters

FAPU Facial Animation Parameter Units

FDP Facial Definition Parameters

FER Facial Expression Recognition dataset

List of Tables

FERET Facial Recognition Technology

GLCM Ggray-Level Cco-occurrence Matrix

HMM Hidden Markov Model

ILSVRC Imagenet Large-Scale Visual Recognition Challenge

JDA Join Cascade Detection and Alignment

KLT Kanade-Lucas-Tomasi

MMI Maja Pantic, Michel Valstar and Ioannis Patras (database)

ML-HMM Multi-Level Hidden Markov Model

NB Naive Bayes

OvA One versus All

PBVD Piecewise Bezier Volume Deformation

PCA Principal Component Analysis

SGLD Second-order Statistical Features

SAE Synchrony Autoencoder

SFEW Static Facial Expressions in the Wild

SSS Stochastic Structure Search

SVM Support Vector Machine

TAN Tree-Augmented Naive Bayes

VC Vapnik-Chervonenkis

1

INTRODUCTION

The present work belongs to the field of affective computing ¹ and enhances a particular knowledge - Facial Expression Recognition.

Through the years, this topic has received remarkable insights, but definitely Darwin's work in 1872 [13], established the ground rules. His research together with other findings, enlightens us with a scientific method of expression categorization, which has been widely explored and considered a major step in the field of automatic facial expressions recognition.

Another major inspiration to the present research belongs to the work of Paul Ekman and his colleagues. Since the 1970s they have developed tools for the interpretation of facial expressions and every contour of it, namely by mapping measurable muscles and emotion space, finding discrete states or prototype emotions.

Several other researchers have attempted to develop new concepts and new paradigms. Concerning the recognition studies prior to the year of 1999, there are a few surveys conducted by Pantic and Rothkrantz, such as [41], that are worth of attention. However, here we will address the most recent developments in the field.

¹ Affective Computing (for interested readers) is well documented by R.W.Picard [43]: "Computers that will interact naturally and intelligently with humans need the ability to at least recognize and express affect."

INTRODUCTION

1.1 CONTEXT

This research arose from a partnership between *EyeSee* and the Department of Computer Science.

The *EyeSee* is a startup, located in Lisbon, which has achieved global visibility due to their intensive work around uncommon issues, mostly related to digital advertising. They intend to challenge consumers to embrace a new perspective focusing on the direct interactions between brands and the end-consumer. In view of these goals *EyeSee* has pursued several research interests, which led them to this project.

This partnership arose from the necessity to deepen their knowledge in the automatic emotion detection, analysis and recognition systems.

1.2 OBJECTIVES

The main objective of this work is to understand the best methodologies to be used for an Automatic Facial Expression Recognition System, using images as an input.

Our research proposes to study the *prototypic* expressions described in the work of Dr. Paul Ekman and his colleagues. We aim to describe these particular facial expressions and study its characteristics in order to better understand and improve the recognition process. Furthermore, we aim to evaluate, elect and fine-tune the relevant data sets that would give us confidence in our detection and classification task.

We intend to detail the entire process of facial expression recognition.

Regarding methodologies, we propose to assess and evaluate past work in this field in order to benchmark our results and findings, thus bridging the gap between past and future work.

For our work, we propose to develop methodologies that would show promising results when compared with the State of the Art.

Finally, we also aim to build a model capable of being scaled to sequence-based systems.

1.3 OUTLINE

The report is structured following a top-down strategy. Chapter 2 describes the most important concepts inherent to an Automatic Face Expression Recognition (AFER) system. It includes a characterization of human neural biological responses related to this field. This chapter also gathers important breakthroughs throughout every single AFER system phase (Face Acquisition, Face Data Extraction, Facial Recognition), reporting a detailed description of their theoretical results and inner concepts.

Chapter 3 acknowledges our motivation to use a deep Convolutional Neural Network (CNN) method. Here we made an overview of the classical CNN to prepare the description of the system model presented in following chapter.

Chapter 4 considers two methods from AFER systems background which were used in preliminary experiments. These solutions are detailed and presented along with a description of the dataset used. We claim that these methods were not enough to accomplish the requirements of the system. Based on distinct results collected from preliminary CNN experiments, we follow up on the hypothesis that CNNs would be the better solution for the problem. Therefore, the following experiments only include CNN architectures which are also detailed in this chapter.

The Chapter 5 presents the results that led us to the CNN model and the final achievements after the refinements. This chapter also depicts the results related to our framework constructed for testing sequence-based images.

Finally, the chapter 6 addresses the main conclusions of the study and delivers particular insights for future work .

FACIAL EXPRESSIONS ANALYSIS-REVIEW

Facial Expressions can be described as a reaction to the presence of an emotion. It fits in the frame set by emotional responses. The challenge in the field is to develop a way to integrate these extended forms of representation.

The more we know about these expressions, the more we can describe ourselves as *mind readers*. That is why, facial expressions analysis plays a higher role in many areas such as clinical psychology and related aspects, such as lie detection, pain assessment, and many other fields whose knowledge depends on information extracted from Human Computer Interfaces.

There have been several attempts to categorize the human expressions despite the existence of contradictory approaches [46], concerning the generalization. Ekman and his colleagues, have proved that at least for the six basic expressions, also known as *prototypic* or *architotypical* expressions is possible to achieve a multi-cultural consensus. **Happiness, sadness, anger, surprise, disgust and fear**, according to Ekman's studies can be "discriminable within any literate culture" [21]. Current affective computing developments have these six prototypic expressions as the final goal for recognition. Different types of representation can be assigned into one of these six [58]. However, our prototypic sets includes another expression - **contempt**, based on its presence in a reliable dataset. This expression was reported to be found above 75% both in Western and non-Western cultures [22]. Expressions, according to some studies [6], can be managed independently of other categories of face analysis, such as age, gender or identity context. Thus, this work follows a structure strictly expression-based, even though these categories are used to provide diversity through datasets used in facial expression analysis systems.

The most widely used structure in Automatic Face Expressions Recognition systems is expressed in figure 1.

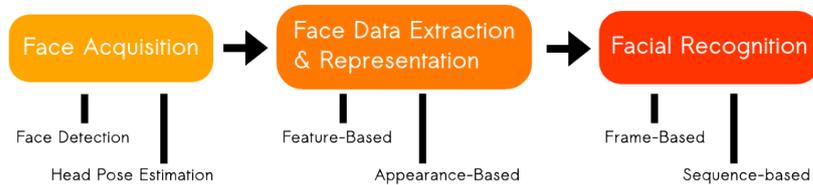


Figure 1: Basic Structure of Facial Expression Analysis Systems based on fig.11.1 at [54]

2.1 EMOTION AND EXPRESSION

Emotion is a state that helps “to deal with important events without our having to think about what to do” [20], it is an important reaction to the environment and helps us to live and survive.

Few scientists agree that humans are permanently in an emotional state, however these emotions are practically unnoticeable sometimes, such that we can consider them inexistent. Taking this into account, only the strikes are considered, and they can even be distinguished from other similar concepts by the short period of time they last. Emotion only takes a few seconds and in particular cases a few minutes. On the other hand, concepts like *moods* take hours or even days.

The emotions can be revealed by visible human changes, in voice, bodily posture and facial expressions, that’s when things get interesting. The changes are the result of some abnormal activity in automatic nervous system which regulates heart rate, breathing, sweating, among other bodily changes, referenced before. They are hints for subsequent actions.

The next seven prototypic expressions will be presented according the Ekman’s point a view [20].

2.1.1 Sadness

Sadness is one of the long-lasting emotions, and like any other emotion, cannot be completely described with words like the feelings linked to it, such as distraught, disappointed, dejected, blue, depressed, discouraged, despairing, grieved, helpless, miserable, sorrowful, as Ekman point out. The emotion must be experienced to fully be understood, and can be revived through others expressions.

We can deceive our brain, making particular face movements and triggering some physiological changes, like force the sadness (or other emotion) to be felt, however this is just a memory exercise, the true feeling (Figure 2) was experienced before.



Figure 2: Expression of Sadness from [20]

People can *have different emotional intensities*, but the baseline movements are similar to each other.

The following movements must be seen as a group, because some points of the expression, if examined separately, may mislead the final emotion. Charles Darwin, referenced that pressing lips tightly together (just like in the expression of Anger) is a common evidence of any physical exertion.

The next group of movements represent Sadness:

- Pull the corner of lips down
- Raise the cheeks
- Drop upper eyelids
- Downward eye look
- Upper inner corners of eyebrows

2.1.2 *Anger*

Anger (Figure 3) may be a reaction to a *disagreement, a challenge, an insult, a minor frustration*, belongs to the negative feelings generated by what the inflicted person designated as offensive.



Figure 3: Expression of Anger from [20]

The next group of movements represent Anger:

- Pull eyebrows down with inner corners down toward the nose
- Open eyes wide
- Upper eyelids against the lowered eyebrows
- Lips pressed together tightly

2.1.3 *Surprise*

Surprise (Figure 4) is a reaction to an unexpected event. It is the briefest of all emotions until the conscience of the situation happened. This emotion is rapidly followed by a different kind of emotion, dependent of the situation.

Surprise features are commonly misled with *fear*, presenting minor differences among each other.

The next group of movements represent Surprise:

- Raised eyebrows
- Drop jaw open
- Open eyes wide



Figure 4: Expression of Surprise from [20]

2.1.4 *Fear*

Fear (Figure 5) is the most researched emotion, probably due to its wide existence in nearly any animal.

It is triggered when there is a threat of harm, due to a clear possibility of physical pain, or just thoughts of danger.



Figure 5: Expression of Fear from [20]

The next group of movements represent Fear:

- Upper eyelids
- Slightly tense lower eyelids
- Drop jaw open
- Staring straight ahead
- Raise eyebrows and draw them together
- Lips stretched back towards the eyes

2.1.5 *Disgust and Contempt*

Ekman's studies (Figure 6) have into account the ideas of the psychologist Paul Rozin, suggesting the existence of two types of disgust. The core one, involving "a sense of oral incorporation of something that is deemed offensive and contaminating" and the interpersonal disgust triggered by the feeling of *strangeness, disease, misfortune and morally tainted*. The last designation is considered the most disgusting in respect to some results conducted in adults.



Figure 6: Expression of Disgust from [20]

The next group of movements represent Disgust:

- Upper lip raised
- Lower lip raised and protruding slightly
- Deep wrinkles from above nostrils downward to beyond the lip corners
- Nostril wings raised
- Cheeks raised
- Lower brows

Disgust is often mistaken by Contempt, an expression exclusive of human interaction which represents a moral judgment and translates a feeling of being superior. Fig. 7a shows that if we represent Disgust with a partial face expression it resembles the expression of Contempt. However, the actual Contempt expression is depicted in Fig. 7b. Normally this expression tightens the lip corners, which are also slightly raised.



(a) Expression of Disgust



(b) Expression of Contempt

Figure 7: Disgust (partial face representation) and Contempt from [20]

2.1.6 Happiness

The word *happiness* (Figure 8) as well as *enjoyment*, are not specific enough according to Ekman, they can express many enjoyable emotions, amusement, *fiero* (a personal satisfaction of accomplish something difficult), *naches* (*beam-immense-pride-and-pleasure*), contentment, excitement, sensory pleasures, relief, wonderment, *schadenfreude* (feeling better comparing to others misfortune), ecstasy, elevation and gratitude. These emotions can be examined in detailed from Ekman's research [20].

Besides their singularity, they all involve smiling, controversial sign of *happiness*. Although smile can be faked, it is acknowledged that there is a method to discover the *true enjoyment smile* or *Duchenne smile*, namely by Ekman, in memory of the French neurologist, Duchenne de Boulogne.

After analyzing photographs activating the *zygomatic major muscle*, Duchenne wrote, "the emotion of frank joy is expressed on face by combined contraction of the zygomatic major muscle and the orbicularis oculi" (referenced in Appendix 6.1). The truth is revealed knowing that the muscle around the eye does not obey the will.

The next group of movements represent happiness:

- Smiling
- Eyebrow and eye cover fold pulled down



Figure 8: *Duchenne Smile* from [20]

2.2 FACE ACQUISITION

Face Acquisition may require face detection, face tracking, segmentation or even geometric normalization. These procedures have to consider some challenges covered in this research, such as translation, rotation or scaling of the head.

The procedure to handle face acquisition is dependent of the input format. Therefore, we expect to improve AFER systems by understanding some neurobiological mechanisms, such as Perception and Recognition (description in Section 2.2.1).

Preliminary experiments had only images as object of study and therefore and only face detectors were considered. Further experiments introduced tracker methods in order to cover video issues.

Face detection is the most widely explored part of Facial Recognition Systems's structure, therefore due to its importance it is well worth to explore its roots.

Table 1 shows a performance analysis concerning several algorithms. This is an incremental version of what was categorized and presented by Ming-Hsuan Yang [60] as having a representative role.

It must be acknowledged that these comparisons, are not using the same datasets, but can be a reference for better methods.

On the other hand, for face tracker methods there is no consensus in the process of categorization, nor even for performance measurements. However, some prominent studies are described in [54].

