Automatic Emotion Recognition From Multimodal Information Fusion Using Deep Learning Approaches

Rubén Darío Fonnegra Tarazona

Instituto Tecnológico Metropolitano
Faculty of Engineering
Medellín, Colombia
2018

Automatic Emotion Recognition From Multimodal Information Fusion Using Deep Learning Approaches

Rubén Darío Fonnegra Tarazona

Thesis submitted as partial requirement to obtain the degree of:
Magister en Automatización y Control Industrial

Advisor:
Ph.D. Gloria Mercedes Díaz Cabrera
Co-Advisor:
Ph.D. Juan Carlos Caicedo Rueda

Research Line:
Máquinas Inteligentes y Reconocimiento de Patrones (MIRP)
Research group:
Automática, Electrónica y Ciencias Computacionales (AEyCC)

Instituto Tecnológico Metropolitano
Faculty of Engineering
Medellín, Colombia
2018

Cherish your visions and your dreams as they are the children of your soul... The blueprints of your ultimate achievements.

– ***Napoleon Hill***

A mi familia por su invaluable apoyo durante todo este tiempo. A mi madre y a mi padre por estar siempre apoyándome ante toda circunstancia.
Este; más que un logro propio... Es también de ustedes.

Acknowledgements

First, I want to express all my thanks to my thesis advisors, Professors PhD. Gloria Díaz and PhD. Juan Carlos Caicedo. All the contributions of this work have been achieved thanks to their invaluable guidance during my constant training as a researcher; and a great growth and personal enrichment have been possible thanks to their incredible values as a people and human beings. Their passion, assistance, patience, advices and collaboration helped me enormously to successfully complete this work, and surpass all the goals proposed when this adventure started.

I also thank the members of my thesis committee, Professors Carlos Madrigal Gonzalez, Wilson Sarmiento Manrique and Álvaro Orjuela Cañón for their valuable contributions to this work. Their comments and accurate suggestions helped me improve the final content and quality of this thesis.

I want to thank all my colleagues and partners I had during these years, besides all the professors who implicitly had contributions to this work. Additionally, I want to thank the collaboration of professors from the Máquinas Inteligentes y Reconocimiento de Patrones (MIRP) laboratory from the Instituto Tecnológico Metropolitano; and the Computer Science and Systems Engineering Laboratory (U2IS) from ENSTA ParisTech; specially professor Ph.D Adriana Tapus for gently sharing their wisdom and knowledge.

Finally, I want to express all my gratitude to my parents and siblings to constantly support me for all these years of efforts. They have been the first witnesses of all the striving I have made to achieve this goal. They won't probably read this document, but I Wouldn't imagine finishing this journey without their valuable company.

Abstract

During recent years, the advances in computational and information systems have contributed to the growth of research areas, including affective computing, which aims to identify the emotional states of humans to develop different interaction and computational systems. For doing so, emotions have been characterized by specific kind of data including audio, facial expressions, physiological signals, among others. However, the natural response of data to a single emotional event suggests a correlation in different modalities when it achieves a maximum peak of expression. This fact could lead the thinking that the processing of multiple data modalities (multimodal information fusion) could provide more learning patterns to perform emotion recognition. On the other hand, Deep Learning strategies have gained interest in the research community from 2012, since they are adaptive models which have shown promising results in the analysis of many kinds of data, such as images, signals, and other temporal data. This work aims to determine if information fusion using Deep Neural Network architectures improves the recognition of emotions in comparison with the use of unimodal models. Thus, a new information fusion model based on Deep Neural Network architectures is proposed to recognize the emotional states from audio-visual information. The proposal takes advantage of the adaptiveness of the Deep Learning models to extract deep features according to the input data type.

The proposed approach was developed in three stages. In a first stage, characterization and preprocessing algorithms (INTERSPEECH 2010 Paralinguistic challenge features in audio and Viola Jones face detection in video) were used for dimensionality reduction and detection of the main information from raw data. Then, two models based on unimodal analysis were developed for processing audio and video separately. These models were used for developing two information fusion strategies: a decision fusion and a characteristic fusion model, respectively. All models were evaluated using the eNTERFACE database, a well-known public audiovisual emotional dataset, which allows compare results with state of the art methods. Experimental results showed that Deep Learning approaches that fused the audio and visual information outperform the unimodal strategies.

Keywords: Emotion Recognition, Deep Learning, Speech emotion recognition, Facial Emotion recognition.

Table of Contents

1 Introduction

Affective Computing is a growing research area that uses the conscious and unconscious actions to determine the emotional user states and modify machine interaction [1]. The recognition of affective content in human actions by computational systems has been a growing field of study during the recent years, since the analysis of expressions, which influence human decisions and behaviors and could be exploited in a large variety of applications to enhance human-computer interactions (HCI); including mood analysis, video games interaction, entertainment fields, emotional advertising, among others [2, 3, 4].

The study of emotional states has been performed from different perspectives such as the analysis of biological signals, voice, and facial expressions. From each point of view, research has been mainly focused on the analysis of single modalities separately with the aim of efficiently performing emotion recognition. Thus, specific techniques of characterization and learning models have been proposed [5, 6, 7]. Nevertheless, the expression of an emotional event implicitly entails a natural connection between all different physical phenomena, since they occurred in sudden moments [8]. This carries an existing correlation of physical data concerning the description of the expression of an emotion that could be considered to perform recognition. Therefore, it could be assumed that the combination or fusion of two or more of these data modalities (multimodal information) instead of single (unimodal information) could take advantage of the implicit correlation to improve the performance of computational models for solving this task [9]. From this point of view, several challenging problems have arisen which are a matter of study; concerning the advantage of multimodal models over unimodal, since some recent unimodal approaches outperform multimodal models [10, 11, 12].

On the other hand, the development of diverse computational systems has made the processing of large amounts of data easier and faster. Those advances have made the academic and industrial community regain interest in machine learning techniques based on Neural Networks developing a new family of algorithms named Deep Learning models. The main drawback of Deep Learning strategies is the high processing capabilities required to adjust parameters of the model (gradient estimation, weights and bias tuning, non-linear operations, among others). However, implementation on high performance dedicated processors (such as graphical processing units - GPU) have shown a significant reduction in the processing time through algorithm parallelization [13]. These models have also been retaken as

a matter of study since 2012, when it demonstrated promising results for solving a classical computer vision problem, consisting in the classification of massive number of classes (1000) in a large-scale database containing more than 1.2 million images (the ImageNet challenge [14]).

In addition, Deep Learning approaches have performed promising results in different applications using diverse kind of data (such as images, temporal data, signals, among others), since they are adaptive models. This means that is possible to implement different architectures to model spatial or temporal patterns depending on the type of data. Information fusion using Deep Learning architectures has been also proposed recently for addressing the emotion recognition challenge [15, 16]. However, the use of these models for fusing other types of data requires to modify and tunning several parameters of the proposed architectures, including its structure according to the fusion strategy employed. For this reason, it is still matter of study in the scientific community.

In this research, a Deep Learning model for identifying emotions from information extracted from voice signals and image sequences (video) is proposed. Initially, unimodal information strategies were implemented using video and audio information separately, which were extracted from the well known and accepted database (the eNTERFACE'05 database); with the aim of establishing a comparative baseline. It is noteworthy to remark that the database selection criteria is based on the selected emotional model for the work (Ekman model), the relevance of the correct psychological content in the samples and the availability of the database. Then, two models for fusing both information sources are proposed, from characteristic and decision fusion perspectives; They were designed to represent information in a similar dimensional space by using convolutional layers. The proposed fusion models are compared with unimodal strategies, and previous state of the art works, with the purpose of validating the hypothesis described above. Main results suggest that Deep Learning is a promising strategy to achieve optimal multimodal information fusion from different perspectives, and its usage demonstrates that multimodal information obtain the best performance for emotion recognition than unimodal strategies, using audio and video sources.

1.1. Research challenge

Automatic emotion recognition is fundamental to advance in the development of computer interaction models. Several computational methods have been developed by pattern identification in audio and physiological signals, body and facial expressions; however, reported results are far from ideal to real scenarios. Taking advantage of the patterns generated in more than one data modality may be the key to improve the performance of these models. On the other hand, machine learning techniques based on Deep Learning have achieved promising results in several learning tasks since they can tune learning parameters according to

kind of information during the training stage. However, the implementation of these strategies aiming multimodal information fusion (different data sources) is challenging due to the combination of architectures for modeling different kinds of data efficiently, taking into account the natural correlation between data, which describes a single phenomenon. Additionally, it requires processing a large amount of data simultaneously to estimate a decision. From this point of view, the problem in this work is the design of a Deep Learning architecture for simultaneously processing different data sources to perform automatic emotion recognition.

1.2. Objectives

1.2.1. Main Objective

Proposing an information fusion model based on combination of multiple Deep Learning architectures, which allows to analyze different data modalities to perform automatic emotion recognition.

1.2.2. Specific Objectives

1. To establish a baseline of state-of-the art techniques for emotion recognition based on unimodal analysis, using Deep Learning techniques.

2. To propose a feature extraction strategy based on Deep Learning approaches for multimodal data representation in a similar dimensional space.

3. To propose a Deep Learning strategy for emotion recognition from multimodal information, using the representation stage developed in specific objective 2.

4. To evaluate the emotion recognition proposed strategy using public and available emotional databases containing multimodal information.

1.3. Hypothesis

The design of a hybrid Deep Learning architecture that efficiently combines multiple specialized structures of unimodal information, will allow simultaneous processing of several data types for emotion classification or recognition, improving performance compared to unimodal data analysis.

1.4. Contributions

The main results of this research project are supported by the following products, publications, related works, undergraduate co-advisory projects and research stage developed from the project proposal.
The first contribution aimed to evaluate several Deep Learning frameworks to establish the use of the best framework to implement all the models in this work. This is aligned with the first objective of this work, corresponding to the developing baseline.

Publications

- **Performance comparison of deep learning frameworks in image classification problems using convolutional and recurrent networks**
  Rubén D. Fonnegra and Bryan Blair and Gloria M. Díaz
  In: 2017 IEEE Colombian Conference on Communications and Computing (COLCOM)
  IEEE Xplorer, 2017

Undergraduate Co-Advisory projects

- **Performance comparison of deep learning frameworks in image classification problems using convolutional and recurrent networks**
  Institution: Instituto Tecnológico Metropolitano (ITM)
  Status: Concluded
  Student: Bryan Blair Álvarez
  Year: 2017

In the following publications, unimodal emotion recognition using speech and video to perform the emotion recognition task are presented. These contributions were products associated to the experimentation during the development of the first objective of this work; in which the baseline is proposed using a common validation strategy (cross-validation).

Publications

- **Speech Emotion Recognition Based on a Recurrent Neural Network Classification Model**
  Rubén D. Fonnegra and Gloria M. Díaz
  In: Cheok A., Inami M., Romão T. (eds) Advances in Computer Entertainment Technology
  Lecture Notes in Computer Science, Springer, Cham. 2018

- **Deep Learning Based Video Spatio-Temporal Modeling for Emotion Recognition**
  Rubén D. Fonnegra and Gloria M. Díaz
  In: Masaaki Kurosu (ed) Human-Computer Interaction: Theories, Methods and Human Issues (Part I)
  Lecture Notes in Computer Science, Springer, Cham. 2018

- **Speech Emotion Recognition Integrating Paralinguistic Features and Autoencoders in a Deep Learning Model**
  Rubén D. Fonnegra and Gloria M. Díaz
  In: Masaaki Kurosu (ed) Human-Computer Interaction: Theories, Methods and Human Issues (Part I)
  Lecture Notes in Computer Science, Springer, Cham. 2018

In the following related works, the authors have significant participation to the individual contributions concerning diverse areas (such as facial analysis and multispectral imaging) using different Deep Learning approaches. In this sense, different publications and undergraduate projects were developed in these scopes.

Publications

- **MSpecFace: A Dataset for Facial Recognition in the Visible, Ultra Violet and Infrared Spectra**
  Rubén D. Fonnegra and Alexander Molina and Andrés F. Pérez-Zapata and Gloria M. Díaz
  In: Botto-Tobar M., Esparza-Cruz N., León-Acurio J., Crespo-Torres N., Beltrán-Mora M. (eds) Technology Trends.
  Communications in Computer and Information Science, Springer, Cham. 2017.

