Automatic risk detection system by audiovisual signal processing

Ilyes Bendjoudi

Thèse de doctorat

Pour obtenir le grade de Docteur de

l'UNIVERSITÉ POLYTECHNIQUE HAUTS-DE-FRANCE et
l'INSA HAUTS-DE-FRANCE
Discipline : **Automatique et traitement du signal**

Présentée et soutenue par : Ilyes Bendjoudi
Le 10 décembre 2021, à Valenciennes.

École doctorale : École Doctorale Polytechnique Hauts-de-France (ED PHF).
Laboratoire : Laboratoire d'Automatique, de Mécanique et d'Informatique Industrielles et Humaines (LAMIH – UMR 8201).

SYSTÈME DE DÉTECTION AUTOMATIQUE DE RISQUES PAR TRAITEMENT DE SIGNAUX AUDIOVISUELS

Président du jury	:	Dornaika Fadi.	Professeur, Université du Pays Basque, St-Sébastien.
Rapporteurs	:	Maaoui Choubaila.	Professeure, Université de Lorraine.
		Raducanu Bogdan.	Professeur, Université Autonome de Barcelone.
Directeurs de thèse	:	Vanderhaegen Frédéric.	Professeur, UPHF.
		Hamad Denis.	Professeur, Université du Littoral Côte d'Opale.

PhD Thesis

Submitted for the degree of Doctor of Philosophy from

UNIVERSITÉ POLYTECHNIQUE HAUTS-DE-FRANCE And
INSA HAUTS-DE-FRANCE
Discipline : **Automation and Signal Processing**

Presented and defended by : Ilyes Bendjoudi.
On December 10^{th} 2021, Valenciennes.

Doctoral School	:	Doctoral School Polytechnique Hauts-de-France.
Research unit	:	Laboratory of Industrial and Human Automation control Mechanical engineering and Computer science (LAMIH – UMR 8201).

AUTOMATIC RISK DETECTION SYSTEM BY AUDIOVISUAL SIGNAL PROCESSING

Jury president	:	Dornaika Fadi	Professor, University of the Basque Country, San Sebastian.
Reporters	:	Maaoui Choubaila.	Professor, University of Lorraine.
		Raducanu Bogdan.	Professor, Autonomous University of Barcelona.
Thesis directors	:	Vanderhaegen Frédéric.	Professor, UPHF.
		Hamad Denis.	Professor, Université du Littoral Côte d'Opale.

To my mother Saliha, my father Yahia, and my brother Anis for their unconditional love and support.

Chapter 0

Résumé

L'analyse automatique du comportement humain connait un intérêt croissant en psychologie, linguistique, neuroscience, informatique et en automatique. Cet intérêt prend encore plus d'ampleur au vu des récents succès des algorithmes d'apprentissage automatique dans les tâches de perception. Comme l'expression du visage et l'intonation de la voix sont des données représentatives de l'état émotionnel d'une personne, notre travail vise à détecter et prédire des situations à risque en analysant l'état cognitif d'un opérateur humain à partir de signaux audio-visuels. Dans ce travail, nous discutons les différentes approches et techniques d'apprentissage automatique pour la reconnaissance d'émotions. Nous montrons ce que les réseaux de neurones à apprentissage profond, en particulier les réseaux de neurones convolutifs, ont apporté à la reconnaissance d'émotions dans un contexte multi-label et multi-tâche. Dans le cadre de la reconnaissance d'émotions à partir d'images en prenant en considération le contexte dans lequel se déroule l'action, nous proposons une architecture originale pour l'extraction des attributs caractéristiques : un module corps, réseau Xception, est dédié à l'extraction d'attributs des émotions de la personne et un module scène, réseau VGG16 modifié, pour l'extraction des attributs de la scène entière. Les sorties de ces deux modules constituent les entrées d'un 3e module, réseau multicouche, composé d'une partie fusion des deux vecteurs de caractéristiques et d'une partie décision pour la reconnaissance d'émotions. Nous présentons aussi une architecture pour la reconnaissance d'émotions à partir de la voix. Nous introduisons le principe du "Fingerprint" de l'état émotionnel et le concept de rupture émotionnelle qui sera un indicateur d'un changement brutal et inattendu de l'état émotionnel. Les résultats obtenus lors d'un processus expérimental sont discutés.

Mots Clés Apprentissage automatique ; Apprentissage profond ; CNN ; Reconnaissance d'émotions ; Détection de risques ; Rupture émotionnelle.

Chapter 0

Abstract

The automatic analysis of human behavior is experiencing an unprecedented interest in psychology, linguistics, neuroscience, computer science and automation. This interest is further enhanced by the recent success of machine learning algorithms in perceptual tasks. Since facial expression and voice intonation are representative of a person's emotional state, our work aims at detecting and predicting risky situations by analyzing the cognitive state of a human operator from audio-visual signals. In this work, we discuss the different approaches and techniques of machine learning for emotion recognition. We show how deep learning neural networks, in particular convolutional neural networks, have contributed to emotion recognition in a multi-label and multi-task context. For image-based emotion recognition, taking into account the context in which the action takes place, we propose an original architecture for the extraction of characteristic attributes: a body module, Xception network, is dedicated to the extraction of attributes of the person's emotions and a scene module, modified VGG16 network, for the extraction of attributes of the whole scene. The outputs of these two modules constitute the inputs of a 3rd module, a multilayer network, composed of a fusion part of the two feature vectors and a decision part for emotion recognition. We also present an architecture for voice-based emotion recognition. We introduce the principle of the "Fingerprint" of the emotional state and the concept of emotional breakdown which will be an indicator of a sudden and unexpected change of the emotional state. The results obtained in an experimental process are discussed.

Keywords Machine Learning; Deep Learning; CNN; Emotion recognition; Risk detection; Emotional Breakdown.

Chapter 0

Acknowledgments

First, I would like to thank my two thesis supervisors: Prof. Fr´ed´eric Vanderhaegen and Prof. Denis Hamad, for trusting me and choosing me to conduct this Ph.D. thesis.

I would also like to thank the jury members: Prof.Choubaila Maaoui and Prof.Bogdan Radacanu, for kindly agreeing to report the thesis.

Then, I would like to thank the LAMIH and its management for welcoming me into its teams and accompanying me throughout my Ph.D. I would also like to thank the Université Polytechnique Hauts-de France and the Hauts-de-France region for jointly funding my Ph.D.

I thank Prof. Fadi Dornaika for welcoming me to his team during my stay at the University of the Basque Country in San Sebastian.

Finally, I would like to take this opportunity to express my gratitude to all the people who have contributed, in any way, to the achievement of this work.

Contents

List of Figures

Chapter 0 List of Figures

List of Tables

Chapter 0 List of Tables

Chapter 1

Motivations and Thesis Objectives

Contents

1.1 Risk detection with emotion recognition

The last decades have seen a constant improvement in the field of human-computer interaction. Several works in this field have aimed to improve the user experience when the latter has to interact with a machine. However, in a classical human-machine interaction system, the machine has no idea of humans' feelings. It, therefore, cannot adapt its functions to human needs. In a world that moves towards full automation, passive feedback from the user is essential for human safety.

A case in point is the tragic accident that occurred in the state of California on March 23, 2018 (see Figure1.1). According to the NTSB (the National Transportation Safety Board of the United States) report [1], a Tesla in Autopilot mode crashed into a low concrete wall at more than 100 km/h causing the death of its passenger. The question "Whose fault was it?" is probably not easy to answer. Even if the system was not perfect, shouldn't the passenger have been paying attention? Was he simply aware? What is sure is that we are facing a lack of information. The car had no information about the passenger's state.

Figure 1.1: Autonomous Tesla crashed in California. Image taken from the NTSB report

Another example is what we have just experienced in the last two years with the Covid19 pandemic. More and more people have started to telecommute, and courses are given online. We can easily imagine a scenario, where we are home alone, and wherein the workload would be so significant that it would inevitably lead to mistakes. Also, for

[1] https://www.ntsb.gov/investigations/AccidentReports/Reports/HWY18FH011-preliminary.pdf

online courses, as it is already difficult for professors to judge students' attention level in class, how could they do it at a distance?

Humans are by nature very expressive and this begins at birth. With 42 facial muscles, screams and cries, we can express very easily, willingly or not, what we feels. This is particularly the case when we find ourselves in a situation of risk, danger, misunderstanding or discomfort, as show in Figure1.2. So the question is, can we design a human operator risk detection system based on his facial and verbal expression?

Figure 1.2: Driver surprised [2]in light of an incident.

1.2 Aim of Thesis

This thesis aims at developing a system for the automatic detection of the cognitive state of a human operator from the analysis of audiovisual signals produced by the voice and the face. The system is non-intrusive and the state is captured from cameras and microphones.

Works on automatic emotion recognition started since computers exist. However, many research challenges still remain. The main challenge to this work is the exploitation of visual and audio signals from cameras and microphones in real-time, to detect environmental or human risks based on facial and verbal characteristics. Many works conducted on emotion recognition traded the temporal aspect with accuracy. This trade-off is not possible for a safety-critical system such as autonomous vehicles for example.

The other constraint for such a system is the outdoor environment. Since the beginning of computer vision, outdoor environments remain the

[2]https://www.dreamstime.com/

most challenging factor. The brightness changes, the occlusion regions and the all unpredictable events that could happen in an uncontrolled environment may highly affect the system accuracy.

1.3 Main Contributions

This thesis focuses on designing an automatic system to detect risks from operators' facial and voice expressions. We designed a bi-modal architecture to detect sudden changes in operator's emotional state. This architecture could be split on three major parts:

- Image and context-based features extractor. The approach is described in Chapter 5 and published in [7]. In this Chapter we worked on the Emotic Database [52] to build a model for image-based multi-task emotion recognition. The motivation behind this choice is that the database offers the possibility to take into consideration the context in which the action takes place. Nevertheless, the labels distribution through the database is not homogeneous. Our first contribution is our categorical loss function named the multi-label focal loss (MFL) that gives much better results (with the same architecture) than the most common categorical loss functions we have compared (Cross-Entropy and the Euclidean loss). Compared to the state of the art, our loss function gave better results on the less frequent labels ($< 5\%$). Our second contribution is our deep learning architecture for context-based emotion recognition. We combined an Xception network [14] with a modified VGG network [96]. We removed the two fully connected layers and added a Conv3D block to get an output of 1024 channels followed by a Global Average Polling layer to get 1024 outputs. We also found out that in a multi-task learning, loss functions must be chosen together instead of choosing the best ones separately regarding previous works. It seems like there is a sort of affinity relationship between loss functions in a multi-task scenario.

- Audio-based features extractor. The approach is also described in Chapter 6. We designed a VGG-based architecture for Audio-based emotion recognition on the RAVDESS database [68]. We transformed the audio signals into images by computing the Mel-Spectrogram of each defined audio segment.

- Audio-visual emotional state Fingerprint. The principle of emotion state Fingerprint published in [5, 6] is described in Chapter 7. We take both of the precedent architectures and prune their outputs assuming that the last layer before the output layer must be considered as the feature layer, i.e. the layer where all the information is combined before a classification or a regression task. By taking the two feature layers of the two architectures and concatenate

them results on a single layer which represents the fingerprint of a unique emotional state (ES).

- EmoRuption: A system for emotional breakdown detection. The process is also described in Chapter 7. Knowing the ES fingerprint at $t-\Delta_t$ and compares it to the actual ES using a distance function (Similarity function) results on the difference between the actual ES and the ES at $t-\Delta_t$, i.e. the temporal derivative of the ES. If the distance exceeds a predefined threshold, an emotional breakdown (EB) happened.

1.4 Outline

The thesis is structured as follows. First we define what is a risk and what is a dissonance. This is presented in Chapter 2 along with a study that shows the relationship between human behaviour through their expressed emotions and audiovisual signals. Chapter 3, presents the basics of image and voice processing and, we overfly some works on visual-based, audio-based and audiovisual-based emotion recognition and give a shot review on machine learning in Chapter 4. We present the main architectures and finally we discuss the impact of machine learning on emotion recognition. In Chapter 5 and 6, we present our architecture for context-based emotion recognition and we show how our loss function improved the recognition on low frequent labels. After that, we present our approach for audio-based emotion recognition and we discuss the audiovisual fusion approaches. Finally we present the principle of Emotional Breakdown (EB) and our validation process. We conclude the thesis with a summary and some research perspectives in Chapter 7. The dependencies between chapters are shown in the flow graph in Fig.1.3.

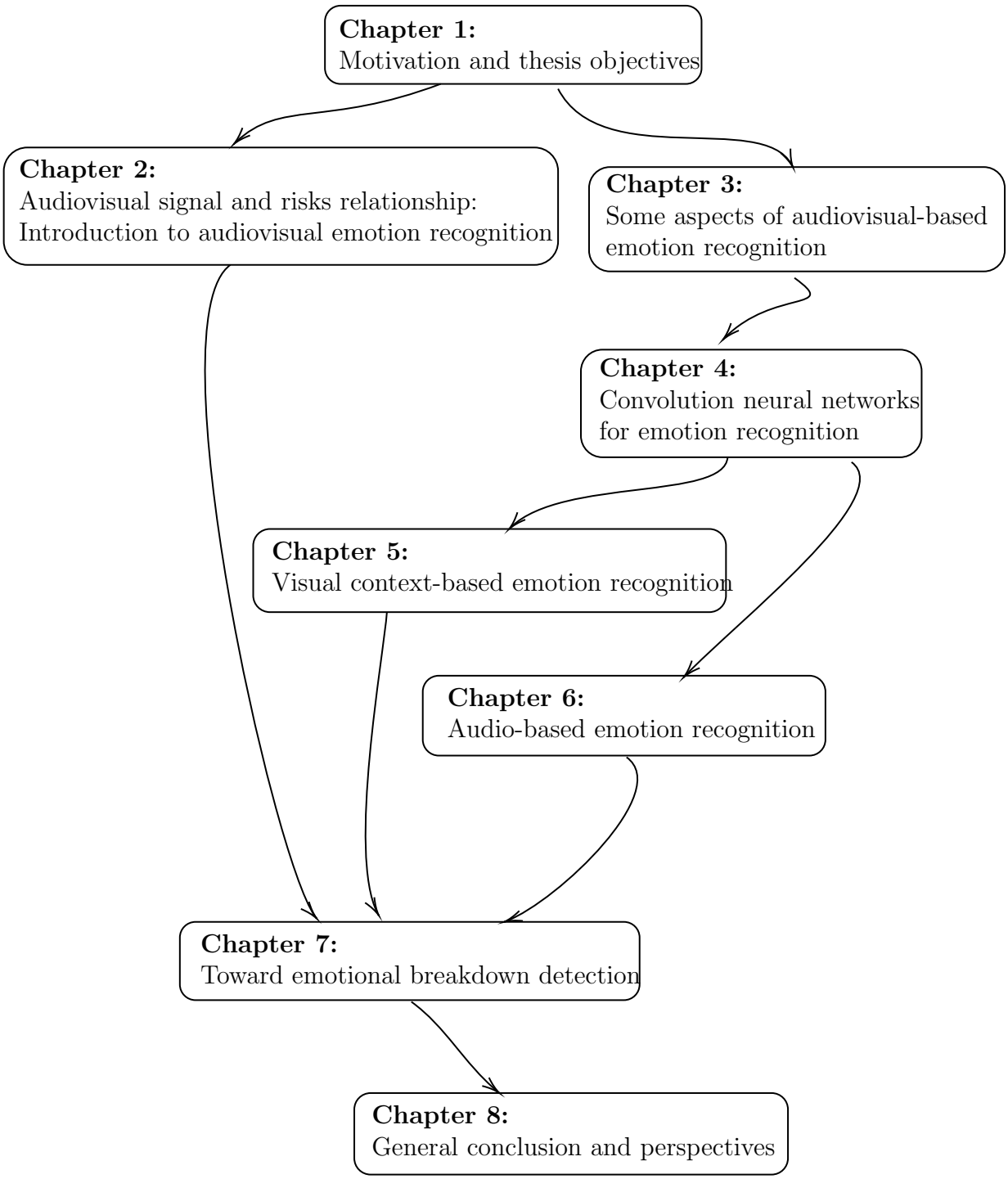

Figure 1.3: The reading dependencies between chapters.

Chapter 2

Audiovisual signal and risks relationship: Introduction to audiovisual emotion recognition

Contents

The technological evolution resulting from the increased development of digital media at work is likely to have a severe impact on human behavior. In this chapter, we are going to highlight the link between the human feeling and the audiovisual signals. To do this we will study the case of dissonances. It relates to suffering or discomfort due to a conflict between personal or collective beliefs with felt or expressed emotions. The study of dissonances comes from work in cognitive science [29] and engineering science [48]. The concept was then explored as part of the analysis of human reliability to extend the study of human error to that of use dissonances of socio-technical systems [111, 113–115]. A dissonance due to a failure of cooperation between a human operator and a robot, for example, may affect factors such as misunderstanding, frustration, embarrassment, or astonishment. Among these factors, there are emotions whose detection is the subject of significant research based on facial or voice recognition systems [51]. For attentional dissonances that are gaps between actual and perceived attention, the synchronization of dynamic events as alarms with heart rate significantly increases the number of errors in their detection [118].

2.1 From human reliability analysis to dissonance engineering

Human reliability can be defined as the human ability to realize the required and the additional tasks, at a given time or during an interval of time, with acceptable consequences regarding criteria such as safety, security, workload, quality, or production of services [114]. Human error is the complement of human reliability. It is the capacity of humans not to realize their requirements or their additional tasks correctly. Off-line human error assessment methods for prospective, retrospective, or cognitive analysis exist to assess qualitatively or quantitatively erroneous human behaviors [8, 12, 18, 38, 49, 54, 82, 84–86, 107, 117]. Many of them consider mainly the first set of tasks, i.e., they study the possible human errors related to what the users are supposed to do. Moreover, they usually focus on the human error impact on system safety or on erroneous behaviors when controlling emergency or safety-critical situations. Prospective methods aim at anticipating human errors during the design of a human-machine system.

Retrospective methods explain human errors that occur on the field and cause incidents or accidents. They produce feedback of experience and can propose modifications of the socio-technical system in order to make it more reliable by preventing it from hazardous human error occurrence. Cognitive model-based methods analyze human error by applying taxonomy of errors, by defining possible factors that facilitate human error occurrence, or by identifying the possible causes and consequences of human errors. Other on-line human reliability ap-

proaches are based on human indicator measures. They relate, for instance, to quantitative measurements or results of subjective evaluation methods [17, 88, 90, 91, 106, 108, 116].

More recently, resilience engineering concepts were developed to study safety management on critical systems. System resilience is then defined as its ability to recover from any instability [114]. Cooperation or learning-based support systems can then be defined in order to make the system more resilient to human errors [1, 26, 46, 77, 82, 109, 119, 120, 134]. Dissonances are interpreted as possible causes of human instability regarding their consequences in terms of discomfort, overload, or stress when they are detected or controlled or of unconsciousness when they are not perceived, and people feel they behave right. The successful control of these dissonances makes the system resilient, while its failure makes it vulnerable.

The concept of dissonance is suitable because it can take into account several subjective or quantitative baselines to identify human disruptions and analyze associated human behaviors, or consider erroneous baseline or a lack of baseline, and treat weak signals [114]. Moreover, different strategies of dissonance control or discovery can be applied and studied to reinforce human knowledge or belief [111, 113].

The architecture model proposed in section 5 is based on the results of two studies presented in sections 3 and 4, respectively. It considers an automated detection of emotion by analyzing audio-visual signals from subjects facing dissonances, and the possible inattention that provokes a lack of their detection and the associated emotion.

2.2 The Reverse Comic Strip for emotion analysis

The Reverse Comic Strip support aims at identifying emotions in the course of human activities by determining a kind of comic strip by reproducing experienced emotions with facial pictures and verbal expressions or thoughts, [110, 112].

The study concerns the discovery of interpretation rules about signaling systems presented in [112]. It consists in discovering the right rule related to the on-lights that are on different panels, Fig. 2.1.

The discovery process involved 36 subjects who were invited to define a rule regarding the lights on or off of the signaling supports of pictures 1, 2, and 3. The first step concerned the rule discovery of picture 1. Before the discovery process of picture 2, the correct rule of picture 1 was given. It was the same process before engaging the discovery process of picture 3: the rule related to picture 2 was given. Therefore, the learning associated with the discovery process was supervised because, at each step, the correct rule was given to the subjects. For pictures 1 and 2, when the lights are on the train will stop at the corresponding

Figure 2.1: Discovery of the meaning of the lights on of the pictures 1, 2 and 3.

Results on Fig. 1 picture 1

(1) Discovery process

Discovery of the right rule	Discovery of the wrong rule	No discovery process
28	7	1

(2) Selected pictures

6	4	10	19	11

(3) Thoughts or words

Positive emotion: "Easy".
Multiple emotions: "Stations served but at the start it was blurry", "What is it? Oh ok!", "What is that?, Ah yes I understood the instructions", "What does this represent? Ah yes!", "I didn't quite understand but I think maybe the cities where we stop", "I didn't understand at first", "Wow, what is all this info? Ah OK, the stations served by the line, PULE18".
Negative emotion: "Not sure!", "What's the point? why not serve all the stations?", "What does this photo mean?", "I didn't really understand", "I don't know", "Didn't understand the meaning of the indicators, late or on time?", "ununderstandable!"
Positive emotion: "Easy", "Photo more readable, more understandable, better ergonomics of the display panel", "Same as before", "I know this principle", "I know, I used to take the metro".
Negative emotion: "Small and unclear display", "I did not understand", "It's small, we don't see anything!", "Not sure, is it the other way around?", "RER A continues to have problems on these lines", "Doubt on the passage of the train when the lights are off".

Figure 2.2: Results on Fig.2.1 picture 1 [110]

stations. In picture 3, there are two metro lines, and the on-lights means the presence of a train at the corresponding station.

Regarding the Reverse Comic Strip parameters, (see Fig.2.2, Fig.2.3,

Results on Fig. 1 picture 2

(1) Discovery process

Discovery of the right rule	Discovery of the wrong rule	No discovery process
25	11	0

(2) Selected pictures

4	1	6	12	21

(3) Thoughts or words

Positive emotion: "Easy", "Photo more readable, more understandable, better ergonomics of the display panel", "Same as before", "I know this principle", "I know, I used to take the metro".

Negative emotion: "Small and unclear display", "I did not understand", "It's small, we don't see anything!", "Not sure, is it the other way around?", "RER A continues to have problems on these lines", "Doubt on the passage of the train when the lights are off".

Figure 2.3: Results on Fig.2.1 picture 2 [110]

Results on Fig. 1 picture 3

(1) Discovery process

Discovery of the right rule	Discovery of the wrong rule	No discovery process
0	29	7

(2) Selected pictures

9	6	12	11	6

(3) Thoughts or words

Multiple emotion: "I don't see much but I think it's the same, what is Simonis Elisabeth?", "I do not understand but I think the RER does not stop when the lights are on", "No idea, I don't know what it is, maybe the train arrival times, the name corresponding to the name of the stop".

Negative emotion: "Hard to read!", "Hard to understand! Not very visible!", " Not visible!", "I did not understand", "I don't understand the picture from afar", "I don't understand the display", "Why doesn't it stop at all stations?", "I have difficulties to understand", "I wonder if I interpret the image correctly", "Not very readable", "Not very understandable", "I do not understand! What's this?", "I did not understand the meaning", "It's weird".

Figure 2.4: Results on Fig.2.1 picture 3 [110]

Fig.2.4), pictures 1 and 3 generate more negative emotions, i.e, the selected emotion pictures, thoughts or words, this is due to the novelty of the situations. On the other hand, the discovery exercise related to picture 2 is less stressful regarding the selected pictures and generate more positive emotions in terms of thoughts or words because it is similar to picture 1.

This study aimed to put a link between human behaviour and audiovisual signal. The results presented motivate the design of an architecture to recognize automatically human operators' emotions from audiovisual signals by analyzing their facial and verbal expressions.

2.3 EmoRuption: Toward a bi-modal architecture for Emotional Breakdown detection

We know that the situations of discomfort expressed through emotions can be very brief but also very intense (state of shock, surprise, discomfort etc). Starting from this postulate, we realize that the very nature of expressed emotions is not really significant when we are interested in the detection of discomfort situations, especially since the neutral state can differ from one person to another. It would be interesting to think that it is rather the brutal and intense changes of the emotional state that are the most likely to describe a discomfort situation.