Approach	Methods	Performance	Notes/Source
Knowledge-Based	Multiresolution Rule-Based Method	83%	60 images, 28 false alarms [59]
Feature Invariant (facial features)	Grouping of edges	85%	110 different images scale, orientation, viewpoint [61]
Feature Invariant (texture)	Space Grey-Level Dependency Matrix of face pattern	98%	30 images, 60 faces, 10% false alarms [11]
Feature Invariant (skin color)	Mixture of Gaussian	S - 55.1% G - 43.6%	specific(S) and generic(G) eigenspace [39]
Template Matching (Predefined Face Templates)	Shape Template	>95%	64 images [10]
Template Matching (Deformable Templates)	Active Shape Model	92%	bad performance, under 60 [31]
Appearance-Based Eigenface	Eigenvector decomposition and clustering	96% average	less performance with variations(orientation, size) [55]
Appearance-Based Distribution-Based	Gaussian distribution and multi-layer perceptron	db1 - 96.3% db2 - 79.9%	71 people : 301 faces (db1) 3 false alarms ; 23 images : 149 faces (db2) 5 false alarms [51]
Appearance-Based Neural Network	Ensemble of neural networks and arbitration schemes	between 77.9% and 90.3%	130 images [45]
Appearance-Based Support Vector Machine (SVM)	SVM with polynomial kernel	A - 97.1% B - 74.2%	313 images: 313 faces, 4 false alarms (A) 23 images : 155 faces 20 false alarms (B) [40]
Appearance-Based Naive Bayes Classifier	Joint statistics of local appearance and position	86.8% average	performance for a different value of a detection threshold, for 125 images (one of many tests) 40 false alarms average [48]
Appearance-Based Hidden Markov Model (HMM)	Discrete Wavelet Transform (DWT) for observation sequence extraction in HMM based face recognition system	98.5% (best result)	With ORL Database [47]
Appearance-Based Information-Theoretic	Kullback relative information	92%	507 faces, low false alarms [9]

Table 1: Face Detection performance completed from Dissertation References

2.2.1 Perception & Recognition

Perception and Recognition were not always seen as separated concepts, the first steps towards this separation came at the end of the nineteenth century [2], and will be succinctly covered.

Perception is a process which works as a reaction to the onset of stimulus, it processes the features of visual images and their configurations. It works like a *database* for human interpretation.

The topic may be oriented through many channels, such as visual, auditory, olfactory, haptic, and gustatory, herein we are interested in visual perception.

The process itself is more complex than every definition that may be constructed. According to some neuroscientists the process of visual perception stimulates one-half of the human cortex[52], and part of this activity depends on how the meaning is assigned. Against the traditional perspective that when we look to an object we just acquire from it specific bits of information including its location, shape, texture, size, and (for familiar

objects) names, the psychologist James Gibson (1979) supports the idea that humans can also acquire object's function[38]. We shall not expect a consensus on how this information is acquired and on how it may be influenced by past and human experiences. The *Classical* approach to Perception definition is always an option when trying to escape all this diversity. Figure 9 exploits this approach. From an object, or **distal stimulus**, the visual system receives and registers the information, develops a **proximal stimulus** and from the back of the eye, the retina creates an image (retinal image). It is displayed inversed, concerning both sides, and displayed upside down, waiting for an interpretation and recognition. The recognition or **percept** is considered the goal of Perception.

As referenced before, Pattern Recognition is linked to Perception, it assembles the environment information (object, event, and so on) and assigns meaning, the percept.

The groundings of **percept** stage are based upon *Gestalt school psychology*, whose principles inspired many points of view related to this subject. Exploiting these principles is beyond the scope of this project, however all of them can fit in a more general law, the *law of Pragmanz* (Koffka, 1935), which realizes that among every attempt to interpret a display, we will tend to organize the information by the simplest and most stable shape and form [38].

Although there is no consensus regarding the perspective used to study Perception definition, there is an agreement on proximal stimulus and percept being distinct phases, as well as the way bits of information reach percept, which is commonly described by the bottom-up/ *conceptually-driven* model.

Bottom-up is an one-way model, and by definition a lower level of processing [38] data, **Template Matching**¹, **Featural Analysis**² are some examples.

Psychologists also assign top-down (or *conceptually-driven*) processes to be part of Perception, which is responsible for managing beliefs, expectations and prior-knowledge over the perceived object.

1 -Template Matching is a model of perception involving a comparison of information between the incoming and the stored - templates, looking for an exactly match

2 -Featural Analysis processes the stimuli as a whole, but as a result of breaking into different components-features. There is evidence that assures the presence of some perceptual detection responsible to scan input patterns, looking for features. As an example, some stimulus may cause some cells to respond strongly to borders between light and dark, the *edge detectors*

Both *bottom-up* and *top-down* are relevant concepts to step into Pattern Recognition's algorithms.

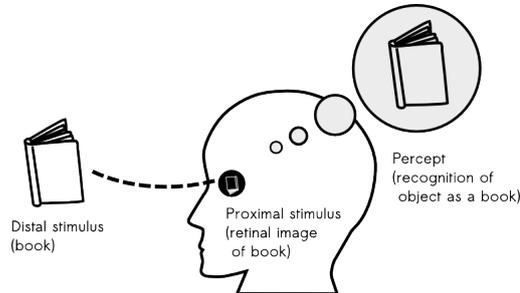


Figure 9: Distal stimulus, proximal stimulus, and percepts adapted from fig.3.1 at [38]

Knowledge-based Top-Down Methods present assumptions related with facial features. These assumptions, are converted into rules to identify possible face candidates. Despite the fact that rules are limited (consequently the model), some proposals have been presented. As an example Yang and Huang [60] show an hierarchical knowledge-based method to detect faces, with three levels of rules. Lower levels find pixel information, such as intensity and gray-scale values. On the other hand, higher level methods are more concerned in delivering general facial features, for example, eyes and mouth location.

Bottom-Up Feature Based Methods are used mainly for face location and aim to find invariant features in the face, they exist even when there are environmental changes. These methods use *facial features* to compute differences between the face and the surroundings, using background properties or adding forms (blobs, streak outlines and others) to accomplishing the task. Another useful information used by these feature based methods is the *texture*, which allows object separation by difference dynamics, instead of using shape properties. As an example, texture is used to exploit face-like surfaces computing second-order statistical features (SGLD) [60]. *Skin Color* is another feature used, relying on color information (histograms) and focusing on intensity rather than their chrominance in order to find differences in images. However, this approach can bring problems with light variation. The major branches are the **Template matching methods** which use a template (standard pattern) to find a correlation between images and the template values, in order to seek face contours, eyes, mouth, nose as an unit.

On the other hand, there are the **Appearance-based methods** which include a large spectrum of methods. Within these we find Neural Networks (NN) which are worth to mention. The NNs replicate an analogy between human biological neurons. The concept of weights introduced computationally and the way they change influenced by the input, mimics the real dynamics and information transmissions reproduced from a human brain.

2.2.2 Tracker methods

Tracker methods are herein slightly explored considering the static images as the scope of this research. In order to test our model we used the Viola-Jones Face Tracker [56] detailed up next. This classifier is still widely used, despite the existence of more effective cases. A recent work [23], proposed a method called Deep Dense Face Detector (DDFD) that represents a remarkable step tackling most of the challenges presented by AFER systems. Without compromising the complexity of the system they propose a method that overcomes orientation problems without the need for landmarks or post-processing schemes. The DDFD uses CNN for classification and feature extraction and can be an option in future experiments.

Viola and Jones Face Tracker has three key contributions. First of all, the concept of “Integral Image” inspired by the Haar features. Each pixel is equal to the entire sum of all pixels above and to the left of the concerned pixel as depicted in Figure 10.

1	1	1
1	1	1
1	1	1

Input image

1	2	3
2	4	6
3	6	9

Integral image

Figure 10: The integral image reprinted from [30]

With this region we can select the specific values for the calculation of the sum of all pixels, allowing constant time results. These values are the pixels in the integral image that coincide with the corners of rectangles defined in the input image. This is demonstrated in Figure 11.

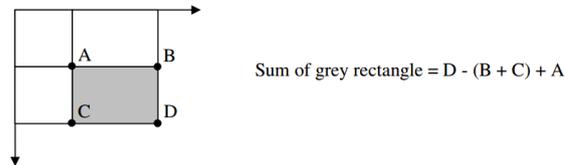


Figure 11: Sum of calculations reprinted from [30]

Since both rectangle B and C include rectangle A, the sum of A has to be added to the calculation.

The Viola-Jones face detector analyzes a given sub-window using features consisting of two or more rectangles. The different types of features are shown in Figure 12. Each

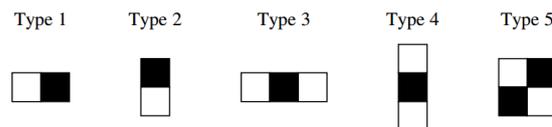


Figure 12: The different types of features reprinted from [30]

feature results in a single value which is calculated by subtracting the sum of the white rectangle(s) from the sum of the black rectangle(s) [30].

This algorithm contributes with a new version for Adaboost classification. Motivated by the large-scale of Haar-like features compared to the number of pixels, this method combines "weak classifiers" ($h(x, f, p, \theta)$) constraining them to one feature. Turning this step into a feature selection process.

A **weak classifier**, $h(x, f, p, \theta)$ is defined as follows:

$$h(x, f, p, \theta) = \begin{cases} 1, & \text{if } pf < p\theta \\ 0, & \text{otherwise} \end{cases}$$

The modified Adaboost by Viola and Jones is represented in the following pseudo-code lines from [56]:

- Given example images $(x_1, y_1), \dots, (x_n, y_n)$ where $y_i = 0, 1$ for negative and positive examples respectively.
- Initialize weights $w_{1,i} = \frac{1}{2m}, \frac{1}{2l}$ for $y_i = 0, 1$ respectively, where m and l are the number of negatives and positives respectively.
- For $t = 1, \dots, T$:
 - Normalize the weights,

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

- Select the best weak classifier with respect to the weight error,

$$\epsilon_t = \min_{f,p,\theta} \sum_i w_i |h(x, f, p, \theta) - y_i|$$

- Define $h_t(x) = h(x, f, p, \theta)$ where $f_t, p_t,$ and θ_t are the minimizers of ϵ_t .
- Update weights:

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

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

- The final strong classifier is:

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

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

This algorithm also introduces a third contribution to improve the computational efficiency. The method progressively includes more complex classifiers with a cascade structure, augmenting the number of features processed based on previous detection rates. An example of the cascade process is depicted in Fig. 13.

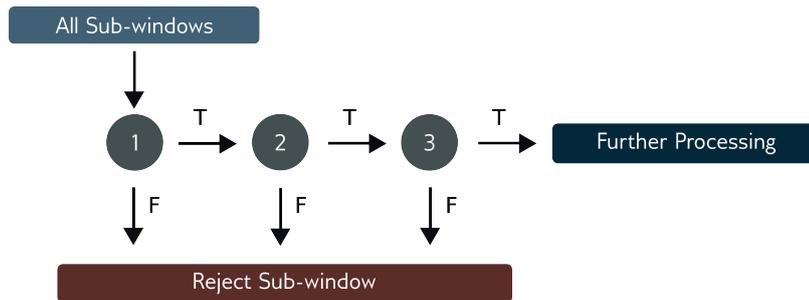


Figure 13: A detection cascade depiction inspired from [56]

2.3 FACIAL EXPRESSION AND REPRESENTATION

“ (...) an initial choice of adequate features is often more an art than a science”

- J.P. Marques de Sá,

Facial Expression Extraction is the crucial turning point in AFER systems. It is an optimized course of action, aligned with the essential characteristics capture, improving the recognition system performance.

A report has referenced an hybrid version as the most qualified to achieve better results [54]. Therefore, its seems reasonable to evaluate some of these hybrid techniques used recently, as shown in the following table 2 built with information from [5].

Methods	Performance	Notes/Source
1- Holistic Spacial Analysis 2- Explicit Measure of local Image 3- Template matching	1 - 89% 2- 57% 3- 85% 1+2+3 - 92%	1- Based on PC of grayscales of images 2- Example wrinkles 3- Usage of motion fields [15]
1- Gabor wavelets 2- Optical Flow 3- Multi-states Models	upper and lower face - 93.3%	[53]
1- Gabor wavelets 2- Facial Regions	1+2 - anger - 92.7% disgust - 85.7% fear - 81.5% sadness - 90.5%	2- Instead of points (weight average) [57]
1- Gabor wavelets 2- Fiducial Points(34)	1- 92.2% 2- 73.3% 1+2 - 92.3%	[63]
1- Gabor wavelets 2- Geometry from Multi-state Models	1+2 - 92,7%	2- one state - brow/cheek two state - eye/furrows three state - lip [36]

Table 2: Facial Feature Extraction based on hybrid algorithms

The following methods represent those with better performance in the range described in the face extraction Table 2.

- Face measurements – Geometric-based methods, are face measurements widely used in traditional approaches and in general they are not self-sustained, preceding, in most cases, another complementary method.

The standard *MPEG-4 FAP* solution ³ is a regular example regarding the face measurements extraction. It compares the original face in video sequence, in a way that progressively the optimal set of Facial Animation Parameters is found.

The MPEG-4 standard [1] includes a parametrization of several feature points known as Facial Definition Parameters (FDPs) on top of a facial skin mesh. Using the corresponding Facial Animation Parameters (FAPs) 14 (in one direction), MPEG-4 is able to proceed through a necessary manipulation.

FAPs are based on the study of minimal facial actions and are related to the muscle actions. Their values parameterize relations between facial components (distance between eyes or lip corners and so on) using Facial Animation Parameter Units (FAPUs), turning possible the use of FAPs on different faces.

Image processing is followed by wavelet transforms, in other words, image is filtered by measurements given by the scale of the wavelet.