- **Automatic Face Recognition in Thermal Images Using Deep Convolutional Neural Networks**
  Rubén D. Fonnegra and Andrés F. Cardona-Escobar and Andrés F. Pérez-Zapata and Gloria M. Díaz
  In: XVII Latin American Conference on Automatic Control CLCA 2016.
  Universidad EAFIT. 2016.

Undergraduate Co-Advisory projects

- **Facial recognition with pose variations in multi spectra images using deep convolutional neural networks**
  Institution: Instituto Tecnológico Metropolitano (ITM)
  Status: Concluded
  Student: Pablo Campaz Úsuga
  Year: 2018

Finally, the following internship were realized in the scope of this work, in which an emotion-based interaction experiment were proposed to investigate the effect of reaction of human stimuli in robotic environments. The experiment helped to better understand the power and elicitation of humans emotions during interactions.

International Research Stage

Beneficiary of the Stages program under resolution N^o 000429 of April 27, 2017 under the 161 agreement subscribed between the Instituto Tecnológico Metropolitano (ITM) and Sapiencia, 2016.

Objective: Propose an experiment at the Autonomous Systems and Robotics from the Engineer and Informatics Systems department (U2IS) in École Nationale Supérieure de Techniques Avancées (ENSTA) ParisTech; for characterizing positive and negative emotions through Human - Human and Human - Robot interactions.
Advisor: Prof. Adriana Tapus.
Stage period: From 15th October to 1st December, 2017.

1.5. Thesis outline

This work is organized as follows. In chapter 2 are described basic concepts required to the development of this work. First, the psychological emotional models proposed in the state of the art to characterize the affective content of human behaviors are shown. Then, the physical points which represent the manifestation of an emotion in humans are presented to understand the nature of the data correlation to describe a single emotional phenomenon. Besides, the automatic emotion recognition works aiming to enhance human-machine interactions are presented to describe the concepts of unimodal and multimodal information, the nature of data used, the computational modeling and feature extraction techniques, the machine learning models used, the experimental framework for the experiments and the results from each unimodal and multimodal data models. Finally, the Deep Learning basics and

operations to understand the proposed models in this work are presented. The design of a hybrid Deep Learning architecture that efficiently combines multiple specialized structures of unimodal information, will allow simultaneous processing of several data types for emotion classification or recognition, improving performance compared to unimodal data analysis.

In chapter 3 are described the unimodal classification models using video and audio data separately. First, the general scheme for each strategy is presented which is basically composed of several stages such as data augmentation, preprocessing and emotion classification. Then, each unimodal model and its specificities is described; concerning the data preprocessing, feature extraction or region of interest detection and classification strategy. Additionally, the description of the database, the performance metrics and the experimental framework are introduced.

In chapter 4 are described the multimodal fusion data approaches at different levels. In this case, each model is introduced according to its strategy, concerning the different stages. The strategy at decision fusion level is described, in which the characterization from unimodal strategies is parallel used before including a classification architecture to process the final decision. On the other hand, the strategy at characteristic fusion uses different parallel architectures for each kind of data besides a relational block to combine data coming from other modalities. The details for each strategy, the results and comparison with unimodal models proposed and previous works in the state of the art, using the same experimental framework are presented.

Finally, in chapter 5 are presented the discussion and conclusions of this work; the considerations and future works. This chapter presents the main discussions of this thesis, how the multimodal models could outperform the results obtained in comparison with unimodal strategies and other works in the state of the art; how multimodal strategies model could be extended to process more than two kinds of data, the advantages of the use of this model, the disadvantages, the limitations of the proposed approaches and the accomplishment of the objectives proposed in this work. Besides, the considerations and future works in terms of new experimentation, the emotional modeling, the use of different data modalities and the variations of this model derived from the development of this thesis.

2 Background and Previous Works

2.1. Emotion Modeling

From the study of the human brain as the center of thinking and feeling, the emotions have been defined as the natural body reactions or human behaviors to brain stimuli, which changes in variation and intensity, depending on the external context [17]. This means that emotions are internal mechanisms triggered by external factors which influence physical variables such as movements, expressions, reactions, among others. These agents are all perceptible, since they allow someone to identify when anyone experiences an specific emotion to see it. Research has shown that the natural reactions of humans are generalized among individuals, which suggests that these patterns can be identified in all humans. For example, a person who feels scared because of a natural disaster will react in a similar way to many other subjects who experience the same act.

A deep discussion about these patterns has focused in the characterization of emotions, in which several authors propose discrete models of "universal" emotions [18, 19, 20]. The argument of the existence of universal emotions was conducted from the study of cross-cultural facial expressions [21]. Participants from a remote area in New Guinea, where members could have not learned the meaning of expressions from the exposure of media depictions of emotions, were asked to show how their face look under different expressiveness situations. Their findings were relevant because they showed that observers from other cultures can identify the emotional context and expression they intended to portray.

Then, the structure of the universal emotions consists of the ones that can be visually identified through facial expressions, no matter the culture or other facts. In this sense, emotions such as anger, happiness, sadness, surprise, disgust, and fear were considered universal emotions by Ekman et. al.[22]. However, other authors have considered emotions not only as universal expressions but as a whole language conditioned for the culture and the environment where people are related. With this purpose, several models have been proposed based on human perception and psychological factors. From this point of view, the study of the other emotional states generated a hypothesis concerning the combination of several universal emotions. The concept of "families of emotions" was illustrated in [23], which does not consider the universal emotions as affective states, but as a group of related states with similar expressive characteristics. However, the manifestation of these characteristics is highly

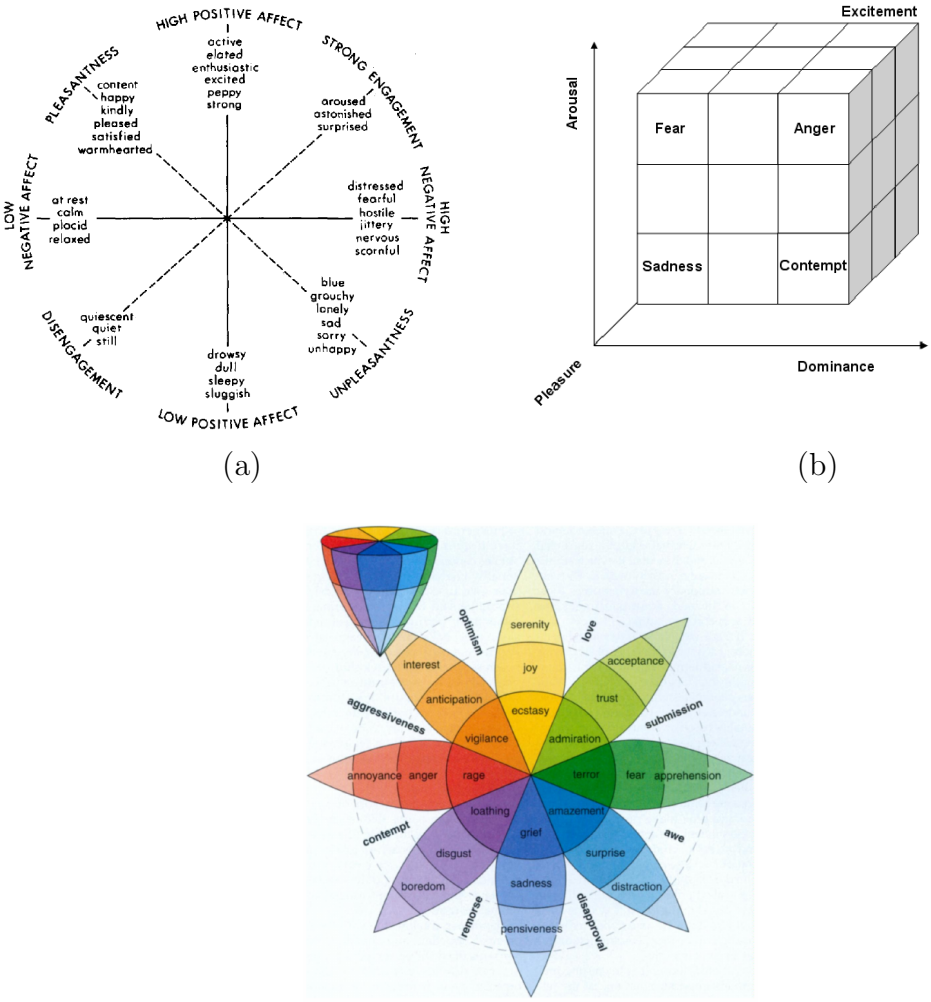
correlated to the cultural bondings of people and the environment.

From the concept of emotion families, this research area was focused on the way to measure and group the expressive content of different emotions, in order to characterize them according to universal affective states. In this sense, emotion characterization models based on diverse variables were proposed. In [24], emotions are defined as communication processes or signals which depend on the individual experience; to finally influence personal relationships. Then, a model based on a circumflex of emotions is proposed, taking into account the assumption of an analogous relationship between emotions and the colors model. To achieve this, the authors mapped the emotions into ten basic groups: love, happiness, mirth, surprise, fear, suffering, anger, determination, disgust, and contempt. Additionally, this model considered an existing overlapping among groups, in which they include other affective states. Finally, to map all the emotions in the wheel (bi-dimensional model) they stacked the states in two axes called unpleasant/pleasant and acceptance/rejection. Additionally, the model could be extended to arrange emotions in concentric circles where inner circles are more basic and outer circles are more complex emotions. Notably, outer circles are also formed by blending the inner circle emotions. In both cases, models are based on circumflex representations, in which emotional words are plotted according to similarity. Figure **2-1(c)** shows a graphical representation of the multidimensional Plutchik model.

On the other hand, a more complex model is shown in [25], called the PAD model which is a multidimensional model based on emotional scales specifically addressed to the connotative meaning of differential emotional-based ratings (evaluation, activity, and potency). The authors proposed a preliminary scheme with nearly orthogonal scales of emotions: pleasure-displeasure as meaning of the evaluation of emotions for the human being (positive/negative emotions); arousal/non-arousal as meaning of intensity of emotions when are expressed (high/low activity); and dominance-submissiveness as meaning of the controlling and dominant nature of the emotion (control/lack of control). A representation of the PAD emotional model is shown in figure **2-1(b)**.

Watson et al.[26, 27] propose to describe emotional content according to the positive affection and negative affection scales (PANAS). The model is a composed of two orthogonal scales to characterize emotions; in which main groups are enthusiasm, active and alertness as positive affect, and aversion, anger, nervousness, contempt, disgust, guilt, and fear as negative affect. The PANAS scales exhibit a significant level of stability through different time intervals; as also represent consistency with a strong dispositional component of affect. This means that even momentary moods are, to a certain extent, reflections of one's general affective group. In figure **2-1(a)** can be seen a representation of the bidimensional Watson model.

From the perspective of the multidimensional emotional models, it can be said that they



(a) (b)

(c)

Figure 2-1: Multidimensional emotional models. Circumflex PANAS [26]; Tridimensional PAD [28] and Multidimensional Plutchik [29]

provide a more complex characterization to describe affective content from human expressions. Nevertheless, from a computational point of view, these models must be handled as multi-label problems, which imply an increase in the complexity in comparison with unidimensional models, such as the universal emotional model. Additionally, these models provide a higher level of interpretability, which is useful in HCI applications. Therefore, this work focuses on the automatic recognition of emotions in a discrete model, specifically the universal model of Ekman.

2.2. Physical Manifestation of Emotions

From the study of [21], the theory of universality of emotions gained force in the academic community to search the physical aspects of the recognition of emotions through facial expressions. However, the expression of emotions is highly correlated to social relationships, cultural environment, psychological activity, stimuli arousal, among others. This does not deny the theory about the universal emotions, but it suggests that expressiveness carry on certain changes according to these aspects.

The expression of an emotion is typically accompanied with speech, body gestures or postural expressions, depending on the social context at moment of the experience. These changes could be considered as messages in which the emotion is expressed. Additionally, these manifestations not only depend on the situation causing the emotion, but their interpretation also depends on the experience of interpreter. Thus, as several traits condition similar aspects for the experimenter-interpreter in the universal emotions, the characterization of those changes (in a cultural group or region) according to an specific emotion model is still a research field of study. Thus, the best way to ensure a correct interpretation is to validate with experts of psychology areas to certify the expression of the emotion in humans when an external factor triggers an emotional event. For this reason, the data used to perform the experiments in this work will not only be based in the use of a database widely explored in the literature, but it will also contain emotional information certified by experts of psychological fields which ensure that the label in a batch of information belongs to its corresponding affective content.