Based on this, we have designed a bi-modal architecture, named EmoRuption (see Fig.2.5) for the detection of these changes in the emotional state. This work has been published in [5, 6]

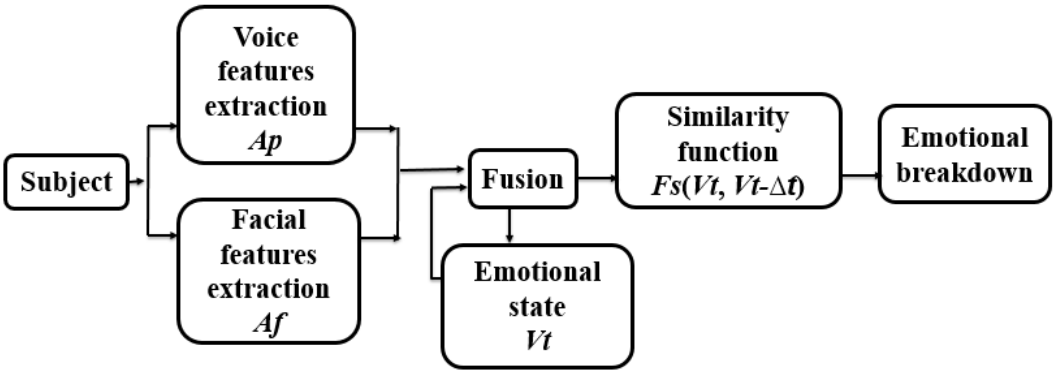

Figure 2.5: EmoRuption's theoretical architecture for emotional breakdown detection

From an audiovisual input signal we extract both voice and facial feature vectors respectively A_p and A_f describing the operator's actual emotional state. Then those feature go through a fusion module where the emotional state vector is built, denoted V_t. On the next iteration, we

compare the actual emotional state V_t with the previous emotional state $V_{t-\delta t}$ using a similarity function. An Emotional Breakdown (EB) occurs when the similarity function becomes greater than a fixed threshold α and is defined as follows:

$$EB(t) = \begin{cases} \text{True if} & S_f\left(V_t, V_{t-1}\right) > \alpha \\ \text{False} & \text{otherwise} \end{cases} \tag{2.1}$$

It is essential to notice that there is no classification here; the very nature of the emotion is not computed. We only focus on the changes in the emotional state.

2.4 Conclusion

This chapter aimed to highlight the relationship between human behaviour through their expressed emotions and audiovisual signals. The results presented in this chapter motivate the design of an architecture to recognize automatically human operators' emotional state from audiovisual signals by analyzing their facial and verbal expressions. After that, we introduced our designed bi-modal architecture to capture human's emotional state and compare it to the previous one so we can calculate the its variations and detect Emotional Breakdowns.

Chapter 3

Some aspects of audiovisual-based emotion recognition

Contents

The automatic analysis of human behavior is experiencing an unprecedented interest in psychology, linguistics, neuro-science, computer science and automation. Emotion recognition has known a broad interest over the past two decades and especially for facial expression recognition, which represents the central axis the literature explored. However, in [72] Mehrabian stated that the facial expression of a message contribute 55% of the overall emotional state information while the vocal and the semantic parts contribute 38 and 7%, respectively. This means that if we only process the facial expression, we miss out on 45% of the available information.

In this chapter, we will firstly introduce some basics of image and audio processing. Next, we will give a global vision of the state of the art in visual-based emotion recognition, audio-based emotion recognition and finally audiovisual-based emotion recognition.

3.1 Some basics of computer vision and image processing

3.1.1 Image matching based area of interest (AOI)

In computer vision, the area of interest designate the interesting parts of the digital image. These areas can be contours, points or regions of interest.

When an area of interest is detected, it is associated with a vector called a features descriptor, which as its name indicates describes this area of interest.

Areas of interest are used for image matching, which consists in finding common elements in two (or more) images, representing the same scene but in different ways. This is very useful in many computer vision tasks such as visual search (finding an image similar to another), or in object recognition in an image, classification and 3D reconstruction. Figure 3.1 shows an example of image matching. The circles on the images represent the areas of interest.

The image matching goes through two steps:

- Detection and description of features for both images.
- Finding pairs of features that match in the two images.

Finding matching pairs is a complex problem that can make matching impossible. In addition, several elements including photometric properties (brightness, contrast ... etc), occlusion, scale and rotation make the task even more difficult.

To facilitate the comparison of features and thus the correspondence, the descriptor must have certain properties of invariance to the elements

[1] `https://openclassrooms.com`

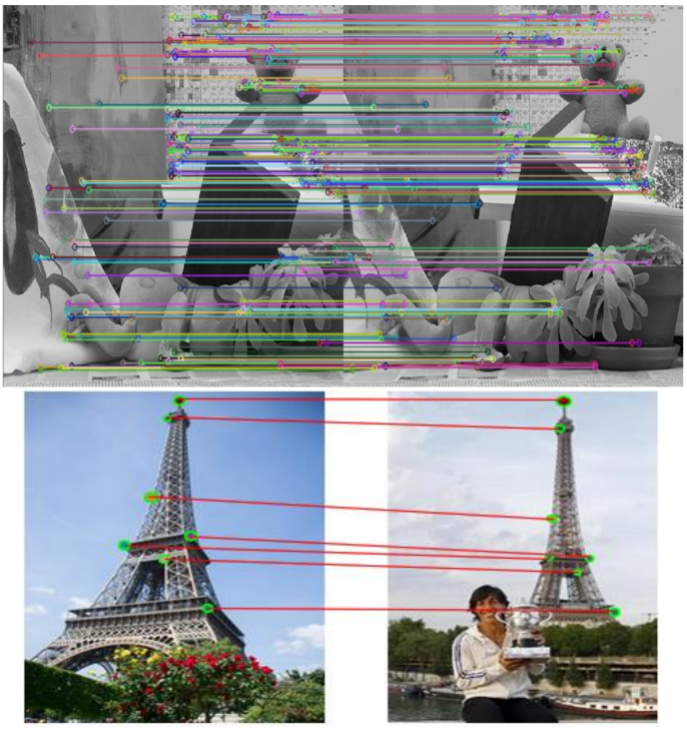

Figure 3.1: Examples of image matching [1]

mentioned above. Among the existing descriptors we can cite the Scale-Invarient Feature Transform (SIFT) which is based on the detection, in the image, of circular areas, each centered around a point called the point of interest, and of a determined radius called the scale factor. SIFT is invariant to scale transformations, rotation and occlusion; but it uses the gradient, which makes it sensitive to photometric changes.

3.1.2 Low and high level features

When we look at the image in Figure 3.2, we can identify a flying bird, the sun and the ocean in the background. This identification is made based on the entire scene. We do not consider the bird, the sun and the ocean separately to classify the image. Instead, we make a description of the visual content: this is called a high-level description (high semantics) of the image.

On the other hand, some algorithms, like SIFT mentioned above, cannot perform this kind of global identification. They rely on local points such as edges and points of interest to detect features, combine them and then classify the image.

Minimizing the gap between high-level representations (interpreted by humans) and low-level features (detected by algorithms), is a key point in vision recognition problems.

Figure 3.2: Low and high level features.

3.1.3 Convolution

Convolution is one of the most important operations in image processing. It is mainly used in feature extraction and is also the basic of convolutional neural networks.

The convolution consists in performing the scalar product between a convolution matrix and the neighborhood of a pixel, according to the formula 3.1. The pixel value is replaced by the resulting value of this product. The convolution matrix, also called convolution kernel or filter, is usually a square and much smaller than the image.

$$f[x,y] * g[x,y] = \sum_{n_1=-\infty}^{\infty} \sum_{n_2=-\infty}^{\infty} f[n_1,n_2] * g[x-n_1, y-n_2] \qquad (3.1)$$

where f represents the numerical image, g the convolution kernel and $*$ is the product convolution. The obtained image can be interpreted as a modified (filtered) version of f.

At first the filter is applied to the upper left part of the image. It is then moved along the image, to calculate the new values of each pixel. This displacement is done by a step of 1 pixel.

3.1.4 Pyramids and image sampling

In the field of computer vision, the pyramid is a multi-resolution or multi-scale representation allowing an analysis of the image on several levels, from the finest detail to the most coarse.

[2] https://qastack.fr/datascience/23183/why-convolutions-always-use-odd-numbers-as-filter-size

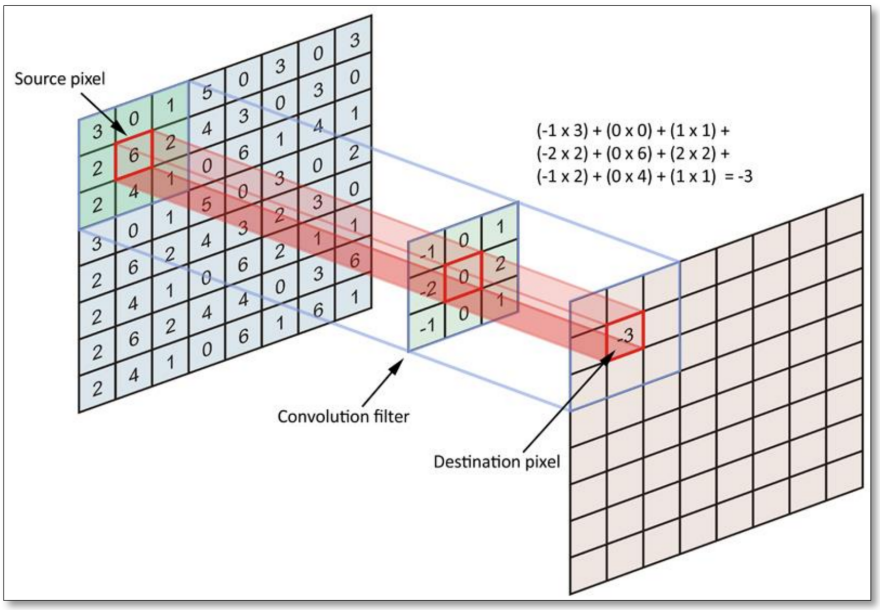

Figure 3.3: Graphical representation of convolution [2]

This theory appeared progressively at the end of the 70's, and was inspired by natural environments which also lend themselves to this multi-level decomposition. Indeed, the real objects, the physical reality, contrary to the mathematical objects, exist in the form of different entities depending on the level of scale which one considers (Example: a leaf, a tree, a forest).

The multi-scale analysis of an image and the extraction of the information contained in it at each level allows a more accurate structural description than that made with a single level analysis.

This technique was introduced for simple image processing, but was later adopted by the field of computer vision. In particular, neural networks with a pyramidal configuration have been used to extract local and global contextual information from images.

In the next section we will explain the basics of audio processing. We will enumerate the most relevant audio features we can extract from a digital audio signal and how they are extracted.

3.2 Some basics in audio processing

As image features, audio features can be categorized in three main categories following their level of abstraction (see Fig.3.5): High-level features as cords, melody or lyrics, mid-level features pitch, fluctuation patters or Mel-Frequency Cepstral Coefficients (MFCC's) and low-level

[3] https://en.wikipedia.org/wiki/Pyramid_(image_processing)

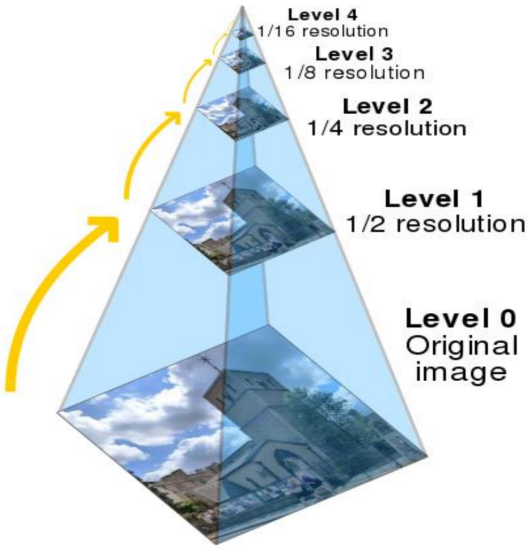

Figure 3.4: Visual representation of a five levels pyramid [3].

features as amplitude envelope, energy or spectral centroid [50]. In the following we will only focus on the mid and low-level features.

The mid and low-level features cited above are either computed in the time domain or frequency domain representation of the audio signal as shown in Fig.3.6. Time domain representation indicates on each point the amplitude of the signal at time t where the frequency domain representation indicates the signal magnitude at different frequencies. This representation is a result of a Fourier Transformation of the time domain representation.

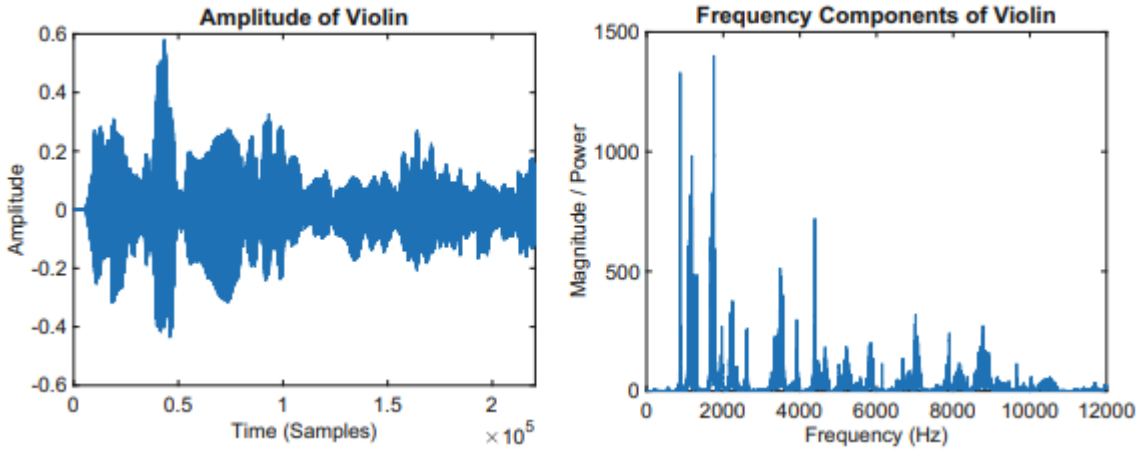

Figure 3.6: Diagrams of time domain representation (left) and frequency representation (right) of a violin [50].

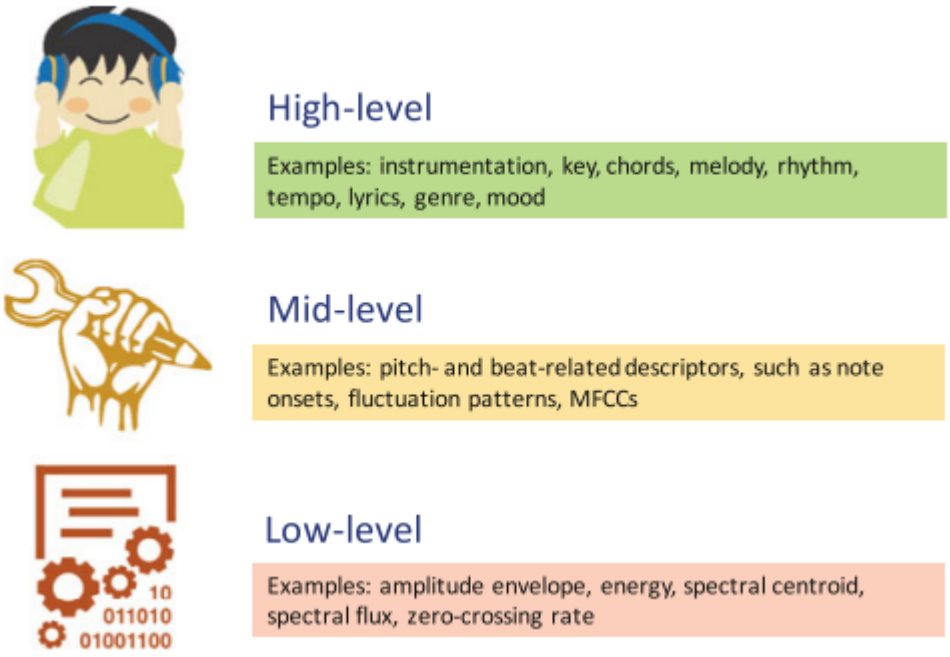

Figure 3.5: Audio features classification following their level of abstraction [50].

A classic and primitive pipeline of acoustic feature extraction is described in [50] and can be represented as follow (Fig.3.7):

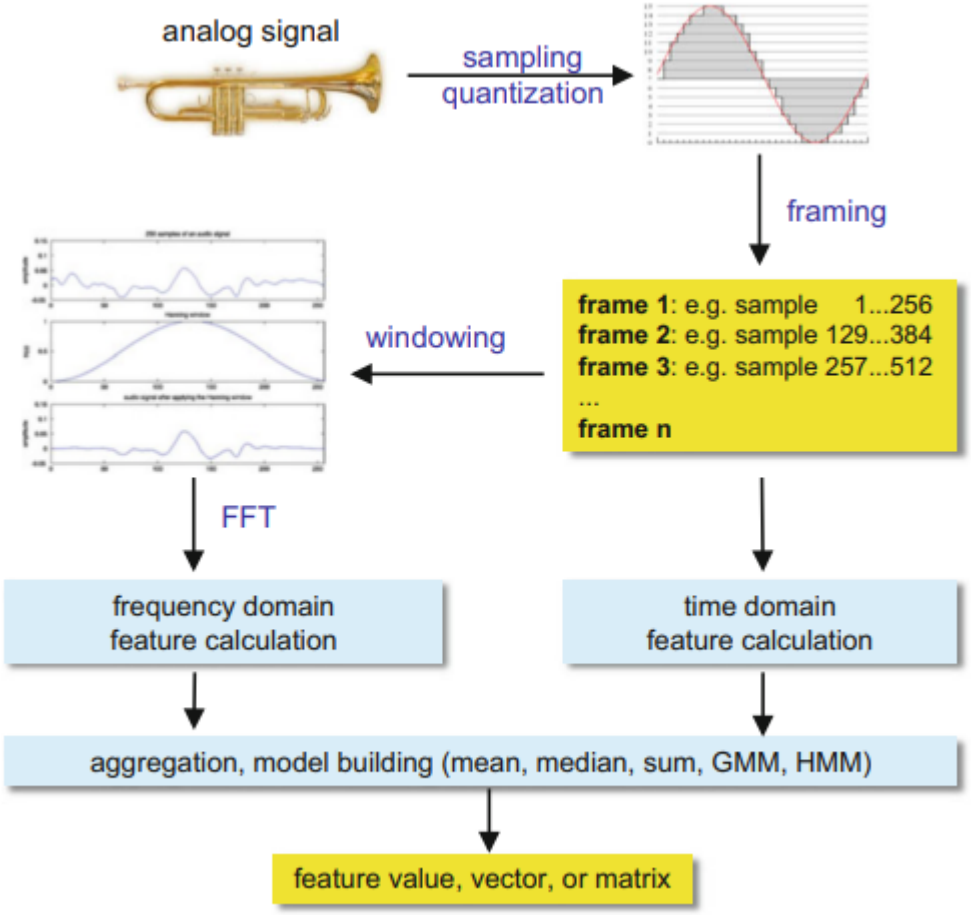

Figure 3.7: Simplified workflow of a typical audio feature extractor [50]

From the analogue signal of an audio source (music instrument or voice for example) we need to convert that signal into a digital one.

After converting the analogue signal of an audio source into a digital signal we will result in a so-called Pulse Code Modulation (PCM) representation of an audio signal. In this representation, an amplitude value is assigned to each sample (a range of time; the smaller the more accurate). However, we cannot use this representation to extract time domain features because the samples are too short, i.e., 1 sample at 44.1 Khz = 0.0.0227 which is way smaller that the ear's time resolution (10 ms). To solve this, we need to perform a framing to the PCM which is basically concatenating samples as shown if Fig.3.7. Frames' size are commonly powers of 2 because of the Fourier Transform which becomes way faster in this case. Typically an overlap of 50% is used. The duration of a frame (in seconds) is computed as follow:

$$d_f = \frac{1}{Sr}.K \tag{3.2}$$

where Sr in the sample rate (how many samples in one second) and K the frame size.

In order to compute the frequency domain representation of sound using the Fourier transformation we need firstly to make a so-called frame windowing. It consists of applying a windowing function to each frame. The Hann function is a common choice and it is defined as follow:

$$w(k) = 0.5.\left(1 - \cos\left(\frac{2.\pi.k}{K-1}\right)\right) \tag{3.3}$$

where, K is the frame size and $k = 1..k$.

Another alternative to the Hann function is the Hamming function and it is defined as follows:

$$w(k) = 0.54 - 0.46.\left(1 - \cos\left(\frac{2.\pi.k}{K-1}\right)\right) \tag{3.4}$$

The windowing process consists of multiplying sample by sample the whole frame with the respective value of the windowing function. Doing so results on a periodic signal and will avoid artefacts in the spectrum when using the Fourier transformation. Fig.3.8. shows the windowing process using Hann function. The upper image represents the frame, the second represents the Hann function the third represents the result of the Hann windowing on the frame. As shown in the third image, the edges of the frame are suppressed, this is why an overlap of 50% is commonly used, to avoid information loss.

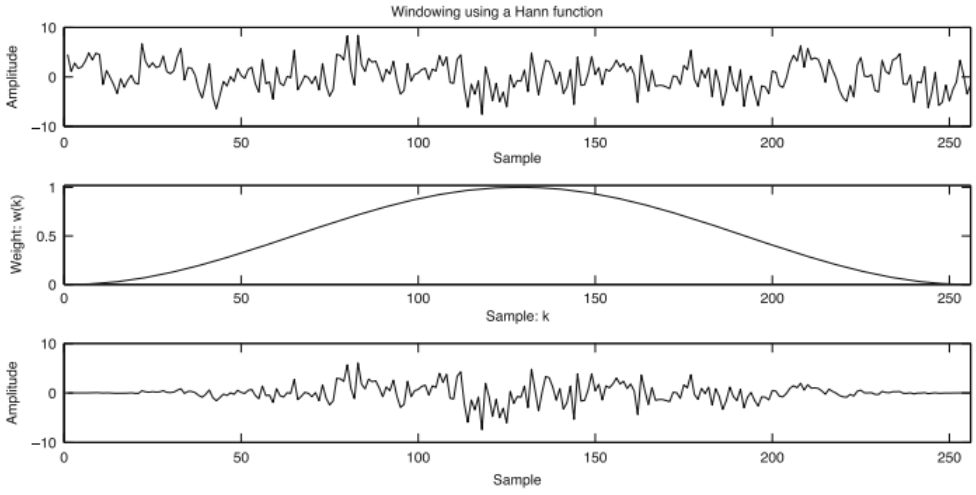

Figure 3.8: Windowing of a 256-sample frame using a Hann function [50]

Now we have our windowed frames we can apply the Fourier transformation to get the frequency domain representation. The fast Fourier Transformation (FFT) can be used if the signal is discrete, which is the

case in digital signal representation with samples and frames. The FFT is defined as follows:

$$X_m = \sum_{k=0}^{K-1} x_k e^{-\frac{2\pi . i}{K} km} \tag{3.5}$$

where, k is the number of samples in the frame, m is an integer from 0 to $K-1$ and x_k is the k^{th} input amplitude. The output X_m is a complex number where the real part $Re(X_m)$ and the imaginary part $Im(X_m)$ represent the cos and sin waves. We then obtain the magnitude:

$$|X_m| = \sqrt{Re(X_m)^2 + Im(X_m)^2} \tag{3.6}$$

which represents the amplitude of the combined sine and cosine waves, and the phase:

$$\phi(X_m) = \tan^{-1} \frac{Im(X_m)}{Re(X_m)} \tag{3.7}$$

which indicates the relative proportion of sine and cosine.

Also there is a third representation combining both time and frequency domains called spectrogram. It is computed by using the Short Time Fourier Transformation (STTF) which consists of using many FFT's over different windowed frames in temporal order. Concatenating the resulting frequency domain representation gives a spectrogram as shown in Fig3.9.

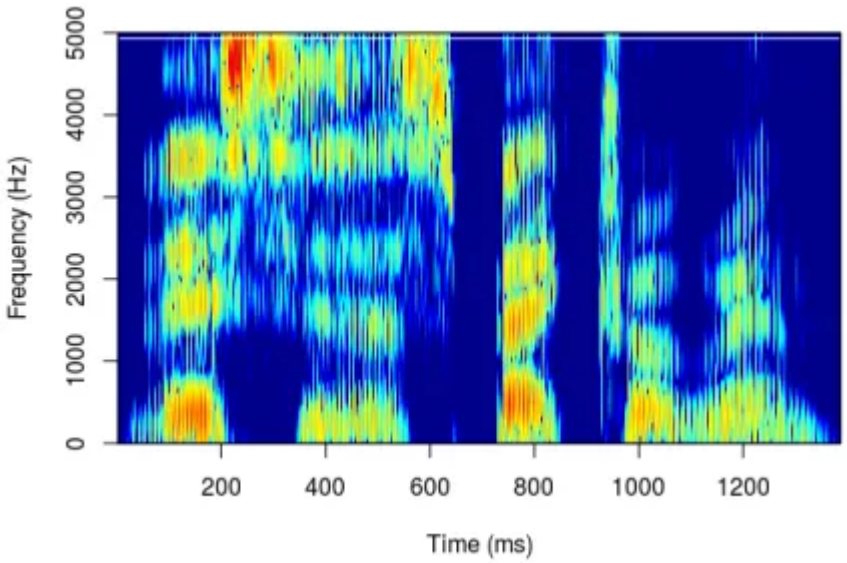

Figure 3.9: Example of spectrogram representation of a sound.

3.2.1 Common low-level audio features

In this section we will present some pertinent audio features. We will firstly present time domain features then frequency domain features and lastly time-frequency domain features.

3.2.1.1 Time domain features

- **Root-Mean-Square Energy (RMS Energy)**: It relates to perceived sound intensity and can be used for loudness estimation and as an indicator for new events in audio segmentation. It is defined as follows:

$$RMS_t = \sqrt{\frac{1}{K}. \sum_{k=t.K}^{(t+1).K-1} s(k)^2} \tag{3.8}$$

  where $s(k)$ is the amplitude of the k^{th} sample and K is the frame size.