³ Available from: http://www.researchgate.net/publication/4027324_MPEG-4_FAP_generation_as_an_optimization_problem (accessed Sept 1, 2016)

2.3 FACIAL EXPRESSION AND REPRESENTATION

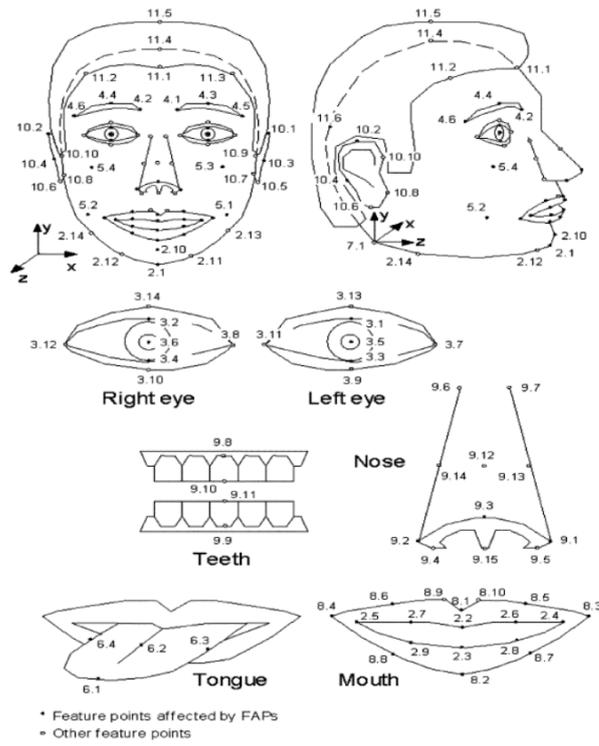


Figure 14: Feature points reprinted from fig.3 at [5]

- **Gabor wavelets** – This method is used frequently on feature extraction, having better performance than geometric-only approaches.

Among other features extraction methods, it has a dimensional reduction process which retains the essential information concerning the highest discrimination power and stability.

Gabor filters detain a rich localization property, their alignment (orientations) combined with the extraction of information in frequency domains may lead to reliable results [30].

Orientation and frequency are the most important parameters in the process, they affect significantly how expressions faces are differentiated.

Gabor filter is modulated as a complex exponential by a Gaussian function represented by the following equation:

$$g(x, y) = \frac{1}{2\pi S_x S_y} \exp\left[-\frac{1}{2}\left(\frac{x'^2}{S_x^2} + \frac{y'^2}{S_y^2}\right)\right] + \exp\left[j\frac{2\pi x'}{\gamma}\right]$$

where

$$x' = x \cos \theta + y \sin \theta; y' = x \sin \theta + y \cos \theta$$

Most cases use a Gabor filter bank with five frequencies and eight orientations (face representation context).

The Gabor features are calculated as a convolution of the input image with the Gabor filter bank function. An example of the Gabor filter bank feature images are shown in Figure 15.

There is contradictory work around which properties are relevant and sufficient to recognize facial expressions. Some studies are confident that geometric-based features are enough, although other perspectives are convinced that expressions are intrinsically linked with context, and that the solution passes by analyzing the eye [2](the reflexion of context).

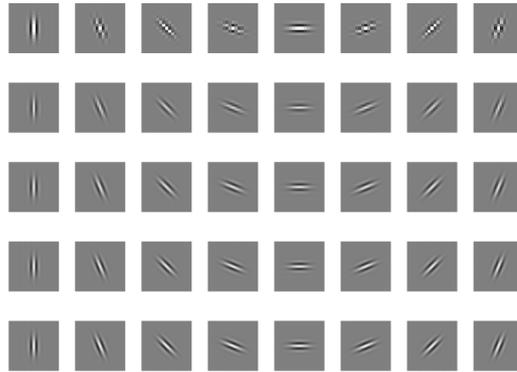


Figure 15: Bank of Gabor filters with 5 frequencies and 8 orientations reprinted from [30]

An attempt to solve the duality problem was provided by the Facial Action Coding System (FACS)⁴.

This standard makes it possible to obtain context-independent results. Letting further assumptions to be considered in a higher level of recognition. The current work does not make use of this coding system, relying on stable features presented in the dataset used (referenced in the chapter 4). However, it can be included in future work.

Facial Action Coding System was developed by Ekman and Friesen in 1977. This is a facial muscle model, which identifies units and groups of muscles in order to detect facial behavior. These changes are categorized by the so called Action Units (AUs) depicted in Appendix 22. They represent most of the non-complex AUs, and their combination is enough to reveal expressions⁵.

2.3.1 Facial Expression Recognition Methods

According to basic AFER systems's structure this is the final stage 1, the stage where all previous efforts will collect relevant data, expressions or AUs or even combination

⁴ FACS are a result of Ekman's studies about the actual influence of context in expressions validation crossing cultures [18]

⁵ For deeper knowledge about FACS and Actions Units presentation readers can follow [19], specially chap.13

of expressions. The recognition phase, is strongly coupled with temporal decisions, in other words the methods are categorized according to time progress dependency. If the groundwork is independent of time, and uses only frames, while having, or not, reference images, we might include this work in frame-based methods, otherwise they assume a sequence-based designation, based on sequence of frames and temporal dynamics as source of information. To assign frame-based methods, mostly static classifiers are used, such as, Naive Bayes (NB), Support Vector Machine (SVM), Tree Augmented Naive Bayes (TAN), Stochastic Structure Search (SSS).

Sequence-based methods are covered by dynamic classifiers, which are the most viable option concerning the person dependency tests and have been widely used to classify expressions, namely the Single Hidden Markov Model (HMM) and the Multi-Level Hidden Markov Model (ML-HMM).

Methods	Performance (Average Rate)	Databases
SVM	76.11%	Cohn-Kanade
NB-L	72.50%	Cohn-Kanade
TAN-L	72.90%	Cohn-Kanade
NB-LUL	69.10%	Cohn-Kanade
TAN-LUL	69.30%	Cohn-Kanade
SSS-LUL	74.80%	Cohn-Kanade
HMM	78.49%	L. Shao-Hsien Chen [7]
ML-HMM	82.46%	L. Shao-Hsien Chen
NB+HMM (Hybrid)	73.22%	Cohn-Kanade

Table 3: Facial Recognition Methods using labeled data (L) and Unlabeled data (UL) or both (LUL)

The Table 3 was built with information retrieved from [5] and the methods presenting the best performance are described hereafter.

- SVM – Support Vector Machines classify data, defining a set of support vectors. The support vectors are a set of data points obtained from the minimization of the Structural Risk. Therefore, the solution given by SVM is sparse which makes it computationally efficient in the recognition phase (recall time). The support vectors from the training inputs outline a l (number of features) dimensional hyperplane in feature space. This large margin hyperplane is responsible for the decision boundary between the distinct classes. The idea behind the Structural

Risk is to reduce the average error of inputs over their targets. Minimizing the Structural Risk allows a trade-off between machine capacity and misclassification errors.

Here, supposing that we are dealing with a binary class problem (not the case of this project), with linearly separable classes with target values $+1$ e -1 , a discriminating hyperplane will satisfy[16]:

$$\begin{cases} w' * x_i + w_0 \geq 0 & \text{if } t_i = +1 \\ w' * x_i + w_0 < 0 & \text{if } t_i = -1 \end{cases} \quad (1)$$

Taking this into account, the margin of a separating hyperplane is defined as the distance $(|w'x_i + w_0| / \|w\|)$ of the closest pattern to hyperplane, the support vector algorithm consists of maximizing this margin of separation.

SVMs also provide a generic mechanism to fit the surface of the hyperplane to the data through the use of a kernel function.

“An exciting property of SVMs is how the ”curse of dimensionality” is avoided by the upper bound on the VC-dimension. The VC (Vapnik-Chervonenkis)- dimension measures the capacity of the machine. This bound does not depend on the dimensionality, but on the separation margin between the classes”[27].

The basic SVMs support only binary classification, however some extensions have been proposed to handle the multiclass classification [3] and will be considered in Chapter 4.

- Multi-Level HMM – Hidden Markov models (Appendix A, Fig. 6.3) have been widely used for many classification and modeling problems. Its use is more common with video input, however it should be referenced for future purposes.

The ability to model non-stationary signals or events is one of the many advantages of HMMs. However, there are some disadvantages and most cases are related with time consuming aspects.

The learning process is based in conditional probabilities, and further assumptions can be seen in a comprehensive tutorial from Rabiner[44].

Returning to the emotion expression topic, the signal can be seen as measurements of the facial motion, strictly speaking the recognition of the expression is done by decoding the active state, according to its time. This signal is non-stationary in nature, since an expression can be displayed at varying rates, with varying intensities across different individuals.

An example of those models is depicted in Appendix 6.3, where the high-level HMM consists of seven states, corresponding to the six basic emotions, including the neutral.

2.4 DEEP LEARNING

During the past few years Computer Vision has brought up deep models into discussion specially for digital and speech recognition.

Recently Convolutional Neural Networks have being revisited based on their significant results in image base classification acknowledged by the annual contest *ImageNet Large-Scale Visual Recognition Challenge (ILSVRC)*.

The use of deep networks for expression recognition remains to be fully explored.

Nowadays, Emotion Recognition in the Wild Challenge (EmotiW) is defining the state of the art focusing on affective sensing with uncontrolled environments⁶. From the EmotiW challenge (2015), we can highlight two versions with expressions as subject: one uses static images from Static Facial Expressions in the Wild (SFEW); and the other uses an acted point of view resorting to an Acted Facial Expressions in the Wild (AFEW) dataset. Among the static-image approaches, the project [62] proposes a 3-way detection of the face with a hierarchical selection from the Joint Cascade Detection and Alignment (JDA), Deep CNN-Based (DCNN) and MoT along with a simple network (11 layers). These detectors are processed in a multiple network framework in order to enhance the performance. It also includes a pre-processing phase to improve accuracy, which might be considered a drawback in the classification response.

⁶ <https://cs.anu.edu.au/few/emotiw2015.html>(accessed Sept 1, 2016)

Resorting to video there is an approach [17] which introduces a Synchrony Autoencoder (SAE) to overcome spacio-temporal issues, by extracting local image features together with an hybrid network - CNN-RNN.

2.5 CHALLENGES/APPLICATION

In Damasio's insights [12] emotion has a major role in decision-making. One of the great challenges of AFER is managing real-time situations influenced by real-time decisions. The perspectives of embracing this knowledge influenced many areas of expertise. This research is not concerned in detailing them, however we can highlight important achievements in the field, like Affectiva⁷ and Emotient⁸ products, or more importantly the modest approach that has motivated our testing framework [49].

Some of the typical challenges faced by these AFER systems are pointed hereafter.

- Different images same expressions
- Different expressions same partial expression
- Controlled environments (data gathering)
- Occlusions
- Tracking partial images

⁷ <http://www.affectiva.com/>(accessed Sept 1, 2016)

⁸ <http://emotient.com/products/emotient-adpanel/>(accessed Sept 1, 2016)

DEEP LEARNING

Deep Learning methods are inspired in the mammalian brain which is also organized in a deep architecture [4] scheme, unlike traditional learning methods with a more shallow network. This machine learning paradigm uses a feature hierarchy way of learning. The learning process of features from higher levels of the hierarchy is done by gathering lower level features. The level of complexity achieved allows it to be completely independent from human-crafted features. A characteristic that made it a requested choice.

Deep Learning methods have the ability “to learn about thousands of objects from millions of images” [29]. These models have been stagnant since the efforts of Kunihiko Fukushima with Convolutional Neural Networks (CNNs) *NeoCognitron* and the advent of *LeNet* which attempts to overcome the lack of relevant results [24] by Yann LeCun and collaborators.

A turning point, occurred in 2006, when Hinton et al. introduced Deep Belief Networks (DBNs) [4]. With a greedy method which would train one layer at a time and explore an unsupervised method - Restricted Boltzmann Machines (RBM). Shortly thereafter, new paradigms emerged in the field, showing success in numerous machine learning tasks (classification, modeling textures, object segmentation, regression and others [4]) motivated several works, including the present one.

Since, one of our goals was to build an AFER system with images and considering that we use the Ekman’s dataset as baseline, we decided to benefit from the included labels from the set. Auto-encoders, RBMs and DBNs are commonly used with unsupervised data and therefore, excluded from this scope.

Alternatively, we used Convolution Neural Networks, a Deep Learning method which according to [4], exploits the synergies between several tasks.

3.1 CONVOLUTION NEURAL NETWORKS

A typical Convolutional Neuronal Network Architecture is built combining three types of layers:

- Convolutional Layer
- Pooling Layer
- Fully-Connected Layer

Unlike SVMs or simple neural networks systems, CNNs learn high-level tasks with complex invariance (shift, scale, distortion) [33], controlled by their depth and breadth. To ensure this, CNNs combine three architectural ideas: *local receptive fields* (filters), *shared weights* and space or temporal *sub-sampling*.

The "elementary feature detectors that are useful on one part of the image are likely to be useful across the entire image" [34]. Thus, the receptive fields of each unit can have identical weight vectors and are organized in planes or *feature maps* which perform the same operation (convolution) over the entire image. These schemes belong to Convolution Layers phase and each dot product uses a kernel with equal size of the input image. This layer intends to reduce spectral variation and model spectral correlation.