A similar problem comes with the measurement of those patterns and body gestures (such as temperatures, responses, movements, among others), since the natural human variables changes when emotion occurs. In this work two of the widely used data to search manifestation of affective content in human behaviors are used, i.e. the facial expression and the audio speech signals. Besides, it is also noteworthy to remark the high accuracy and synchronization during the acquisition time; since most of the modern camera devices contain microphones and visual sensors to capture information, synthesized in a single package.

2.3. Automatic Emotion Recognition

The development of diverse interaction devices for enhancing the experience between humans and machines has made the recognition of emotional states of humans a growing field of study in academic and industrial research. This field, called Affective Computing covers wide amount of applications such as human-computer interactions [30, 31], the emotional marketing and specialized market programs [32, 33] analysis of behavior disorders [34, 35, 36, 37], among others. The recognition of emotions as a research field has been strongly bounded to

the analysis of generated changes in people's physiology such as facial expressions, pattern body movements, voice tone and pitch alterations, body temperature changes or biological signals modifications [38, 39].

Among them, several techniques in the state of the art suggest analysis of facial expressions for emotion recognition, consisting in the searching of patterns in 2D or 3D images, for static [40] and sequential images (video) [41], and for different spectra (visible and infrared [42]). Additionally, the characterization stages for analyzing the data involve dynamic muscle description of movements in regions of interest such as nose, eyes, eyebrows or lips [43]. Likewise, regarding body movement change analysis, studies demonstrates the use of pattern search for lie detection [44, 45] in which techniques based on inertial sensors analysis or capture movement systems have been used to characterize certain body movements expressions [46, 47]. Besides, previous research for analysis of biological signal for emotion recognition involves electroencephalography (EEG), electromyography (EMG), galvanic skin response (GSR), where used algorithms consist in Fourier and Wavelet analysis, power spectral density, mel-frequency cepstrum, prosodic, mathematical momentums or Hilbert-Huang transform [48, 49].

Among the physical changes, speech, audio signals and facial regions are the most widely explored data types in the state of the art. In case of speech, several strategies had been proposed [50, 51, 52, 53, 54]; e.g. in [51] a transfer learning model is proposed, in which 16 low-level descriptors (LLDs) and 12 functionals audio features are extracted using the OpenEAR toolkit. A transfer learning model, which includes auto-encoders based technique for feature transfer, maps a general structure of input characteristics by moving them from source to target to train a support vector machine (SVM). The main drawback of this model is that it is highly dependent on training reconstruction of data for knowledge transfer. An SVM learning model was also used in [52] to differentiate between the six different emotions included in the eNTERFACE database, but the feature vector was composed of 7 short time-based features and three long-time based features extracted from the speech audio signals using JAudio toolkit. Likewise, in [53] an audio emotion recognition system based on extreme learning machine (ELM) is proposed. Initially, a signal processing stage extracts multi-directional regression features (MDR) by pre-emphasize audio signals and frame them using hamming windows. Then, Fourier transform based spectral analysis and filter is applied using 24 Mel-scale Frequencies, obtaining 24 values per frame. At this point, a four-directional three-point linear regression is carried out to extract 96 features. Finally, other works such as [54] has proposed SER systems based on a recurrent deep learning strategy. However, the feature description model is focused on the analysis of utterances where emotions can reach their highest expression peak, besides suppressing silence in sentences, or non-expressive words. That is to say, that only verbal features are used for characterizing the signals.

Strategies based on the analysis of facial expressions also represents important challenges such as illumination changes, pose variations (in still images [55]) and spatiotemporal modeling (in video or image sequences [56]). A comprehensive review of these approaches can be consulted in [57, 58]. In [59] an emotion recognition framework from video sequences is presented, which is composed of different subsystems: the first subsystem implements a preprocessing stage, in which a face is automatically detected and the image pixels are normalized using a histogram equalization approach. In the second subsystem, a dimensional subspace from preprocessed images to construct a prototypical template for the N emotions is performed by a simple eigen-decomposition of each emotion scatter matrix. In the third subsystem, a matching stage is applied for comparing encoding representations from faces, using a canonical correlation analysis, via singular value decomposition (SVD). Performance obtained using this strategy is highly dependent on the dimensionality of subspaces created for encoding the images, which increases the computational complexity of the algorithm. In [52] the authors propose a model for emotion classification from visual data using an extreme learning machine. In their strategy, they first make a manual annotation of every frame in the video, to subtract those frames with no relevant emotional content (neutral). Then, they perform a facial extraction of relevant points of the face (eyes, mouth, nose, eyebrows and chin) using the Luxand FSDK software to obtain a vector of features per frame. The feature vector for each video is obtained then, using coordinate-wise averaging from feature vectors of individual frames. Finally, they use an extreme learning machine (ELM) to perform the classification task of video clips. However, the main drawback lays on the manual annotation of the frames because it might elicit a significant loss of frames with relevant emotional content to achieve the recognition.

On the other hand, recent works suggest that the combination of different modalities of data could increase the description of an emotional event due to the existing implicit correlation during their acquisition. The combination of different modalities of data is known in the state of the art as multimodal analysis. Nevertheless, the complexity of the multimodal analysis lays in the way the data is processed and characterized; considering that the techniques for describing the different modalities are not the same. To aim this problem, authors suggest to evaluate samples using different models for each kind of data and then, propose a strategy to determine a decision based on the combination of predictions of each model. This strategy for fusion information is called decision level fusion. However, these strategies do not take advantage of the internal correlation of the data. On the other hand, other strategies proposed different characterizations of multiple kinds of data, to use a method to combine extracted features and perform classification. The strategy is called characteristic level fusion, and is the most promising in the literature review. In this sense, the combination of machine learning techniques has been proposed, but especially Deep Learning has gained attention in research and academic communities, given its characteristic of adaptive learning

depending on the data, no matter their provenance which allows integrating a different kind of information for the multimodal analysis.

In [60], a multimodal strategy based on the selection of parts from video and audio in which the emotion expressiveness achieves the highest point. Then, those frames are processed as still images. Authors achieve the task by grouping all the video frames into K clusters based on dissimilarity between the local phase quantization (LPQ) features in each frame. They use a complete agglomerative link clustering algorithm, named dendrogram clustering algorithm (DEND-CLUSTER) to group the frames and then, select a frame whose average distance from the rest of the frames in the cluster is minimum. Thus, they compute the ideal selection measure (ISM) score based on the gradient of each pixel, to arrange the selected frames in descending order. The peak frames features are used for training a SVM with a linear kernel to classify emotions. This approach did not consider dynamic motion change of the face, which it is an expected condition in a real context. Likewise, in [61] a sparse local discriminant canonical correlation analysis for multimodal information fusion was presented. In the case of audio emotion recognition, authors propose to apply a feature extraction stage to convert time domain signals into spectrograms with a 20ms window and 10ms overlap. The spectrograms are processed using the Principal Components Analysis (PCA) method to obtain 60 components, which are then considered as inputs of a sparse auto-encoder (400 units) to create a subspace representation, which is also used to train an SVM model.

Recently, information fusion based on Deep Learning models have shown promising results[15, 16]. Sun et al. [16] propose a model for feature representation and fusion from video data containing face and body gestures. With this aim, several stages were implemented: first, a data preprocessing for extract aligned faces and image normalization; then a Feature extraction and representation stage to extract body pose and faces contours, including a deep network, which combines convolutional and recurrent layers, and a principal component analysis (PCA) to select the most relevant features extracted from the Deep Learning model; finally, a fusion-based classification model, which combines feature-level and decision-level fusion model using the representation found by the PCA analysis and applying a weighted fusion network to select most suitable decision. On the other hand, Yan et al. [15] propose a Hybrid Deep Learning model that combines bidirectional recurrent layers and convolutional layers to characterize faces, facial landmarks and audio. The objective was to generate trajectories to characterize facial movements with a convolutional architecture, parallel to a SVM and a bidirectional recurrent network; and convolutional layers for audio recognition based on the INTERSPEECH paralinguistic features. Finally, the fusion strategy is based on a weighted rule of decision estimated from all architectures. These fusion models use characteristic and decision-level fusion from different perspectives; the use of architectures to effectively represent information before classification stage; the use of feature selection algorithms for selecting best features; and the use of temporal modeling using different layers to model

temporal condition.

Other works using either only audio speech or fusing it with other information provided by visual information of facial, body expressions and other physiological signals, as can be found in some recent literature reviews [62, 63]. However, the problem of multimodal information fusion is still a challenging task in the emotion recognition task for various reasons:

1. The implementation of algorithms that take advantages of internal correlation of data has not achieved enough performance to solve the challenges of emotion recognition problem, such as resolution, dimensionality, balancing, heterogeneity, among others.

2. The featuring extraction algorithms for emotional content in multimodal data is still a challenge since their analysis is highly correlated to the nature of data (such as video, speech, text, signals, among others) and the data processing.

3. Despite the machine learning techniques for multiple data fusion have increased during last five years, the problem is still open from the amount of applications in which emotion recognition is pertinent. In this sense, the adaptive models have achieved the most promising results since they aim to adjust to data variation to outperform results.

The main objective of this work aim to propose a model to effectively fuse multimodal data, using adaptive machine learning algorithms (Deep Learning) to perform the emotion recognition task. The contribution main of this work lies in the architecture of the fusion model and its performance evaluation using public available data in the state of the art works.

2.4. Deep Learning Basics

Deep Learning is a machine learning technique based on computational models that adjust automatically according to input type and amount of data to be processed [64]. Methodologies based on Deep learning obtain inner data representations at different abstraction levels from non linear transformations applied in multiple layers. These computational techniques based on artificial neural networks, implement core functions (activation function) inside different units (neurons) contained in multiple groups (layers) across the entire network. Each layer is composed by a defined number of neurons with equal or not activation functions; and every neuron output is connected to its own input or other neuron's input related to the same or other layer. This condition (called non linear connections between layer) depends on kind of data and activation function of neurons. The data is introduced to the network through a single level or group of neurons (input layer) which is connected to other non linear layers (hidden layers); and they apply required transformations in order to extract relevant information from input data depending on activation functions and connections. Information

from characterization of data is extracted through last group of neurons (output layer). The whole amount of layers (input, hidden and output), connections and neurons are known as network architecture [65, 66].

Because of the model's capacity for extracting deep relations from data sources at diverse abstract levels, it requires large amounts of data to achieve efficient learning during training stage. This characteristic implies a high computational cost [67, 68]. However, Deep Learning architectures have been retaken in in different applications due to increase of available information generated during last 20 years (Big Data) as training examples; and the technological advances of dedicated processing units (GPU) which allows significant reduction in computational cost and memory consumption through parallel programming strategies [69].

State-of-the-art Deep Learning architectures have shown promising results in applications involving abstract relationships in input data; applications with time analysis dependence or with the need of complex feature extraction stages. Among them are audio signal [70], voice signal [71], physiological signal processing [72], medical images processing [73] and video or image sequences processing [74].

Nowadays, there are several Deep Learning architectures specialized according to input data source. The literature describes architectures that use convolution operations (Convolutional Neural Networks - CNN); architectures based on analysis of changing information through time (Recurrent Neural Networks - RNN), which take into account parameters from information within a long time lapse (Long - Short Term Memory Networks - LSTM); architectures with the ability to estimate predictions from similarities in data input sources (Deep Belief Networks - DBN); and architectures with the aim of characterizing data through encoding and decoding of information (Autoencoders - AE, Autodecoders - AD). From this specialized architectures, hybrid models for Deep Learning architectures have been proposed in the literature in order to improve results of machine learning tasks, such as Convolutional Autoencoders (ConvAE), Autoencoders with Memory cells (AE-LSTM), recurrent Autoencoders (RAE), Convolutional Deep Belief Networks (Conv-DBN), Convolutional Recurrent Neural networks (Conv-RNN) among others [75]. According to this, theoretical frameworks for architectures proposed in this work are presented in subsections 2.4.1, 2.4.3, 2.4.4 and 2.4.5.