- **Zero-Crossing Rate (ZCR)**: It measures the number of times the amplitude value changes its sign. It is defined as follows:

$$ZCR_t = \frac{1}{2} \sum_{k=t.K}^{(t+1).K-1} |sgn(s(k)) - sgn(s(k+1))| \tag{3.9}$$

3.2.1.2 Frequency domain features

- **Fundamental frequency (F_0)**: The fundamental frequency denoted F_0 is the lowest frequency component in a complex sound wave (Fig.3.10). It is important in the understanding of intonation and closely corresponds to pitch [58].

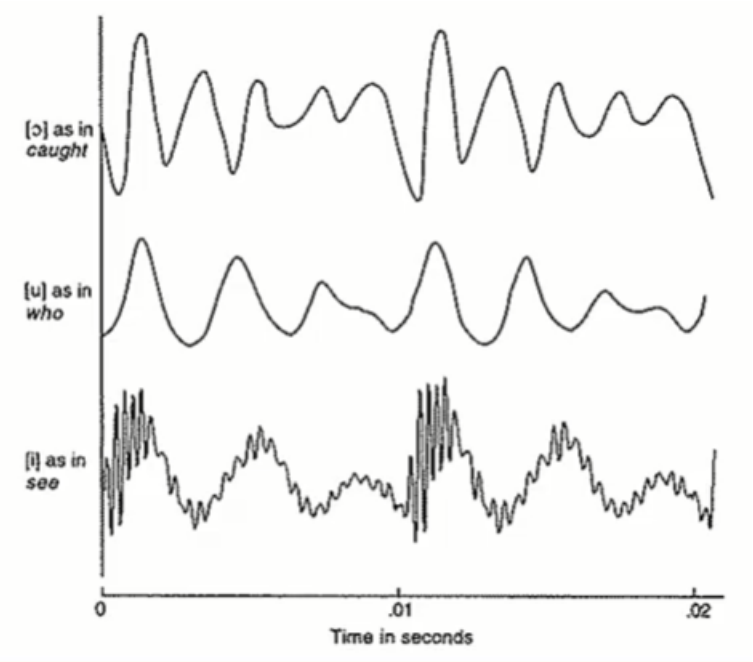

Figure 3.10: Wave form representation of the sounds [o], [u] and [i] from up to down, respectively. We can see that the lowest frequency repeats itself each 0.01s so $F_0 = 100Hz$ [58].

- **Band Energy Ratio (BER)**: BER relates the energy in the lower frequency bands to the energy in the higher bands and in this way measures how dominant low frequencies are. It is defined as follows:

$$BER_t = \frac{\sum_{n=1}^{F-1} m_t(n^2)}{\sum_{n=F}^{N} m_t(n^2)} \tag{3.10}$$

  where F denotes the split frequency band and $m_t(n)$ the magnitude of the signal in the frequency domain at frame t in frequency band n. The choice of F highly influences the resulting range of values.

- **Spectral Centroid (SC)**: Represents the frequency band where most of the energy is concentrated. It is related to the sound timber and defined as follows:

$$SC_t = \frac{\sum_{n=1}^{N} m_t(n).n}{\sum_{n=1}^{N} m_t(n)} \tag{3.11}$$

- **Spectral Spread (SS)**: Is derived from the spectral centroid. It can be interpreted as variance from the mean frequency in the signal. It is defined as follows:

$$SS_t = \frac{\sum_{n=1}^{N} |n - SC_t| \, .m_t(n)}{\sum_{n=1}^{N} m_t(n)} \tag{3.12}$$

- **Spectral Flux (SF)**: Describes the change in the power spectrum between consecutive frames, it is often used as speech detector. It is defined as follows:

$$SF_t = \sum_{n=1}^{N} (D_t(n) - D_{t-1}(n))^2 \tag{3.13}$$

where $D_t(n)$ is the frame-by-frame normalized frequency distribution in frame t.

3.3 Emotion conceptualization and structures

In order to work on emotion recognition, we firstly need to understand how emotions have been conceptualized and structured. In psychology, three primary ways of conceptualization have been retained.

The most long-standing way that conceptualize emotions is drawn from everyday life. Discrete categories of emotions [22, 23, 105]. The most popular example is the 6 basic emotions, happiness, sadness, anger, disgust, surprise and fear as shown in Fig.3.11.

Figure 3.11: Graphical representation of the 6 basic emotions. Anger, Happiness, Surprise, Disgust, Sadness, Fear. Image taken from the grimace project[5]

[5] http://www.grimace-project.net/

The notion of basic emotions was introduced by Ekman [22] in 1977. Following his cross-cultural study where he states that no matter the religion or culture, everyone agrees on the 6 basic emotions. This interpretation is very simple and useful when it comes to labeling databases. Most people agree on the emotions conveyed by a facial expression. However, this discrete representation of emotions quickly finds its limits because it fails to describe the wide spectrum of emotions in scenarios where communication is natural and sometimes complex.

The second way of interpreting emotions is through a multidimensional space by choosing a small number of dimensions. Among these dimensions we can mention: evaluation, activation, power, arousal, dominance etc. The categorical representation allows to choose one emotion among a set of emotions, on the other hand, the continuous representation (see Fig.3.12) of emotions allows the raters to choose an interval of emotions. However, this way of describing emotions is very sensitive to the choice of dimensions. A too small number of dimensions will result in a loss of information, a too high number of dimensions will result in an exponential complexity. Furthermore, this representation is not intuitive, a training for the raters is required.

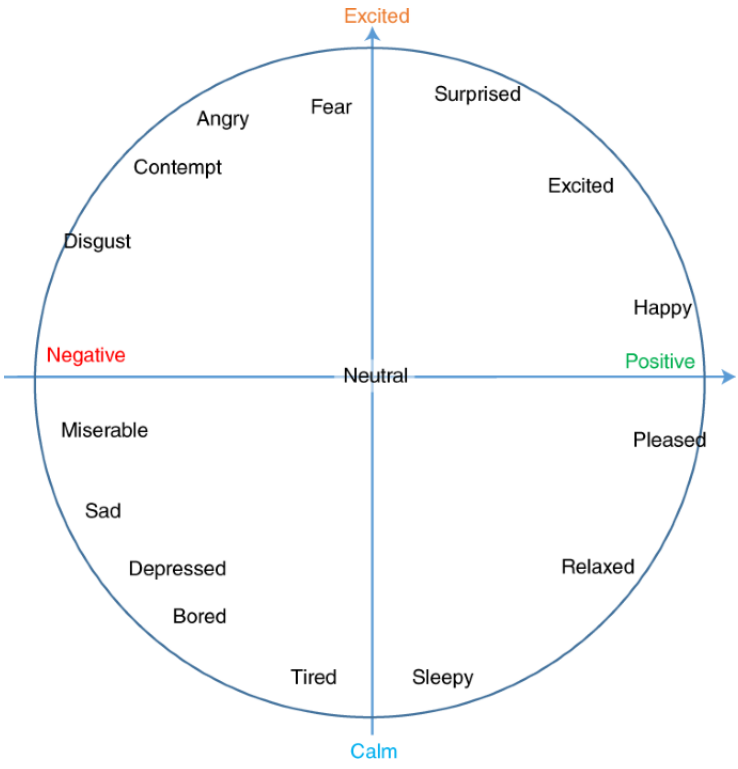

Figure 3.12: Graphical representation of emotions representation through a 2D continuous space. Dimensions are Valence and Arousal [104]

3.4 Databases

There are hundreds on audio, visual and audio-visual databases for emotion recognition. In the table.3.1 we will only focus on the most famous/used **accessible and free** databases. For each database we will detail its proprieties regarding the type (audio, visual or both), the number of samples, the number of subjects and the emotion description.

Table 3.1: Databases for emotion recognition.

Name	**A/V**	**Language**	**Number of samples**	**Number of subjects**	**Emotions description**	**Year**
Cohn-Kanade (CK) [47]	V		480 videos	210	6 basic emotions	2000
MMI [80]	V		1200 videos 600 images	61	6 basic emotions	2005
FABO face and body gesture [35]	V		210 videos	23	6 basic emotions + uncertainty, anxiety and boredom	2006
EMOTIC database [51]	V		23,571 images	34,320	26 categories and 3 continious dimensions: Arousal, valance and dominance	2017
Danish Emotional Speech database [25]	A	Danish	10 audio tracks	4	Categories: neutral, surprise, sadness, anger, hapiness	1997
ISL meeting corpus [9]	A	English	18 audio tracks	5	Categories: Negative, neutral, positive	2002
RAVDESS [68]	A/V	English	2880 videos	24	Categories: Calm, happy, sad, angry, fearful, surprise, and disgust	2018
SEMAINE [71]	A/V	English	959 conversations	150	5 continious dimensions: Valence, activation, power, expectation, intensity	2010
RML [121]	A/V	6 languages	500 videos	8	6 basic emotions	2008

3.5 Vision-based emotion recognition

Facial expressions represent the major part of the information on the emotional state conveyed by the human, this is why most of the work on emotion recognition is based on facial expressions. There are two ways to process facial-based emotion recognition: extracting facial features then classification or extract so-called facial unites which have a rule-based system to determine emotions.

3.5.1 Action Units and the FACS (Facial Action Coding System)

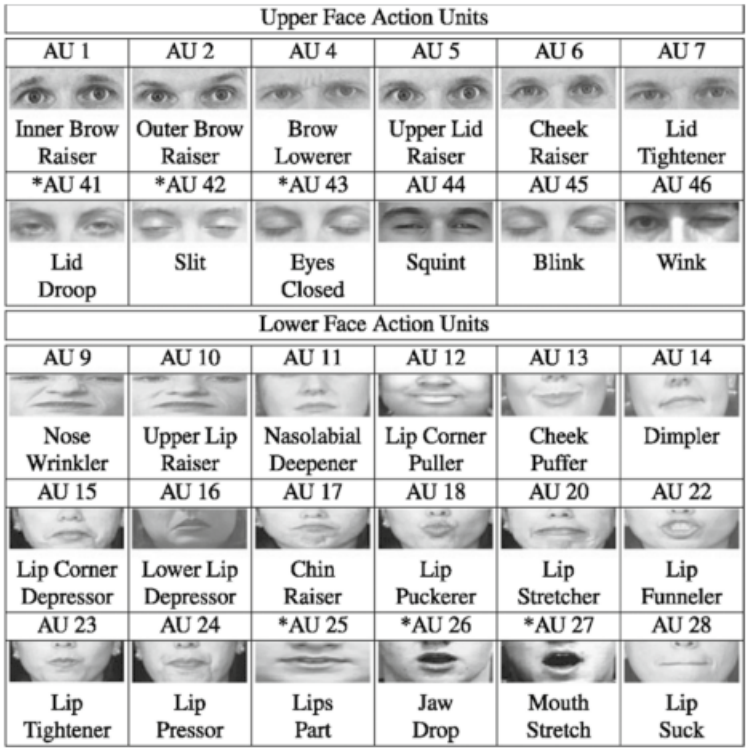

Figure 3.13: FACS action units [135]

Instead of focusing on the direct interpretation of facial expressions, FACS [22] proposes to codify muscular movements into so-called action units (AU's) as shown in Fig.3.13. Since these action units are independent of any interpretation, they can be used to describe any emotion. Ekman [22] proposed a rule-based system to find the 6 basic emotions from the AU's and the recognition of other emotional states such as depression [24] or pain [123] with other systems of high-order interpretation categories.

3.5.2 Straight emotion recognition

Instead of using the FACS codification, some works have focused on a straight way to detect emotion by directly analysing facial movements and then classify them (see Fig.3.14. In this exercise, we can distinguish two different feature classes: Geometric and appearance features.

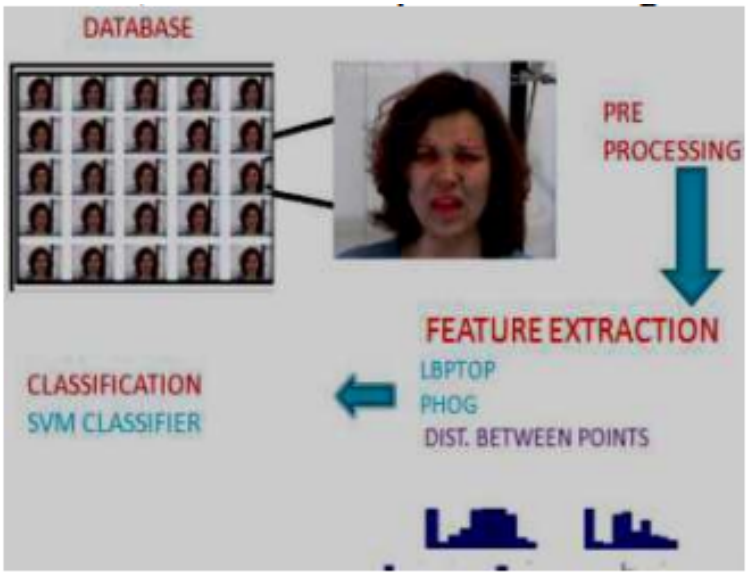

Figure 3.14: Basic system for emotion recognition [21].

3.5.2.1 Geometric features

The geometric features represent the shape of the facial components such as eyes or mouth and salient facial points such as the corners of the eyes. Some approaches in the literature as Kanade Lucas Tomasi (KLT) [53] or Elastic Bunch Graph Matching (EBGM) [31] have been explored to extract geometric features.

Figure 3.15: Geometric points tracking [21]

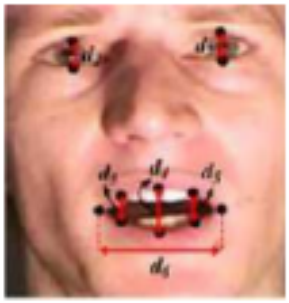

Figure 3.16: Distance between geometric points [21]

3.5.2.2 Appearance features

Appearance features represent the facial texture, wrinkles, bulges, and furrows. Descriptors as Local Binary Pattern (LBP) [94], Local Mean Binary Pattern (LMBP) [32], Local Gabor Binary Pattern (LGBP) [75] or Local Phase Quantization (LPQ) [122] are the most effective descriptors for an appearance-based features extraction.

As suggested in different studies [69, 102, 103], a hybrid approach combining both appearance and geometric features might be the best choice.

3.6 Audio-based emotion recognition

Audio-based emotion recognition is also based on basic emotions. However, there are some works that have targeted some emotions such as disappointment [33, 39], certainty [66] or empathy [98].

Research for audio-based emotion recognition often focuses on the acoustic characteristics of the sound signal. These acoustic features can be classified into two categories: Prosodic features and spectral features. However, there are some works that have used linguistic features (language and speech) to improve the performance of emotion recognition from audio signals [4, 62, 67].

3.6.1 Prosody features

From the Cambridge dictionary, prosody is defined as "the rhythm and intonation (the way a speaker's voice rises and falls) of language". In computer science prosody features are described by the pitch, energy and speech rate (duration). The pitch baseline and topline, as well as the pitch range, are commonly computed based on the mean and variance of the logarithmic F_0 values. Energy features are computed based on the intensity contour. Similar to the F_0 features, a variety of energy related range features, movement features, and slope features are computed using various normalization. Speech rate is obtained by computing

the pause duration between each word.

Many studies show that energy and pitch features contribute the most to audio emotion recognition [56].

3.6.2 Spectral features

Spectral features are obtained by converting the time based signal into the frequency domain using the Fourier Transform, like: fundamental frequency, frequency components, spectral centroid, spectral flux, spectral density, spectral roll-off, etc.

3.7 Audiovisual-based emotion recognition

Besides the audio and visual feature extraction discussed above, audiovisual-based emotion recognition can be seen as a fusion problem. Where and how do we merge the two modalities ?

We can distinguish four major data fusion strategies: Feature level, decision level, model level and hybrid fusion strategy. Feature level fusion consists on concatenating prosodic features and facial features to build a feature vector. This vector in then propagated into a classifier. The main issue with this approach is the different time scale and metrics from both modalities [11, 92, 129].

Decision feature level is the most used approach in the literature. It consists of processing the two modalities separately then combine the decision of each single-modal decision (as a voting system). This approach assume that there is no dependencies between the two modalities, which is a bit naive and more likely incorrect [3, 40, 78, 126, 128]

To solve the dependency problem, a couple of model level fusion methods have been proposed to make use of the correlation and dependencies that both modalities might have. Methods like Multi-stream fused Hidden Markov Models (HMM) [129] or tripled HMM [97] for prosodic and upper and lower face features have been proposed. Neural networks [30] and Bayesian Networks [93] also have been explored. As each strategy has its advantages and its disadvantages, hybrid or multistage fusion strategies seems to confer the best compromise.

3.8 Conclusion

This chapter has been devoted to explain some basic notions in image and audio processing and the basics of emotion recognition. We explained how emotions are represented in psychology and how they have been conceptualized in computer science. We covered the main aspects and techniques of visual-based, audio-based and audiovisual-based emotion recognition.

In the next chapter we will introduce the main aspects of Artificial Intelligence and especially machine learning and how it did impact the field of emotion recognition.

Chapter 4

Convolution neural networks for emotion recognition

Contents

> *There is no intelligence without learning - Yan LeCun, head of artificial intelligence (AI) at Facebook.*

In general, machine learning allows a machine to perform tasks that would be impossible to perform using conventional algorithms and for which it is not explicitly programmed beforehand, by extracting and exploiting information present in a data set. This process allows machines to evolve through a systematic process, so that they can make independent decisions and react appropriately to unknown situations.

Machine learning techniques are used in many different fields. For example, it can be used in the medical field where machines help diagnose tumors, in the banking field to estimate a person's ability to repay a loan, or in the transportation industry to develop driverless navigation systems.

> *A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E* - Tom MITCHELL, in his book "Machine Learning" [74].

According to this definition, a program learns from an experiment E to perform a task T, if its performance measure P, measured for task T, improves with experiment E. The book [74] details each aspect as follow:

1. **The task T:**

   The goal of machine learning is to teach a machine to perform a specific task T, which would be very difficult or impossible to perform with a classical algorithm. These tasks can be for example tasks of:

   - **Classification:** This task consists in determining the class to which an input X belongs, with C the set of classes. As an example, we can consider the classification of flowers, where we have to find the class to which a flower belongs knowing the number of petals, the color ... etc.
   - **Regression:** This task is similar to classification, except that the set of outputs is not discrete but continuous. An example is learning to guess the prices of houses knowing some criteria such as the area.
   - **Transcription:** It consists in having as output a sequence of symbols corresponding to an input, as in speech recognition, where it is a matter of generating a sequence of words from sound waves.
   - **Machine translation:** Here the input and output are sequences of symbols. It is a matter of converting a sequence of words written in one language into another language.

- **Structured output:** This category includes any task whose output is a vector (or other structure with multiple values), whose elements are linked. Transcription and machine translation are therefore also part of this task. Another example is low-level semantic segmentation of images (pixel-wise semantic segmentation), where the program assigns each pixel of the image to a specific category.
- **Anomalies detection:** The program analyzes and examines a series of input data and tries to find an unusual event in order to report it. This is used for example in the detection of fraud in the use of credit cards.
- **Sampling and synthesis:** A program performing this task is called to generate new examples similar to the ones it has learned. This is useful for example in video games to automatically generate textures for large objects or landscapes, instead of labeling each pixel manually.
- **Denoising:** This task consists in cleaning or reconstructing an input data that is corrupted, and returning the original data.

2. **The performance P**

   P-performance is a quantitative measure used to evaluate the machine learning algorithm. This performance is measured on data that has not been used for training, which is called "test data". This will determine the ability of the algorithm to perform task T on real-world data that it has never encountered before. The measures differ from one task to another. For example, for classification we often use Accuracy, which represents the number of correctly classified examples. The same information can be obtained by measuring the Error Rate, which gives the number of misclassified examples.

3. **The experience E**

   Machine learning includes three main methods: supervised learning, unsupervised learning and reinforcement learning.

   - **Reinforcement learning** This learning technique is based on a control strategy. The program is seen as an agent that interacts with its environment and receives in return (feedback), reward values that tell it whether the decision it has taken is good or bad. This establishes a policy that aims at maximizing its gains (rewards), and thus encourages the model to make better decisions. In practice, this type of learning is widely used in logic games, which are defined as a sequence of decisions (Poker, Chess, AlphaGo...). It is also used in logistics, scheduling, as well as in the control of robotic arms in

order to find the most efficient motor combination for robot navigation or to learn obstacle avoidance behavior through the negative feedback that accompanies hitting obstacles.

- **Unsupervised learning** In this method, the program receives a set of input data, and it tries to learn by itself the correlations or features linking these data in order to group them into clusters. The goal is to maximize the coherence of the data belonging to the same cluster (inter-class), and minimize it between the classes (intra-class).
- **Supervised learning** The program is provided with a set of labeled data. It therefore knows the output (class or value) for each input example. The goal is to compare its returned value with the true value, if it conforms then it is on the right track, otherwise modifications on the learned function are performed.

4.1 Some examples of machine learning algorithms

There is a wide range of algorithms in the field of machine learning including support vector machines (SVM), linear regression, logistic regression, decision trees, Bayesian classifiers, etc. In this paper, we will focus on a method, which is widely used for supervised and unsupervised learning: Artificial neural networks.

4.1.1 Artificial Neural Networks (ANN)

An artificial neural network, also called a multilayer perceptron, is an information processing model inspired by the functioning of the biological nervous system. Like the human brain, a neural network learns by example. Each artificial neural network is configured for a specific application such as object recognition or data classification, through a learning process. Its architecture is composed of several layers, each of which contains computing units called perceptrons (neurons) interconnected and working together to solve a specific task.

4.1.2 Perceptron (Neuron)

A perceptron has several inputs connected to a single output as shown in Fig.4.1. In its most simplified version, the inputs and the output are Boolean. More generally, the inputs can be real numbers. The perceptron [74] computes the output O as a function of input variables $I_1, ..., I_n$, each of which has a weight w, according to the formula: $s = b + \sum_{i=1}^{n} w_i I_i$ where b represents the bias. This sum is then passed through an

activation function. This function returns a value depending on s. If the emitted value exceeds a certain threshold, then the neuron is activated.

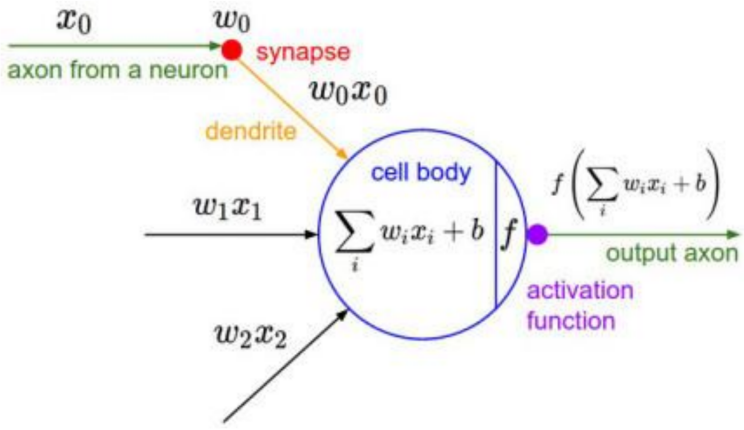

Figure 4.1: Mathematical representation of a perceptron [74].

4.1.3 Multi-layer Neural Networks

As its name indicates, a neural network is composed of several layers of neurons as shown in Fig.4.2. The input layer does not perform any computation, it is only in charge of receiving data from the outside and transmitting them to the next layer. The neurons of a hidden layer receive the data from the previous layer, and each one makes the weighted sum of all the inputs as explained previously. The output layer is responsible for generating the results. A loss function is associated with the final layer to calculate the error.

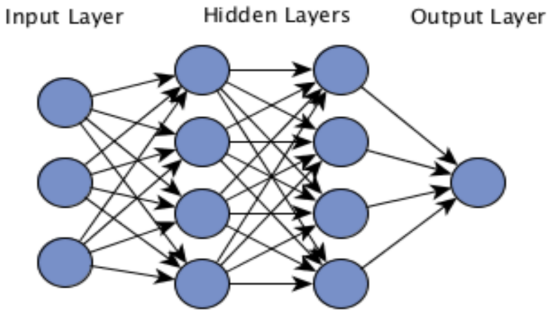

Figure 4.2: 2 hidden layers neural network architecture

Synaptic connections between neuronal cells in the human brain are malleable and constantly evolving through learning: this is also the case in artificial neural networks. For a neural network to learn, the weights

associated with the neural connections must be updated after data runs. Adjusting the weights helps to reconcile the difference between the actual and predicted results for subsequent runs (minimizing the error). Since in a neural network with many layers and many neurons it would make no sense to test a large number of weights to find the combination that does the best job. The weights are therefore updated methodically. They are learned by back-propagation of the gradient: The parameters that minimize the loss function are calculated progressively (for each layer, starting from the end of the network). The optimization is done with a stochastic gradient descent.