Furthermore, subsequent layers are included for higher-order feature extraction. *Sub-sampling* is an example responsible for reducing the sensitivity to shift and distortion, by decreasing the resolution of feature maps. This downsampling scheme defines the *Pooling Layer* and changes the volume of the output. Fig. 16 describes an operation of downsampling on one depth slice. This example represents a *Max Pooling* from which only the maximum values remain when using the information retrieved with a 2 kernel and a stride (filter step) of 2.

Traditional CNNs learn with Back-Propagation¹, the same method presented in preliminary experiments. However there are other methods worth to mention, including the

¹ For inside notes about Back-Propagation algorithm Chapter 6 from Duda's[16] may be interesting

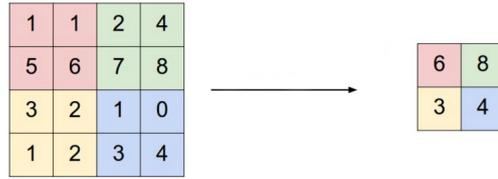


Figure 16: Max Pooling Layer (2×2 filter and stride 2) detailed with one depth slice

Softmax Loss used in further experiments. The softmax loss layer computes the multinomial logistic loss after the inputs run over a softmax function.

$$f_j(z) = \frac{\exp^{z_j}}{\sum_k \exp^{z_k}}$$

The softmax classifier is a generalization of Logistic Regression to a multi-class problem. It delivers as an output a normalization of the classe's probabilities.

Multinomial Logistic Loss

$$L_i = -\log\left(\frac{\exp^{f_{y_i}}}{\sum_j \exp^{f_j}}\right)$$

or equivalent

$$L_i = -f_{y_i} + \log\left(\sum_j \exp^{f_j}\right)$$

This has a probabilistic interpretation defined as $P(y_i|x_i, W)$ where y_i is the correct label, x_i is extracted from the image and W is the weight matrix.

3.1.1 Lenet-5

The network - Lenet-5 (Appendix 6, Fig. 6.4) - combines seven layers represented by a convolutional layer with 16 feature maps with a 5×5 kernel size, followed by a sub-sampling by half. This sub-sampling (pooling layer) and the next convolutional layer (16 feature maps) are not connected in order to break symmetry. The network includes also another pooling stage, maintaining the 16 feature maps and the kernel to 5×5 . It also includes another convolutional full sized (kernel 1×1), mapping 120 units before the last full connected layers with 84 and 10 neurons ("LeNet digits"), respectively.

4

EXPERIMENTAL SETUP

Before we delve any deeper into the model experiments, we should notice, from the state of the art, some methods using Gabor Wavelets have great results considering the traditional approaches. Therefore, they have motivated two solutions from the preliminary experiments which we mention and name as *Background* models. Running against these models, we also present Deep Convolutional Networks due to their recent breakthroughs.

Most of the experiments used the Cohen-Kanade Extended dataset as input, mainly motivated by reasons that we enumerate in the following section.

4.1 DATASETS

Quality of AFER systems relies significantly on dataset choices. Face recognition field longed to realize the need of a standard dataset, however the mindset changed facing the necessity of a comparison tool for algorithms evaluation.

The first prominent attempt towards a de-facto dataset came from *Facial Recognition Technology* (FERET) dataset¹ which is no longer reachable. This was followed by many others such as *Cohn-Kanade dataset*, *AR Face dataset*².

Due to the expressions complex nature, some constraints might be present in face-based dataset. To achieve more robustness, datasets have to pursue a versatile result, thus they

¹ FERET - Sponsored by the Department of Defense's Counterdrug Technology Development Program through the Defense Advanced Research Products Agency (DARPA)

² AR Face dataset - created by Aleix Martinez and Robert Benavente in the Computer Vision Center (CVC) at the *Autonoma de Barcelona*

must possess a high level of variations. The goal is to reduce the distance from real context, exhibiting spontaneous expressions as much as possible, and being aware of lighting changes, rotations, occlusions and other aspects that may reinforce the recognition system.

Following this direction, some datasets emerged and spread out. MMI Facial Expression Dataset is one of them, with a complete documentation and freeness in researching environment. According to this work [42] "MMI dataset aims to deliver large volumes of visual data of facial expressions to the facial expression analysis community". They proved that this dataset is a comprehensive and well prepared group of large data suited for analysis, despite some limitations concerning expression phases. A more recent collection of data, the Candid Image Facial Expressions dataset (CIFE) [35] and the Facial Expression Recognition (FER) dataset [26], should be highlighted due to their significant attempt to deliver a more "candid" or uncontrolled facial expressions.

There are still two aspects that may require some attention when considering facial datasets, first the need for labeled data and second the absence of one of the three phases of face expression (onset, apex and offset³).

Labeling data related to expressions, is a complex field regarding the psychological principals behind the concepts to be recognized. Additionally, context observations and non-linearity can be a concern.

According to some studies the above issues are minimized by a reliable and consistent source, the Cohen-Kanade dataset. This dataset is widely used despite some concerns in presence of temporal dependent classifiers, such as ML-HMM (lacking for a facial phase).

Our experiments during this research used the Cohen-Kanade Extended as the main dataset and the FER dataset for refinement options.

³ Expressions present an evolution in their nature, starting with an onset phase, the time just before the peak of expression - apex, ending with a closing expression, the offset

4.1.1 *Cohen-Kanade Extended Dataset*

Cohn-Kanade dataset, also known as CMU-Pittsburg AU coded dataset, was elaborated by members of the Affect Analysis Group, an interdisciplinary research group located in the Clinical Psychology Program of the Department of Psychology at the University of Pittsburgh.

The **Cohn-Kanade Extended** dataset [37] (CKP) is labeled between 0–7 corresponding to *Neutral, Anger, Contempt, Disgust, Fear, Happiness, Sadness, Surprise* and contains 593 sequences across 123 subjects with posed and non-posed (candid) expressions captured into a $640x \times 490px$ or $640 \times 480px$ frame, depending on the channel. The class Contempt despite being excluded from the range of the six basic expressions, was used mainly because it was reported to be found above 75% both in Western and non-Western cultures [22]. On the other hand, the neutral face was not considered in the training stage since it is hardly present in video-based classification. Only images with labels and in the peak of expression (apex state) were considered (1631 images), cropped as depicted in Fig. 18. The CKP dataset was split into 70% for training; the remaining were taken for the validation and test phases. In order to feed properly the network, the **CKP** set of images were **augmented** with random perturbations, based on the expressive results [62] from an experiment over a lower resolution dataset (**FER dataset**). The perturbation set, skew, translation, scale, rotation and horizontal flip worked separately in order to achieve a wider set, instead of the proposed overlapping method. Skew parameters were randomly selected from $\{-0.1, 0, 0.1\}$, translation parameters were sampled from $\{0, \delta\}$, where δ is a random sample from a $[0, 4]$ set, scaling uses a δ value to define a random parameter $c = 47/(47 - \delta)$ and the rotation is dependent on the angle sampled randomly from $\{-\pi/18, 0, \pi/18\}$. The final augmentation version has 978, 288 and 132 images per class in training, validation and test phases, respectively. We included a set of images to the test phase, populated with frames (Fig.17) from a real-time video framework composed by a Viola-Jones *OpenCV*⁴ face tracker [56] along with our classifier. Images were resized to the classifier input shape

⁴ <http://opencv.org/>

EXPERIMENTAL SETUP

($224 \times 224\text{px}$), captured within 35 frames per second, and classified in 0.250s (average including cropping process) into an expression displayed in the command line.



Figure 17: CKP static images vs our frame sequence, both cropped to 224×224 pixels

4.2 METHODS AND ALGORITHMS

The following sections describe the algorithms involved in the experiments to deploy the AFER system and the reasons behind those options.

Face acquisition is not included in model's experiments. We settled an algorithm for this phase based on state of the art and decided to focus on the two main challenges of an AFER system - Facial Expression Extraction and Recognition. The algorithm chosen, Viola-Jones algorithm, is described as a reliable source even in the presence of significantly input variance.

4.2.1 Facial Expression Extraction - Background Model

The first approach was developed taking the best results, of table 2, into account. It uses a combination of Gabor filters and Landmarks of the face. The choice of filter was based in the minimization of the absolute value of the spectral difference (E_{SD}) between the original image filter output ($O(u, v)|_{m,n}$) and the noisy image filter output ($O(u', v')|_{m,n}$) in the classification error rate.

$$E_{SD} = \arg \min_{m,n} \{O(u, v)|_{m,n} O(u', v')|_{m,n}\} \quad (2)$$

The best subset of filters is selected from F_i classification accuracy in training set, as shown in the following algorithm :

Algorithm 1 from [30]

- Create filter bank (G) by generating filters with different scales and orientations ($G = \{g_{0,0}, g_{0,1}, \dots, g_{m,n}\}$).
- Apply spectral difference E_{SD} for each member (g) $g \in G_{m,n}$.
- Create subsets (F_i) with different size (*i.e.* 4, 5, ...).
- Calculate the error rate (ϵ_i) for each subset.

$$\epsilon = \frac{\text{number of misclassified samples}}{\text{Total number of samples}}$$

- Select the subset with minimum error: $arg\,min\,\epsilon_i$
-

The landmarks were computed with the *connected components* or *blobs* using `Matlab-regionprops` function, expecting a label matrix containing contiguous regions. The centroids of these regions were assigned as features.

4.2.2 Facial Expression Extraction - Background Branch Model

Considering the flaws of the previous model we decided to come up with a better solution. Here we introduce a competitor to Gabor filter. The analysis of texture is made by creating a gray-level co-occurrence matrix (GLCM). The GLCM, uses a `graycomatrix` function for building the matrix by calculating how often a pixel with the intensity (gray-level) value i occurs in a specific spatial relationship to a pixel with the value j ⁵.

⁵ <http://www.mathworks.com/help/images/gray-level-co-occurrence-matrix-glcm.html>(accessed Sept 1, 2016)

The whole input face was partitioned into groups or regions (eyes, nose and mouth) discarding unnecessary features during model computation of the energy.

$$\sum_{ij} p(i, j)^2 \quad (3)$$

According to the feature extraction review, every model benefits with multi-algorithm combination, a reason to include the landmarks features from previous model.

Hereafter, we described the Deep Learning versions developed to overpass the results retrieved by the Background Methods.

4.2.3 Facial Expression Extraction and Recognition - CNNs

Herein we discriminate the steps towards the final model using Convolutional Neural Networks. The preliminary experiments include the first attempts to study CNNs while running a comparison with traditional methods. These experiments also helped to outline our baseline network.

Later in this section, we describe our model and explore the choices behind the network architecture and also its hyperparameters refinements.

Preliminary Experiments

Our preliminary deep model presents a final outputmap architecture, with 12 feature maps in convolutional layer (7×7), followed by a half-size downsampling layer and another convolutional layer (7×7) with six feature maps (12c-2s-6c), ending with a fully connected layer sized according to our class problem. The convolutional layers work with a 7×7 kernel size, which the product with input maps represents the kernel weights for the specific layer in process.

The parameters variation is limited to the image size ((28×28) - cropped version of the original (640×490), also responsible for the 3D input array ($28 \times 28 \times 327$)).

This preliminary experiments used a toolbox from Matlab - Deep Learning Toolbox created by *Rasmus Berg Palm* ⁶.

Once defined that the Deep Learning was the method to deploy the final AFER system, we focused the next experiments in the refinement of the most suitable network parameterizations.

Fixing the network, involved testing some prominent networks from the state of the art: the AlexNet from classifications of the ILSVRC2012 challenge [28], a more recent GoogLeNet and the classic LeNet-5 to report a fair judgment.

The non-augmented dataset was considered small enough for a CNN input and therefore a candidate to make a sanity check on the hyperparameters. In this context, we used GoogleNet and LeNet-5 and both passed the test, overfitting with an high accuracy between [0.9,0.92] in training, whereas the validation loss computed by summing the total weighted loss over the network (for the outputs with non-zero loss), reached a value of 0.8. In order to follow the right network in place, we conducted some preliminary classification experiments over a short amount of training time and the best gains in the test set came from the LeNet-5 as depicted in Fig. 19.

Since the results over the training (Table 4) and validation set had a retarded loss decay, no further experiments on GoogleNet and Alexnet were developed because training memory and processing time would become an issue.

Data	Solver	Net	Loss Train
CKP + Augm	Ada	Alexnet	0.0082
CKP + Augm	SGD	Alexnet	0.0004
CKP + Augm	SGD	Googlenet	0.0001
CKP + Augm	SGD	Lenet Ov	0.0001

Table 4: Train loss results used to select the shallow network

Considering these marks, the classic LeNet-5 architecture of CNNs was used as baseline for this experiments.

⁶ <http://www.mathworks.com/matlabcentral/fileexchange/38310-deep-learning-toolbox>(accessed Sept 1, 2016)

Lenet-5, Our version

The model (Fig. 18) is composed by an initial convolutional set with a 5×5 kernel size and 20 feature maps plus a shared bias ending up with 520 parameters. The next layer or Pooling Layer performs a downsampling with a maximum value of a 2×2 kernel size. This process is repeated except for the pooling stride which changed to 2 and the convolution process expecting 50 instead of 20 feature maps, augmenting the parameters to 1300. We used in our work Krizhevsky [28] alternative model neurons output. Instead of the standard functions $f(x) = \tanh(x)$ or $f(x) = (1 - e^{-x})^{-1}$, we use a faster version $f(x) = \max(0, x)$ designated as *Rectified Linear Units* (ReLU). The convolutional neural networks with ReLU proved to be 6 times faster than an equivalent network using saturating neurons, reaching 25% of training error rate. Taking this into account, the full connected layer that follows, containing 500 filter numbers, is connected with an ReLU. So far, the structure is similar to Lenet-5, however we introduced a *Dropout* between the full connected layers, reducing between 0.4 and 0.5 percent of their connectivity by dropping randomly some units or neurons which do not contribute to the forward pass. This procedure will overcome overfitting. Dropout prevents co-adaptation showing a significant improvement by 10% of accuracy, namely in the ILSVRC 2012 validation and test sets [50]. Finally, the last full connected layer is responsible for shrinking the feature maps to our class problem - (seven expressions).