2.4.1. Convolutional Neural Networks (CNN)

Convolutional neural networks (CNN) are special networks in which one or more layers contain units implementing convolutional operations for transforming the input data. The purpose of convolving the information through the network layers is to analyze the data taking into account certain regions of a signal to create different recognition patterns. Similar

to conventional neural networks, the convolutional layers also have trainable weights and biases. However, their size and dimensionality depend on the type of convolution used for the layer (1D conv, 2D conv or 3D conv). The convolutional architectures in Deep Learning have gained a lot of attention since 2012, when this strategy outperformed classical computer vision algorithms in a very challenging task i.e., classification 1,2 million images-within 1000 categories [14]. The equations 2-1, 2-2 and 2-3 describe the response to the convolution in 1D, 2D and 3D, of a function $f(f[x], f[x,y], f[x,y,z])$ to a filter $h(h[x], h[x,y], h[x,y,z])$ when it is displaced $u([u], [u,v], [u,v,w])$ spaces. A graphical representation of 2D convolution is shown in figure **2-2**.

$$f[x] \times h[x] = \sum_{u} f[u] \times h[x-u] \tag{2-1}$$

$$f[x,y] \times h[x,y] = \sum_{u} \sum_{v} f[u,v] \times h[x-u, y-v] \tag{2-2}$$

$$f[x,y,z] * h[x,y,z] = \sum_{u} \sum_{v} \sum_{w} f[u,v,w] \times h[x-u, y-v, z-w] \tag{2-3}$$

The convolutional layers are commonly implemented along with Pooling layers, with the aim of reducing the size of the characteristic subspace (computational cost) and avoid the overfitting problem in the network layers. The pooling layers contain 2 hyper parameters corresponding to the stride value (S_p) and the pooling size (F_p). The pooling operation commonly uses the MAX operation across the input regions for preserving the most relevant sections of the input. The pooling operation with size x_o, y_o and z_o for an input array of sizes x_i, y_i and z_i in terms of hyper parameters S_p and F_p is shown in 2-4 (a), (b) and (c) respectively

$$(a)x_o = \frac{x_i \times F_p + 1}{S_p} \qquad (b)y_o = \frac{y_i \times F_p + 1}{S_p} \qquad (b)z_o = z_i \tag{2-4}$$

2.4.2. Recurrent Neural Networks

The recurrent neural networks, first proposed by Elman in [76] are structures for facing the problem of sequence modeling across time. In traditional neural networks, the inputs and outputs are completely independent from each unit. However, in several problems (such as natural language processing) the output predictions are highly dependent on the previous states or inputs, in which the estimation of a word would better achieve higher performance considering previous states. So, recurrent neural networks perform same operation over all elements in a sequence considering previous states, which is an approximation of natural behavior of time series. The behavior of the internal state could be associated to a "memory", since it modifies its value according to the series in the sequence.

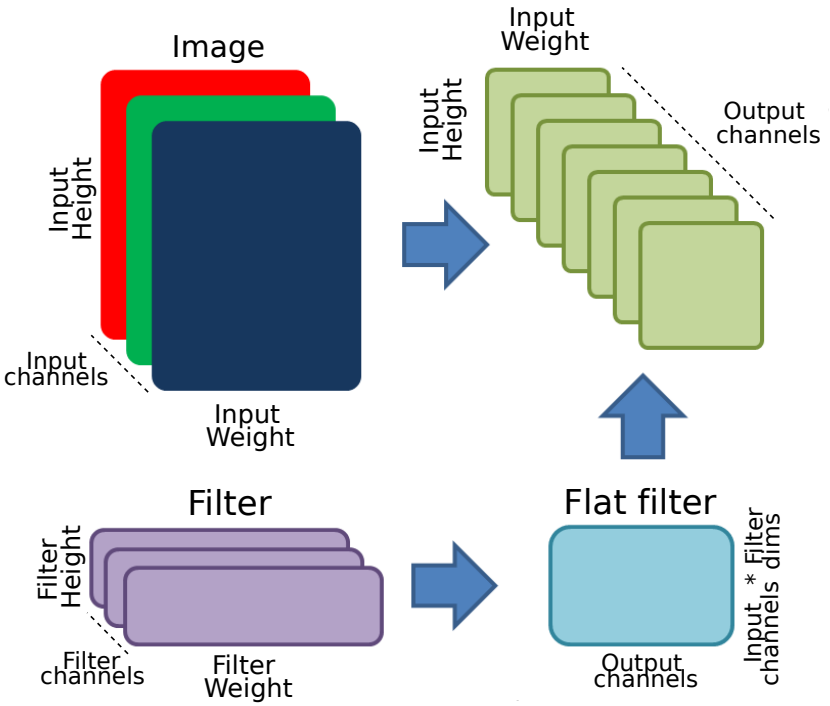


Figure 2-2: Convolution operation for 2D images

The estimation of parameters of a recurrent neural network consist updating the internal state of the cell s_t, which for an input x_t at a time step t is calculated as it is shown in equation 2-5. The function f is usually a non-linearity activation (commonly $tanh$ or $ReLU$) of the state to regularize the parameter increase of the network. Additionally, the corresponding parametric weights of a traditional neural network are preserved (b, W). However, a difference with traditional models lies in the parameter sharing during the whole sequence. This represents an important reduction on the parameter calculations and time complexity. Despite of this advantages, the recurrent neural networks have a considerable limitation since it only could pick information on the short-term sequence elements across time steps.

$$s_t = f(W_{s_{t-1}} + b_{s_t}) \tag{2-5}$$

2.4.3. Long Short Term Memory Networks

Long Short Term memory networks (LSTM) are a specific kind of architectures composed by groups of neurons where outputs are connected to neurons in the same input, the output or the hidden layers in order to create a redundant analysis through time. These characteristics allow to solve a classical problem of computational machine learning, consisting in dependence of analysis for long time interval differences of data [77], mentioned in the recurrent neural networks section (2.4.2). LSTM networks associate information, "remembering" behavior of

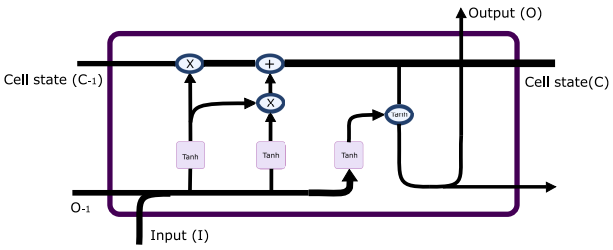


Figure 2-3: Internal architecture for the LSTM recurrent unit.

data for long and different time periods. Due to this, the learning algorithm for these architectures does not initialize from zero state, but it takes into account particularities in data for generating behavior patterns. LSTM networks are basically composed by a memory cell (which are considered as neurons) with an activation functions, and also receives feedback information for considering past and present states. In addition, LSTM cells are connected to other memory cells for propagating information. The whole memory cells assembly composes a link which is considered as the network architecture, and is shown in figure **2-3**.

The learning function during training stage generally involves an optimizer along with the back-propagation algorithm for tuning parameters while minimizing error; and usually optimizing connections in the memory cells of the network to achieve better performance (Fine Tuning algorithm [78]). LSTM networks have been used in state-of-the-art applications such as automatic speech recognition systems (ASR), musical composition, handwriting recognition (HWR), or other applications where time modeling is necessary.

$$i_t = \sigma(W_i \cdot (h_{t-1}, x_t) + b_i) \tag{2-6}$$
$$f_t = \sigma(W_f \cdot (h_{t-1}, x_t) + b_f) \tag{2-7}$$
$$o_t = \sigma(W_o \cdot (h_{t-1}, x_t) + b_o) \tag{2-8}$$
$$\bar{C}_t = tanh(W_C \cdot (h_{t-1}, x_t) + b_C) \tag{2-9}$$
$$C_t = f_t \times C_{t-1} + i_t \times \bar{C}_t \tag{2-10}$$
$$h_t = o_t \times tanh(\bar{C}_t) \tag{2-11}$$

The LSTM state is calculated considering an input gate which defines how much of the new state for the current input is getting through (i_t); and propagates it through output gate which defines how much of the internal state will be exposed (o_t); taking into account an additional forget gate (f_t) which determines how much past information from new state must be preserved during learning stage. It is notable to remark than equations 2-6, 2-7 and 2-8 have the same structure, however the weight parameters change corresponding to its respective gate (W_i, W_f and W_o). Additionally, these networks make use of the sigmoid function which squashes the parametric values between 0 and 1 (to avoid gradient vanishing), and the regularization of gradient growing by introducing a parameter to determine the maximum value

of gradients (gradient clipping). In figure **2-3** is shown a general architecture for a LSTM cell. The $\bar{C}_t$ in equation 2-9 is the update state which is estimated based on the current input and the previous state. The update rule is based on the classical recurrent unit, so the redundant analysis is preserved. The C_t in equation 2-10 is the internal memory parameter of the cell. It computes how much combination will happen between the forget gate (considering the previous state) and the new input (considering the updated state). Finally, the output state (h_t) 2-11 in equation is calculated considering the output gate and the memory parameter.

2.4.4. Gated Recurrent Units (GRU Cells)

A modification of LSTM cells is the gated recurrent unit (GRU) [79]. This change introduces a combination of the forget and input gates (called the reset gate) to determine how much of the new input will be preserved according to the internal memory; and a new update gate which determines how much of the previous state must be updated according to the new state. This modification of the reset gate (r_t) and the update gate (z_t) make the GRU cells have fewer parameters in comparison with LSTM since they do not have a internal memory which is implicit in the reset gate. Additionally, the implicit internal memory allows the system to bypass signals across several time steps, which makes back propagation easier than in LSTMs. It is notable to remark that GRU cells do not have output gate, nevertheless the output state is exposed without squashing values as the LSTM. In figure **2-4** is shown a general architecture for a GRU unit. The reset gate (r_t), update gate (z_t), update state ($\bar{h_t}b$) and output state (h_t) for the GRU cell are calculated as shown in equations 2-12, 2-13, 2-14 and 2-15 respectively.

$$r_t = \sigma(W_r \cdot [h_{t-1}, x_t]) \tag{2-12}$$

$$z_t = \sigma(W_z \cdot [h_{t-1}, x_t]) \tag{2-13}$$

$$\bar{h_t} = tanh(W \cdot [h_{t-1} \times r_t, x_t]) \tag{2-14}$$

$$h_t = (1 - z_t) \times h_{t-1} + z_t \times \bar{h_t} \tag{2-15}$$

2.4.5. Adam Optimizer for parameter tuning

The stochastic gradient-descent based algorithm named Adaptive Moment Estimation (Adam) optimizer [80] is a momentum based learning algorithm (using mean and variance). Its main feature is the allowing of single parameter tuning (such as Adagrad [81] and RMSprop [82]) considering gradients initialization and small decaying rates. These conditions significantly improve parameters optimization to increase accuracy and to avoid divergence during the training stage. Adam moment estimation and optimization rules are described by the equations (2-16), (2-17) and (2-18) respectively, in which the optimal constant values suggested

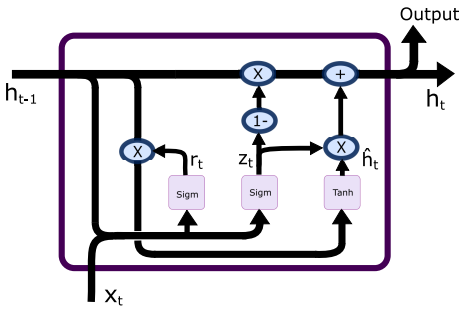


Figure 2-4: Architecture for GRU units.

for the authors (used in this work) are $\beta_1 = 0{,}9, \beta_2 = 0{,}999, \alpha = 0{,}001, \varepsilon = 10^{-8}$ (being $\alpha =$ Learning rate) during training.

$$m_t = \beta_1 m_{t-1} + (1 - \beta_1) g_t \tag{2-16}$$

$$v_t = \beta_2 v_{t-1} + (1 - \beta_2) g_t^2 \tag{2-17}$$

$$\theta_{t+1} = \theta_t - \frac{\alpha}{\sqrt{v_t} + \epsilon} m_t \tag{2-18}$$

3 Classification Models Based on Unimodal Information Analysis

In state of the art review, emotion recognition has been approached using several information sources, such as images ([83, 84]), physiological signals ([85, 86]), speech signals ([87, 88]), and thermography ([89, 90]), among others data modalities (such as facial patterns analysis, body pattern analysis, electroencephalography signal alterations, electrocardiography signals changes, body temperature changes, speech and vocalization variations, among others). For each modality, several works have proposed techniques to recognize emotions from single data sources described above, obtaining results that have not achieved enough performance to completely solve the emotion recognition problem. According to each kind of data, several works have proposed techniques to recognize affective content from unimodal information.