4.2 Deep Learning

There are several methods for doing machine learning, but the most promising approach that has literally "changed the world" recently is deep learning. The principles of this approach have been known since the 1980s, thanks to the work of some researchers such as Yann Lecun, one of the pioneers of deep learning. But it did not evolve in terms of use until 2012. Indeed, following the evolution of processing and storage technologies as well as the growth of data, deep learning was successfully applied in the "ImageNet Large Scale Visual Recognition challenge" [19] in 2012. During this competition, a team from the University of Toronto, Canada, trained a convolutional neural network, a type of network using deep learning, to classify a dataset containing 1.2 million high-resolution images into 1000 different classes, with a minimal error rate compared to those generated by other techniques After this victory, deep learning was brought back to the forefront with resounding success. Today, it is at the heart of all scientific issues. This attention is clearly deserved, given the enormous progress it has made. It has clearly turned the world of artificial intelligence upside down.

But what exactly is deep learning?

Yann Lecun explains the concept of deep learning as follows:

> *The idea is very simple: the trainable system consists of a series of modules, each representing a processing step. Each module is trainable, with adjustable parameters similar to the weights of linear classifiers. The system is trained from start to finish: at each example, all the parameters of all the modules are adjusted in order to bring the output produced by the system closer to the desired output. The advantage of deep architectures is their ability to learn to represent the world in a hierarchical way. Since all layers are trainable, there is no need to build a feature extractor by hand. The training will take care of that. Moreover, the first layers will extract simple*

> *features (presence of contours) that the following layers will combine to form more and more complex and abstract concepts: assemblies of contours into patterns, of patterns into parts of objects, of parts of objects into objects, etc.*

To summarize, deep learning is a subset of machine learning, using an architecture with a high level of abstraction (multiple layers of non-linear neurons). The idea behind this approach is that all layers are trainable, in order to allow feature extraction through a hierarchical learning process; at each level, abstractions are learned and sent to the next level. More complex concepts will be trained at that level, based on the abstractions received from the previous level. Deep learning models can achieve a very high level of accuracy. This accuracy does not only depend on the architecture of the network or the machine, but also on the data samples used. The quality of these samples has a direct impact on the efficiency of the learning: a small volume of data for example is likely to reduce the performance even if the machine used for learning is powerful, while a large volume of data and a good representation of it can lead to very accurate results. Some models require thousands or even millions of data examples to achieve good accuracy. There are several architectures of deep neural networks such as Deep Belief networks [43], Recursive neuronal networks, Long short-term memory, Deep Q-networks and Convolutional neural networks. A specific architecture is more efficient in some domains than in others. For example, convolution neural networks are widely used in the field of computer vision; while recursive neural networks are, for example, used for combinatorial optimization problems.

4.2.1 Convolutional Neural Networks

Convolutional neural networks (CNN) are a class of neural networks that have proven to be very effective in areas such as image recognition and classification: identification of objects, faces, traffic signs, vision feeding of robots and autonomous cars...etc. They represent the most widespread deep architecture. Developed in the late 80's by the pioneer of deep learning, Yann Lecun, these networks were inspired by the composition of the visual cortex. The visual cortex is composed of small cells, sensitive to certain regions of the visual field. In 1962, Hubel and Wesley demonstrated that certain neurons in the visual cortex respond only to certain contours, with a specific orientation. For example, some neurons act when exposed to vertical contours, others to horizontal or diagonal contours. The two researchers found that these neurons were distributed in columns, and that together they were able to produce visual perception.

The idea of components within a system, each of which is specialized in a specific task (neural cells seeking particular features), is the basis of convolutional neural networks.

There are several architectures of convolutional neural networks, ranging from the most basic to the most complex. "LeNet" was one of the very first convolutional neural networks that helped propel the field of deep learning. This pioneering work by Yann LeCun was named LeNet-5. At that time, the LeNet architecture was mainly used for character recognition tasks such as reading postal codes, numbers, etc. As shown in Figure 4.3, the LeNet stacks different convolution and subsampling layers, followed at the end by a fully connected layer.

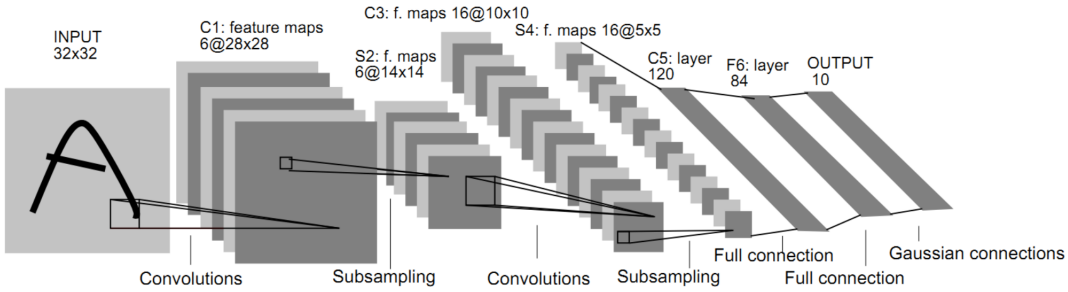

Figure 4.3: LeNet-5 architecture [60].

Several new architectures, which are enhancements of the LeNet, have been proposed in recent years. But, as complex as an architecture can be, it still has some basic elements which are:

- Convolution Layer
- Pooling /Subsampling layer
- Fully Connected Layer
- Activation Function
- Loss Function

Convolution Layer: It is the main component of a convolutional neural network and constitutes at least the first layer. As explained previously, the principle of convolution in the field of image processing, consists in scrolling a convolution filter of size (n*n) by a fixed step, on each region of size (n*n) of a digital image, and to calculate at each step the new value of the pixel located at the center of this region. The result of the image browsing is a map called features map. In a convolutional neural network, each perceptron of each convolution layer is linked to a subset of perceptrons of the previous layer, which represents a rectangular region of the image. This image can be the original, or the resultant of the operations of the previous layers. Unlike classical methods, where

the filter values are predefined, in a convolutional neural network, these values are learned during the training phase. Nevertheless, some parameters, notably the convolution step (stride) and the padding to zero (padding) are fixed beforehand. A stride represents the number of pixels with which the filter is moved on the image between two convolutions, Fig.4.4 shows a convolution with a stride of 2. Since the application of a convolution cannot be done on the edges of the image, the padding or zero-padding technique is used. It consists in adding zeros on the edges of the image. Fig.4.5 shows the application of a padding of size 1.

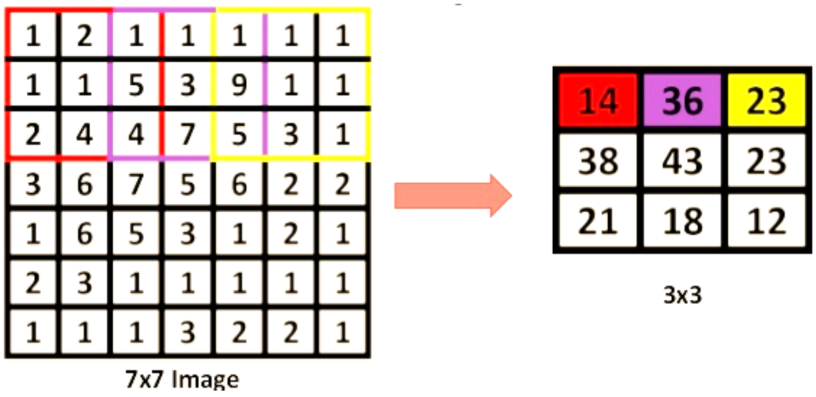

Figure 4.4: 3x3 convolution with a stride of 2[1]

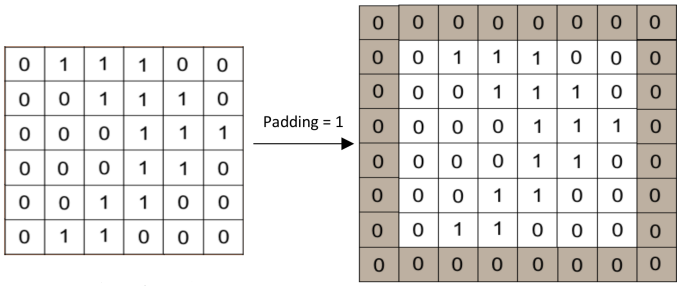

Figure 4.5: Padding of 1

Pooling / Subsampling: The goal of this layer is to reduce the size of the feature map, while keeping the most relevant information. There are several subsampling methods such as max-pooling, proposed by LeCun in his LeNet [61], and average pooling. Max-pooling consists in dividing a matrix (image) into small non-overlapping fragments and

[1] https://shangeth.com/post/gan-4/

taking the maximum value of each fragment, in order to produce a new reduced matrix. Fig.4.6 shows an example of this process.

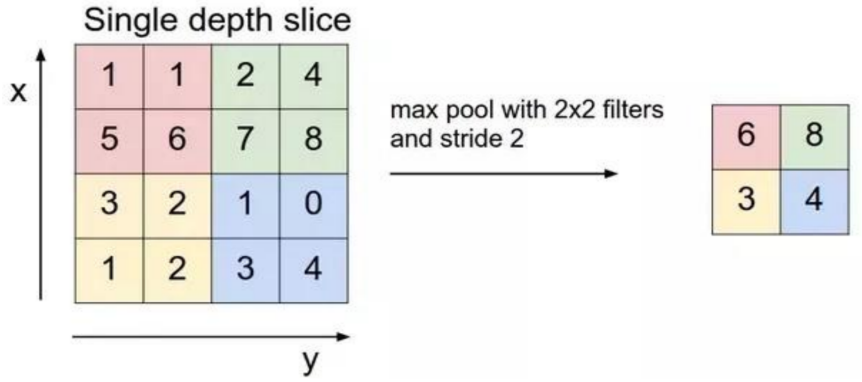

Figure 4.6: Max-Pooling application [2]

Average pooling proceeds in exactly the same way, except that the values in the new matrix represent the means of the fragments and not the maximum values.

Fully connected layer: This type of layer is used at the end of the neural network, after all the convolution and subsampling layers. It performs the final processing, not feature extraction.

In a fully connected layer, each perceptron is connected to all neurons in the previous layer as shown in Fig.4.7.

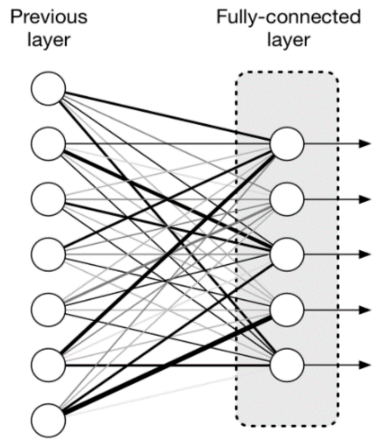

Figure 4.7: Fully-connected layer representation

[2] https://datascientest.com/convolutional-neural-network

These layers (Convolution, Pooling and Fully-connected) are, as mentioned before, the basic layers of a convolutional neural network. There are other types of layers, some of which are designed for very specific tasks.

Activation function All convolutional layers as well as the fully connected layer have an activation function. There are several activation functions, the choice of the function differs according to the architecture of the network. In CNNs, generally the ReLU (Rectified Linear Unit) function is used for the convolutional layers and the Softmax function for the output layer (for a classification task).

- **ReLu:** It is defined as follows:

$$\begin{cases} ReLU(x) = 0 & if \quad x < 0 \\ ReLU(x) = x & else \end{cases} \tag{4.1}$$

  Given the remarkable results obtained [16] from the use of this function, it has recently become the most widely used in practice.

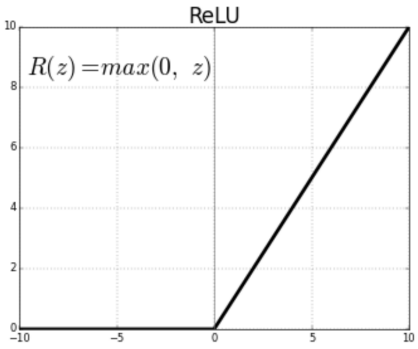

Figure 4.8: ReLu function

- **Softmax:** It is a function programmed to perform a multi-class classification from a K-output network. This function is often used in the final layer of a neural network based classifier. It is defined as follows

$$\sigma(z)_j = \frac{e^{Z_j}}{\sum_{k=1}^{K} e^{Z_k}} \quad forj = 1, ..., K \tag{4.2}$$

Loss fucntion: The loss function is used to "guide" the training process of a neural network by calculating the error committed by the latter. The choice of the error function has an effect on the performance

of the model. In most learning networks, the error is calculated as the difference between the actual output and the predicted output. Different loss functions are used to handle different types of tasks, namely regression and classification. Multi-task learning applications use a combination of different loss functions.

4.2.2 Some Deep Learning architectures

This section presents some deep learning architectures either we used or are fundamental to understanding other deep learning architectures.

4.2.2.1 VGG

VGG [95] has a simple feed-forward back-propagation architecture (see Fig.4.9) but is nevertheless powerful to learn complex patterns. Its architecture goes as follow:

- Input: VGG takes in a 224x224 pixel RGB image. For the ImageNet competition, the authors cropped out the center 224x224 patch in each image to keep the input image size consistent.
- Convolutional Layers: The convolutional layers in VGG use a very small receptive field (3x3, the smallest possible size that still captures left/right and up/down). There are also 1x1 convolution filters which act as a linear transformation of the input, which is followed by a ReLU unit. The convolution stride is fixed to 1 pixel so that the spatial resolution is preserved after convolution.
- Fully-Connected Layers: VGG has three fully-connected layers: the first two have 4096 channels each and the third has 1000 channels, 1 for each class.

Figure 4.9: VGG architecture [95]

4.2.2.2 Res-Net

Residual neural networks (Res-Net) [37] are also feed-forward back-propagation network with the particularity that some blocks can skip some layers.

Thus creating a contribution of so-called residues of the preceding layers. Doing so avoids the problem of vanishing gradient [41] and speeds up the learning problem by simplifying the network as it skips layers.

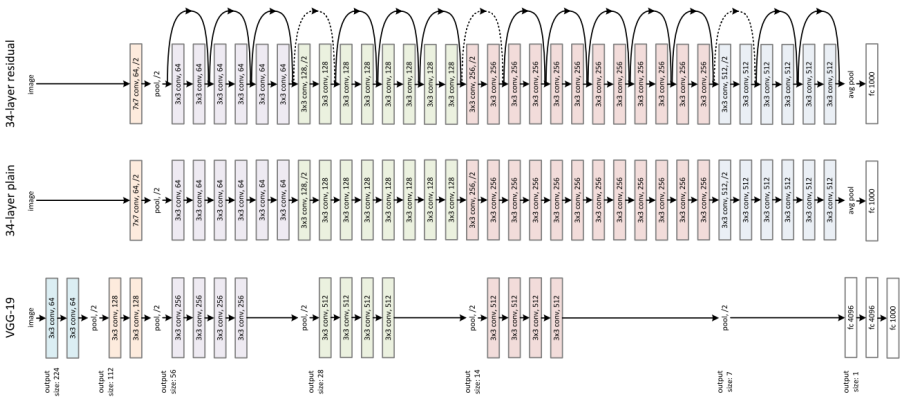

Figure 4.10: Res-Net architecture compared to VGG network [3].

4.2.2.3 Inception

Instead of classic CNN's, Inception networks [100] work not with one but different kernel size for convolution (commonly three different filters, 1x1, 3x3 and 5x5). After the max-pooling layer, the obtained results are concatenated and sent to the next layer. By structuring the CNN to perform its convolutions on the same level, it goes wider instead of going deeper. The network was designed to solve computational expense and overfitting issues.

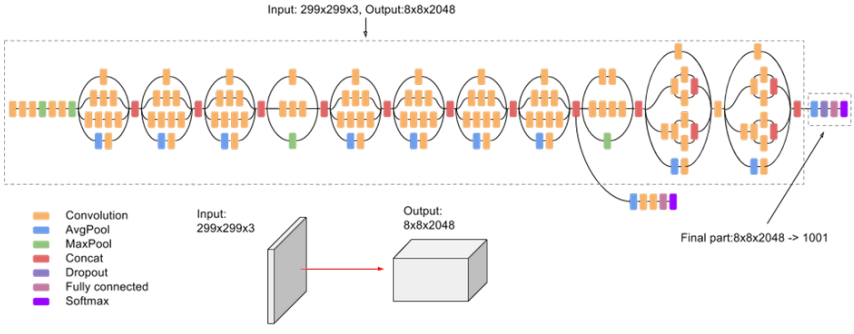

Figure 4.11: Inception-V3 architecture [101].

[3] https://fr.acervolima.com/reseaux-residuels-resnet-deep-learning/

4.2.2.4 Xception

Xception networks [14] are typically the combination of residual networks and Inception network, it goes wider with residues. We will detail its architecture in Chapter 5.

4.3 Artificial Intelligence for emotion recognition

AI and more specifically machine learning, has considerably changed the way information in proceeded. We do not need descriptors and feature extractors anymore as long as we have enough data to learn from them how to extract features. Nevertheless, machine learning algorithms highly suffer from unbalanced datasets. Many works have been conducted on machine learning's loss functions, on purpose of learning the very less frequent labels in inhomogeneous datasets. The most classic approach is to weight output labels in terms of their distribution as made in [34, 52, 87]. Another approach was introduced by [64] and based on the Cross-Entropy loss function weights up the hard samples regardless of their distribution.

4.3.1 Vision-based emotion recognition

As same as image processing, facial expression recognition achieved a point of mature due to two main factors: The availability of massive databases and the significant increase of computing power with GPU's which both allowed the use of sophisticated algorithms such as deep learning [59]. In the last decade, convolutional neural networks (CNN) [59] showed their superiority to extract features from static and dynamic images comparing to hand-crafted descriptors. Many works have been conducted using deep learning for affective computing. In [13], authors proposed a new loss function to enhance the discriminative power of deeply learned features. It reduces the intra-class variations while amplifying the inter-class variations. In the same manner, the work presented in [63] adopted a deep learning architecture which aims to enhance the discriminative power of deep features by preserving the locality resemblance while maximizing the inter-class scatters. They also proposed the first multi-labeled database for emotion recognition. In [73], authors proposed a new deep learning architecture and a new loss function in order to alleviate the inter-subject variation. They proposed to learn simultaneously expression and identity features with two CNNs. The work described in [10] proposed a new CNN architecture to learn facial expressions. They used a parallel feature extraction block (FeatEx) which consists of Convolutional, Pooling, and ReLU Layers. Besides the ability to learn by themselves how to extract image features, CNN's or rather 3D-

CNN's have allowed the take into consideration the temporal dimension by processing videos as consecutive blocks of images [55, 124, 132, 133].

4.3.2 Audio-based emotion recognition

In the last decade prosodic features have shown their efficiency to describe voice carried emotions as discussed in Chapter.3. In the other hand, CNN's have shown their power to learn by themselves how to extract image features and easily reach 98% of accuracy in some classification problems, this is why the major part of the recent works have explored different techniques to transform sound into images. In [28] authors proposed a CNN approach for audio-based emotion recognition where the input was a spectrogram of the audio signal. [76] made a breakthrough using deep retinal CNN's (DRCNN's) which was successful to recognize emotion from speech. [42, 57, 130] also used Mel-spectrogram transformation to generate CNN's input.

4.3.3 Audiovisual-based emotion recognition

Many modality fusion strategies have been developed in the recent years. As introduced and partially covered in Chapter.3 there are four main fusion strategies: Feature-level, decision-level, model-level and hybrid fusion. Machine learning has improved the most interesting but also challenging model-level fusion [44] but nevertheless this strategy remains widely unexplored. Most of the recent works focuses on hybrid multistage fusions and seem to have reach a point of mature [15,36,70]. In [2] authors made a three stage fusion pipeline using SVR's to combine early and late fusion strategies.

4.4 Conclusion

This chapter has been devoted to some basic notions of machine learning and deep learning, necessary for the understanding of the rest of the chapters. Also, we reviewed the major uses of machine learning in audio-based, visual-based and audio-visual based emotion recognition.Thus, we can start the next chapter with the necessary background.

Chapter 5

Visual context-based emotion recognition

Contents

Emotional state and human behavior analysis is knowing an increasing interest in psychology, neural science, and especially in computer science. This interest is mainly due to our need to make machines take the lead while interacting with humans, moving from computer-centered to human-centered human-computer interaction designs (HCI) [79, 81]. In human-human interactions, users' affective state and its changes are fundamental components. However, most current HCI designs are made to work with explicit information while ignoring the implicit ones. For instance, an autonomous car needs to know if the driver feels uncomfortable in case of tiredness without the driver explicitly informing the car.

Over the past four decades, researchers from many fields developed systems to automatically recognize the human emotional state. Starting from Ekman in 1967 [22], who encoded facial emotional information using Action Units (AU's) into the Facial Action Units System (FACS), to the explosion of Deep Neural Networks (DNN) due to the availability of the significant amount of data and computing power.

In this chapter we will tackle the first component of our solution presented in Chapter 2: The facial feature extraction module as shown in Fig.5.1. We will describe our solution to exploit visual signals in order to recognize emotions. As explained previously, the context is a must have modality for emotion recognition. Therefor, we designed an architecture composed of three main modules. A module for body features extraction, a module for scene features extraction and finally a module for the fusion of the two modalities.

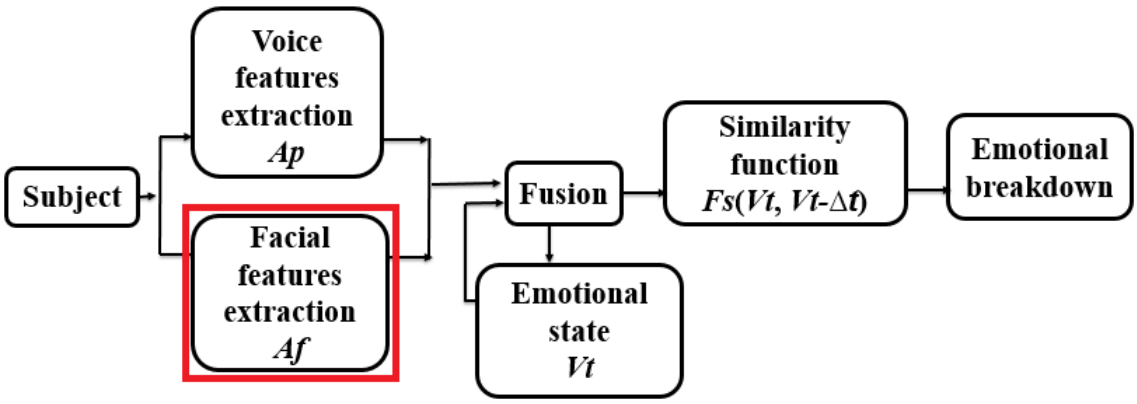

Figure 5.1: Our proposed solution in which the part tackled in this chapter, i.e. facial fractures extraction, is framed in red

5.1 Motivations

The classic emotion recognition problem, which was the classification of the six primary emotions (neutral, fear, surprise, disgust, sadness, and

happiness), was solved. However, the scientific community agrees that the emotional state cannot be reduced to six discrete categories due to two main reasons [127]. First of all, our emotions are complex and cannot be naively resumed to only six of them. On the other hand, emotions must be seen as overlapping clusters in a multi-dimensional continuous space and not as a set of discrete classes.

Figure 5.2: From EMOTIC database [52] : A child looking as he is chocked or surprised.

Figure 5.3: The same child in Fig.1 with the whole scene blowing out his birthday candles.

Even if many studies have pointed out the importance of the context (fig.5.3 and fig.5.3) in emotion recognition [27, 45, 89, 125], no approach has been explored in-depth. This is mainly due to the unavailability of data and also the difficulty of representing the context with classical approaches. Recently, EMOTIC database has been released [52]. This database contains 34320 people in 23571 images annotated on 26

emotion categories and their degree of arousal, valence, and dominance. Note that, emotions are complex expressions that are confusing and the dataset is so unbalanced that it is challenging to classify these many of categories in an accurate manner.

5.2 Contributions

In this work, we advance two main contributions. The first contribution show that in a highly correlated multi-task learning, loss functions must be chosen by groups instead of choosing them separately. To this end, we compare three categorical loss functions and three regression ones.

The second contribution is the proposal of a new loss function based on the binary Focal loss [64] that gives better results when dealing with unbalanced data.

In the following we will present our approach for a context-based multi-task emotion recognition on the Emotic dataset. We will firstly present the datasets we used, then present our CNN-based architecture and our new loss function called the Multi-label Focal Loss (MFL). We will compare our loss function to the most used loss functions in categorical classification combined with three mainly used loss functions for regression problems. Also, we will explain how we implemented our solution and the tools we used to this end. Finally, we will present and discuss our results. This work has been published in [7].

5.3 Databases

In this section we will present and describe the datasets invoved in this work. Those datasets are : ImageNet, Places and Emotic Datasets.

5.3.1 ImageNet dataset

ImageNet is, at this date, the largest image dataset. It has more than 14 million images labeled in 20,000 categories. A subset of ImageNet is used in ILSVRC, the most notorious challenge of image classification. In the challenge, only 1000 classes are considered. The most accurate models of this challenge are considered as the basis models for transfer learning. In the following we used a VGG network pre-trained on ImageNet.