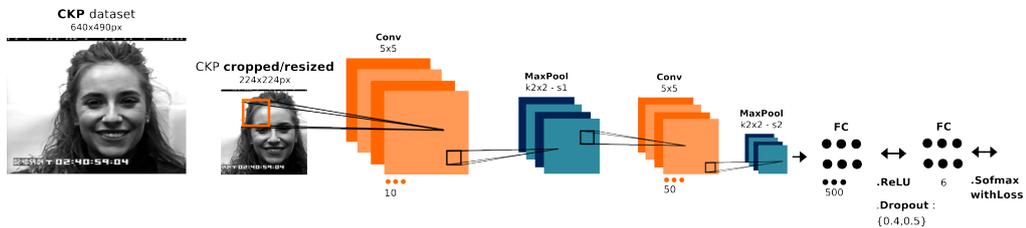


Figure 18: Flowchart of the different stages of our CNN model, adapted from the classic LeNet-5 - Lenet Ov (Our version)

We specified several heuristics for the model configuration and ran the experiments in a standalone version of *Caffe*⁷ running over an *Amazon EC2* instance, in order to access GPU advanced hardware.

Moreover, LeNet-5 baseline parameters were addressed from the *Caffe* standards.

The network was prepared taking into consideration the dependency on the convolutional neural networks to the feature extraction process, and how first layers and their high level of information are determinant to the success of the classification. Therefore, we tested different ways of adjusting the weights involved. Our network was tested with three types of fillers, namely, *gaussian*, *positive unitball* and *xavier*. The *gaussian* filler only chooses values according a gaussian distribution, limiting non-zero inputs up to 3, and the standard deviation assume a 0.01 value (increased from the default 0.005). The *positive unitball* fills a blob with values between $[0, 1]$ such that $\forall i \sum_j x_{ij} = 1$. Finally, the *xavier* type (weight filler) initializes the incoming matrix with values from an uniform distribution within $[-\sqrt{\frac{3}{n}}, \sqrt{\frac{3}{n}}]$, where the n is the number of the input neurons. This *Caffe* version of Xavier differs from what was initially introduced by Glorot [25], removing the output information. Our version of the network is optimized with a stochastic gradient descent solver, since the Alexnet trained with Adagrad (see Fig. 19) was not expressive. The solver hyperparameters were highly influenced with the results [14] from an automatic hyperparameters optimization over the MNIST dataset. Our parameters fit their hyperspace, with 0.09 for the momentum, 0.0005 of weight decay and a initial learning rate of 0.001, dropping a factor of 10 in the last 10% of iterations.

Our model produces a set of discrete class labels (0-6) or predicted classes which can be distinguished from the actual classes with the information provided by the confusion matrices, as shown in Fig. 19.

From the model above it was possible to extract after test set classification (Table 5), three types of outcomes, as depicted in Table 7, and infer at least two scores, Precision

⁷ The *Caffe* is a public C++ deep learning library, with contributions at GitHub (<http://github.com/BVLC/caffe>) and was used in our experiments, considering that it enables Python and MATLAB bindings, and has modularity properties granting that any part of the model assembled can be easily exchanged. These facts were highlighted by a Berkeley Vision and Learning Center (BVLC) group (<http://caffe.berkeleyvision.org/> (accessed Sept 1, 2016)).

EXPERIMENTAL SETUP

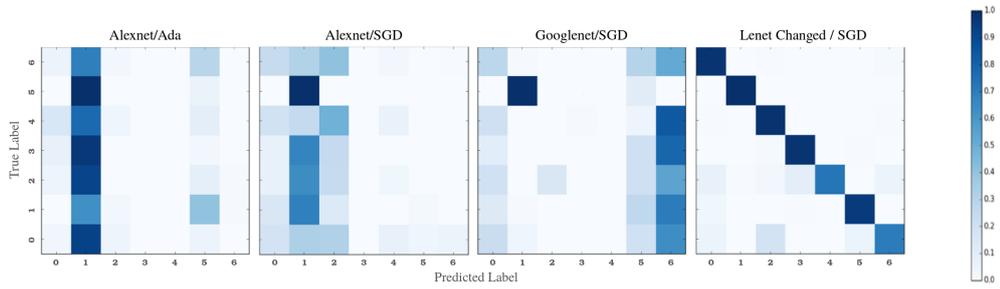


Figure 19: Normalized Confusion Matrices - (a) Alexnet/Ada; (b) Alexnet/SGD; (c) Googlenet/SGD; (d) Lenet changed/SGD.

(P) and Recall (R). In order to determine if the model was actually a reliable source, a video framework was built. From the framework used to test the model (Fig. 20) a confusion matrix with 30 frames per class is presented. Ideally we should gather with 132 different subjects to test our framework, equal to the number used to test CKP dataset, however this simple test is enough to infer the learning stage. During this test, we tried to mimic the CKP dataset then expecting same labels in return.

Weight/Class	Anger	Contempt	Disgust	Fear	Happiness	Sadness	Surprise
Gaussian (a)	0.98	1	0.89	0.86	0.59	0.79	0.48
UnitBall (b)	0.98	1	1	0.91	0.78	0.95	0.70
Xavier(c)	0.98	1	1	0.92	0.75	0.92	0.70

Table 5: Test Accuracy with three versions of the CNN model (a), (b) and (c).

 EXPERIMENTAL RESULTS AND DISCUSSION

In this section we discuss the results from the preliminary experiments, as well as the final model achievements comparing to our video framework.

Table 6 makes a comparison of the best recognition results regarding the background models and preliminary tests with a CNN model. The background models did not performed as good as the Deep Learning solution which achieved **86.8%** of accuracy rate. A low number of patterns were presented to the neural network and it was expected less accuracy then the alternatives, however the type of the neural network used - convolutional - overcame this issue. This fact guided us to further experiments only considering Convolutional Neural networks.

Expression/ Models	Conventional#1 (SVM)	Conventional#2 (SVM)	CNN (BP)
Anger	92%	66,7%	64,7%
Disgust	37,5%	100% (4 false alarms)	100% (1 false alarms)
Fear	21%	100% (4 false alarms)	55,6%
Happiness	90%	100% (37 false alarms)	95,4%
Sadness	37%	23%	100% (1 false alarms)
Surprise	20%	0%	92%

Table 6: Methods for comparison

Fig. 19 shows the density of the first classification tests (trained over 20000 iterations) after the preliminary experiments. This demonstration crossed with the Table 4 was used to select the best network shape for the problem, filtered by the most accurate results over a short period of training. From Table 4 we can also infer that the poor

expressiveness between methods is representative of the lack of some augmentation diversity.

As explained earlier in this section, the most complex networks (AlexNet and GoogLeNet) were discarded due to their poor performance which points to an unbalance relation within the low dimension dataset (even when augmented). The best results from the 3-type version of our assembled model used the *UnitBall* configuration trained over 100000 iterations and achieved a top recognition with F1-score of 0.906, representing an average value retrieved from Table 7. The version accuracy, also 90% on test set, is close to the current state of the art, with the CKP dataset represented by a 93% of accuracy.

	Gaussian (a)			UnitBall (b)			Xavier (c)		
	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1
Anger	0.747	0.985	0.850	0.949	0.985	0.967	0.833	0.985	0.903
Contempt	0.949	1	0.974	1	1	1	1	1	1
Disgust	0.652	0.894	0.754	0.820	1	0.901	0.830	1	0.907
Fear	0.820	0.864	0.41	0.924	0.917	0.920	0.897	0.924	0.910
Happiness	1	0.591	0.743	0.972	0.780	0.866	0.943	0.750	0.835
Sadness	0.840	0.795	0.817	0.881	0.955	0.916	0.884	0.924	0.904
Surprise	0.716	0.477	0.573	0.861	0.705	0.775	0.949	0.705	0.809

Table 7: Precision, Recall and F1 from three above versions (a)(b)(c).

According to Table 5 the model with the best results (Unitball(b)) was the same used to perform the tests in the video framework that captured 30 frames per class. Fig. 20 shows that in our framework, the expression of Contempt is the most expressive, misleading the recognition of others. Moreover, the expression of Anger, Happiness and Sadness were not properly learned. Despite, the results not matching the desirable performance of the static test version, there are some significant achievements comparing to the state of the art, where some of the "emotions in the wild" [62] were also misclassified or inexistent (such as Disgust in the test set). In this case study, difficulty of video environment constraints were compared with the complexity of their dataset, which presents more candid images.

EXPERIMENTAL RESULTS AND DISCUSSION

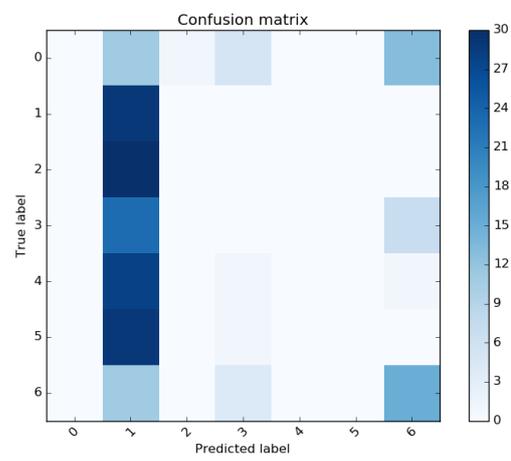


Figure 20: Confusion Matrix for the classification of the Framework in the test set

CONCLUSION AND FUTURE WORK

The main goal of this dissertation was to define the best method to tackle an AFER system with images as an input. To achieve this, we conducted an extensive research on the alternatives for each AFER system stage.

Deep Learning proved to overcome the conventional problems.

Deep convolutional neural network method was selected from preliminary results in order to integrate the final deployment.

We developed a CNN model tailored for the problem of static facial expressions recognition using the CKP dataset and customized hyperparameters.

After several enhancements, the results of our crafted network design are significant, attaining 90% of accuracy in the test set. Both the inclusion of a Dropout layer in architectural decisions and the augmentation performed on the dataset with several transformations led to a superior performance with the seven facial expressions. Despite limitations, the deep inference model is able to perform also in sequence-based video frames, at least for some of the facial expressions (a problem that also exists in "wilder" datasets).

This dissertation proved that we can start from static baselines to build a sequence model relying on deep emotion recognition models. This work should not be seen as a discouragement of the conventional methods, but instead an acknowledgment of the barriers related to feature extraction refinement.

CONCLUSION AND FUTURE WORK

In the final model, we could improve the accuracy with a threshold value to infer the classes with best confidence. However, if further work includes sequence base systems, the use of a threshold is questionable, since time constraints are an issue during recognition.

Future work should focus on the improvement of the training set by increasing transformations and promoting different lighting and positions.

Moreover, the real-time face tracker should also be improved with recent methodologies that could tackle the partial face challenge within frame correlations. Another possibility is to include active learning with human judgment.

The work developed in this dissertation resulted in an article [32] (Appendix B) that it will be presented at the International Conference of Pattern Recognition and will be published in the Conference Proceedings endorsed by Springer.

All in all, this dissertation constitutes an interesting "stepping stone" for future work in Facial Expression Recognition.