In this chapter, the unimodal analysis of video and speech (which are included into most widely explored modalities in the state of the art) will be addressed. A general scheme of the proposed unimodal analysis is shown in figure **3-1**. For each proposal, it can be seen that the first stage includes a preprocessing according to each data modality (which could be significantly different for its provenance), the second stage consists of a data augmentation technique, the third consist of feature extraction from samples, and the fourth stage involves the Deep Learning strategy. The unimodal analysis will be considered to accomplish the objective 1 of this work, consisting of establish a baseline for comparing performance with multimodal emotion classification.

3.1. Emotion Classification Based on Audio Signals

Due to main characteristics of temporal audio signals, a classification stage for non-stationary data should be achieved in order to extract patterns depending on the emotion while it is expressed. The main problem of temporal analysis of signals is the dynamic treatment for correctly describing the changes of data. In this sense, several works have been proposed with the aim of extracting efficient characteristics for describing dynamic changes of signals. Some of these works, are specifically focused on signal features for emotion recognition. In this section, we will proposed an audio emotion recognition system with temporal analysis and characterization; as well as a classification model using Deep Learning techniques. The

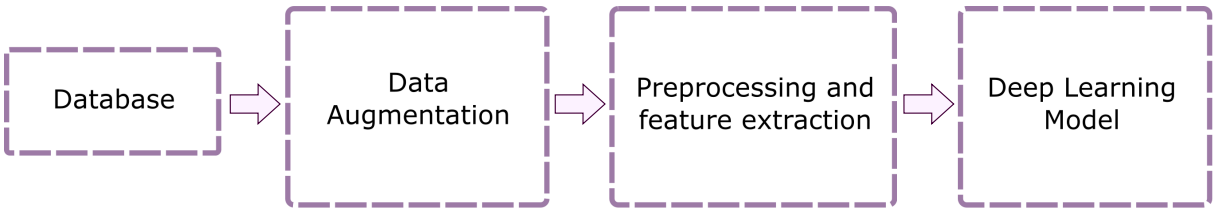


Figure 3-1: General scheme of unimodal strategies proposed in this chapter

general structure of the audio emotion recognition proposal is shown in figure **3-2**.

3.1.1. Data augmentation

Deep learning strategies have shown promising results in several applications, however one of its limitation lies in the amount of required data to achieve generalization when training the models. To solve this issue, and improve the learning capabilities of the models, a data augmentation strategy is used. The augmentation consist in separating every audio sample and extracting overlapping windows (subsamples) of data with specific sizes. In this case, the subsamples will be considered as independent inputs with its corresponding label (according to its respective sample) during the training stage. As a consequence of this augmentation, the predictions for the testing stage will also be overlapping windows. Nevertheless, the predictions for the full sample in the testing stage will be a weighted combination of the predicted values in all windows for the sample. The weighted combination is better described in 3.3.2. The specific values for the size (2 seconds) and overlap (0,9) are selected, taking into account the minimum amount of data required to identify a significant variance in the expression of emotions under real circumstances ($\approx 500ms$ for audio samples, and $\approx$ 1,5 seconds for facial expressions). These parameters are invariant for all the experiments performed in this work.

3.1.2. Feature Extraction

Due to the main characteristic of signals, which is data changes across time; a temporal modeling of data was considered as processing and feature extraction stage, as proposed in [91]. In this case, each subsample extracted after data augmentation is processed by the OpenSmile extractor [92], which generates a vector of 1582 low-level features, corresponding to the well known INTERSPEECH 2010 Paralinguistic challenge features [93]. The INTERSPEECH 2010 features are a set of descriptors especially estimated to extract the emotional content from utterances. As they have been widely used in the state of the art, these descriptors are designed, not only to extract linguistic information from audio signals but also identify non-verbal patterns which could tell an emotion in order to improve performance.

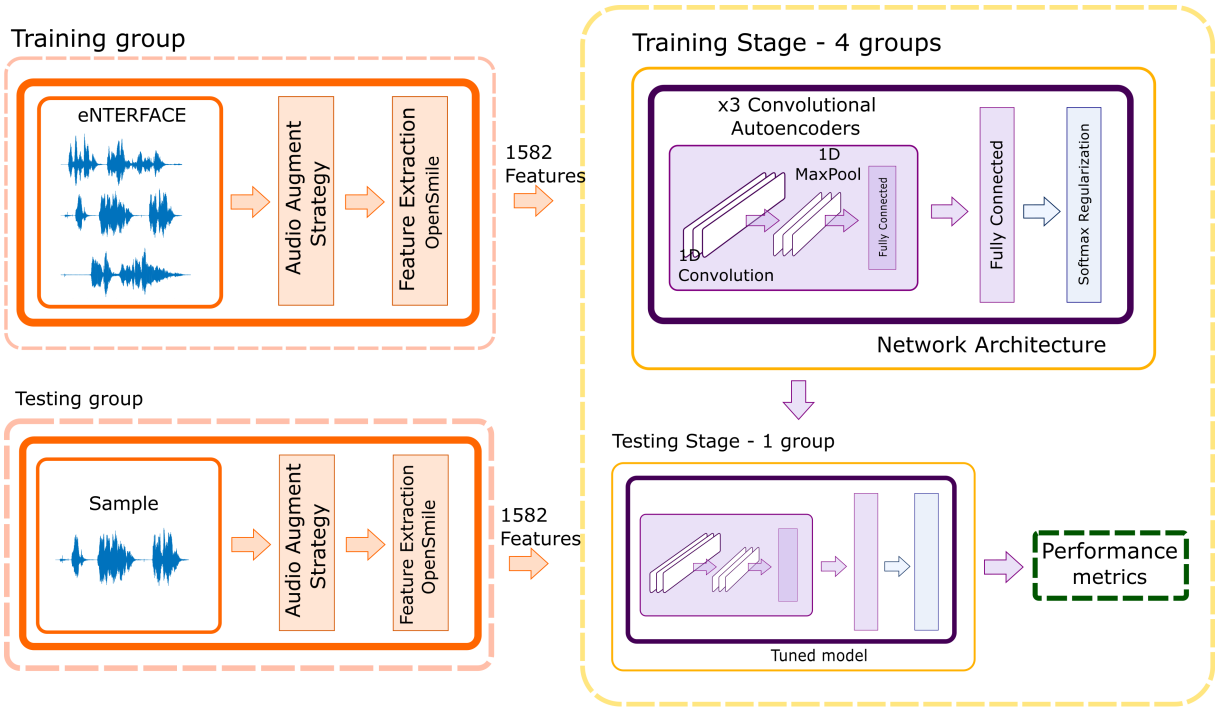


Figure 3-2: Architecture of the model for audio emotion classification

Additionally, the use of these descriptors is considered to reduce the amount of parameters in the network, which directly reduces computational cost.

The 1582 extracted features correspond to a set of 34 low-level descriptors (LLDs), with its corresponding delta coefficients namely: loudness raised to a power of 0,3; Mel Frequency Cepstral coefficients (MFCCs), logarithmic power of Mel-frequency bands, 8 line spectral pair frequencies from 8 linear prediction coding (LPC) coefficients, envelope of fundamental frequency contour, voicing probability of fundamental frequency, maximum and minimum value absolute positions, contour mean, slope of the contour linear approximation and, offset of the contour linear approximation. Besides, a set of 21 functionals were applied to 68 LLDs (1428), and 19 additional functionals were applied to the 4 pitch-based LLDs (152), such as standard deviation of the values in the contour, skewness, kurtosis, the smoothed fundamental frequency contour, among others. In table **3-1** are shown features obtained using the OpenSmile toolkit. The full description of those features can be found in OpenSmile on line documentation [92].

3.1.3. Deep Learning Model for Classifying Audio Signals

Modeling of signal characteristics was performed by a deep learning strategy with encoding instances; i.e. a convolutional encoder-based neural network. Every encoder unit is a structure included in a neural network, that use convolution operations to encode inputs for creating

Descriptor	Functional
PCM loudness	Max/Min (position)
Mel Freq Cepstral Coefficients [0-14]	Arith, mean, std dev
Log. Mel Freq. Band [0-7]	Skewness, kurtosis
LSP [0-7]	Lin. regression slope, offset
F0 sub-harmonics	Lin.regression error
F0 envelope	Quartile 1/2/3
Voicing prob.	Quartile range 2-1/3-2/3-1
Jitter local	Percentile 1/99
Jitter DDP	Percentile range 99-1
Shimmer local	Up-level time 75/90

Table 3-1: INTERSPEECH challenge 2010 descriptors and functionals extracted with the OpenSmile toolbox. LSP = Linear spectral pairs, DDP = Double delta of jitter

a higher-level nonlinear combination of the audio data. In this way, we preserve the most relevant patterns from a chain of events in the audio data. The encoding network implementation allows to determine an output sequence according to an input in the network, and the structure of the units allows to store information from the context of each sample.

The proposed network is composed of multiple convolutional encoding layers which allow to encode the inputs as a combination of the components using convolutions. With the convolution, we can take advantage of the operation to express the inputs as the combination of multiple values, in order to create a higher level model for recognizing emotions. The convolution for the input $(f(x))$ with a filter $(h(x))$ is described in the equation (2-1). Additionally, each convolutional encoder unit is composed of a convolutional layer, a max-pooling layer and a fully connected layer. We use different amount of convolutional encoder layers $(3, 4, 5, 6)$ before selecting the amount obtaining the best performance. Besides, a fully connected network is included to perform classification between the six emotions. A complete description of the model and the composition of a single convolutional encoder is shown in figure **3-3**.

3.2. Emotion Classification Based on Video Analysis

Unimodal processing systems for emotion recognition have taken as one of main options the image processing techniques in order to identify patterns which make possible differentiation between emotions. With this aim, the objective of proposing an unimodal video processing

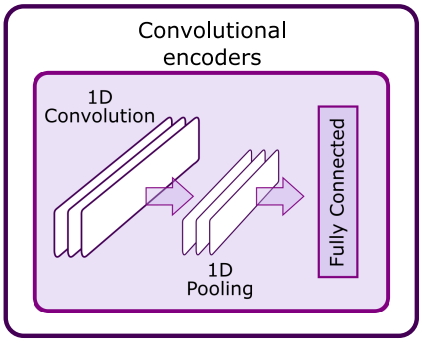


Figure 3-3: Structure of single convolutional encoder

strategy look for the detection and analysis of local patterns in images for characteristic expressions when emotions are expressed. With this aim, there have been works where frame to frame characterization in video processing has been considered for emotion recognition obtaining promising results. However, this strategy is highly expensive in computational terms for real life implementations. Additionally, this strategy does not consider image sequence modeling which is an important concept as information source in video. In figure **3-4** is denoted the strategy we used for emotion recognition in video.

3.2.1. Video preprocessing and data augmentation

According to previous works in the state of the art, in which video processing was approached with frame-to-frame strategies (missing temporary modeling), it was decided to use a framework where dynamic characterization of facial expressions is considered, through the analysis of short time periods from the videos. This conveniently contributes to the data augmentation strategy for improve models. First, all video frames are converted to gray-scale, and are processed using face detection algorithm (Viola and Jones) [94]. This is used to discard all irrelevant information from frames, such as background, body parts, among others. Then, the faces extracted across the video are separated, labeled and predicted using parallel frame-samples as overlapping windows extracted from a short period of time, such as described in 3.1.1. Besides, for comparative purposes, the timestamps of video subsamples are completely synchronized with audio subsamples. This means that training and testing stages will have same information from samples to fit and predict. With this approach, the spatial information along with temporary sequences of fix duration from frames across the video is preserved to train the model.