5.3.2 Places Dataset

Places dataset [136] is a repository of more than 10 million scene photographs, labeled in 434 scene semantic categories (Office, Hotel room, Valley, Nursery ...). A subset of places called Places365-Challenge is used as a benchmark in the ILSVRC Challenge. In this subset only 365 categories are retained. The training set has 8 millions images, the

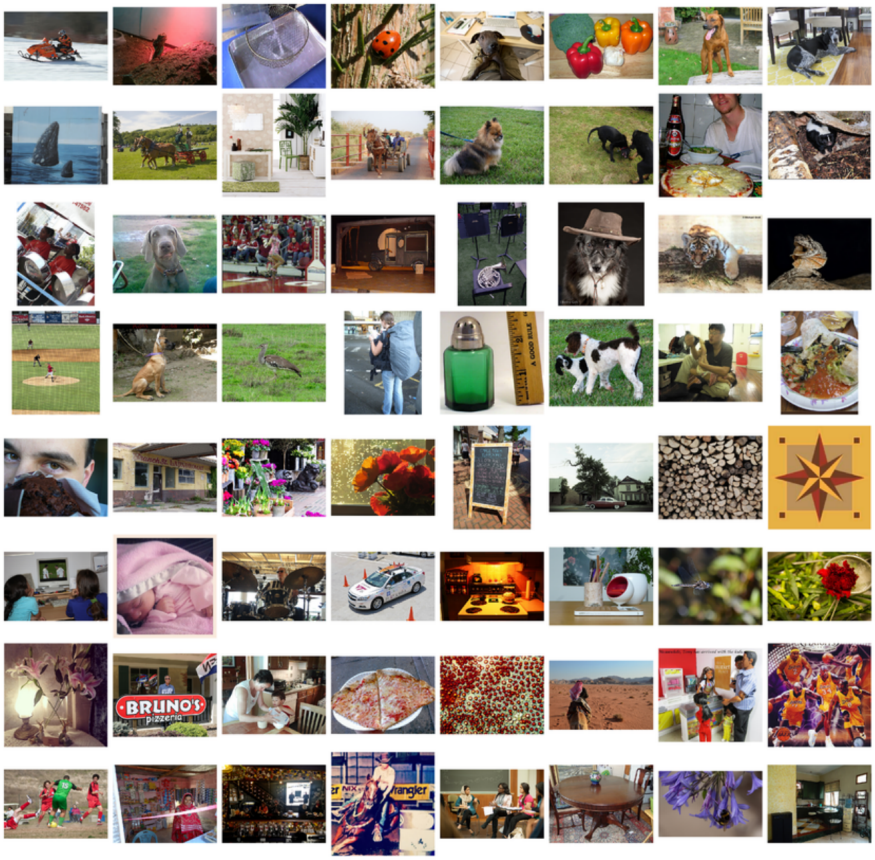

Figure 5.4: Images from ImageNet [19].

validation set has 50 images per class and the test set has 900 images per class. In the following we used the Xception newtork pre-trained on Places365-Challenge.

Figure 5.5: Some examples from Places Dataset [136]. The dataset contains three macro-classes: Indoor, Nature, and Urban.

5.3.3 Emotic dataset

The Emotic dataset is a set of images of people in uncontrolled environment. It contains 23,571 images and 34,320 annotated people. A part of the images was collected from Google search engine and the rest of the images is from two public datasets : COCO [65] and Ade20k [137].

One image can contain one or many people and each person is labeled with one or many emotion categories listed in table.5.2 and has its dominance, arousal and variance values from 0 to 10. Some examples are shown in fig.5.6 and fig.5.7.

EMOTIC dataset contains 34,320 annotated people, where 66 percent of them are males and 34 percent of them are females. There are 10 percent children, 7 percent teenagers and 83 percent adults amongst them.

Table 5.1: List of categorical emotions with explanation in EMOTIC dataset (part 1) [52]

1. Affection: fond feelings; love; tenderness
2. Anger: intense displeasure or rage; furious; resentful
3. Annoyance: bothered by something or someone; irritated; impatient; frustrated
4. Anticipation: state of looking forward; hoping on or getting prepared for possible future events
5. Aversion: feeling disgust, dislike, repulsion; feeling hate
6. Confidence: feeling of being certain; conviction that an outcome will be favorable; encouraged; proud
7. Disapproval: feeling that something is wrong or reprehensible; contempt; hostile
8. Disconnection: feeling not interested in the main event of the surrounding; indifferent; bored; distracted
9. Disquietment: nervous; worried; upset; anxious; tense; pressured; alarmed
10. Doubt/Confusion: difficulty to understand or decide; thinking about different options
11. Embarrassment: feeling ashamed or guilty
12. Engagement: paying attention to something; absorbed into something; curious; interested
13. Esteem: feelings of favourable opinion or judgement; respect admiration; gratefulness
14. Excitement: feeling enthusiasm; stimulated; energetic
15. Fatigue: weariness; tiredness; sleepy
16. Fear: feeling suspicious or afraid of danger, threat, evil or pain; horror
17. Happiness: feeling delighted; feeling enjoyment or amusement
18. Pain: physical suffering
19. Peace: well being and relaxed; no worry; having positive thoughts or sensations; satisfied

Table 5.2: List of categorical emotions with explanation in EMOTIC dataset (part 2) [52]

20. Pleasure: feeling of delight in the senses
21. Sadness: feeling unhappy, sorrow, disappointed, or discouraged
22. Sensitivity: feeling of being physically or emotionally wounded; feeling delicate or vulnerable
23. Suffering: psychological or emotional pain; distressed; anguished
24. Surprise: sudden discovery of something unexpected
25. Sympathy: state of sharing others emotions, goals or troubles; supportive; compassionate
26. Yearning: strong desire to have something; jealous; envious; lust

Figure 5.6: Some examples from the Emotic dataset with their categorical labels

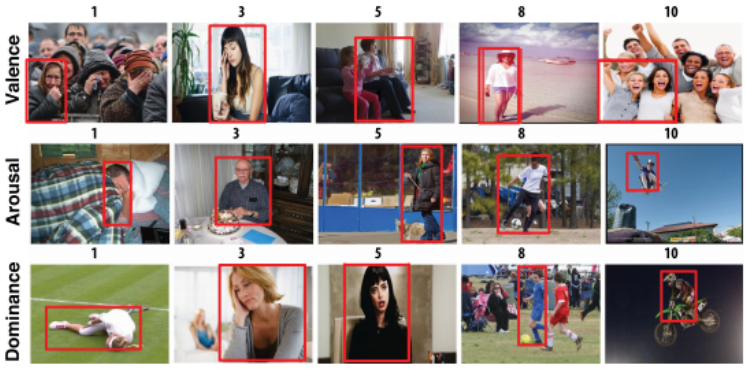

Figure 5.7: Some examples from the Emotic dataset with their valence, arousal and dominance values

Fig. 5.8 shows the number of annotated people for each of the 26 emotion categories, sorted by decreasing order. Notice that the data is unbalanced, which makes the dataset particularly challenging. An interesting observation is that there are more examples for categories

associated to positive emotions, like Happiness or Pleasure, than for categories associated with negative emotions, like Pain or Embarrassment. The category with most examples is Engagement. This is because in most of the images people are doing something or are involved in some activity, showing some degree of engagement. Fig.5.8b, 5.8c and 5.8d show the number of annotated people for each value of the 3 continuous dimensions. In this case we also observe unbalanced data but fairly distributed across the 3 dimensions which is good for modelling

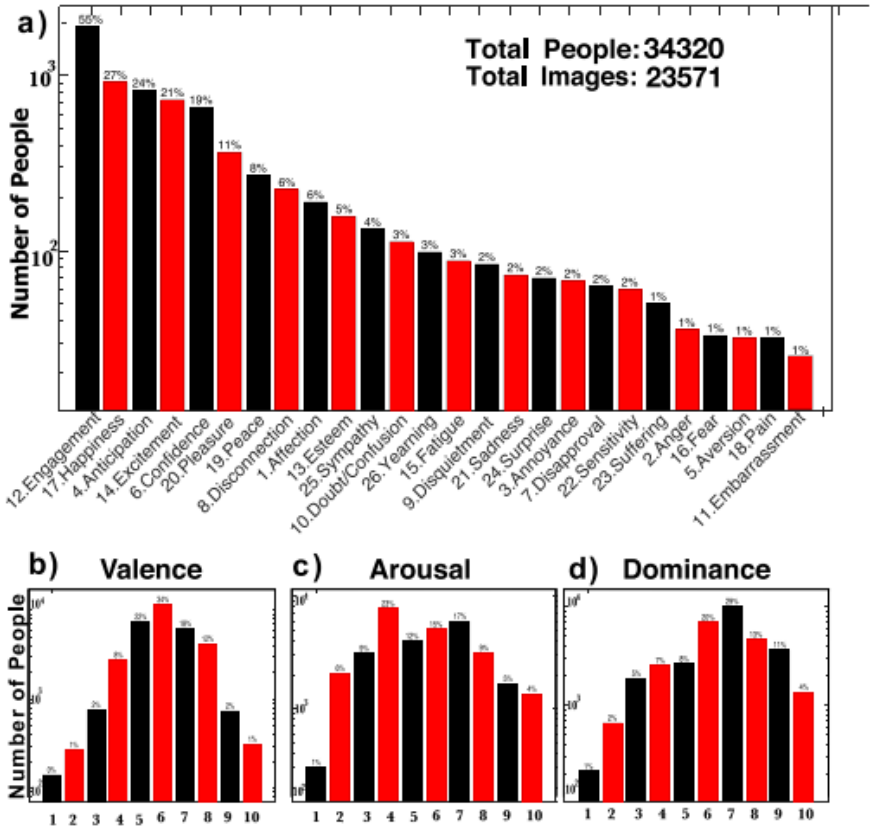

Figure 5.8: Data distribution in Emotic database

5.4 Approach

In this section, we describe our architecture that receives in input two images for categorical emotion classification over 26 classes and continuous spacial emotion regression over three values. We also present a new categorical loss function called the multi-label focal loss, two other categorical and three regression loss functions to compare with.

5.4.1 CNN architecture

As shown in [52], the scene (context) seems to improve the final classification results. We followed the same idea and designed an end to end architecture (Fig.5.9) built by three main modules: scene features extraction, body features extraction, and fusion-decision network. The first module takes as input the entire image, the second only the body delimited by a bounding box, and the third one is used to merge the scene and body features and output decisions.

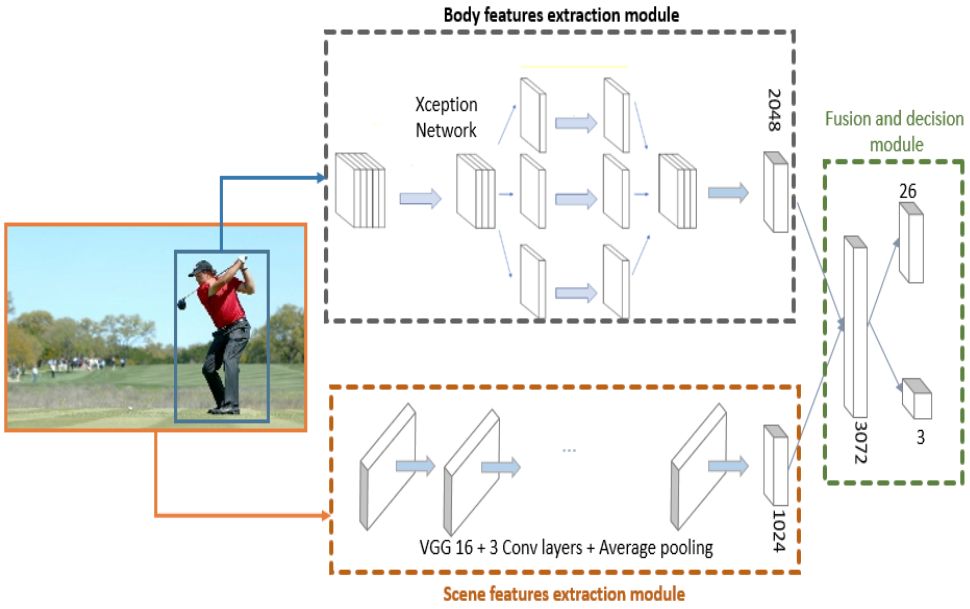

Figure 5.9: The proposed architecture for emotion recognition on EMOTIC database. The architecture takes in input two images, the whole image is propagated into the scene module and the cropped image is propagated into the body module.

The details architecture of our solution is shown in Fig.5.10. The left part represents the Xception network, i.e. the body feature extraction module, it takes as input images with size 299x299x3 (the last dimension if for RGB colors). As explained in Chapter4 Xception networks are a combination of Inception networks and residual networks, so except for the two first Conv blocks the architecture repeats itself (with different weights) as blocks of two SeparableConv and a residual Conv to the next block. At the end a Global Average Pooling is applied to the last SeparableConv wich contains 2048 filters. In top right corner of the image is shown the scene feature extraction module. Until the 5^{th} layer, the architecture is that of VGG16. We truncate the architecture at this point and added a new block if three Conv layers of 1024 filters and then applied a Global Average Pooling on the last layer. Doing so, we were able to lighten the scene feature extraction module from 134 million to 38 million trainable parameters.

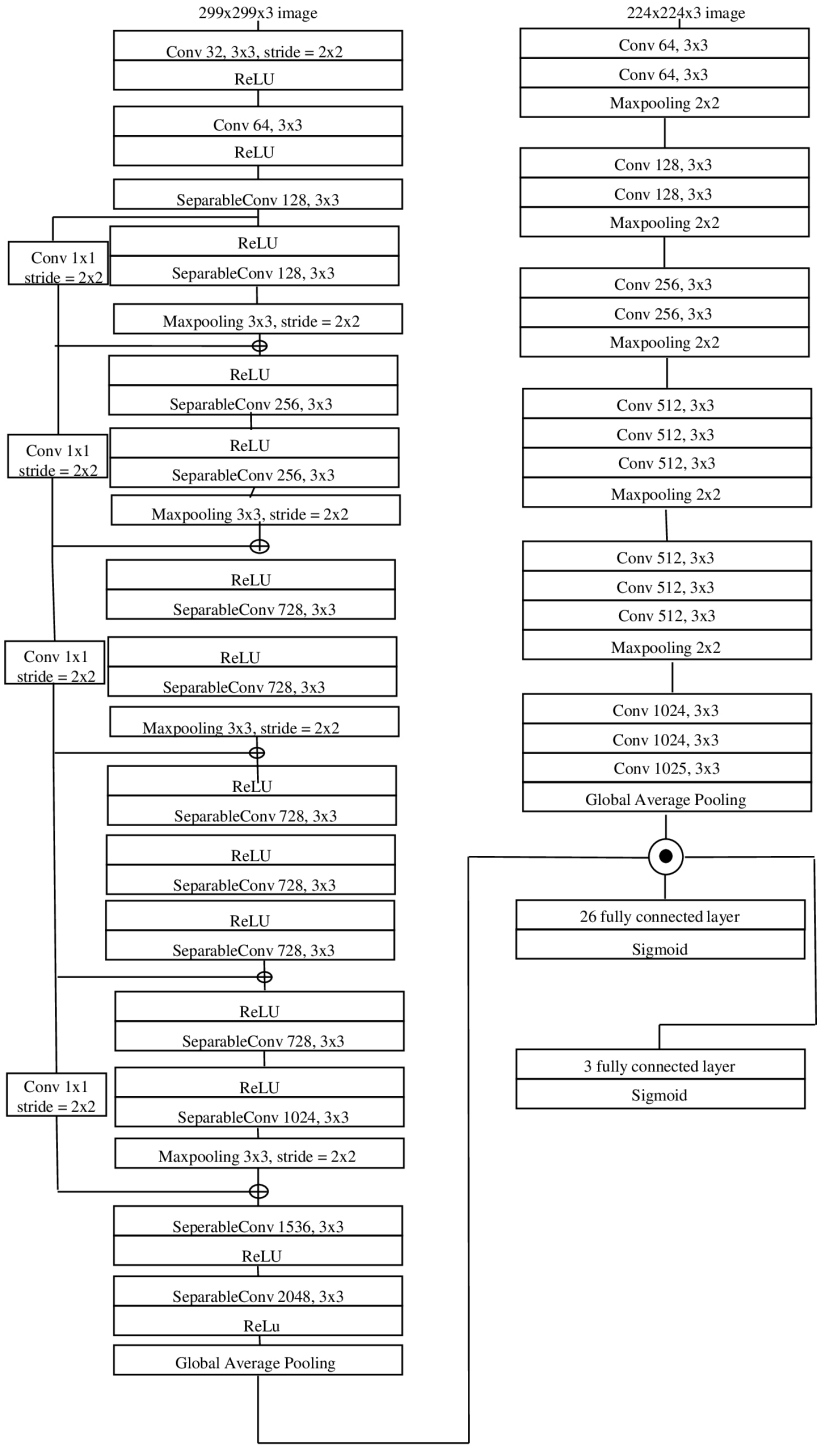

Figure 5.10: Detailed version of our proposed architecture. It combined the Xception network and a modified version of VGG network.

The last part in the bottom right of the image in the fusion module. We concatenate the two resulting layers with dimension 1x2048

and 1x1024 for, respectively, body feature and scene features extraction modules resulting in 1x3072 vector layer and then fully connected to both decision layers, i.e. 1x26 layer for categorical classification and 1x3 layer for regression prediction. Both layers have Segmoid function as activation function.

5.4.1.1 The scene features extraction module

The scene feature extraction module takes as input the entire image, which may contain more than one person. Those features reflect and codify the main aspects of the image. We used only the convolutional layers of VGG16 [95] pre-trained on Places Database [136]. To reduce computation complexity and overfitting issues, we remove the fully connected layer, and we replace it with a convolutional block of three layers and a Global Average Pooling layer. So, in the end, we compute 1024 scene features.

5.4.1.2 The body features extraction module

The body feature extraction module takes as input the body part in the image that implicitly contains information like facial expressions, head position, and body gesture are extracted. We used Xception network [14] pre-trained on ImageNet [20], which outputs 2048 body features.

5.4.2 The fusion-decision module

The fusion-decision module concatenates the body feature vector and the scene feature vector (3072 features). The output layers of classification (26 outputs) and regression (3 outputs) are fully connected to the concatenated layer. To avoid any eventual overfitting, except for the output layer, no additional fully connected layers were used.

5.4.2.1 Loss functions

The global loss function is a weighted sum of two distinct loss functions, $L_{global} = \lambda L_{cat} + L_{reg}$ where L_{cat} and L_{reg} represent respectively, the multi-label classification and the continuous variable regression loss functions. However, both problems, classification, and regression are highly correlated. We can say that the two problems are, to a certain point of view, a unique problem represented as a classification and a regression problem. Thus said, the back-propagation on regression output effects and may improve the categorical classification and vice versa. Therefore, L_{cat} and L_{reg} must be seen and chosen as a couple instead of two independent loss functions. We tried several couples of loss functions, which will be detailed in the following.

5.4.2.2 Classification loss functions

Since the categorical labels are unbalanced, we define a multi-label focal loss function from the binary focal loss [64], which has better results while dealing with unbalanced data. The binary focal loss is defined as:

$$FL(\sigma_t) = -\alpha\,(1-\sigma_t)^{\gamma}\log(\sigma_t) \tag{5.1}$$

where γ is the focusing parameter and

$$\sigma_t = \begin{cases} \sigma \text{ if } y = 1 \\ (1-\sigma) \text{ otherwise} \end{cases} \tag{5.2}$$

where, σ and y are, respectively, the categorical model output and its groundtruth label for the positive class.

We define the multi-label focal loss (MFL) as follows:

$$\begin{aligned} MFL_{\alpha,\gamma}(y,\sigma) = &-\sum_{i=1}^{Ne} \alpha\,(1-\sigma_i)^{\gamma}\, y_i \log(\sigma_i) \\ &+ (1-\alpha)\,\sigma_i^{\gamma}\,(1-y_i)\log(1-\sigma_i) \end{aligned} \tag{5.3}$$

where Ne is the number of classes. σ_i and y_i denote respectively the categorical model output and its groundtruth label for the i^{th} class. α and γ are two empirical parameters. γ, the focusing parameter, still has the same purpose as in the binary focal loss. It down-weights the easy classification samples while up-weights the hard ones, which in this case might be the less frequent ones. On the other hand, α here does not serve to balance the two binary classes but to promote either recall or precision costs. Also, we experimented with three other loss functions in order to compare our results.

- The Cross-Entropy loss (CE) is defined by:

$$CE(y,\sigma) = -\sum_{i=1}^{Ne} y_i \log(\sigma_i) + (1-y_i)\log(1-\sigma_i) \tag{5.4}$$

Note that the CE is a particular case of MFL when α and γ in Eq.5.3 are set to 0.5 and 0, respectively.

- And finally, the Euclidean loss (EUC) is defined as:

$$EUC(y,\sigma) = \sum_{i=1}^{Ne} (\sigma_i - y_i)^2 \tag{5.5}$$

Table 5.3: Macro-Precision for MFL+Huber for different values of γ.

γ values	$\gamma = 1$	$\gamma = 2$	$\gamma = 3$	$\gamma = 4$
Macro-Precision	26.83	26.28	**28.33**	26.62

5.4.2.3 Regression loss functions

We tested the three main loss functions for regression problem with each of the above mentioned multi-label loss functions.

- The Huber loss is defined by:

$$Huber_{\delta}(z,s) = \sum_{i=1}^{3} \begin{cases} \frac{1}{2}(z_i - s_i)^2 & \text{for } |z_i - s_i| \leq \delta \\ \delta\left(|z_i - s_i| - \frac{1}{2}\delta\right), & \text{otherwise} \end{cases} \tag{5.6}$$

where, s_i and z_i represent respectively the continuous model output and its groundtruth value for the i^{th} dimension. δ is an empirical parameter.

- The Mean Squared Error (L2) is defined as:

$$L2(z,s) = \frac{1}{3}\sum_{i=1}^{3}(z_i - s_i)^2 \tag{5.7}$$

- And the Mean Absolute Error (L1) is defined as:

$$L1(z,s) = \frac{1}{3}\sum_{i=1}^{3}|z_i - s_i| \tag{5.8}$$

5.5 Training setup and experimental results

We trained the whole architecture on EMOTIC database which contains only one partition with all loss function combinations. After several attempts, we empirically found the most suitable parameters for loss functions and the training parameters.

In all the results below, regarding the multi-label focal loss, γ is set to 3. We tried four values for γ in the set $\{1, 2, 3, 4\}$ the corresponding performance is depicted in Table 5.3. As it can be seen, when γ was set to 3 the performance was the best. α is set to 0.5 to give the same weight to precision and recall. Regarding Huber loss, δ is set to 0.1 because output values are in $[0, 1]$ and δ is usually around 10%.

For the training parameters, λ in the global loss function is set to 5. We made a dichotomic search on $[0.5, 10]$ and found that $\lambda = 5$ gave the best results. Due to a lack of available memory, the batch size was set to 10 samples and the learning rate starts from 10^{-4} with a decay of 10^{-5}.

The training, validation and test sets are already provided by [52]. These three sets contain respectively 12957, 3334, and 7280 images.

We tested all the combinations of the above classification and regression loss functions. We trained nine models of the proposed architecture, which takes in input the body and scene image and outputs the categorical emotion and the three continuous emotional values (Valence, Arousal, Dominance). The body feature extraction module is pre-trained on ImageNet, the scene feature extraction module is pre-trained on Places, and the fusion-decision module is trained from scratch. We denote in the following MFL for multi-label focal loss, CE for cross-entropy, EUC for euclidean loss, L2 for mean squared error, and L1 for mean absolute error. Our comparative metrics for the categorical classification are the precision (where Precision = Recall) for each class. Note that, in this case the Macro-Precision over all classes is equal to Maco-F1. For the regression problem, we use the Mean Average Error.

For each loss function combination, the models convergence time was almost similar. It took around 25 epochs to the models to converge and each epoch lasted more than 3 hours which made it very time consuming to adjust all the hyper-parameters. The algorithm was running on a Nvidia Titan V GPU.

5.5.1 Categorical classification

In Tables 5.4, 5.5 and 5.6 we show our results when combining all loss functions discussed above.

Table 5.4 shows results when combining MFL, CE, and EUC with L2. In this case, CE outperformed MFL and EUC with a precision of 27.08%.

In Table 5.5, CE outperformed MFL and EUC when all of them were combined with L1. Also, CE and EUC gave better results with L1 than with L2. However, MFL gave worse results with L1 than with L2. We notice that the results did not increase homogeneously, which means that there is a synergy between loss functions.

In table 5.6, we show the results of combining MFL, CE, and EUC with the Huber loss. We can clearly remark that the combination of MFL with the Huber loss outperforms all the other combinations. Moreover, it is important to mention that this performance is not due to the Huber loss alone. Indeed, using the Huber loss instead of L2 and L1 does not necessarily improve the performance. For example, EUC with L1 gave a Macro-Precision of 26.68% (Table 5.5) when it decreased to 26.06% (Table 5.6) with the Huber loss.

The obtained experimental results showed that the performance obtained by the focal and cross-entropy losses were better than that obtained by the Euclidean loss. It is well known that, in general, the cross entropy loss functions are better then EUC in classification problems while EUC loss is more suitable for regression problems.

Table 5.4: Precision on the test set.