APPENDIX A

6.1 FACE ANATOMY

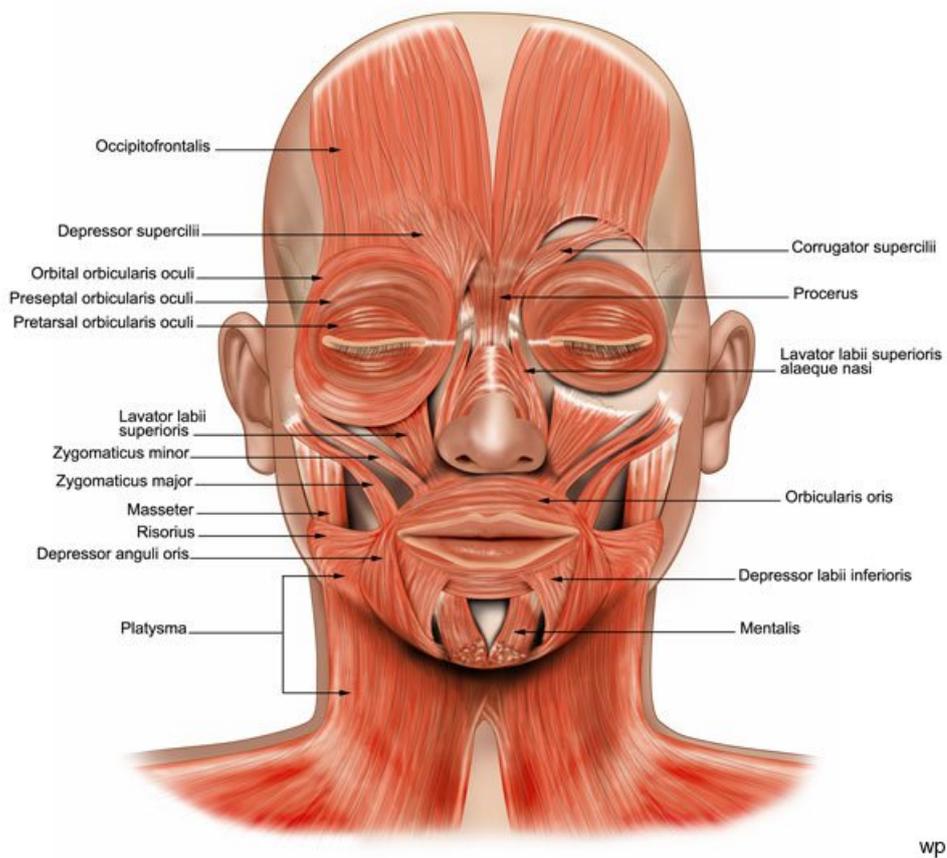


Figure 21: http://medicalart-work.co.uk/wordpress/?page_id=8(accessed gallery: Sept 1, 2016)

CONCLUSION AND FUTURE WORK

6.2 FACS ACTION UNITS

Upper Face Action Units					
AU 1	AU 2	AU 4	AU 5	AU 6	AU 7
					
Inner Brow Raiser	Outer Brow Raiser	Brow Lowerer	Upper Lid Raiser	Cheek Raiser	Lid Tightener
*AU 41	*AU 42	*AU 43	AU 44	AU 45	AU 46
					
Lid Droop	Slit	Eyes Closed	Squint	Blink	Wink
Lower Face Action Units					
AU 9	AU 10	AU 11	AU 12	AU 13	AU 14
					
Nose Wrinkler	Upper Lip Raiser	Nasolabial Deepener	Lip Corner Puller	Cheek Puffer	Dimpler
AU 15	AU 16	AU 17	AU 18	AU 20	AU 22
					
Lip Corner Depressor	Lower Lip Depressor	Chin Raiser	Lip Puckerer	Lip Stretcher	Lip Funneler
AU 23	AU 24	*AU 25	*AU 26	*AU 27	AU 28
					
Lip Tightener	Lip Pressor	Lips Part	Jaw Drop	Mouth Stretch	Lip Suck

Figure 22: FACS Action Units Table.11.1 at [54]

6.3 HMM

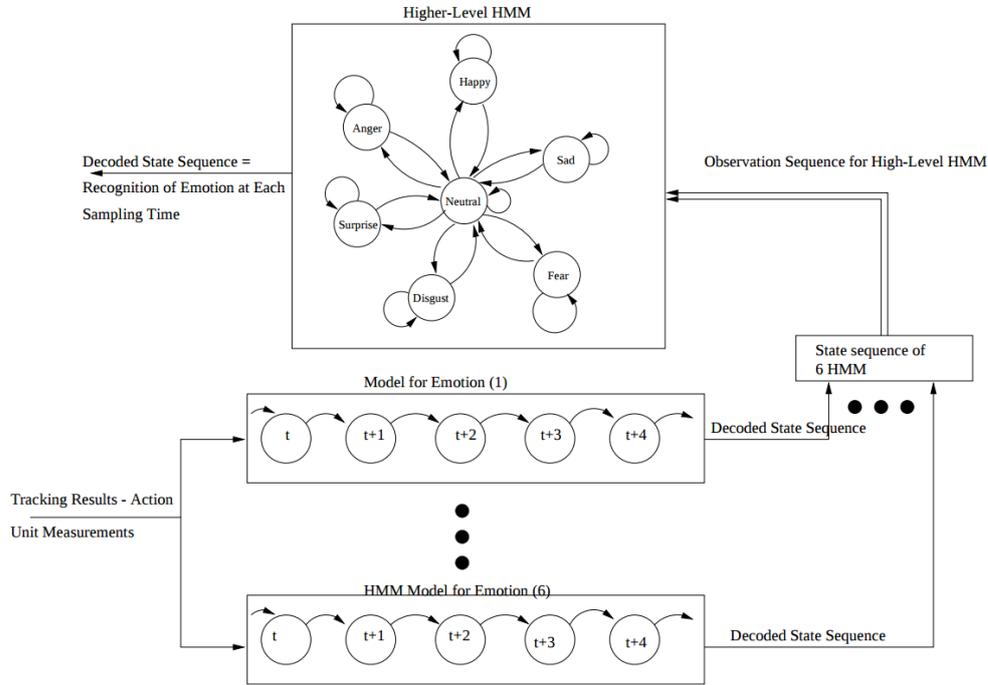


Figure 23: Multilevel HMM architecture for automatic segmentation and recognition of emotion reprinted from [8]

6.4 LENET-5

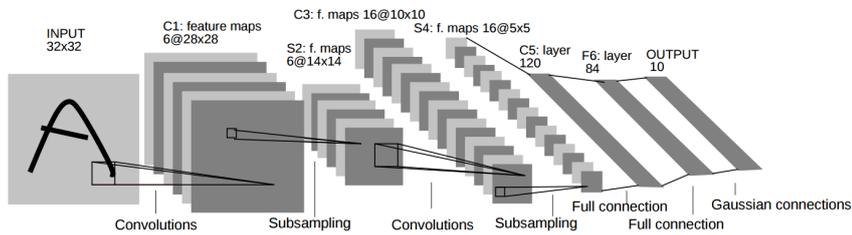


Figure 24: Architecture of LeNet-5 representing Figure 2 from [34]

APPENDIX B

How Deep can we Rely on Emotion Recognition

Ana Laranjeira¹, Xavier Frazão², André Pimentel², and Bernardete Ribeiro¹

¹CISUC - Department of Informatics Engineering, University of Coimbra, Portugal

²EyESee Solutions, Lisboa, Portugal

bribeiro@dei.uc.pt, afolgado@student.dei.uc.pt

Abstract. The emerging success of digital social media has had an impact on several fields ranging from science to economy and business. Therefore, there is an invested interest in emotion detection and recognition technology from facial expressions, in order to increase their market competitiveness. This area still presents many challenges, namely the difficulty in achieving real-time facial recognition. Herein we tackle this problem by crossing methods targeting both static images and active images. In this work, we explore the recent technological breakthroughs in deep learning and develop a system based on automatic recognition of human face expressions using Convolutional Neural Networks (CNN). We use the Cohn-Kanade Extended (CKP) dataset for testing our proposed CNN model along with an augmented version, which demonstrated effectiveness in seven basic expressions. In order to enhance the quality of the results instead of the overlapping method for building the augmented dataset we propose random perturbations from a wide set including: skew, translation, scale, and horizontal flip. Moreover, we built a real-time video framework using our model (a version of LeNet-5) which is fed with frames detected with Viola-Jones face tracker that reproduce the CKP dataset. The results are promising.

Keywords: Emotion recognition, Convolutional Neural Networks

1 Introduction

Facial Expressions are the source of Human Emotion Recognition and a deep understanding on these emotional responses gives a major advantage to any dependent field. The information retrieved from emotion detection and recognition technology will leverage a wide range of applications such as lie detections, pain assessment, surveillance, healthcare, consumer electronics, law enforcement and many others dependent on Human Computer Interfaces. The challenge in the field is to develop a way to integrate a large spectrum of expressions intrinsically related to the capacity of humans to express feelings. However, there are six basic (or prototypic) expressions that can be proven consensual within any culture, namely *Anger*, *Disgust*, *Fear*, *Happiness*, *Sadness*, *Surprise* [4]. These have been used by the majority of Facial Expression Recognition (FER) systems.

A classic FER system is structured according to three stages, starting with facial detection, followed by the feature extraction and the final stage, expression recognition. The first stage is the most explored and can be considered as an exhausted topic. A few surveys have identified and categorized many proposed improvements and extensions [1]. Viola and Jones made a boost in facial detection [16] and their classifier is still widely used despite the recent appearance of more effective cases [6]. The CIFE dataset survey [12] includes a descriptive table of the algorithms used for the above stages, and concludes that feature-based approaches perform better. Most of the drawbacks come from the dataset construction, performed under a controlled

environment (positions, lightning, no occlusions), therefore jeopardizing the robustness of the system in presence of invariance. Despite these constraints, in the past few years with hardware technological innovations and the consequent development of deep learning hierarchical models it was possible to extract complex data interactions from large scale datasets and build advanced solutions on demand. Within this context, an important and possible solution has been revisited - **Convolutional Neural Networks**, leading to significant results in image base classification, as demonstrated in the annual contest *ImageNet Large Scale Visual Recognition Challenge*(ILSVRC). There are some works that stand out for their performance in this large image classification problem, the *AlexNet* [10] from 2012 and a more recent, *GoogLeNet* [15] inspired in a deeper concept - "Network inside Network".

In this paper, we propose and develop a Convolutional Neural Network for emotion facial recognition. Our proposed model was trained with seven expressions, introducing also the expression of *Contempt* to the prototypic set, based on its presence in a widely explored dataset, the **Cohn-Kanade Extended** [13] and therefore considered as a reliable source. In order to enhance the quality of the results we have augmented the dataset based on random perturbations from a wide set including: skew, translation, scale, and horizontal flip. We specified several heuristics for the model configuration and ran the experiments in a standalone version of *Caffe* [9], supported by an instance from *Amazon EC2* which allows access to GPU advanced hardware. We are aware that in the literature the methods are dependent on the type of dataset, falling into static-images or dynamic images sequences. In the case of sequence-based approaches, the temporal information could be a concern due to real time constraints. In this work, we look at how the static methods can be applied to sequence events without compromising the reliability and even improving speed of recognition. We put forward, a real-time video framework composed by Viola-Jones *OpenCV*¹ face tracker. This framework shows promising preliminary results with our deep approach and matches the question we try to answer: How deep can we rely on emotion recognition? Future work should fully answer this question.

This paper is organized as follows. In the next section we present the related work highlighting inspiring studies in the field of emotion recognition. In Section 3 our proposed model is described from both static image data and sequence based data for real-time application. In Section 4 the datasets, the experimental setup, the technicalities of the approach and the evaluation metrics of our model classifier are described. In Section 5 the results are discussed. Finally, in Section 6 the conclusions are presented with the future work.

2 Related Work

An annually scientific contest *Emotion Recognition in the Wild Challenge* (EmotiW) is defining the state of the art focusing on affective sensing with uncontrolled environments². From the latest year contest (EmotiW 2015), two versions of dealing with expressions can be selected: one uses static images from Static Facial Expressions in the Wild (SFEW); and the other uses an acted point of view resorting to an Acted Facial Expressions in the Wild (AFEW) dataset. Among the static-image approaches the project [17] proposes a 3-way detection of the face, with a hierarchical selection from the Joint Cascade Detection and Alignment (JDA), Deep CNN-Based (DCNN) and MoT along with a simple network (11 layers). These detectors are processed in a multiple network framework in order to enhance the performance. It also includes a pre-processing phase to improve accuracy, which might be considered a drawback in the classification response. An approach resorting to video is strongly dependent on spatio-temporal issues. In [3]

¹ <http://opencv.org/>

² <https://cs.anu.edu.au/few/emotiw2015.html>

a Synchrony Autoencoder (SAE) was introduced for local motion feature extraction together with assembling an hybrid network, CNN-RNN.

3 Proposed Emotion Recognition Model

Our proposed model was constructed with LeNet-5 as baseline and has been progressively developed. Its current stage is depicted in Fig. 1.

The classic LeNet-5 architecture [11], is a combination of seven layers represented by a convolutional layer with 16 feature maps with a 5×5 kernel size, followed by a sub-sampling by half. This sub-sampling (pooling layer) and the next convolutional layer (16 feature maps) are not connected in order to break symmetry. The network includes also another pooling stage, maintaining the 16 feature maps and the kernel to 5×5 . As opposed to our version of the LeNet-5, it includes another convolutional full sized (kernel 1×1), mapping 120 units before the last full connected layers with 84 and 10 neurons, respectively (since LeNet is addressed to digit recognition).

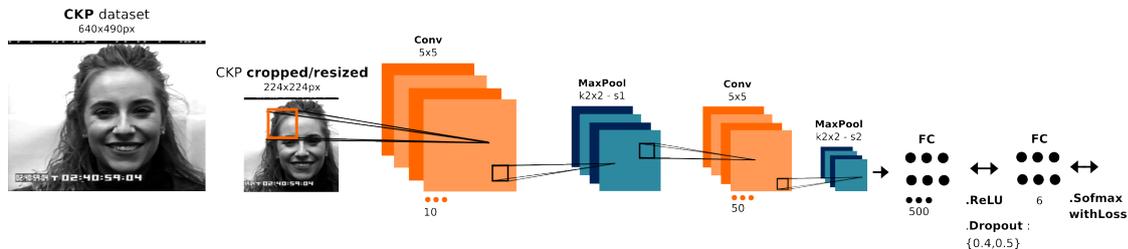


FIG. 1: Flowchart of the different stages of our CNN model, adapted from the classic LeNet-5 - Lenet Ov (Our version)

Our model is composed by an initial convolutional set with a 5×5 kernel size and 20 feature maps plus a shared bias ending up with 520 parameters. The next layer or Pooling Layer performs a downsampling with a maximum value of a 2×2 kernel size. This process is repeated except for the pooling stride which changed to 2 and the convolution process expecting 50 instead of 20 feature maps, augmenting the parameters to 1300. We used in our work Krizhevsky [10] alternative model neurons output. Instead of the standard functions $f(x) = \tanh(x)$ or $f(x) = (1 - e^{-x})^{-1}$, we use a faster version $f(x) = \max(0, x)$ designated as *Rectified Linear Units* (ReLU). The convolutional neural networks with ReLU proved to be 6 times faster than an equivalent network using saturating neurons, reaching 25% of training error rate. Taking this into account, the full connected layer that follows, containing 500 filter numbers, is connected with an ReLU. So far, the structure is similar to Lenet, however we introduced a *Dropout* between the full connected layers, reducing between 0.4 and 0.5 percent of their connectivity by dropping randomly some units or neurons which do not contribute to the forward pass. This procedure will overcome overfitting. Dropout prevents co-adaption showing a significant improvement by 10% of accuracy, namely in the ILSVRC 2012 validation and test sets [14]. Finally, the last full connected layer is responsible for shrinking the feature maps to our class problem - 7 (expressions). The weights presented to the net follow the *xavier* type except for the first convolutional layer which is set between *gaussian*, *unitball*, *xavier*.