3.2.2. Deep Learning Model for Classifying Video Data

In order to propose a model which can model spatial and temporal data from samples described in subsection 3.2.1, a Deep Learning architecture is proposed, combining convolutional

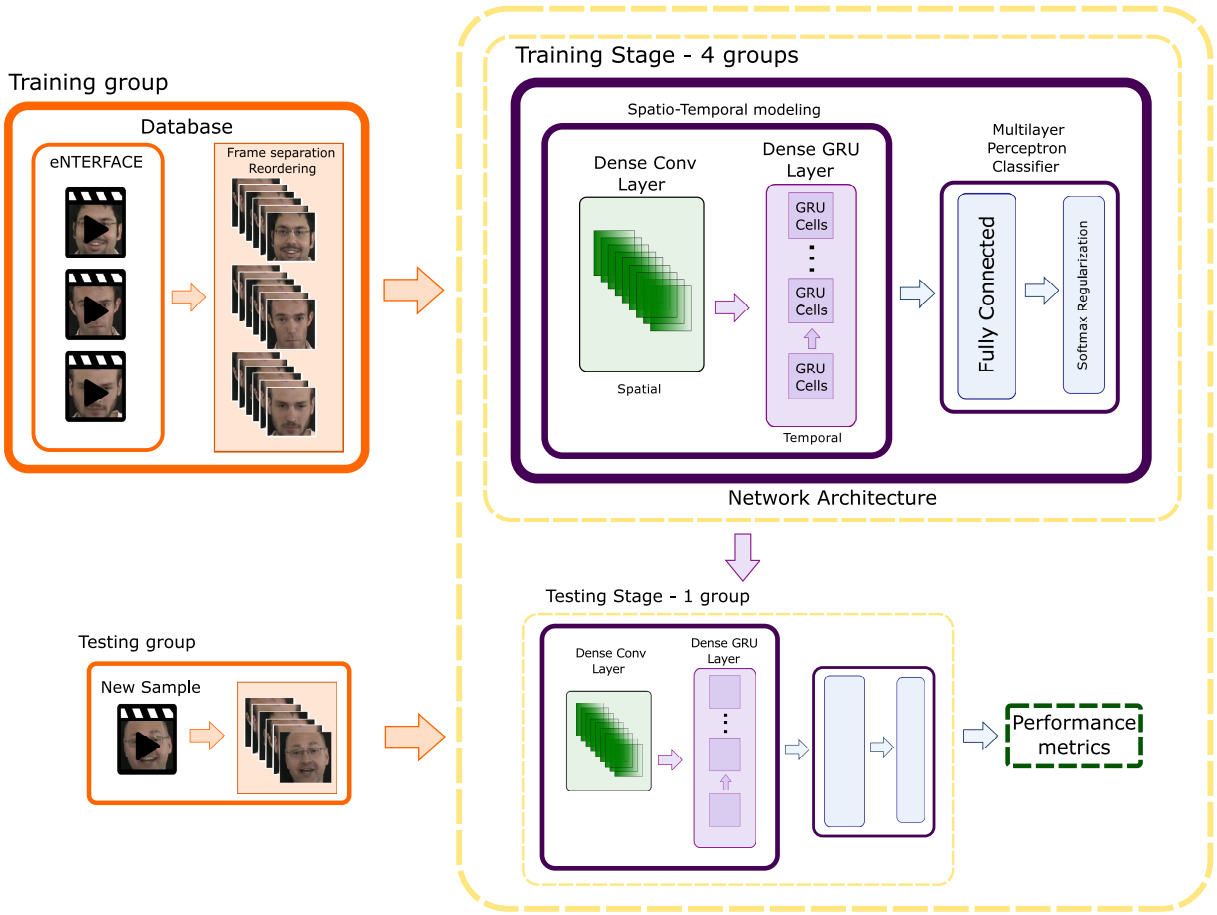


Figure 3-4: Spatio-temporal network model for video emotion classification

layers (for spatial information analysis) and Recurrent units (for temporal modeling of data). The convolutional layers allow to extract spatial features from every sample. From the way the samples are organized (described in subsection 3.2.1), convolutions will learn the main patterns of movements across the frames when emotions are expressed, and propagate them through the network layers. The convolutional network contains 3 convolutional layers with 32, 64 and 128 filters with size 3×3, full padding and stride size of 2×2. The selected activation function for the convolutional layer is the rectified linear units (ReLU) [95] in order to maximize extraction of spatial information in the network units during training stage. Several works in the state of the art demonstrates more expressive models using ReLU units than classical activation function ([96, 97]).

After the convolutional network, a single dense recurrent layer for modeling time dimension of video frames is included. With this aim, the spatial information extracted from convolutional layers is considered in order to model patterns of movements as sequence of temporal data which change across the time. To accomplish this task, a gated recurrent unit (GRU

cells) layer is included to model temporal dimensions in the spatial characteristics. The GRU cells are used over Long-Short Term Memory units (LSTM cells - widely used in sequential models [77]) due to their modular flow of information inside the unit, without having separate memory cells. This characteristic makes the GRU cell computationally more efficient compared to LSTM. Besides, they avoid vanishing gradient problem since the cell bypass multiple time steps, allowing the error to back-propagate easily [98]. The dense GRU layer contains 128 units with backwards sequence processing (this means the cell take the sequence backwards and then, reverse the output again), and a gradient clipping of 1 (to minimize computational cost) and a dropout regularization to avoid overfitting.

Finally, after spatio-temporal modeling described in previous layers, a multilayer perceptron (MLP) is implemented to perform classification task. In this sense, the MLP is composed by 4 layers, with 256, 128, 64 and 32 units each, including a ReLU activation function. The output layer contains the number of 6 emotions, with a softmax activation function, in order to create a probabilistic density function of classes. The proposed approach is graphically described in figure **3-4**; where is shown the proposal of the network architecture for the Deep Learning strategy.

3.3. The Database and Performance Metrics

In this section will be described the database, the prediction model and the performance metrics used to evaluate the proposed experiments. These criteria will be considered as well for comparison purposes with previous works in the state of the art.

3.3.1. The eNTERFACE'05 database

The eNTERFACE'05 database is free and online available bimodal emotion dataset proposed in [99], labeled considering the Ekman emotional model. The database contains information from audio (speech modality) and videos (facial expression modality) in English language concerning 42 different actors coming from 14 different countries in which 19 % are women and 81 % are men. During data acquisition stage, people were asked to express emotions through facial movement and specific sentences (5 sentences) for six different emotions (anger, fear, disgust, happiness, surprise and sadness). It in total gathers 1230 samples, corresponding to 5 videos (one per emotion) where 42 people said the sentences (one sentence per video) and expressed 6 different emotions. The videos for the dataset were taken using a 30fps camera, and the microphone sampling rate used was $48KHz$. Figure **3-5** shows random video and audio samples extracted from the eNTERFACE'05 database.

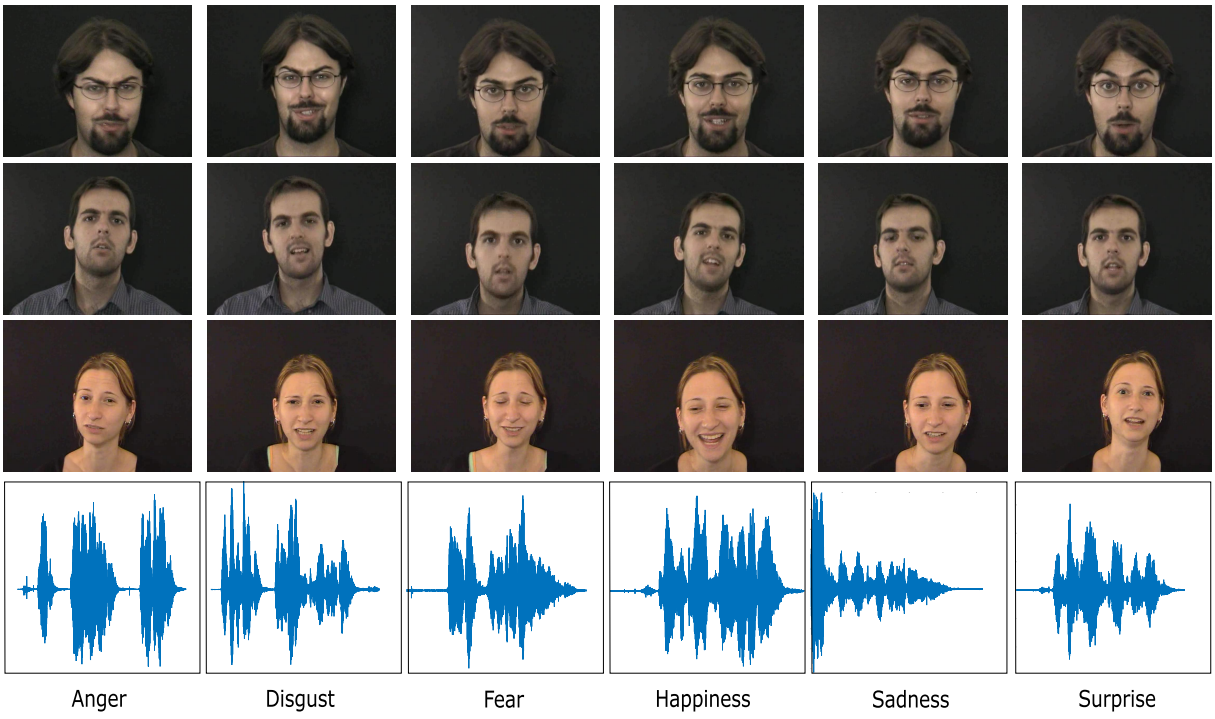


Figure 3-5: Video and audio samples extracted from eNTERFACE database

To perform all the experiments, a gender-independent analysis for emotion recognition is used. This is considered due to the database corpora which is considerably unbalanced with respect to gender (19 % are women and 81 % are men). Additionally, 4 more subjects were not considered, due to lack of files in the database. Besides, the authors of the database evaluated the emotional content of every video to ensure that every sample contains the emotional content it was intended to express. However, they suggested that several subjects were not determined to express the expected emotional content they meant. These subjects were considered to realize the experiments with the aim of increase generality of the model by adding the lack of affective content. To summarize, a total of 40 subjects were taken from the database, to which there were extracted 5635 samples (after data augmentation described in 3.1.1) to perform all experiments. It is noteworthy to remark that the number of samples could increase or decrease depending on the sampling size and the overlapping parameters.

3.3.2. Gaussian weighted prediction

With the aim of increasing the performance validation of the network, a Gaussian weighted prediction for the audio and video sequence is employed. It consists in assign a weight to each prediction for the subsamples extracted after the data augmentation stage, according to their specific order in the sequence of data before estimating an emotional content in the

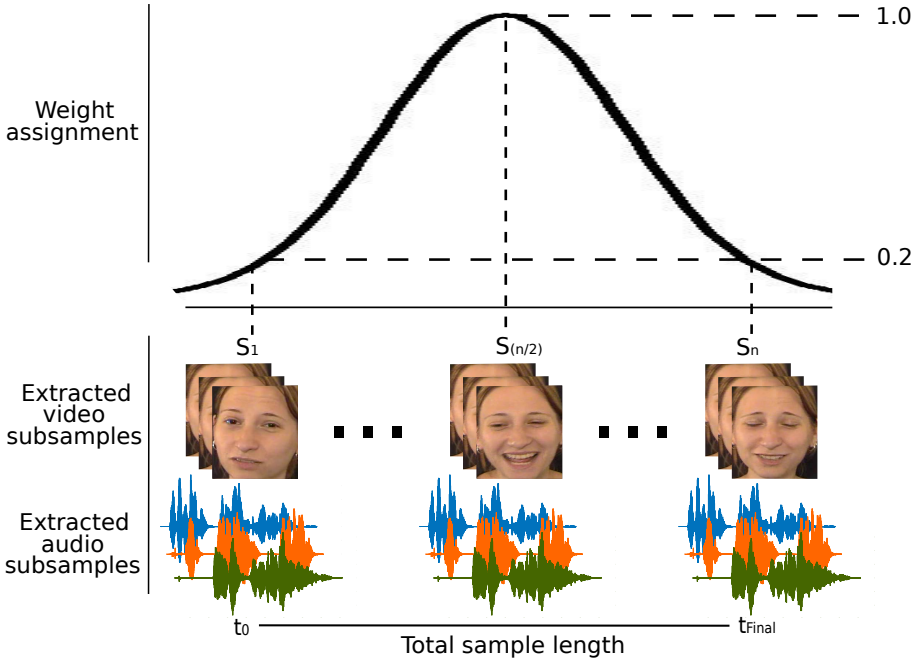


Figure 3-6: Gaussian weights distributions for subsamples extracted from a video.

whole sample. To the n extracted subsamples from a single utterance in the database after data augmentation, a weight is assigned according to a Gaussian bell curve with a standard deviation (σ) of 1. n weights are multiplied with n subsample predictions in every stage. This allowed to assign more importance to predictions of subsamples extracted from the middle of the sequence (which obtained a weight assignation near to 1) in comparison with the ones extracted from the farthest segments in the sample (which obtained a weight assignation near to 0). The Gaussian weights for n subsamples extracted from single utterance are calculated as described in equation (3-1).