Categorical Emotions	MFL + L2	CE + L2	EUC + L2
1. Affection	**29.44**	28.30	26.13
2. Anger	**13.65**	11.02	10.66
3. Annoyance	16.36	**17.58**	16.48
4. Anticipation	**57.31**	57.27	56.85
5. Aversion	08.10	**08.93**	07.05
6. Confidence	73.08	**76.07**	73.40
7. Disapproval	14.30	**14.90**	13.20
8. Disconnection	**27.10**	26.74	26.45
9. Disquietment	**18.76**	18.67	17.81
10. Doubt/Confusion	20.66	**20.83**	19.84
11. Embarrassment	02.26	**02.84**	02.52
12. Engagement	**85.90**	**85.89**	84.79
13. Esteem	15.78	**15.82**	14.85
14. Excitement	69.04	**71.54**	69.28
15. Fatigue	12.83	**13.09**	11.75
16. Fear	05.79	**05.82**	04.62
17. Happiness	**76.63**	74.67	74.64
18. Pain	**09.93**	06.91	08.26
19. Peace	23.40	**23.86**	22.21
20. Pleasure	**46.25**	44.65	44.64
21. Sadness	19.44	19.39	**20.15**
22. Sensitivity	05.20	**06.60**	06.48
23. Suffering	21.60	**22.21**	19.96
24. Surprise	07.55	**08.53**	07.56
25. Sympathy	**13.68**	12.16	11.11
26. Yearning	08.58	**09.67**	08.29
Macro-Precision	27.02	**27.08**	26.12

Table 5.5: Precision on the test set.

Categorical Emotions	MFL + L1	CE + L1	EUC + L1
1. Affection	29.14	28.84	**29.23**
2. Anger	11.24	**15.75**	08.36
3. Annoyance	13.75	**17.69**	15.10
4. Anticipation	56.98	57.34	**57.49**
5. Aversion	07.72	**08.54**	06.98
6. Confidence	74.90	74.81	**75.69**
7. Disapproval	12.64	13.49	**13.76**
8. Disconnection	27.34	27.25	**27.86**
9. Disquietment	17.88	**19.11**	17.49
10. Doubt/Confusion	20.32	20.82	**20.86**
11. Embarrassment	02.45	**03.15**	02.66
12. Engagement	85.09	84.68	**85.87**
13. Esteem	15.40	15.43	**16.20**
14. Excitement	69.53	70.26	**70.34**
15. Fatigue	11.63	**13.37**	12.41
16. Fear	**06.31**	05.99	04.58
17. Happiness	74.86	74.61	**76.07**
18. Pain	**11.33**	09.53	07.42
19. Peace	23.27	**24.00**	22.75
20. Pleasure	44.55	44.52	**47.05**
21. Sadness	19.52	**20.61**	20.40
22. Sensitivity	05.95	**06.50**	05.54
23. Suffering	20.11	**24.20**	22.53
24. Surprise	08.03	**08.76**	07.71
25. Sympathy	11.23	**13.18**	11.29
26. Yearning	**09.50**	09.09	08.12
Macro-Precision	26.56	**27.27**	26.68

Table 5.6: Precision on the test set.

Categorical Emotions	MFL + Huber	CE + Huber	EUC + Huber
1. Affection	**31.92**	29.63	29.91
2. Anger	**13.94**	11.81	11.78
3. Annoyance	**17.42**	16.78	15.09
4. Anticipation	**57.73**	57.35	57.28
5. Aversion	08.18	**08.73**	07.13
6. Confidence	75.29	**77.78**	74.53
7. Disapproval	**14.88**	13.97	12.56
8. Disconnection	**28.32**	26.11	25.79
9. Disquietment	**19.72**	19.11	17.43
10. Doubt/Confusion	**23.11**	22.57	19.77
11. Embarrassment	**02.84**	02.67	02.00
12. Engagement	**85.83**	84.98	85.51
13. Esteem	16.72	**16.81**	15.81
14. Excitement	70.43	**70.66**	69.52
15. Fatigue	**14.43**	12.46	10.92
16. Fear	**08.27**	05.62	04.69
17. Happiness	**76.61**	73.94	75.18
18. Pain	**09.38**	08.64	07.17
19. Peace	**24.31**	22.42	23.23
20. Pleasure	**46.89**	45.07	42.76
21. Sadness	**23.94**	19.69	17.00
22. Sensitivity	06.28	**07.41**	05.06
23. Suffering	**26.24**	20.87	19.80
24. Surprise	**10.07**	09.30	07.87
25. Sympathy	**13.98**	12.35	11.74
26. Yearning	**09.71**	09.09	07.96
Macro-Precision	**28.33**	27.15	26.06

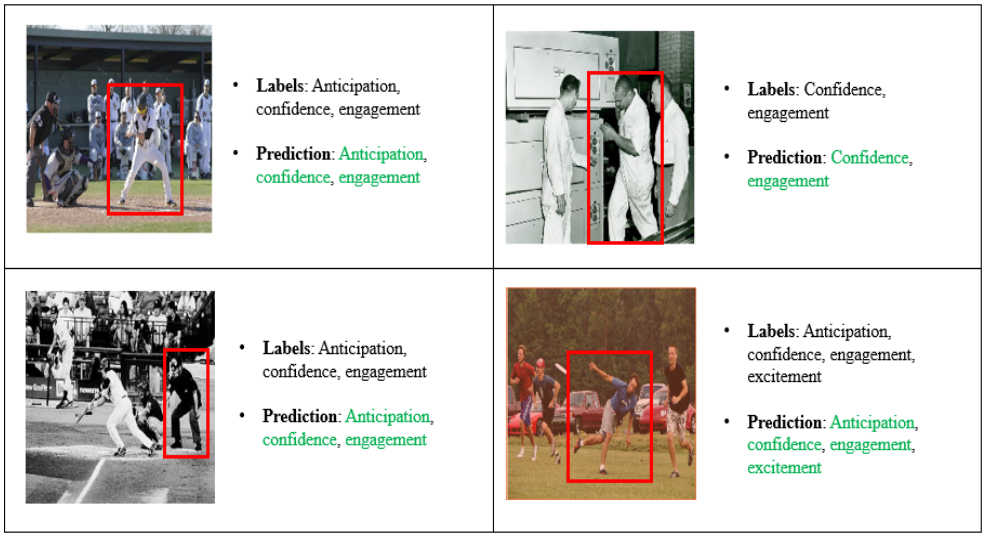

Figure 5.11: Some well predicted examples from the Emotic dataset

5.5.2 Regression

Tables 5.7, 5.8 and 5.9 show the Mean Average Error of the nine models discussed above on Valence, Arousal and Dominance.

Table 5.7: Mean Average Error for each continuous variable.

Dimensions	MFL + L2	CE + L2	EUC + L2
Valence	**0.085**	0.090	0.088
Arousal	**0.101**	0.105	0.106
Dominance	**0.094**	0.095	0.102
Mean	**0.094**	0.097	0.099

Table 5.7 shows the errors when L2 is combined with MFL, CE, and EUC. We note that the best results were obtained when L2 was combined with MFL.

Table 5.8: Mean Average Error for each continuous variable.

Dimensions	MFL + L1	CE + L1	EUC + L1
Valence	**0.084**	0.088	0.095
Arousal	**0.103**	0.106	**0.103**
Dominance	0.096	0.100	**0.095**
Mean	**0.094**	0.098	0.098

Table 5.8 shows the error when L1 is combined with MFL, CE, and

EUC. We notice that the best results were also when L1 was combined with MFL.

Table 5.9: Mean Average Error for each continuous variable.

Dimensions	MFL + Huber	CE + Huber	EUC + Huber
Valence	**0.083**	0.085	0.089
Arousal	**0.099**	**0.099**	0.103
Dominance	**0.092**	0.093	0.097
Mean	**0.091**	0.092	0.096

Lastly, Table 5.9 shows the errors when we combine Huber with MFL, CE, and EUC. We also notice that the best results were given by MFL.

These results show that the Huber loss is better than L2 and L1. However, combining the Huber loss with another loss function may straighten or weaken its results.

5.5.3 Comparing with the state of the art

In table 5.10, we compare our best model (MFL + Huber loss) with the current state of the art (as far as we know). We had better results than that of [52] where authors used an EUC loss weighted by labels frequency over the train set. Also, for the labels with a frequency less than 6%, our model is better in 8 labels out of 11 as shown in Figure 5.12.

Table 5.10: Comparing our model to the current state of the art.

Categorical Emotions	Our Model	Results in [52]	Results in [131]
1. Affection	31.92	27.85	**46.89**
2. Anger	**13.94**	9.49	10.87
3. Annoyance	**17.42**	14.06	11.23
4. Anticipation	57.73	58.64	**62.64**
5. Aversion	**08.18**	07.48	05.93
6. Confidence	75.29	**78.35**	72.49
7. Disapproval	14.88	**14.97**	11.28
8. Disconnection	**28.32**	21.32	26.91
9. Disquietment	**19.72**	16.89	16.64
10. Doubt/Confusion	23.11	**29.63**	18.68
11. Embarrassment	02.84	**03.18**	01.94
12. Engagement	85.83	87.53	**88.56**
13. Esteem	16.72	**17.73**	13.33
14. Excitement	70.43	**77.16**	71.89
15. Fatigue	**14.43**	09.7	13.26
16. Fear	08.27	**14.14**	04.21
17. Happiness	**76.61**	58.26	73.26
18. Pain	**09.38**	08.94	06.52
19. Peace	24.31	21.56	**32.85**
20. Pleasure	46.89	45.56	**57.46**
21. Sadness	23.94	19.66	**25.42**
22. Sensitivity	06.28	**09.28**	05.99
23. Suffering	**26.24**	18.84	23.39
24. Surprise	10.07	**18.81**	09.02
25. Sympathy	13.98	14.71	**17.53**
26. Yearning	09.71	08.34	**10.55**
Macro-Precision	28.33	27.39	**28.42**

Our model did not outperform [131] regarding Macro-Precision over all classes. However, with our proposed MFL, we had better results over the first nine less frequent classes (embarrassment, pain, anger, sensitivity, aversion, fear, suffering, disapproval, and annoyance) as shown in Figure 5.13. Also, for the labels with a frequency less than 6%, our model in better in 10 labels out of 11.

Our model gave better results on the less frequent labels due to how the MFL loss considers unbalanced data. The MFL loss does not consider the bias between classes, it up weights the error of the hard samples, which generally have the less frequent labels.

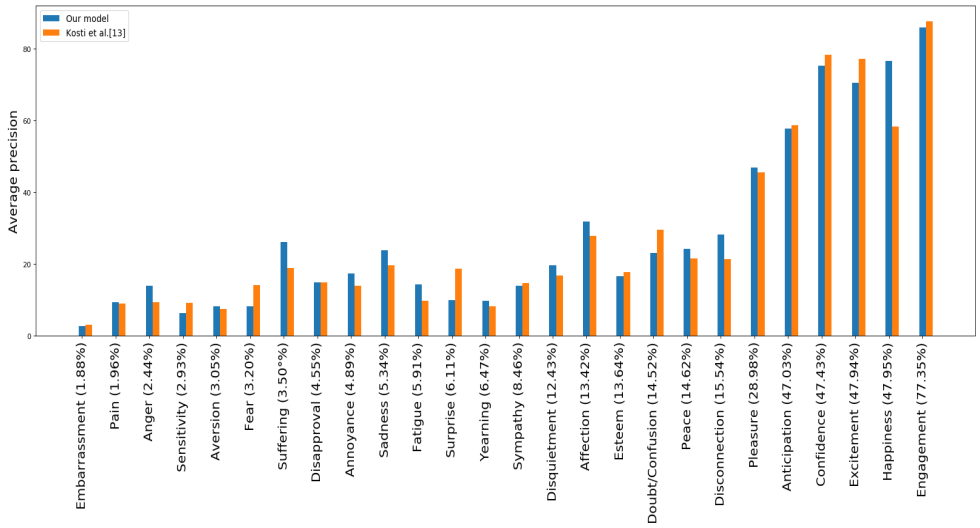

Figure 5.12: Comparing our model's precision evolution along all classes sorted by their distribution over the test set with results in [52]. (From the less frequent to the most frequent)

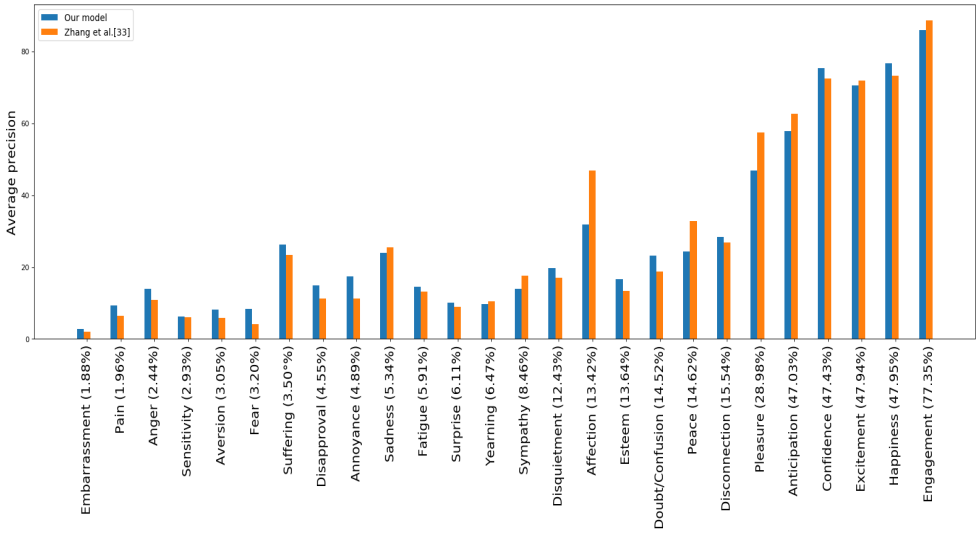

Figure 5.13: Comparing our model's precision evolution along all classes sorted by their distribution over the test set with results in [131] (From the less frequent to the most frequent)

5.6 Conclusion

This chapter introduced a new deep learning architecture and a new loss function, namely the multi-label focal loss (MFL). The objective is to deal with unbalanced emotion classes. We described our architecture, its components, and the training process. We compared our MFL loss function with two other loss functions known as the standard loss functions for categorical classification and studied their behavior when combined

with three regression loss functions. The proposed MFL outperformed the binary cross-entropy and the euclidean loss. It did also point out the synergy between loss functions in a multi-task learning problem, especially when dealing with high correlated multi-task problems. The comparison of our results with the current state of the art showed that MFL gave the best results on less frequent labels on the EMOTIC database.

Chapter 6

Audio-based emotion recognition

Contents

As explained previously, audio features contain 38% of the global emotional information. In this chapter we will tackle the second component of our solution presented in Chapter 2: The voice feature extraction module as shown in Fig.6.1. We will describe our solution to exploit audio signals in order to recognize emotions.

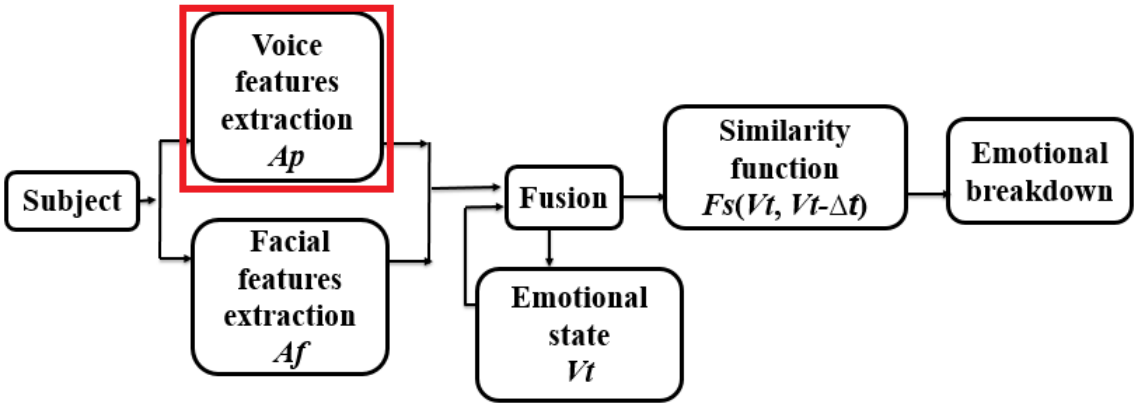

Figure 6.1: Our proposed solution in which the part tackled in this chapter, i.e. voice features extraction, is framed in red

6.1 RAVDESS database

The RAVDESS database [68] is an audio-visual database for emotional speech and song among 7356 files. The database contains 24 professional actors (12 female, 12 male), vocalizing two lexically-matched statements in a neutral North American accent. Speech includes calm, happy, sad, angry, fearful, surprise, and disgust expressions, and song contains calm, happy, sad, angry, and fearful emotions. Each expression is produced at two levels of emotional intensity (normal, strong), with an additional neutral expression. Note, there are no song files for Actor18.

In the following we will only focus on audio-only speech file. The 24 actors say "Kids are talking by the door" and "Dogs are siting by the door" with each previously stated emotion at two levels of intensity. The duration of sound tracks is between 3 to 5 seconds and 1440 files, non-divided on train/test sets, compose the dataset. The data distribution is homogeneous as shown in Fig.6.2.

6.2 Approach

As show previously, CNN's proved their efficiency to extract features from images and use those features for classification or regression tasks. It is known that the 2-D spectrogram contains low-level spectral information related to the speaker's emotion expression, such as energy and pitch which is suitable as input for a CNN.

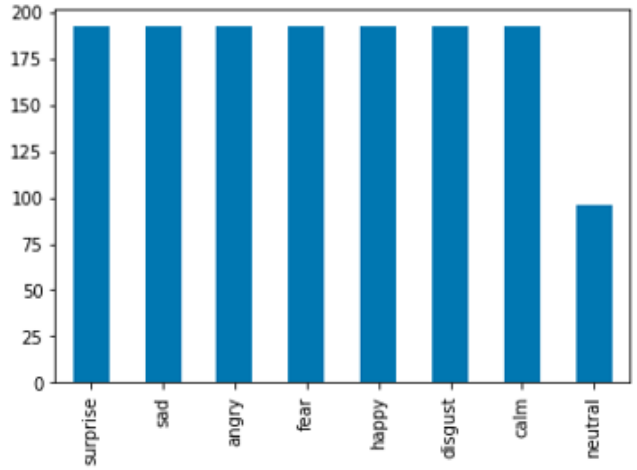

Figure 6.2: RAVDESS data distribution

6.2.1 Data preprocessing

In order to transform sound tracks into images, both frequency and time domain representation must be conserved so CNN's could extract both domain features. Thus, the spectrogram representation could be seen as an image carrying both frequency and time domain features of sound track. In the following we will use the Mel-spectrogram representation, a spectrogram converted to the mel scale [99] which is which is closest to the human ear, i.e. we can easily distinguish two sounds at 500Hz and 1000Hz but hardly distinguish between 10 000Hz and 10 500Hz.

Two data representation are presented in this Chapter: computing the whole Mel-spectrogram of a sound track and use it for classification and computing the whole Mel-spectrogram and split it into multiple time-frames and then use theses segments for classification. Both data representation have their advantages and disadvantages as in the first all the features describing the sound track are fully accessible for the CNN which may improve the recognition accuracy. However, for real-time running, such an approach might be too greedy because the system will compute one Mel-Spectrogram each 4 or 5 seconds while there could be many expressed emotion such a lapse of time. The second approach fixes this issue but the training will be more difficult and the accuracy might be lower as few features will be propagated into the network.

6.2.1.1 The entire Mel-Spectrogram as input

- From the digital sound track input we remove the empty signal at the beginning and the end of the track as show in Fig.6.3.

- We compute the frequency domain representation using Short Time Fourier Transformation (STFF).

- For a single sound track we adopt 64 Mel-filter banks from 20 to 8000 Hz (see Fig.6.4)
- Windowing with a Hann window of 23 and 12 ms overlapping.
- Each segment has a size of $64xN$ where N satisfies the equality $\frac{23}{2}(N-1) =$ sound track duration.

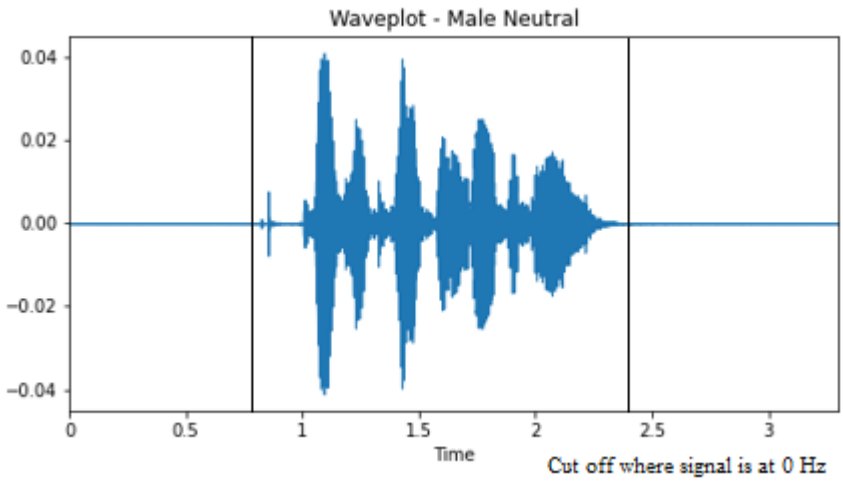

Figure 6.3: 1-D audio time domain representation cut off of the edges where the signal is at 0 Hz

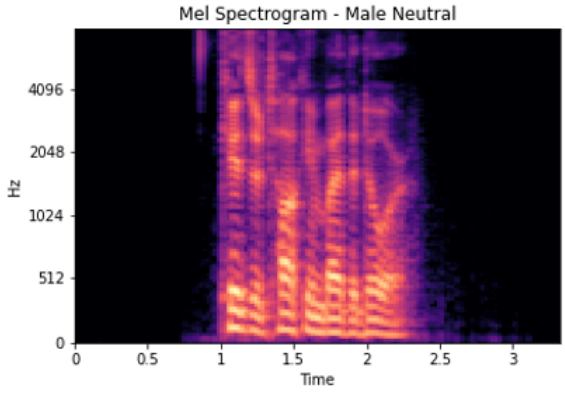

Figure 6.4: Mel-Spectrogram representation

6.2.1.2 Framing the Mel-Spectrogram using a context window

In the following we will compute Mel-spectrogram frames with size 64x64 using a 64x64 context window.

- From the digital sound track input we remove the empty signal at the beginning and the end of the track.

- We compute the frequency domain representation using Short Time Fourier Transformation (STFF).
- For a single sound track we adopt 64 Mel-filter banks from 20 to 8000 Hz.
- Windowing with a Hann window of 11.6ms and 5.8 ms overlapping.
- Parsing the whole Mel-spectrogram with a context window of 64 frames to obtain audio segments with size of 64x64 with an overlap of 10 frames.
- Each segment has a size of 64x64 and its time duration is $\frac{11.6}{2}(64-1) \approx 356ms$ which is higher than the minimum recommended duration ($250ms$) to recognize emotions [83].

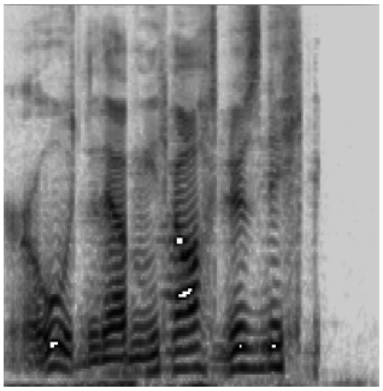

Figure 6.5: The resulting image from the audio track

This extracted Mel-spectrogram can be regarded as the gray-scale image feature representation of audio data as shown in Fig.6.5. In order to use a pre-trained VGG16 model (Fig.6.6) which has as input size 299x299x3, the Mel-spectrogram frames are resized to the suitable shape.

6.2.2 Architecture and training

To train our model, we used a pre-trained VGG 16 version on Imagenet. It contains 6 convolution blocks and two fully connected layers (see Fig.6.7). The output layer contains 8 neurons for the 8 classes with Softmax as an activation function.

For both approaches, we trained our models on 80% of the available data, randomly chosen. For each data segmentation approach, the resulting train and test sets contain, respectively, 63864 and 15966 samples for splited Mel-spectrograms and 1440 and 288 of whole spectrogram as shown in Table.6.1. After several tests, the learning rate was initially set at 10^{-4} with a decay of 10^{-5}.