4 Experimental Design

The **Cohn-Kanade Extended** dataset [13] (CKP) set is labeled between 0–7 corresponding to *Neutral, Anger, Contempt, Disgust, Fear, Happiness, Sadness, Surprise* and contains 593 sequences across 123 subjects with posed and non-posed (candid) expressions captured into a $640x \times 490px$ or $640 \times 480px$ frame, depending on the channel. The class Contempt despite being excluded from the range of the six basic expressions, was used mainly because it was reported to be found above 75% both in Western and non-Western cultures [5]. On the other hand, the neutral face was not considered in the training stage since it is hardly present in video-based classification. Only images with labels and in the peak of expression (apex state) were considered (1631 images), cropped as depicted in Fig. 1. The CKP dataset was split into 70% for training; the remaining were taken for the validation and test phases. In order to feed properly the network, the **CKP** set of images were **augmented** with random perturbations, based on the expressive results [17] from an experiment over a lower resolution dataset (**FER dataset** [8]). The perturbation set, skew, translation, scale, rotation and horizontal flip worked separately in order to achieve a wider set, instead of the proposed overlapping method. Skew parameters were randomly selected from $\{-0.1, 0, 0.1\}$, translation parameters were sampled from $\{0, \delta\}$, where δ is a random sample from a $[0, 4]$ set, scaling uses a δ value to define a random parameter $c = 47/(47 - \delta)$ and the rotation is dependent on the angle sampled randomly from $\{-\pi/18, 0, \pi/18\}$. The final augmentation version has 978, 288 and 132 images per class in training, validation and test phases, respectively. We included a set of images to the test phase, populated with frames (Fig.2) from a real-time video framework composed by a Viola-Jones *OpenCV*³ face tracker [16] along with our classifier. Images were resized to the classifier input shape ($224 \times 224px$), captured within 35 frames per second, and classified in 0.250s (average including cropping process) into an expression displayed in the command line. The experiments

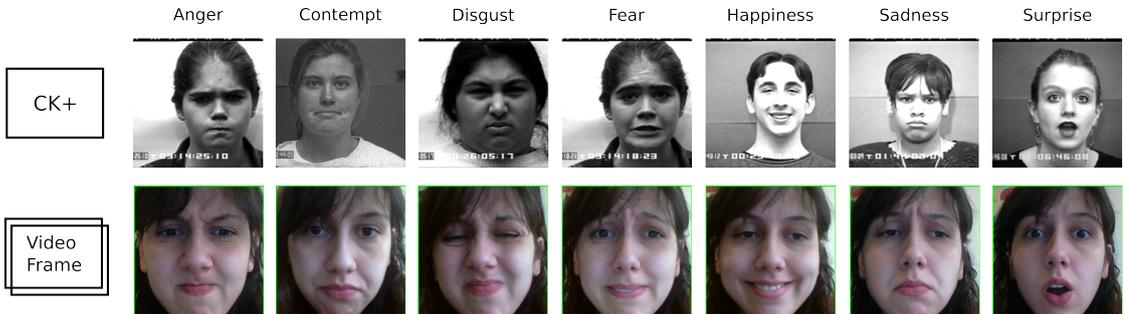


FIG. 2: CKP static images vs our frame sequence, both cropped to 224×224 pixels

started from the detriment of the network, questioning the relevant inner parameters comparing with an appropriate solver. Fixing the network, involved testing some prominent networks from the state of the art: the AlexNet from classifications of the ILSVRC2012 challenge [10], the recent GoogLeNet and the classic LeNet-5, to report a fair judgment. The non-augmented dataset was considered small enough for a CNN input and therefore a candidate to make a sanity check on the hyperparameters. In this context, we used GoogLeNet and LeNet-5 and both passed the test, overfitting with an high accuracy between $[0.9, 0.92]$ in training, whereas the validation loss

³ <http://opencv.org/>

computed by summing the total weighted loss over the network (for the outputs with non-zero loss), reached a value of 0.8. In order to follow the right network in place, we conducted some preliminary classification experiments over a short amount of training time and the best gains in the test set came from the LeNet-5 as depicted in Fig. 3. Since the results over the training (Table 1) and validation set had a retarded loss decay, no further experiments on GoogleNet and Alexnet were developed because training memory and processing time would become an issue. Considering these marks, a new version of LeNet-5 (LeNet - Ov) was explored. Moreover, LeNet-5 baseline parameters were addressed from the *Caffe* standards.

The network was prepared taking into consideration the dependency on the convolutional neural networks to the feature extraction process, and how first layers and their high level of information are determinant to the success of the classification. Therefore we tested different ways of adjusting the weights involved. Our network was tested with three types of fillers, namely, *gaussian*, *positive unitball* and *xavier*. The *gaussian* filler only chooses values according a gaussian distribution, limiting non-zero inputs up to 3, and the standard deviation assume a 0.01 value (increased from the default 0.005). The *positive unitball*, fills a blob with values between $[0, 1]$ such that $\forall i \sum_j x_{ij} = 1$. Finally, the *xavier* type (weight filler), initializes the incoming matrix with values from an uniform distribution within $[-\sqrt{\frac{2}{n}}, \sqrt{\frac{2}{n}}]$, where the n is the number of the input neurons. This *Caffe* version of Xavier differs from what was initially introduced by Glorot [7], removing the output information. Our version of the network is optimized with a stochastic gradient descent solver, since the Alexnet trained with Adagrad (see Fig. 3) was not expressive. The solver hyperparameters were highly influenced with the results [2] from an automatic hyperparameters optimization over the MNIST dataset. Our parameters fit their hyperspace, with 0.09 for the momentum, 0.0005 of weight decay and a initial learning rate of 0.001, dropping a factor of 10 in the last 10% of iterations. Our model produces a set of discrete class labels (0-6) or predicted classes which can be distinguished from the actual classes with the information provided by the confusion matrices, as shown in Fig. 3. It is possible then to extract from the test set classification (Table 2), three types of outcomes, as depicted in Table 3, and infer at least two scores, Precision (P) and Recall (R). From the framework used to test the model (Fig. 4) a confusion matrix with 30 frames per class is presented. Ideally we should gather with 132 different subjects to test our framework, equal to the number used to test CKP dataset, however this simple test is enough to infer the learning stage. During this test, we tried to mimic the CKP dataset then expecting same labels in return.

5 Results and Discussion

In this section we discuss the results from the preliminary experiments, as well as the final model achievements comparing to our video framework. Fig. 3 shows the density of the first classification tests, trained just over 20000 iterations. This demonstration crossed with the Table 1 was used to select the best network shape for the problem, filtered by the most accurate results over a short period of training. From Table 1 we can also infer that the poor expressiveness between methods is representative of the lack of some augmentation diversity. As explained in Section 3, the most complex networks (AlexNet and Googlenet) were discarded due to their poor performance which reinforces the relation unbalance with the low dimension dataset (even augmented). The best results from the 3-type version of our assembled model used the *UnitBall* configuration trained over 100000 iterations and achieved a top recognition with F1-score of 0.906, representing an average value retrieved from Table 3. The version accuracy, also 90% on test set, is close to the current state of the art, with the CKP dataset represented by a 93% of accuracy. We could improve the accuracy with a threshold value to infer the classes with best confidence, however the

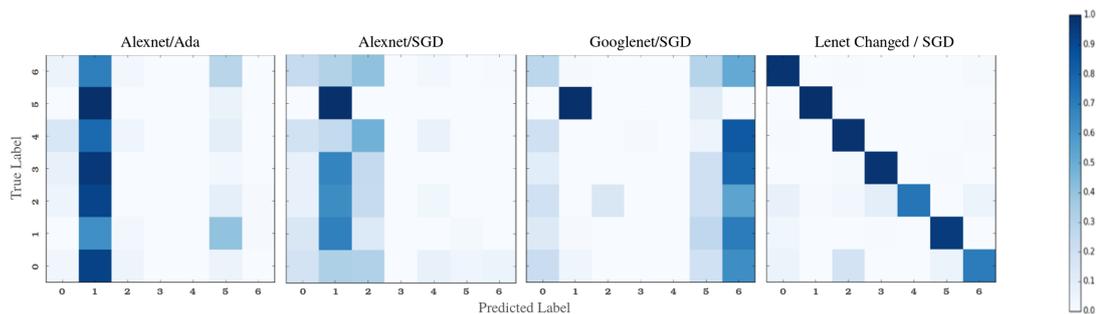


FIG. 3: Normalized Confusion Matrices - (a) Alexnet/Ada; (b) Alexnet/SGD; (c) Googlenet/SGD; (d) Lenet changed/SGD.

purpose of this paper is to evaluate a sequence base system which is temporal dependent, due to this, we made an effort to reduce recognition time. According to Table 2 the model with the best results was the same used to perform the tests in the video framework that captured 30 frames per class. Fig. 4 shows that in our framework, the expression of Contempt is the most expressive, misleading the recognition of others. Moreover, the expression of Anger, Happiness and Sadness were not properly learned. Despite, the results not matching the desirable performance of the static test version, there are some significant achievements comparing to the state of the art, where some of the “emotions in the wild” [17] were also misclassified or inexistent (such as Disgust in the test set). In this case, difficulty of video environment constraints were compared with the complexity of their dataset, which presents more candid images.

Data	Solver	Net	Loss Train
CKP + Augm	Ada	Alexnet	0.0082
CKP + Augm	SGD	Alexnet	0.0004
CKP + Augm	SGD	Googlenet	0.0001
CKP + Augm	SGD	Lenet Ov	0.0001

TABLE 1: Train loss results used to select the shallow network

Weight/Class	Anger	Contempt	Disgust	Fear	Happiness	Sadness	Surprise
Gaussian (a)	0.98	1	0.89	0.86	0.59	0.79	0.48
UnitBall (b)	0.98	1	1	0.91	0.78	0.95	0.70
Xavier(c)	0.98	1	1	0.92	0.75	0.92	0.70

TABLE 2: Test Accuracy with three versions of the CNN model (a), (b) and (c).

	Gaussian (a)			UnitBall (b)			Xavier (c)		
	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1
Anger	0.747	0.985	0.850	0.949	0.985	0.967	0.833	0.985	0.903
Contempt	0.949	1	0.974	1	1	1	1	1	1
Disgust	0.652	0.894	0.754	0.820	1	0.901	0.830	1	0.907
Fear	0.820	0.864	0.41	0.924	0.917	0.920	0.897	0.924	0.910
Happiness	1	0.591	0.743	0.972	0.780	0.866	0.943	0.750	0.835
Sadness	0.840	0.795	0.817	0.881	0.955	0.916	0.884	0.924	0.904
Surprise	0.716	0.477	0.573	0.861	0.705	0.775	0.949	0.705	0.809

TABLE 3: Precision, Recall and F1 from three above versions (a)(b)(c).

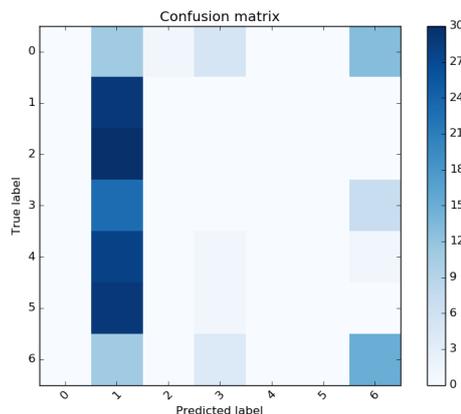


FIG. 4: Confusion Matrix for the classification of the Framework in the test set

6 Conclusion

In this paper, we present a deep inference model for emotion recognition from facial expressions. We developed a convolutional neural network tailored for the problem of emotion recognition from static facial expressions using the CKP dataset. After several enhancements, the results of our crafted network design are significant, attaining 90% of accuracy in the test set. Both the inclusion of a Dropout layer in architectural decisions and the augmentation performed on the dataset with several transformations, led to a superior performance with the seven facial expressions. Although the results are preliminary, the deep inference model is able to perform also in a sequence of video frames, in real-time, at least for some of the facial expressions (a problem that also exists in “wilder” datasets). This paper proved that we can start from static baselines to build a sequence model, thus answering in a positive way our initial question that we can rely on a deep emotion recognition model. Future work should focus on the improvement of the training set by increasing transformations and promoting different lighting and positions. Moreover, the real-time face tracker should also be improved by including active learning with human judgment.