This assignation is considered since the maximum peak of emotion expressiveness is approximately achieved in the middle data of the utterance. Then, more importance should be assigned to those subsamples in comparison with the ones from the start and end of the complete sample. A graphical representation of the assignation of Gaussian weights for n subsamples extracted from a single utterance is shown in figure **3-6**.

$$w(n) = e^{-\frac{1}{2}(\frac{n}{\sigma})^2} \tag{3-1}$$

3.3.3. Performance metrics

To evaluate methods described before, a 5-fold cross validation algorithm is used. However, the separation of fold samples is made according to the subjects; and not the total number

of samples. This means that each fold will contain the samples extracted from 8 subjects, no matter the number of subsamples extracted after data augmentation. This evaluation model (also known as Leave-One Subject Group) guarantees the subject-independence of the data for validation, which is more similar to realistic scenarios.
From every trained model in the cross validation experiment, we obtained confusion matrices (comparing correct predicted emotions per class with amount of samples per class) and overall accuracy (comparing correct predicted emotions with amount of samples). Overall accuracy corresponding to the mean value of all folds was estimated according to equation (3-2).

$$Overall\ Accuracy = \sum_{i=1}^{folds} (\frac{Correct\ predicted\ samples\ (hits)}{Amount\ of\ samples}) \tag{3-2}$$

3.4. Results

In this section, we are willing to describe obtained results for models proposed above.

3.4.1. Audio-based emotion classification experiment

The proposed model for emotion classification, several experiment were evaluated, varying the number of convolutional encoders $(3, 4, 5, 6)$ with same number of units (64 and 32 units pairwise subsequently). The results of these experimentations is shown in table **3-2**, in which the model with best performance is selected, in order to report best results. The model consists of 3 convolutional encoder layers, containing 64, 32, and 64 1D convolutional filters and same number of units per layer. Each encoding layer contains a filter size of 3, stride 2, full padding, max-pooling layer with pooling area 2 and ReLU activation function. After encoding layers, a MLP is used to perform classification task, in which 2 layers are used containing 384 ReLU units each. Additionally, to avoid overfitting, a dropout function before the output layer with 0,5 regularization value was included.

Number of Convolutional Encoders	**Recognition rate**
3	0.54
4	0.48
5	0.41
6	0.38

Table 3-2: Performance comparison of multiple convolutional encoder layers.

Emotion	Predictions					
	Anger	**Disgust**	**Fear**	**Happiness**	**Sadness**	**Surprise**
Anger	**0.492**	0.107	0.040	0.061	0.050	0.000
Disgust	0.127	**0.536**	0.120,	0.061	0.100	0.071
Fear	0.159	0.071	**0.680**	0.030	0.000	0.143
Happiness	0.063	0.036	0.040	**0.515**	0.100	0.143
Sadness	0.063	0.143	0.080	0.061	**0.650**	0.179
Surprise	0.095	0.107	0.040	0.273	0.100	**0.464**

Table 3-3: Confusion Matrix for audio emotion recognition.

Strategy	**Recognition rate**
Datcu et al. [100]	0.559
Huang et al.[101]	0.523
Audio classification (our)	0.547

Table 3-4: Performance comparison of audio strategies with previous works

Table **3-3** shows the results obtained for the model in the classification task, with an overall accuracy of 0,54. For the audio emotion recognition, Fear (0,680) and Sadness (0,650) obtained similar results. It is notable to remark that Fear scored similar confusion rates to Anger (0,159) and Surprise (0,143) and Sadness scored similar confusion rates to Disgust (0,143) and Surprise (0,179). On the other hand, Surprise (0,464) obtained lower results, achieving high confusion rate with Happiness of 0,273. Additionally, Surprise obtained the highest confusion scores in comparison with other emotions. On the other hand, table **3-4** shows the comparative results of the audio recognition model with other found in the state of the art. It is important to remark that the comparison criteria is the experimental framework; in which the use of LOGSO validation and the same database was considered. The comparative results portray that the proposed model is compatible with methods used in the state of the art.

3.4.2. Video-based emotion classification experiment

The convolutional network contains 3 convolutional layers with 32, 64 and 128 filters with size 3×3, full padding and stride size of 2×2 with ReLU units. Then, a dense GRU layer contains 128 units with backwards sequence processing with a gradient clipping of 1 (to minimize computational cost) and a dropout regularization of 0,5 to avoid overfitting. Finally, after spatio-temporal modeling described previously, a multilayer perceptron (MLP) is implemented to perform classification. In this sense, the MLP is composed by 4 layers,

with 256, 128, 64 and 32 units each, including a ReLU activation functions.

Table 3-5 shows the results obtained for the model, which achieved an overall accuracy of 0,46; obtaining highest accuracy rates in Disgust (0,594). However, it is along with Anger, the one obtaining the highest confusion rates compared with the other emotions. On the other hand, the Fear (0,282) obtained the lowest results from the rest of the emotions; comparable with its confusion rate with Disgust (0,230) and Surprise (0,205). Besides, table 3-6 shows a comparison between the results obtained from the video recognition model with previous ones find in the literature. The comparison criteria is the experimental framework; in which the use of LOGSO validation and the same database was considered.

Emotion	**Predictions**					
	Anger	**Disgust**	**Fear**	**Happiness**	**Sadness**	**Surprise**
Anger	**0.522**	0.204	0.136	0.0	0.090	0.045
Disgust	0.108	**0.594**	0.108	0.027	0.135	0.027
Fear	0.153	0.230	**0.282**	0.076	0.051	0.205
Happiness	0.157	0.157	0.052	**0.5**	0.052	0.078
Sadness	0.162	0.135	0.162	0.0	**0.459**	0.081
Surprise	0.114	0.057	0.257	0.01	0.2	**0.371**

Table 3-5: Confusion Matrix for video emotion recognition.

Strategy	**Recognition rate**
Datcu et al. [100]	0.377
Huang et al.[101]	0.564
Video classification (ours)	0.468

Table 3-6: Performance comparison of video strategy with previous works.

4 Classification Models Based on Multimodal Information Fusion

The emotion recognition task performed in previous works from unimodal perspectives. However, the combination of multiple data sources could increase performance to achieve recognition rates from data correlations. With the aim of comparing the results obtained in the baseline stage in chapter 3; in which unimodal perspective is performed, in this chapter we will describe the emotion recognition from both, video and speech modalities together, using Deep Learning strategies. The deep learning based unimodal processing stages described in the previous chapter are used as baselines to compare against the proposed multimodal information models. The two multimodal models aim to address the problems described in objectives 2 and 3 of chapter 1, corresponding to both data representation in a similar space and classification according to previous representation. To achieve this, each characterization will use the same kind of layers (Convolutional), with the aim of obtain similar abstract patterns from the response of filters to input data. This will guarantee a similarity among the representation of different data modalities before fusion. The first model fuses information at the decision level and the second model fuses information at the features level. This chapter describes both architectures and compares performance.

4.1. Deep Learning based classification model for decision fusion

To fuse multimodal information from audio and video, we used similar strategies proposed in previous chapters, i.e., based on deep learning architectures. Thus, similar to what we described in section 3.1 from chapter 3; for the audio signals we use convolutional encoders as in previous architecture. Additionally, we propose an MLP for achieving a higher abstractive representation of the characterization task. Consequently, similar as we described in section 3.2 from same chapter; for video characterization, we proposed a spatio-temporal modeling involving convolutional and recurrent layers.

After characterization stage using Deep Learning, we proposed an additional stage in order to combine both results of characterization. At first, we concatenate features obtained by

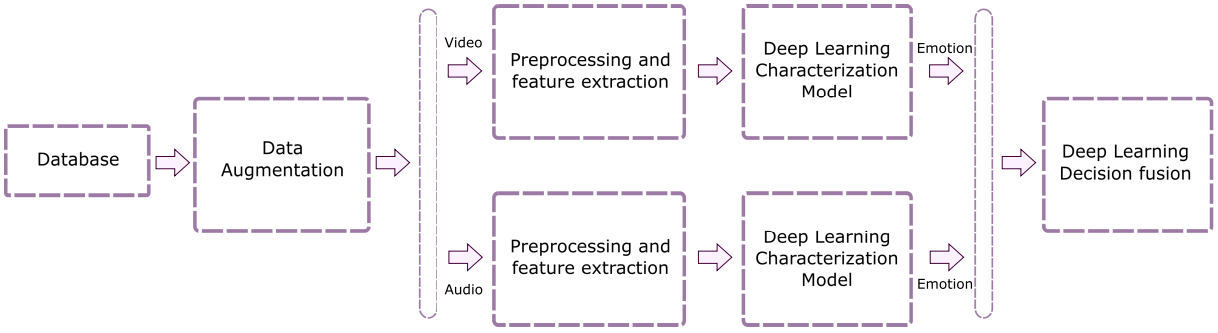


Figure 4-1: General scheme for the multimodal information fusion at decision level

video and audio; then, we use MLP layers for encoding representations obtained for both kinds of data.

The selected amount of layers for the MLP depend on how deep the representations will encode the features. This means that as many layers as the MLP has, they will encode features from both data types, relating them in the same space, before performing the classification task. In figure **4-1** is shown an example of structure for the classification model of multimodal information at decision level.

In this case, the model for classification using the strategy fusion, consists of separated inputs for each data source (audio and video). The audio input is connected to an encoder containing 3 layers with similar number of units (32, 64 and 32), with the aim of obtaining a higher abstract representation of the data from audio signals. Parallel, another input for video is placed in the network for handle spatio-temporal representations. The network is composed by 3 convolutional layers; the first one contains 32 filters, the second one with 64 and the third one with 128. All layers uses 3×3 kernel sizes, stride size of 2×2, full padding and ReLU activation functions. Additionally, with the purpose of concatenating both Deep Learning representations, a flatten layer must be included in order to make both layers outputs to be computationally compatible.

The concatenated layers are then connected to a 4 layer MLP, including "bottlenecks" with the purpose of achieving higher encoding capabilities. In this way, it is considered to use one in the second layer, being then $64, 128, 64$ and 32 units from 1st to 4th layers, consecutively. As it was mentioned before, the layers of the MLP will allow to encode features from both video and audio representations into a same space which let us separate emotions. Besides, to avoid overfitting during classification stage for the characterization task, several dropout layers were included (after video, audio characterization and before output) with 0,2 probability of dropping units. Finally, the output layer contains 6 units (number of predicted emotions) consecutively, applying softmax to establish probabilistic density function