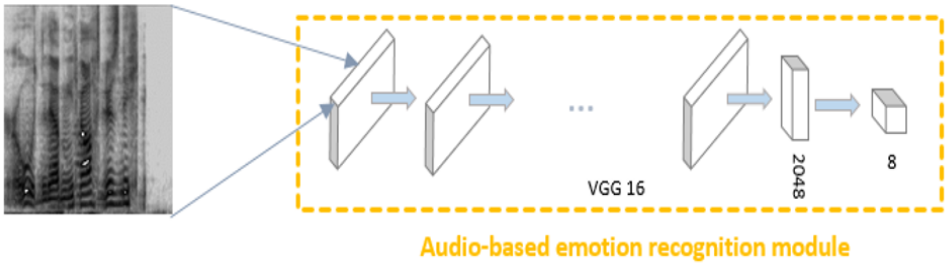

Figure 6.6: Proposed architecture for audio-based emotion recognition

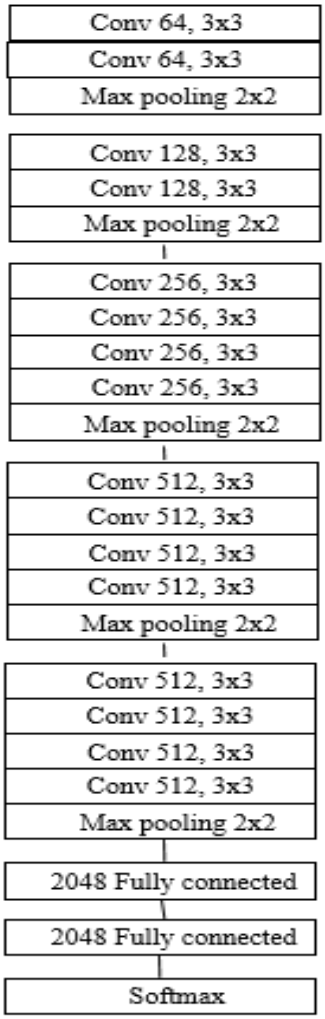

Figure 6.7: Detailed VGG16 architecture for audio-based emotion recognition

Table 6.1: Number of samples in train and test sets.

	Whole Mel-Spectrogram	Mel-Spectrogram split into frames
Train set	1152	63 864
Test set	228	15 966

In the next section we will discuss the results of each model.

6.3 Results

We tested our two models on the remaining 20% of their data. As shown in Table.6.2 model A gave better results in overall, this is due the availability of data and features describing the emotional state, contained in the whole Mel-spectrogram which makes it easier to classify emotions.

Table 6.2: Accuracy of model A and B in their test sets.

	Model A	**Model B**
Neutral	**0.68**	0.36
Calm	**0.9**	0.79
Happy	**0.58**	0.51
Sad	0.47	**0.49**
Angry	**0.74**	0.67
Fearful	**0.72**	0.41
Disgust	**0.74**	0.58
Surprised	**0.82**	0.58
Mean	**0.70**	0.56

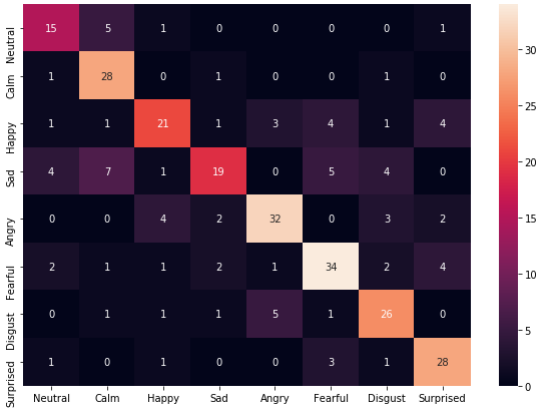

Figure 6.8: Confusion matrix through the test set dor Model A

From the confusion matrices shown in Fig.6.8 and 6.9 we can conclude that from both models and approaches, the confusions between classes are relatively the same (at different scale).

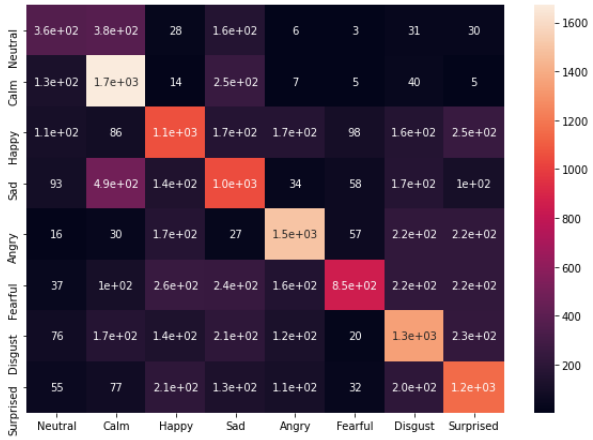

Figure 6.9: Confusion matrix through the test set for Model B

Even if model A gave better results than model B, for our system, the model B, i.e. which has been trained on split frames, will be chosen. Even if the accuracy is lower than model A, the temporal flexibility offered by such an approach matters more, especially for real time audio-visual synchronization and the detection of brief events.

6.4 Conclusion

In this chapter, we presented how we designed our audio-based emotion recognition model. We presented the RAVDESS dataset on which we trained our architecture and how we preprocessed raw audio tracks to transform them into images. Then we presented the model's results on RAVDESS dataset.

Chapter 7

EmoRuption: Towards emotional breakdowns detection

Contents

As firstly presented in Chapter.2, our objective is to build an architecture to fingerprint the current emotional state and compare it with the previous one. We will, in the following, introduce and explain the principle of "fingerprinting" an emotion state so we can calculate the difference between two emotional states.

7.1 Architecture construction

It is known that neural networks work like big feature extractors with a final decision layer. Then, if we prune the decision layer of a neural network, we will end-up with a feature extractor. A feature extractor that extract features from an image and return a 1-D vector. This is what we called emotional state fingerprinting (see Fig.7.1).

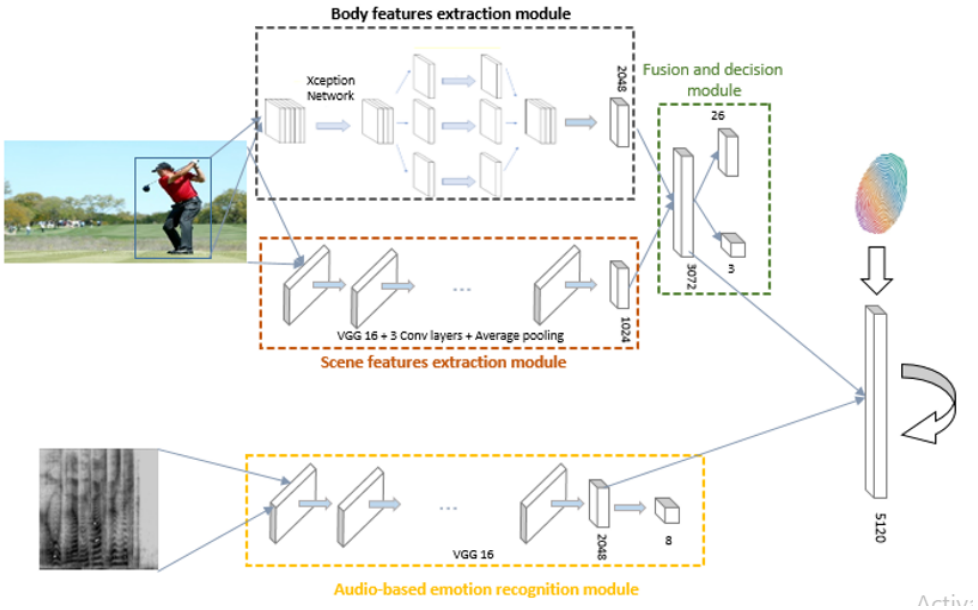

Figure 7.1: EmoRuption architecture

Now that we can put a fingerprint on an emotional state, without knowing the very nature of emotions, we can calculate the difference between two emotional states. It comes down to calculate the distance between two vectors.

Fig.7.1 shows our architecture to detect Emotional Breakdowns (EB). As firstly introduced in chapter.2 an EB occurs when the difference between two emotional state is superior to a threshold.

As Algo.1 shows, the facial features extraction is computed with the image-based emotion recognition module in addition of an open-source facial recognition algorithm to fingerprint the facial expressions as a vector of features at time t. We denote A_f the facial features vector.

The audio features extraction is computed with the audio-based emotion recognition module to fingerprint the audio expressions as a vector

Algorithm 1 EmoRuption pseudo code

Require: $facialModel$ & $voiceModel$
  $facialModel \leftarrow PruneOutput(facialModel)$
  $voiceModel \leftarrow PruneOutput(voiceModel)$
  $i \leftarrow 0$
  loop
    if $i\%10 = 0$ **then**
      $voice \leftarrow MicrophoneCapture()$
      $spectrogram \leftarrow PreprocessingVoice(voice)$
      $A_p \leftarrow feed(voiceModel, spectrogram)$
      $i \leftarrow 0$
    end if
    $i \leftarrow i + 1$
    $image \leftarrow CameraCapture()$
    $faceImage, sceneImage \leftarrow PreprocessingImage(image)$
    $A_f \leftarrow feed(facialModel, [faceImage, sceneImage])$
    $A_f \leftarrow 2.A_f$
    $V_t \leftarrow Concatenate(A_f, A_p)$
    if $V_{t-1} \neq \emptyset$ **then**
      $d \leftarrow Distance(V_t, V_{t-1})$
      if $d > \alpha$ **then**
        print "EMOTIONAL BD!"
      end if
    end if
    $V_t \leftarrow V_{t-1}$
  end loop

of features at time t. We denote A_p the facial features vector.

In the fusion module, the two vectors A_f and A_p are concatenated. The resulting vector V_t describes the emotional state at time t. To detect an EB, we compute the distance between V_t and V_{t-1} as follows:

$$EB(t) = \begin{cases} \text{True if} & S_f\left(V_t, V_{t-1}\right) > \alpha \\ \text{False} & \text{otherwise} \end{cases} \tag{7.1}$$

where α is an empirical threshold, and

$$S_f = d\left(V_t, V_{t-1}\right) \tag{7.2}$$

where, for instance, the similarity function is represented by the Euclidean distance and defined as follows:

$$d(p,q) = \frac{1}{N}\sqrt{\sum_{i}^{N}\left(p_i - q_i\right)^2} \tag{7.3}$$

where p and q are two vectors of length N.

7.2 Synchronization and modalities weights

As our audio-based emotion recognition model has been trained to extract features from segments with a duration of 325 ms and our image-based emotion recognition model uses image by image recognition we need to synchronize the two models so they can work together. Let fr be the camera frame rate and approximately 3 audio frames represent 1s, i.e. $1000/325 \approx 3$. Thus, for each audio frame $fr/3$ frames need to be propagated into the network. As previously mentioned, Mehrabian [72] stated that the facial expressions contribute 55% of the overall emotional state information while the vocal and the semantic parts contribute 38 and 7%, respectively. In order to respect those proportions, weights have to be applied to both vectors A_f and A_p, and since verbal semantics is not treated by the system, A_f and A_p are weighted by, respectively, $\frac{2}{3}$ and $\frac{1}{3}$.

7.3 Results and discussion

In the following the camera frame rate acquisition was set to 30 frames per second, so for 10 consecutive pair of images, the same audio image was propagated to ensure the synchronization.

Few tests have been conducted, the results over the tests are similar in overall. Fig.7.2 represents a signal where each point represents the difference between two consecutive emotional states (30 frames per

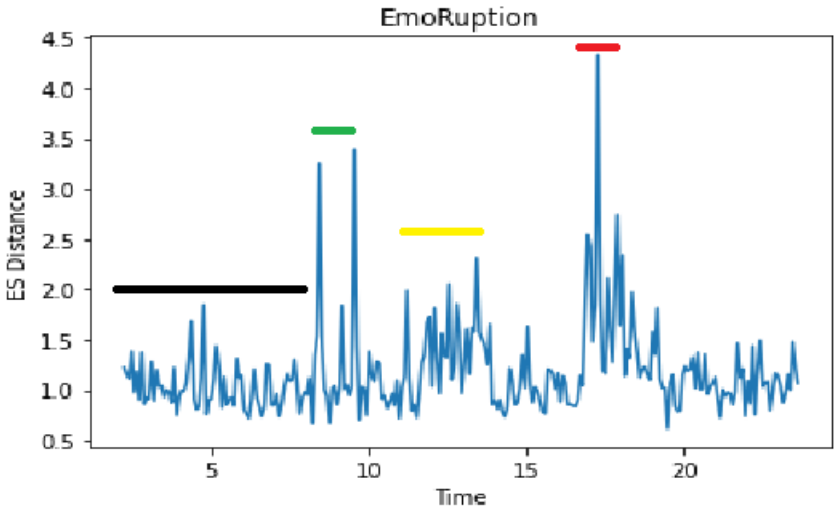

Figure 7.2: EmoRuption output signal representing the distance between two consecutive emotional states.

second). The black segment represents the period when nothing was happening, clearly those oscillations are due to camera and microphone noise, i.e. light variations, body and facial movements and noisy sounds. The green segments represents two emotional breakdowns when smiling, because the network has been trained to detect happiness, which is mostly characterized by a smile.

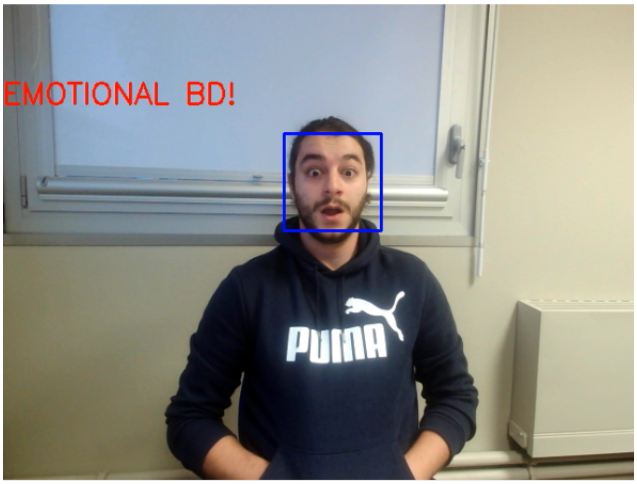

Figure 7.3: An emotional breakdown occurred when feeling chocked

We can distinguish two peaks: the first one is when the smile occurred, then when kept smiling the distance decreased (the same emotional state) and then a second peak when going from smiling to neutral.

The yellow segments also represents a period when nothing was happening and the small variations are due to noise. The red segment represents an emotional breakdown, it occurred when the human operator felt surprised/chocked as shown in Fig.7.3.

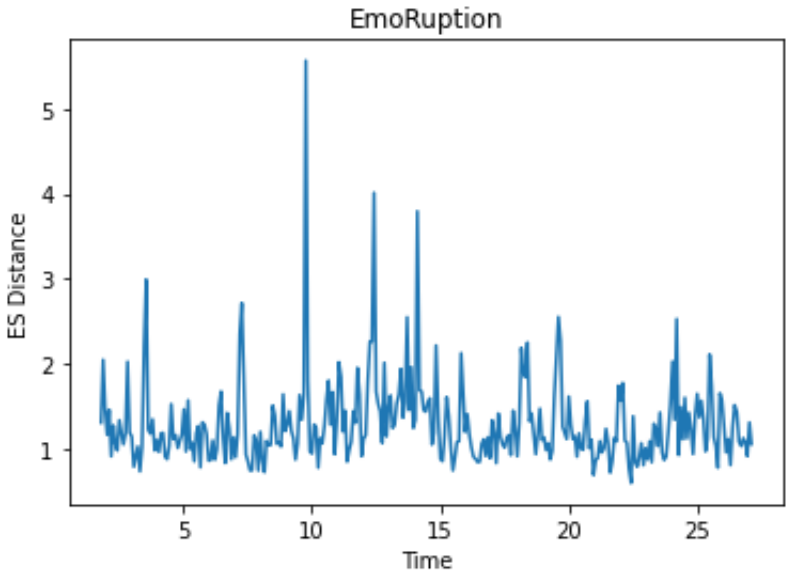

Figure 7.4: Some other examples of peaks representing emotional breakdowns

After few tests and iterations, and as it can be seen in Fig.7.2 and Fig.7.4 a threshold $\alpha = 2.8$ is likely a reasonable value to detect emotional breakdowns.

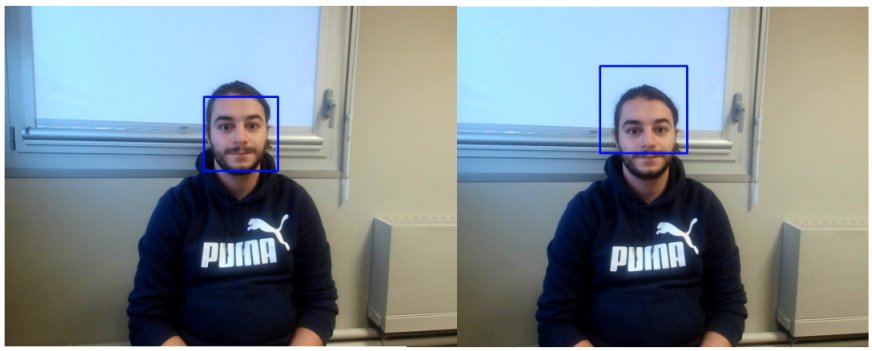

Figure 7.5: Facial detection missing the face

As detailed in Chapter 5, the facial feature extraction module works with pairs of images: One for the face/body and an other one for the context. However, face detection algorithms are sometime not stable and

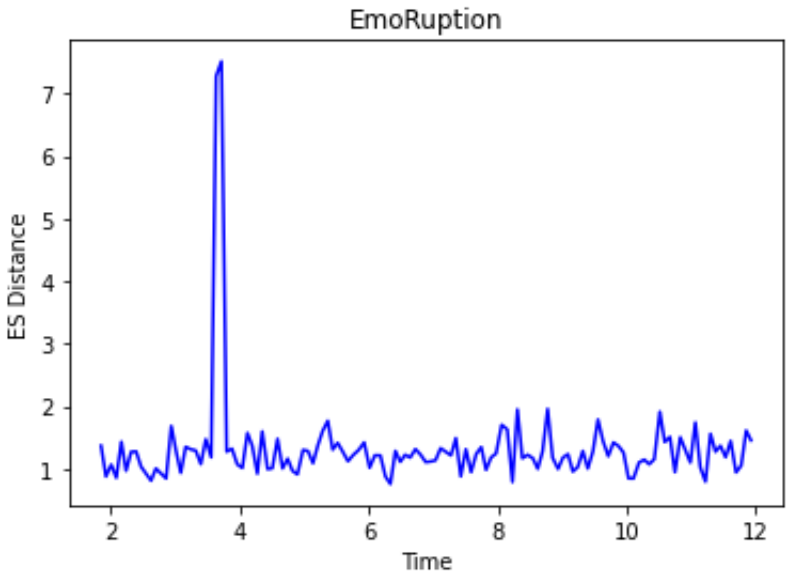

Figure 7.6: False positive emotional breakdown caused by the instability of face detection

can generate a jump (false positive EB) in the output signal as shown in Fig.7.5 and Fig.7.6. An upper body detection has been tested but it did not fix the issue, some instabilities still occurred. Also, in overall using face detection gave better results than upper body detection even if the facial features extraction module has been trained on full bodies.

7.4 Conclusion

This chapter presented EmoRuption. An architecture for emotional breakdowns detection with the image-based emotion recognition model presented in Chapter 5 and the audio-based emotion recognition model presented in Chapter 6. We introduced the principle of emotional state fingerprinting and how to calculate the distance between two emotional states. Finally, we presented our results and discussed them. It is important to notice that EmoRuption is a modular architecture and does not depend on its components. Whether it is the similarity function, the audio module or the image module, they can all be independently changed.

Chapter 8

Conclusion and perspectives

Contents

8.1 General challenges

The rapid development of automatic systems during the last years has not taken into account the way human-machine interaction should evolve. Since the last two decade human-machine interaction became trapped by the supremacy humans have on machines. In the current definition, humans have to give explicit tasks to machine to execute them while machines are becoming more and more autonomous and intelligent. An imperative change must occur where we need to redefine the human-machine interaction, machines have to be aware of their environment and must understand humans needs without those lasts explicitly ordering it and moreover machines have to collect non-explicit information about their environment. Automatic emotion recognition seems to be the solution to this issue. By continuously analysing human's emotions, machine would be able to receive information from humans in order to improve their interaction. This thesis aim to develop a system able to analyze human emotion from audiovisual signals.

We firstly resumed a study based on the Reverse Comic Strip that showed the relationship between human expressed emotional state and audiovisual signals. We concluded that audiovisual signals, i.e. facial expressions and voice, are relevant to identify human emotional discomfort when they are facing undesirable situations. Then, we exposed the theoretical architecture we designed motivated by the idea that the very nature of emotions might be not very relevant if we want to detect briefs but intense reactions. Rather we headed to the idea that the brief changes in the emotional state are more likely relevant to describe such a feeling. Our proposed architecture aims to capture and fingerprint the emotional state at time t and compare it to the previous one. The resulted value can be considered as the difference between two emotional states.

The two next chapters, 3 and 4, were dedicated to the state of the art. On the first one, we firstly introduces some basics in computer vision and audio signal processing, then, we presented how emotions are conceptualized in human science, i.e. psychology and cognitive science, and computer science with a discussion about the advantages and disadvantages of the categorical and continuous representations. After that, we presented the mainly used, visual, audio and audiovisual datasets in emotion recognition. We concluded the chapter 4 by a short review on the main techniques and features for image-based, audio-based and audiovisual-based emotion recognition. On the second chapter, we gave an overview about artificial intelligence and more precisely machine learning. Starting from a single perceptron to complex deep learning architectures, we described the functioning of some of artificial networks components and introduced some recent architectures. After that, we highlighted the role that machine learning and deep learning played in improving emotion recognition in overall.

8.2 Main contributions

The fifth chapter is allocated to our contributions in context-based multi-task emotion recognition where we present our motivations and underline the importance the context has in emotion recognition. We firstly introduce the EMOTIC Database which represents labeled persons in the wild. Emotions are labeled in their both categorical and continuous representations and the context in which the action takes place is available. However, the categorical labels are highly unbalanced which make the dataset very challenging. In order to solve this issue, we developed a new loss function we called the Multi-Label Focal loss (MFL) based on the Binary Focal Loss which gave better results than the commonly used categorical loss function. We designed a new architecture composed of an Xception network for the body feature extraction and a modified VGG-16 for the scene feature extraction (the context in which the action takes place). We compared our results with the state of art and our solution gave the best results on the less frequent labels, i.e less that 5%, (the hardest classes to learn).

In the Chapter 6, we presented our solution for audio-based emotion recognition on the RAVDESS database. Firstly, we presented the RAVDESS database and how it is structured. Then, we detailed our data preprocessing and our training setup and lastly, we presented and discussed our results. After that, in Chapter 7 we moved to the realisation of our system EmoRuption, where we used our model for context-based multi-task emotion recognition as the image feature extractor and our model for audio-based emotion recognition. We pruned the decision layers of both models and kept the last layers before considering them as the feature vectors of both modalities. We then concatenated the two vectors which resulted in a fingerprint of the emotional state. Following our first idea, we compare the actual emotional state with the previous emotional state by computing the difference between the two vectors, the resulting value indicate how fare the two emotional states are. We proposed then to compare this value to a threshold which must be empirically chosen and if the value exceeds the threshold then an emotional breakdown has occurred. We finally end this chapter by a short discussion about such an architecture and the advantages that it may presents.

8.3 Perspectives

The architecture presented is this thesis aims to simplify the data representation, starting from images and voice to end up with a scalar. Nevertheless, instead of considering the distance between two emotional states as a single scalar, we can consider it as an input signal for another architecture which aims to filter, smooth and process this signal to extract relevant features about the emotional state derivations. More-

over, instead of computing the mean of values differences between two emotional state, computing the vector representing the gradient of the emotional state over time might be interesting. It will be more than piratical to figure out which features fluctuate the most and which does not.

In order to reduce noise, computing the difference between the actual emotional state with a sort of mean of the previous emotional states over a limited time window can be interesting. Also, the implementation of a speech recognizer in mandatory so we do not compute noisy audio input such as ambient noise when the operator is not speaking.

Lastly, an end-to-end fine tuning on a home made dataset which contains sequences of frames representing the presence of emotional breakdown would get rid of the empirical threshold and by design improve the detection of emotional breakdowns. Such a dataset will also be helpful to make a feature selection and remove the potential useless features.