References

1. Bettadapura, V.: Face Expression Recognition and Analysis: The State of the Art. CoRR abs/1203.6722 (2012)

2. Domhan, T., Springenberg, J.T., Hutter, F.: Speeding up Automatic Hyperparameter Optimization of Deep Neural Networks by Extrapolation of Learning Curves. In: Proc. of the 24th Int Joint Conf on Artificial Intelligence (IJCAI) (2015)
3. Ebrahimi, S., Michalski, V., Konda, K., Memisevic, R., Pal, C.: Recurrent Neural Networks for Emotion Recognition in Video. In: Proc. of the ACM on Int Conf on Multimodal Interaction. pp. 467–474. ACM (2015)
4. Ekman, P., Friesen, W.: Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology* 17(2), 124–129 (1971)
5. Ekman, P., Heider, K.G.: The universality of a contempt expression: A replication. *Motivation and Emotion* 12(3), 303–308 (1988)
6. Farfadi, S., Saberian, M., Li, L.J.: Multi-view Face Detection Using Deep Convolutional Neural Networks. In: Proc. of the 5th ACM on Int Conf on Multimedia Retrieval. pp. 643–650 (2015)
7. Glorot, X., Bengio, Y.: Understanding the difficulty of training deep feedforward neural networks. In: *Int Conf on Artificial Intelligence and Statistics* (2010)
8. Goodfellow, I.J., et al.: Challenges in Representation Learning: A report on three machine learning contests. In: *Int Conf On Neural Information Processing* (2013)
9. Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T.: Caffe: Convolutional Architecture for Fast Feature Embedding. In: Proc. of the ACM Int Conf on Multimedia. pp. 675–678 (2014)
10. Krizhevsky, A., Sutskever, I., Hinton, G.: ImageNet Classification with Deep Convolutional Neural Networks. In: Pereira, F., et al. (eds.) *Advances in Neural Information Processing Systems 25*, pp. 1097–1105. Curran Associates, Inc. (2012)
11. Lecun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. *Proc. of the IEEE* 86(11), 2278–2324 (1998)
12. Li, W., Li, M., Su, Z., Zhu, Z.: A deep-learning approach to facial expression recognition with candid images. In: 14th IAPR Int Conf on Machine Vision Applications. pp. 279–282 (2015)
13. 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: *IEEE Conf. on Computer Vision and Pattern Recognition*. pp. 94–101 (2010)
14. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *J. Mach. Learn. Res.* 15(1), 1929–1958 (2014)
15. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going Deeper with Convolutions. In: *IEEE Conf. on Computer Vision and Pattern Recognition*. pp. 1–9. IEEE (Jun 2015)
16. Viola, P., Jones, M.J.: Robust Real-Time Face Detection. *Int. J. Comput. Vision* 57(2), 137–154 (2004)
17. Yu, Z., Zhang, C.: Image based Static Facial Expression Recognition with Multiple Deep Network Learning. In: Proc. of the ACM on Int Conf on Multimodal Interaction (2015)

BIBLIOGRAPHY

- [1] G.A. Abrantes and F. Pereira. MPEG-4 facial animation technology: survey, implementation, and results. *Circuits and Systems for Video Technology, IEEE Transactions on*, 9(2):290–305, Mar 1999.
- [2] Ralph Adolphs. *Recognizing Emotion From Facial Expressions: Psychological and Neurological Mechanisms*, 2002.
- [3] Mohamed Aly. *Survey on Multiclass Classification Methods*, 2005.
- [4] Yoshua Bengio. Learning Deep Architectures for AI. *Found. Trends Mach. Learn.*, 2(1):1–127, January 2009.
- [5] Vinay Bettadapura. Face Expression Recognition and Analysis: The State of the Art. *CoRR*, abs/1203.6722, 2012.
- [6] V Bruce and A Young. Understanding face recognition. *Br J Psychol*, 77 (Pt 3):305–27, August 1986.
- [7] Lawrence Shao-hsien Chen. Joint processing of audio-visual information for the recognition of emotional expressions in human-computer interaction. Technical report, 2000.
- [8] Ira Cohen, Ashutosh Garg, and Thomas S. Huang. Emotion Recognition from Facial Expressions using Multilevel HMM. In *Neural Information Processing Systems*, 2000.
- [9] A.J. Colmenarez and T.S. Huang. Face detection with information-based maximum discrimination. In *Computer Vision and Pattern Recognition, 1997. Proceedings., 1997 IEEE Computer Society Conference on*, pages 782–787, Jun 1997.
- [10] Ian Craw, David Tock, and Alan Bennett. Finding Face Features. In *European Conference on Computer Vision*, pages 92–96, 1992.
- [11] Ying Dai and Yasuaki Nakano. Face-texture model based on {SGLD} and its application in face detection in a color scene. *Pattern Recognition*, 29(6):1007–1017, 1996.
- [12] Antonio Damasio. *Descartes' error*. Quill, 1995.
- [13] Charles Darwin. *The expression of the emotions in man and animals / by Charles Darwin*. New York ;D. Appleton and Co., 1916. <http://www.biodiversitylibrary.org/bibliography/4820>.

Bibliography

- [14] T. Domhan, J. T. Springenberg, and F. Hutter. Speeding up Automatic Hyperparameter Optimization of Deep Neural Networks by Extrapolation of Learning Curves. In *Proc. of the 24th Int Joint Conf on Artificial Intelligence (IJCAI)*, 2015.
- [15] G. Donato, M.S. Bartlett, J.C. Hager, P. Ekman, and T.J. Sejnowski. Classifying facial actions. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 21(10):974–989, Oct 1999.
- [16] Richard O. Duda, Peter E. Hart, and David G. Stork. *Pattern Classification*. Wiley, New York, 2 edition, 2001.
- [17] S Ebrahimi, V Michalski, K Konda, R Memisevic, and C Pal. Recurrent Neural Networks for Emotion Recognition in Video. In *Proc. of the ACM on Int Conf on Multimodal Interaction*, pages 467–474. ACM, 2015.
- [18] P. Ekman and W. Friesen. Facial action coding system: A technique for the measurement of facial movement: Investigator’s guide 2 parts, 1978.
- [19] P. Ekman, W.V. Friesen, and J.C. Hager. *Facial Action Coding System (FACS) Manual*. A Human Face, Salt Lake City (USA), 2002.
- [20] Paul Ekman. *Emotions revealed : recognizing faces and feelings to improve communication and emotional life*. H. Holt, New York (N. Y.), 2007.
- [21] Paul Ekman and Wallace V. Friesen. Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology*, 17(2):124–129, 1971.
- [22] Paul Ekman and Karl G. Heider. The universality of a contempt expression: A replication. *Motivation and Emotion*, 12(3):303–308, 1988.
- [23] Sachin Farfade, Mohammad Saberian, and Li-Jia Li. Multi-view Face Detection Using Deep Convolutional Neural Networks. In *Proc. of the 5th ACM on Int Conf on Multimedia Retrieval*, pages 643–650, 2015.
- [24] Kunihiko Fukushima. Neocognitron: A Self-Organizing Neural Network Model for a Mechanism of Pattern Recognition Unaffected by Shift in Position. *Biological Cybernetics*, 36:193–202, 1980.
- [25] X Glorot and Y Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Int Conf on Artificial Intelligence and Statistics*, 2010.
- [26] Ian J. Goodfellow et al. Challenges in Representation Learning: A report on three machine learning contests. In *Int Conf On Neural Information Processing*, 2013.
- [27] M. Arfan Jaffar and Eisa Al Eisa. Classification of Facial Expression Using Transformed Features. *International Journal of Information and Electronics Engineering*, 4(4):269–273, 2014.

- [28] A Krizhevsky, I Sutskever, and G Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In F. et al.. Pereira, editor, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. 2012.
- [29] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In F. Pereira, C.J.C. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012.
- [30] Seyed Mehdi Lajevardi and Zahir M. Hussain. A Novel Gabor Filter Selection Based on Spectral Difference and Minimum Error Rate for Facial Expression Recognition. In *2010 International Conference on Digital Image Computing: Techniques and Applications*. IEEE, dec 2010.
- [31] A. Lanitis. Automatic face identification system using flexible appearance models. *Image and Vision Computing*, 13(5):393–401, June 1995.
- [32] Laranjeira, A. Frazao X. Pimentel A. and Ribeiro B. How Deep can we Rely on Emotion Recognition. *LNCS*, December 2016.
- [33] Y. LeCun, S. Chopra, M. Ranzato, and Fu-Jie Huang. Energy-Based Models in Document Recognition and Computer Vision. In *Document Analysis and Recognition, 2007. ICDAR 2007. Ninth International Conference on*, volume 1, pages 337–341, Sept 2007.
- [34] Yann Lecun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. In *Proceedings of the IEEE*, pages 2278–2324, 1998.
- [35] W. Li, M. Li, Z. Su, and Z. Zhu. A deep-learning approach to facial expression recognition with candid images. In *14th IAPR Int Conf on Machine Vision Applications*, pages 279–282, 2015.
- [36] Ying li Tian, Takeo Kanade, and Jeffrey F. Cohn. Evaluation of Gabor-Wavelet-Based Facial Action Unit Recognition in Image Sequences of Increasing Complexity. In *FGR*, pages 229–234. IEEE Computer Society, 2002.
- [37] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews. The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 94–101, 2010.
- [38] Dawn M. McBride. *BUNDLE: McBride: Cognitive Psychology + McBride: Cognitive Psychology Interactive eBook*. SAGE Publications, Inc, 2015.
- [39] Stephen J. Mckenna, Shaogang Gong, and Yogesh Raja. Modeling Facial Colour and Identity with Gaussian Mixtures. *Pattern Recognition*, 31(12):1883–1892, 1998.

Bibliography

- [40] Edgar Osuna, Robert Freund, and Federico Girosi. Training Support Vector Machines: an Application to Face Detection. In , pages 130–136, 1997.
- [41] Maja Pantic and Leon J. M. Rothkrantz. Automatic Analysis of Facial Expressions: The State of the Art. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(12):1424–1445, 2000.
- [42] Maja Pantic, Michel Valstar, Ron Rademaker, and Ludo Maat. Web-based database for facial expression analysis. In *Proc. IEEE Int’l Conf. Multimedia and Expo*, pages 317–321, 2005.
- [43] Rosalind W. Picard. *Affective Computing*. MIT Press, Cambridge, MA, USA, 1997.
- [44] L.R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE*, 77(2):257–286, 1989.
- [45] Henry A. Rowley, Shumeet Baluja, and Takeo Kanade. Neural Network-Based Face Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(1):23–38, 1998.
- [46] James A. Russell. Is there universal recognition of emotion from facial expression? A review of the cross-cultural studies. *Psychological Bulletin*, 115:102–141, 1994.
- [47] Ferdinando Samaria and Steve Young. HMM-based architecture for face identification. *Image and Vision Computing*, 12(8):537–543, 1994.
- [48] H. Schneiderman and T. Kanade. Probabilistic Modeling of Local Appearance and Spatial Relationships for Object Recognition. *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, 0:45–51, 1998.
- [49] I. Song, H. J. Kim, and P. B. Jeon. Deep learning for real-time robust facial expression recognition on a smartphone. In *2014 IEEE International Conference on Consumer Electronics (ICCE)*, pages 564–567, Jan 2014.
- [50] N Srivastava, G Hinton, A Krizhevsky, I Sutskever, and R Salakhutdinov. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *J. Mach. Learn. Res.*, 15(1):1929–1958, 2014.
- [51] Kah K. Sung and Tomaso Poggio. Example-Based Learning for View-Based Human Face Detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 20(1):39–51, January 1998.
- [52] Michael J. Tarr and Thomas J. Palmeri. *Visual Memory (Oxford Series in Visual Cognition)*. Oxford University Press, 2008.

- [53] Ying-Li Tian, T. Kanade, and J.F. Cohn. Recognizing action units for facial expression analysis. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 23(2):97–115, Feb 2001.
- [54] Ying-Li Tian, Takeo Kanade, and Jeffrey F. Cohn. Facial Expression Analysis. In *Handbook of Face Recognition*, pages 247–275. Springer New York.
- [55] M. Turk and A. Pentland. Eigenfaces for Recognition. *Journal of Cognitive Neuroscience*, 3(1):71–86, January 1991.
- [56] Paul Viola and Michael J. Jones. Robust Real-Time Face Detection. *Int. J. Comput. Vision*, 57(2):137–154, 2004.
- [57] Zhen Wen and T.S. Huang. Capturing subtle facial motions in 3D face tracking. In *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, pages 1343–1350 vol.2, Oct 2003.
- [58] W.G.Parrot. Emotion in Social Psychology. October 2000.
- [59] Guangzheng Yang and Thomas S Huang. Human face detection in a complex background. *Pattern Recognition*, 27(1):53–63, 1994.
- [60] M. Yang, D. Kriegman, and N. Ahuja. Detecting Faces in Images: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(1):34–58, 2002.
- [61] Kin Choong Yow and Roberto Cipolla. Feature-based human face detection. *Image and Vision Computing*, 15(9):713–735, 1997.
- [62] Z Yu and C Zhang. Image based Static Facial Expression Recognition with Multiple Deep Network Learning. In *Proc. of the ACM on Int Conf on Multimodal Interaction*, 2015.
- [63] Zhengyou Zhang, Michael J. Lyons, Michael Schuster, and Shigeru Akamatsu. Comparison Between Geometry-Based and Gabor-Wavelets-Based Facial Expression Recognition Using Multi-Layer Perceptron. In *IEEE International Conference on Automatic Face and Gesture Recognition*, pages 454–461, 1998.