Bibliography

[1] Felipe Aguirre, Mohamed Sallak, Frederic Vanderhaegen, and Denis Berdjag. An evidential network approach to support uncertain multiviewpoint abductive reasoning. *Information Sciences*, 253:110–125, 2013. 21

[2] Bagus Tris Atmaja and Masato Akagi. Multitask learning and multistage fusion for dimensional audiovisual emotion recognition. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4482–4486. IEEE, 2020. 63

[3] Anja Austermann, Natascha Esau, Lisa Kleinjohann, and Bernd Kleinjohann. Prosody based emotion recognition for mexi. In *2005 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 1138–1144. IEEE, 2005. 47

[4] Anton Batliner, Kerstin Fischer, Richard Huber, Jörg Spilker, and Elmar Nöth. How to find trouble in communication. *Speech communication*, 40(1-2):117–143, 2003. 46

[5] Ilyes Bendjoudi, Denis Hamad, Frédéric Vanderhaegen, and Fadi Dornaika. Audio-visual and heart signals for attention and emotion analysis. In *Proceedings of the 30th European Safety and Reliability Conference and the 15th Probabilistic Safety Assessment and Management Conference*, pages 2795–2801, 2020. 16, 24

[6] Ilyes Bendjoudi, Denis Hamad, Frédéric Vanderhaegen, and Fadi Dornaika. Vers une méthode d'analyse de signaux audio-visuels et cardiaques pour la détection de dissonances éthiques basée sur l'émotion et l'attention. In *Actes de la conférence sur la maîtrise des risques et de la sûreté de fonctionnement, LambdaMu22*, 2020. 16, 24

[7] Ilyes Bendjoudi, Frederic Vanderhaegen, Denis Hamad, and Fadi Dornaika. Multi-label, multi-task cnn approach for context-based emotion recognition. *Information Fusion*, 76:422–428, 2021. 16, 68

[8] Maurizio Bevilacqua and Filippo Emanuele Ciarapica. Human factor risk management in the process industry: A case study. *Reliability Engineering & System Safety*, 169:149–159, 2018. 20

[9] Susanne Burger, Victoria MacLaren, and Hua Yu. The isl meeting corpus: The impact of meeting type on speech style. In *Seventh International Conference on Spoken Language Processing*, 2002. 43

[10] Peter Burkert, Felix Trier, Muhammad Zeshan Afzal, Andreas Dengel, and Marcus Liwicki. Dexpression: Deep convolutional neural network for expression recognition. *arXiv preprint arXiv:1509.05371*, 2015. 62

[11] Carlos Busso, Zhigang Deng, Serdar Yildirim, Murtaza Bulut, Chul Min Lee, Abe Kazemzadeh, Sungbok Lee, Ulrich Neumann, and Shrikanth Narayanan. Analysis of emotion recognition using facial expressions, speech and multimodal information. In *Proceedings of the 6th international conference on Multimodal interfaces*, pages 205–211, 2004. 47

[12] Pietro Carlo Cacciabue. Human error risk management for engineering systems: a methodology for design, safety assessment, accident investigation and training. *Reliability Engineering & System Safety*, 83(2):229–240, 2004. 20

[13] Jie Cai, Zibo Meng, Ahmed Shehab Khan, Zhiyuan Li, James O'Reilly, and Yan Tong. Island loss for learning discriminative features in facial expression recognition. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 302–309. IEEE, 2018. 62

[14] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1251–1258, 2017. 16, 62, 76

[15] Yucel Cimtay, Erhan Ekmekcioglu, and Seyma Caglar-Ozhan. Cross-subject multimodal emotion recognition based on hybrid fusion. *IEEE Access*, 8:168865–168878, 2020. 63

[16] George E Dahl, Tara N Sainath, and Geoffrey E Hinton. Improving deep neural networks for lvcsr using rectified linear units and dropout. In *2013 IEEE international conference on acoustics, speech and signal processing*, pages 8609–8613. IEEE, 2013. 59

[17] Joost CF de Winter, Yke Bauke Eisma, CDD Cabrall, PA Hancock, and Neville A Stanton. Situation awareness based on eye movements in relation to the task environment. *Cognition, Technology & Work*, 21(1):99–111, 2019. 21

[18] Sidney WA Dekker. The danger of losing situation awareness. *Cognition, Technology & Work*, 17(2):159–161, 2015. 20

[19] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 8, 54, 69

[20] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 76

[21] S Dhall and P Sethi. Geometric and appearance feature analysis for facial expression recognition. *International Journal of Advanced Engineering Technology*, 7(111):01–11, 2014. 7, 45, 46

[22] Paul Ekman. Facial action coding system. 1977. 40, 41, 44, 66

[23] Paul Ekman and Dacher Keltner. Universal facial expressions of emotion. *Segerstrale U, P. Molnar P, eds. Nonverbal communication: Where nature meets culture*, 27:46, 1997. 40

[24] Paul Ekman and Erika L Rosenberg. *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*. Oxford University Press, USA, 1997. 44

[25] Inger S Engberg, Anya Varnich Hansen, Ove Andersen, and Paul Dalsgaard. Design, recording and verification of a danish emotional speech database. In *Fifth European conference on speech communication and technology*, 1997. 43

[26] Simon Enjalbert and Frédéric Vanderhaegen. A hybrid reinforced learning system to estimate resilience indicators. *Engineering Applications of Artificial Intelligence*, 64:295–301, 2017. 21

[27] Beat Fasel, Florent Monay, and Daniel Gatica-Perez. Latent semantic analysis of facial action codes for automatic facial expression recognition. In *Proceedings of the 6th ACM SIGMM international workshop on Multimedia information retrieval*, pages 181–188, 2004. 67

[28] Haytham M Fayek, Margaret Lech, and Lawrence Cavedon. Evaluating deep learning architectures for speech emotion recognition. *Neural Networks*, 92:60–68, 2017. 63

[29] L Festinger. A theory of cognitive dissonance: Stanford univ pr. *Fornell, C., & Larcker, DF (1981). Evaluating structural equation models with*, 1957. 20

[30] Nickolaos Fragopanagos and John G Taylor. Emotion recognition in human–computer interaction. *Neural Networks*, 18(4):389–405, 2005. 47

[31] Deepak Ghimire and Joonwhoan Lee. Geometric feature-based facial expression recognition in image sequences using multi-class adaboost and support vector machines. *Sensors*, 13(6):7714–7734, 2013. 45

[32] Mahesh M Goyani and Narendra Patel. Recognition of facial expressions using local mean binary pattern. *ELCVIA: electronic letters on computer vision and image analysis*, 16(1):54–67, 2017. 46

[33] Martin Graciarena, Elizabeth Shriberg, Andreas Stolcke, Frank Enos, Julia Hirschberg, and Sachin Kajarekar. Combining prosodic lexical and cepstral systems for deceptive speech detection. In *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, volume 1, pages I–I. IEEE, 2006. 46

[34] Fidel A Guerrero-Pena, Pedro D Marrero Fernandez, Tsang Ing Ren, Mary Yui, Ellen Rothenberg, and Alexandre Cunha. Multiclass weighted loss for instance segmentation of cluttered cells. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 2451–2455. IEEE, 2018. 62

[35] Hatice Gunes and Massimo Piccardi. A bimodal face and body gesture database for automatic analysis of human nonverbal affective behavior. In *18th International conference on pattern recognition (ICPR'06)*, volume 1, pages 1148–1153. IEEE, 2006. 43

[36] Salam Hamieh, Vincent Heiries, Hussein Al Osman, and Christelle Godin. Multi-modal fusion for continuous emotion recognition by using auto-encoders. In *Proceedings of the 2nd on Multimodal Sentiment Analysis Challenge*, pages 21–27. 2021. 63

[37] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 60

[38] EM Hickling and JE Bowie. Applicability of human reliability assessment methods to human–computer interfaces. *Cognition, technology & work*, 15(1):19–27, 2013. 20

[39] Julia Bell Hirschberg, Stefan Benus, Jason M Brenier, Frank Enos, Sarah Friedman, Sarah Gilman, Cynthia Girand, Martin Graciarena, Andreas Kathol, Laura Michaelis, et al. Distinguishing deceptive from non-deceptive speech. 2005. 46

[40] Stefan Hoch, Frank Althoff, Gregor McGlaun, and Gerhard Rigoll. Bimodal fusion of emotional data in an automotive environment. In *Proceedings.(ICASSP'05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, volume 2, pages ii–1085. IEEE, 2005. 47

[41] Sepp Hochreiter, Yoshua Bengio, Paolo Frasconi, Jürgen Schmidhuber, et al. Gradient flow in recurrent nets: the difficulty of learning long-term dependencies, 2001. 61

[42] M Shamim Hossain and Ghulam Muhammad. Emotion recognition using deep learning approach from audio–visual emotional big data. *Information Fusion*, 49:69–78, 2019. 63

[43] Yuming Hua, Junhai Guo, and Hua Zhao. Deep belief networks and deep learning. In *Proceedings of 2015 International Conference on Intelligent Computing and Internet of Things*, pages 1–4. IEEE, 2015. 55

[44] Jian Huang, Jianhua Tao, Bin Liu, Zheng Lian, and Mingyue Niu. Multimodal transformer fusion for continuous emotion recognition. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3507–3511. IEEE, 2020. 63

[45] Qiang Ji, Peilin Lan, and Carl Looney. A probabilistic framework for modeling and real-time monitoring human fatigue. *IEEE Transactions on systems, man, and cybernetics-Part A: Systems and humans*, 36(5):862–875, 2006. 67

[46] David Jouglet, Sylvain Piechowiak, and Frederic Vanderhaegen. A shared workspace to support man–machine reasoning: application to cooperative distant diagnosis. *Cognition, Technology & Work*, 5(2):127–139, 2003. 21

[47] Takeo Kanade, Jeffrey F Cohn, and Yingli Tian. Comprehensive database for facial expression analysis. In *Proceedings Fourth IEEE International Conference on Automatic Face and Gesture Recognition (Cat. No. PR00580)*, pages 46–53. IEEE, 2000. 43

[48] Georges Yves Kervern. *Eléments fondamentaux des cindyniques*. Economica, 1995. 20

[49] Barry Kirwan. Validation of human reliability assessment techniques: part 1—validation issues. *Safety Science*, 27(1):25–41, 1997. 20

[50] Peter Knees and Markus Schedl. *Music similarity and retrieval: an introduction to audio-and web-based strategies*, volume 9. Springer, 2016. 7, 33, 34, 35, 36

[51] Ronak Kosti, Jose Alvarez, Adria Recasens, and Agata Lapedriza. Context based emotion recognition using emotic dataset. *IEEE transactions on pattern analysis and machine intelligence*, 2019. 20, 43

[52] Ronak Kosti, Jose M Alvarez, Adria Recasens, and Agata Lapedriza. Context based emotion recognition using emotic dataset. *arXiv preprint arXiv:2003.13401*, 2020. 8, 11, 16, 62, 67, 71, 72, 74, 78, 84, 85, 86

[53] Irene Kotsia and Ioannis Pitas. Facial expression recognition in image sequences using geometric deformation features and support vector machines. *IEEE transactions on image processing*, 16(1):172–187, 2006. 45

[54] R Kubota, K Kiyokawa, M Arazoe, H Ito, Y Iijima, H Matsushima, and H Shimokawa. Analysis of organisation-committed human error by extended cream. *Cognition, Technology & Work*, 3(2):67–81, 2001. 20

[55] S Kumari, U Kowsalya, R Preethi, R Theepa, J Edward Paulraj, and S JeyaAnusuya. Audio-visual emotion recognition using 3dcnn and dbn techniques. 2018. 63

[56] Oh-Wook Kwon, Kwokleung Chan, Jiucang Hao, and Te-Won Lee. Emotion recognition by speech signals. In *Eighth European conference on speech communication and technology*, 2003. 47

[57] Soonil Kwon et al. A cnn-assisted enhanced audio signal processing for speech emotion recognition. *Sensors*, 20(1):183, 2020. 63

[58] Peter Ladefoged. *Elements of acoustic phonetics*. University of Chicago Press, 1996. 7, 38, 39

[59] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015. 62

[60] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. 8, 56

[61] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. 57

[62] Chul Min Lee and Shrikanth S Narayanan. Toward detecting emotions in spoken dialogs. *IEEE transactions on speech and audio processing*, 13(2):293–303, 2005. 46

[63] Shan Li, Weihong Deng, and JunPing Du. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2852–2861, 2017. 62

[64] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 62, 68, 77

[65] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 70

[66] Jackson Liscombe, Julia Bell Hirschberg, and Jennifer J Venditti. Detecting certainness in spoken tutorial dialogues. 2005. 46

[67] Diane Litman and Kate Forbes-Riley. Predicting student emotions in computer-human tutoring dialogues. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pages 351–358, 2004. 46

[68] Steven R Livingstone and Frank A Russo. The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. *PloS one*, 13(5):e0196391, 2018. 16, 43, 90

[69] Michael Lyons, Shigeru Akamatsu, Miyuki Kamachi, and Jiro Gyoba. Coding facial expressions with gabor wavelets. In *Proceedings Third IEEE international conference on automatic face and gesture recognition*, pages 200–205. IEEE, 1998. 46

[70] Ziyu Ma, Fuyan Ma, Bin Sun, and Shutao Li. Hybrid mutimodal fusion for dimensional emotion recognition. In *Proceedings of the 2nd on Multimodal Sentiment Analysis Challenge*, pages 29–36. 2021. 63

[71] Gary McKeown, Michel Valstar, Roddy Cowie, Maja Pantic, and Marc Schroder. The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *IEEE transactions on affective computing*, 3(1):5–17, 2011. 43

[72] Albert Mehrabian. Communication without words. *Communication theory*, pages 193–200, 2008. 29, 100

[73] Zibo Meng, Ping Liu, Jie Cai, Shizhong Han, and Yan Tong. Identity-aware convolutional neural network for facial expression

recognition. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, pages 558–565. IEEE, 2017. 62

[74] Thomas M. Mitchell. *Machine Learning*. McGraw-Hill, Inc., New York, NY, USA, 1 edition, 1997. 8, 50, 52, 53

[75] Stephen Moore and Richard Bowden. Local binary patterns for multi-view facial expression recognition. *Computer vision and image understanding*, 115(4):541–558, 2011. 46

[76] Yafeng Niu, Dongsheng Zou, Yadong Niu, Zhongshi He, and Hua Tan. A breakthrough in speech emotion recognition using deep retinal convolution neural networks. *arXiv preprint arXiv:1707.09917*, 2017. 63

[77] Kiswendsida Abel Ouedraogo, Simon Enjalbert, and FréDéRic Vanderhaegen. How to learn from the resilience of human–machine systems? *Engineering Applications of Artificial Intelligence*, 26(1):24–34, 2013. 21

[78] Pritam Pal, Ananth N Iyer, and Robert E Yantorno. Emotion detection from infant facial expressions and cries. In *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, volume 2, pages II–II. IEEE, 2006. 47

[79] Maja Pantic, Alex Pentland, Anton Nijholt, and Thomas S Huang. Human computing and machine understanding of human behavior: A survey. In *Artifical Intelligence for Human Computing*, pages 47–71. Springer, 2007. 66

[80] Maja Pantic, Michel Valstar, Ron Rademaker, and Ludo Maat. Web-based database for facial expression analysis. In *2005 IEEE international conference on multimedia and Expo*, pages 5–pp. IEEE, 2005. 43

[81] Alex Pentland. Socially aware, computation and communication. *Computer*, 38(3):33–40, 2005. 66

[82] Philippe Polet, Frédéric Vanderhaegen, and Stéphane Zieba. Iterative learning control based tools to learn from human error. *Engineering Applications of Artificial Intelligence*, 25(7):1515–1522, 2012. 20, 21

[83] Emily Mower Provost. Identifying salient sub-utterance emotion dynamics using flexible units and estimates of affective flow. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 3682–3686. IEEE, 2013. 93

[84] Siqi Qiu, N Rachedi, Mohamed Sallak, and Frédéric Vanderhaegen. A quantitative model for the risk evaluation of driver-adas systems under uncertainty. *Reliability Engineering & System Safety*, 167:184–191, 2017. 20

[85] Subeer Rangra, Mohamed Sallak, Walter Schön, and Frédéric Vanderhaegen. A graphical model based on performance shaping factors for assessing human reliability. *IEEE Transactions on Reliability*, 66(4):1120–1143, 2017. 20

[86] James Reason. *Human error*. Cambridge university press, 1990. 20

[87] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 62

[88] Jonathan L Rosch and Jennifer J Vogel-Walcutt. A review of eye-tracking applications as tools for training. *Cognition, technology & work*, 15(3):313–327, 2013. 21

[89] James A Russell, Jo-Anne Bachorowski, and José-Miguel Fernández-Dols. Facial and vocal expressions of emotion. *Annual review of psychology*, 54(1):329–349, 2003. 67

[90] Shabnam Samima, Monalisa Sarma, Debasis Samanta, and Girijesh Prasad. Estimation and quantification of vigilance using erps and eye blink rate with a fuzzy model-based approach. *Cognition, Technology & Work*, 21(3):517–533, 2019. 21

[91] Penelope Sanderson, Jennifer Crawford, Annyck Savill, Marcus Watson, and W John Russell. Visual and auditory attention in patient monitoring: a formative analysis. *Cognition, Technology & Work*, 6(3):172–185, 2004. 21

[92] Bjöern Schuller, Ronald Müeller, Benedikt Höernler, Anja Höethker, Hitoshi Konosu, and Gerhard Rigoll. Audiovisual recognition of spontaneous interest within conversations. In *Proceedings of the 9th international conference on Multimodal interfaces*, pages 30–37, 2007. 47

[93] Nicu Sebe, Ira Cohen, Theo Gevers, and Thomas S Huang. Emotion recognition based on joint visual and audio cues. In *18th international conference on pattern recognition (ICPR'06)*, volume 1, pages 1136–1139. IEEE, 2006. 47

[94] Caifeng Shan, Shaogang Gong, and Peter W McOwan. Robust facial expression recognition using local binary patterns. In *IEEE International Conference on Image Processing 2005*, volume 2, pages II–370. IEEE, 2005. 46

[95] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 8, 60, 76

[96] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 16

[97] Mingli Song, Jiajun Bu, Chun Chen, and Nan Li. Audio-visual based emotion recognition-a new approach. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, volume 2, pages II–II. IEEE, 2004. 47

[98] Stefan Steidl, Michael Levit, Anton Batliner, Elmar Noth, and Heinrich Niemann. " of all things the measure is man" automatic classification of emotions and inter-labeler consistency [speech-based emotion recognition]. In *Proceedings.(ICASSP'05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, volume 1, pages I–317. IEEE, 2005. 46

[99] Stanley S Stevens and John Volkmann. The relation of pitch to frequency: A revised scale. *The American Journal of Psychology*, 53(3):329–353, 1940. 91

[100] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015. 61

[101] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016. 8, 61

[102] Usman Tariq, Kai-Hsiang Lin, Zhen Li, Xi Zhou, Zhaowen Wang, Vuong Le, Thomas S Huang, Xutao Lv, and Tony X Han. Emotion recognition from an ensemble of features. In *Face and Gesture 2011*, pages 872–877. IEEE, 2011. 46

[103] Usman Tariq, Kai-Hsiang Lin, Zhen Li, Xi Zhou, Zhaowen Wang, Vuong Le, Thomas S Huang, Xutao Lv, and Tony X Han. Recognizing emotions from an ensemble of features. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 42(4):1017–1026, 2012. 46

[104] Antoine Toisoul, Jean Kossaifi, Adrian Bulat, Georgios Tzimiropoulos, and Maja Pantic. Estimation of continuous valence and arousal levels from faces in naturalistic conditions. *Nature Machine Intelligence*, 3(1):42–50, 2021. 7, 41

[105] Silvan Tomkins. *Affect imagery consciousness: Volume I: The positive affects*. Springer publishing company, 1962. 40

[106] Frédéric Vanderhaegen. Toward a model of unreliability to study error prevention supports. *Interacting With Computers*, 11(5):575–595, 1999. 21

[107] Frédéric Vanderhaegen. A non-probabilistic prospective and retrospective human reliability analysis method—application to railway system. *Reliability Engineering & System Safety*, 71(1):1–13, 2001. 20

[108] Frédéric Vanderhaegen. Human-error-based design of barriers and analysis of their uses. *Cognition, Technology & Work*, 12(2):133–142, 2010. 21

[109] Frédéric Vanderhaegen. Cooperation and learning to increase the autonomy of adas. *Cognition, Technology & Work*, 14(1):61–69, 2012. 21

[110] Frédéric Vanderhaegen. Toward a reverse comic strip based approach to analyse human knowledge. *IFAC Proceedings Volumes*, 46(15):304–309, 2013. 7, 21, 22, 23

[111] Frédéric Vanderhaegen. Dissonance engineering: a new challenge to analyse risky knowledge when using a system. *International Journal of Computers Communications & Control*, 9(6):776–785, 2014. 20, 21

[112] Frédéric Vanderhaegen. Dissonance engineering for risk analysis: a theoretical framework. *Risk Management in Life-Critical Systems*, pages 157–181, 2015. 21

[113] Frédéric Vanderhaegen. A rule-based support system for dissonance discovery and control applied to car driving. *Expert Systems With Applications*, 65:361–371, 2016. 20, 21

[114] Frédéric Vanderhaegen. Towards increased systems resilience: New challenges based on dissonance control for human reliability in cyber-physical&human systems. *Annual Reviews in Control*, 44:316–322, 2017. 20, 21

[115] Frédéric Vanderhaegen and O Carsten. Can dissonance engineering improve risk analysis of human–machine systems?, 2017. 20

[116] Frédéric Vanderhaegen and Victor Jimenez. The amazing human factors and their dissonances for autonomous cyber-physical & human systems. In *First IEEE conference on industrial cyber-physical systems, Saint-Petersbourg, Russia*, pages 14–18, 2018. 21

[117] Frédéric Vanderhaegen, David Jouglet, and Sylvain Piechowiak. Human-reliability analysis of cooperative redundancy to support diagnosis. *IEEE Transactions on Reliability*, 53(4):458–464, 2004. 20

[118] Frédéric Vanderhaegen, Marion Wolff, and Régis Mollard. Synchronization of stimuli with heart rate: a new challenge to control attentional dissonances. *Automation Challenges of Socio-technical Systems*, pages 1–28, 2019. 20

[119] Frédéric Vanderhaegen and Stéphane Zieba. Reinforced learning systems based on merged and cumulative knowledge to predict human actions. *Information Sciences*, 276:146–159, 2014. 21

[120] Frédéric Vanderhaegen, Stéphane Zieba, Simon Enjalbert, and Philippe Polet. A benefit/cost/deficit (bcd) model for learning from human errors. *Reliability Engineering & System Safety*, 96(7):757–766, 2011. 21

[121] Yongjin Wang and Ling Guan. Recognizing human emotional state from audiovisual signals*. *IEEE Transactions on Multimedia*, 10(5):936–946, 2008. 43

[122] Zhen Wang and Zilu Ying. Facial expression recognition based on local phase quantization and sparse representation. In *2012 8th International Conference on Natural Computation*, pages 222–225. IEEE, 2012. 46

[123] Amanda C de C Williams. Facial expression of pain: an evolutionary account. *Behavioral and brain sciences*, 25(4):439–455, 2002. 44

[124] Matthias Wimmer, Björn Schuller, Dejan Arsic, Bernd Radig, and Gerhard Rigoll. Low-level fusion of audio and video feature for multi-modal emotion recognition. In *Proc. 3rd Int. Conf. on Computer Vision Theory and Applications VISAPP, Funchal, Madeira, Portugal*, pages 145–151, 2008. 63

[125] Martin Wöllmer, Angeliki Metallinou, Florian Eyben, Björn Schuller, and Shrikanth Narayanan. Context-sensitive multimodal emotion recognition from speech and facial expression using bidirectional lstm modeling. In *Proc. INTERSPEECH 2010, Makuhari, Japan*, pages 2362–2365, 2010. 67

[126] Zhihong Zeng, Yuxiao Hu, Glenn I Roisman, Zhen Wen, Yun Fu, and Thomas S Huang. Audio-visual spontaneous emotion recognition. In *Artifical intelligence for human computing*, pages 72–90. Springer, 2007. 47

[127] Zhihong Zeng, Maja Pantic, Glenn I Roisman, and Thomas S Huang. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE transactions on pattern analysis and machine intelligence*, 31(1):39–58, 2008. 67

[128] Zhihong Zeng, Jilin Tu, Ming Liu, Tong Zhang, Nicholas Rizzolo, Zhenqiu Zhang, Thomas S Huang, Dan Roth, and Stephen Levinson. Bimodal hci-related affect recognition. In *Proceedings of the 6th international conference on Multimodal interfaces*, pages 137–143, 2004. 47

[129] Zhihong Zeng, Zhenqiu Zhang, Brian Pianfetti, Jilin Tu, and Thomas S Huang. Audio-visual affect recognition in activation-evaluation space. In *2005 IEEE International Conference on Multimedia and Expo*, pages 4–pp. IEEE, 2005. 47

[130] Bin Zhang, Changqin Quan, and Fuji Ren. Study on cnn in the recognition of emotion in audio and images. In *2016 IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS)*, pages 1–5. IEEE, 2016. 63

[131] Minghui Zhang, Yumeng Liang, and Huadong Ma. Context-aware affective graph reasoning for emotion recognition. In *2019 IEEE International Conference on Multimedia and Expo (ICME)*, pages 151–156. IEEE, 2019. 8, 85, 86

[132] Shiqing Zhang, Shiliang Zhang, Tiejun Huang, Wen Gao, and Qi Tian. Learning affective features with a hybrid deep model for audio–visual emotion recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 28(10):3030–3043, 2017. 63

[133] Tong Zhang, Wenming Zheng, Zhen Cui, Yuan Zong, and Yang Li. Spatial–temporal recurrent neural network for emotion recognition. *IEEE transactions on cybernetics*, 49(3):839–847, 2018. 63

[134] Zhicheng Zhang, Philippe Polet, Frédéric Vanderhaegen, and Patrick Millot. Artificial neural network for violation analysis. *Reliability Engineering & System Safety*, 84(1):3–18, 2004. 21

[135] Ruicong Zhi, Mengyi Liu, and Dezheng Zhang. A comprehensive survey on automatic facial action unit analysis. *The Visual Computer*, 36(5):1067–1093, 2020. 7, 44

[136] Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. Learning deep features for scene recognition using places database. In *Advances in neural information processing systems*, pages 487–495, 2014. 8, 68, 69, 76

[137] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *International Journal of Computer Vision*, 127(3):302–321, 2019. 70