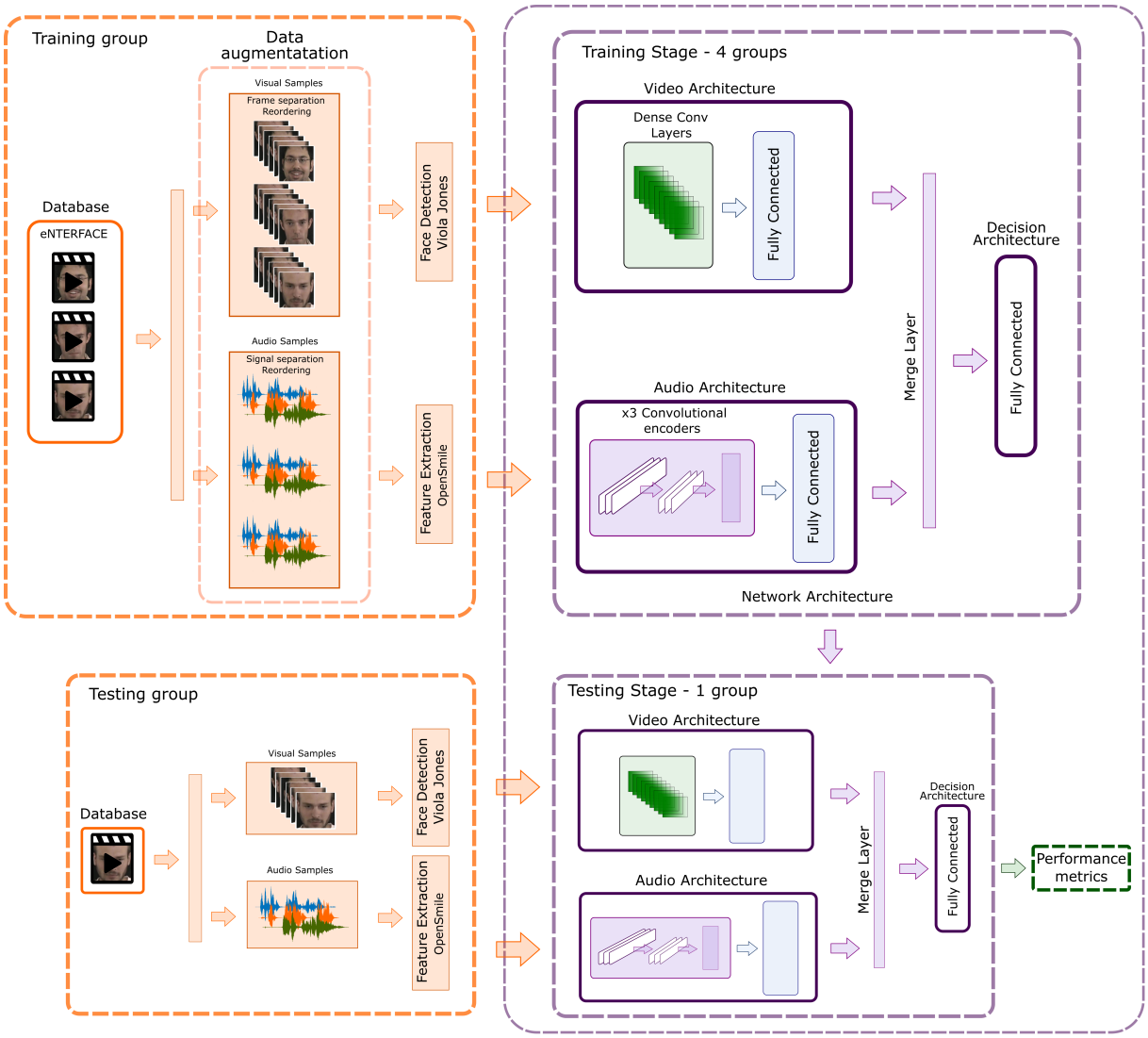


Figure 4-2: Multimodal decision fusion strategy using Deep Learning

of classes. Also, the selected optimizer was Adaptive Moment Estimation (Adam) due to the parameter considerations of the algorithm to avoid divergence and maximizing performance. The full proposed architecture for emotion recognition with decision fusion is shown in Figure 4-2.

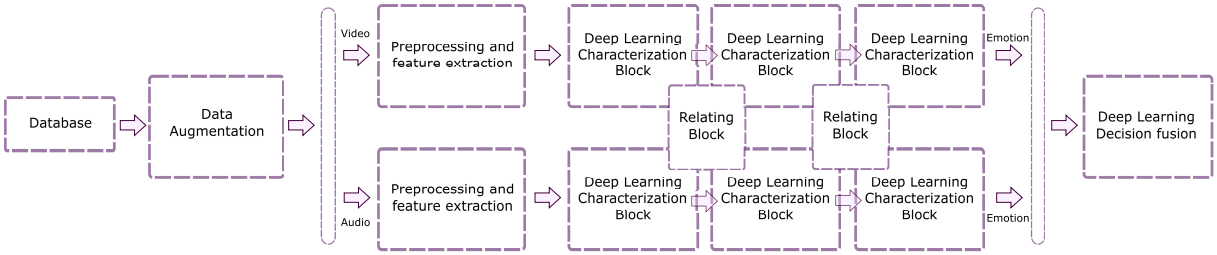


Figure 4-3: General scheme for the multimodal information fusion at characteristic level

4.2. Deep Learning based classification model for characteristic fusion

Additionally to the multimodal fusion model described above, it was considered as well the implementation of a similar strategies with characteristic fusion approach for comparative purposes. However, in this case is required to encode the incoming data using set of layers to relate the data from different kinds of information. The proposed model consists of using parallel architectures for each type of data (audio and video); then introduce a first layer for each independent architecture to encode features obtained and a second layer to encode the representation obtained by one architecture with the other one. The output of the two set of layers will continue feeding the characterization architecture, which prioritize the same data input but considering characterization of the other model.

Finally, the output of this characterization will be connected to an MLP to perform classification task. As it was mention in the subsection 4.1, the selected number of layers for the MLP depend on how deep the representations will encode the related features. In figure **4-3** is shown a scheme of structure for the classification model of multimodal information at characteristic fusion level.

The model for classification using the fusion strategy consists of separated inputs for each data type (audio and video). The audio input is connected to a convolutional network consisting of one $1D$ convolutional layer with 64 filters of size 3, stride of 2, full padding and ReLU activation function. Additionally, a max-pooling layer with pooling area and stride of 2 is used, along with a fully connected layer with same number of units and activation function (64 units and ReLU, consecutively). This combination of layers is considered as a convolutional encoder. Differently from proposed models described above, the convolutional encoders are employed in this model due to the intrinsic relation of the mathematical operation (convolution) to extract similar patterns from video and audio; allowing to relate them in a similar dimensional space to perform classification.

Parallel to audio architecture, another convolutional encoder is employed for video data, in

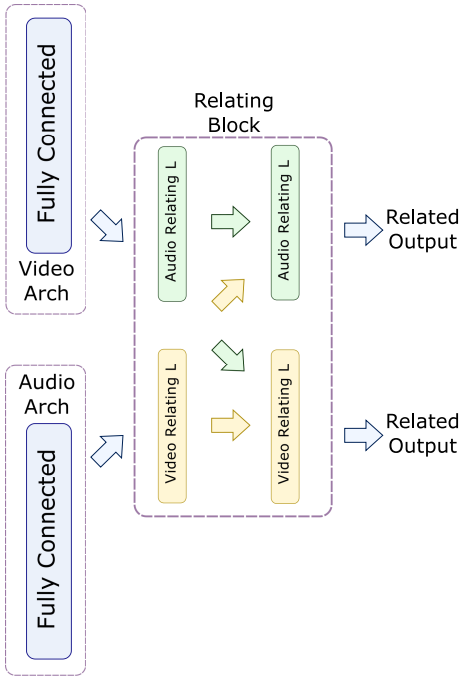


Figure 4-4: Relational block to combine extract video and audio features

which the $2D$ convolutional layer (given input of gray scale images) use 64 filters of size 3×3, stride of 2×2, full padding and ReLU activation function. Besides, the max-pooling layer uses pooling area and stride of 2×2, and the fully connected layer with same number of units and activation function (64 units and ReLU, consecutively).

After the first block of convolutional encoders for each architecture, a set of two blocks of layers is connected. The purpose of these layers is to relate the information from an architecture with the information of the other across epochs during the training stage. Each set of layers is considered as the relational block of layers. Thus, the first layer of the relational block for the audio architecture takes as input the characterization obtained from its respective convolutional encoder, plus the concatenated output of the convolutional encoder for video; and same for the video architecture. These inputs of the relational blocks are connected to a full connected layer, which is the output of the block. Then, the previous output for each block is connected to the corresponding video/audio convolutional encoder architecture. A graphical representation of relational block is shown in Figure 4-4. In case of the proposed model, the first relational block contains 128 and 256 ReLU units in first and second layer consecutively, for both architectures. The idea of including same number of units each relational block come from the purpose of give equal importance to both output relation in the architectures.

Subsequently, 2 additional combination of these blocks (convolutional encoder and relational blocks) are used to increase the depth of the network to get higher abstract representations.

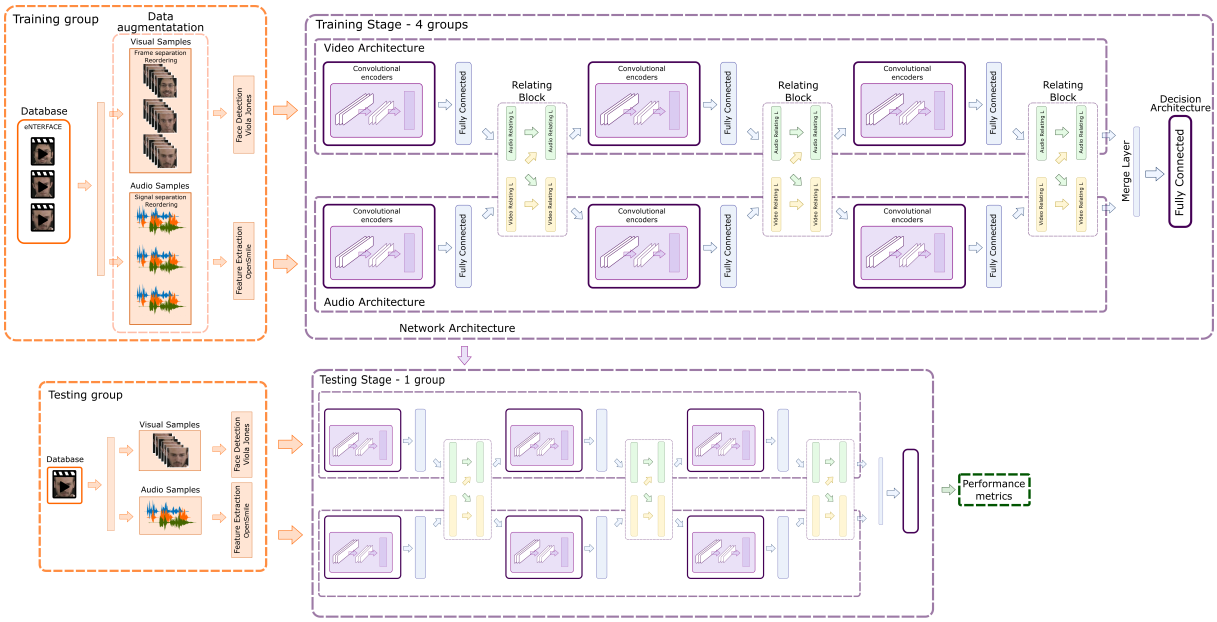


Figure 4-5: Multimodal characteristic fusion strategy using Deep Learning

For the audio processing architecture, a first convolutional encoder is set with 64 filters of size of 3, stride of 2, full padding and ReLU activation function; max-pooling with 2 pool size and fully connected layer of 64 ReLU units. For video, the convolutional encoder is set with 32 filters of size of 3, stride of 2, full padding and ReLU activation function; max-pooling with 2 pool size and fully connected layer of 32 ReLU units. After this stage, other relational block is included to combine abstract data from the network. The second relational unit is set with 64 and 128 ReLU units from first an second layer in both architectures.

Finally, a third combination of these blocks is used. This time, the audio convolutional encoder is set with 32 filters of size of 3, stride of 2, full padding and ReLU activation function; max-pooling with 2 pool size and fully connected layer of 32 ReLU units. On the other hand, the video convolutional encoder is set with 16 filters of size 3, stride of 2, full padding and ReLU activation function; max-pooling with 2 pool size and fully connected layer of 16 ReLU units. The output of this block is connected to a MLP with 3 layer containing 256 and 128 ReLU units respectively. From this point, two more experiments were carried out in which amount of relational blocks is changed with the aim of evaluate the influence of the amount of blocks in the network. As a consequence, it implies to increase the processing layers when a relational block is included. However, the parameters of the additional relational blocks are selected according to the third combination of previous layers described above. At the end, the output of the MLP is a layer containing 6 softmax units concerning to the belonging probability of an input sample to an emotional utterance. A full illustration of the network is shown in Figure 4-5.

4.3. Results

In this section are described the main results for both models described above. To evaluate the models, the metrics described in 3.3.3 are extracted. The main purpose of the experiments is testing the performance of the multimodal models at decision level (through classification architecture) and characteristic level (through relational block), in comparison with unimodal strategies and previous works in the state of the art.

4.3.1. Multimodal decision fusion strategy

In table 4-1 are shown the results obtained for the model in the emotional task, obtaining an overall accuracy of 0,62. The model obtained a recognition rate for Anger (0,795) and Happiness (0,842). This could be caused for the duality of the emotional content of them (negative / positive) which could be physically manifested in significantly different ways. On the other hand, emotions such as Disgust (0,432) and Fear (0,461) obtained lower results; where it is noteworthy to remark that highest confusion occurred with Happiness and Anger respectively. This could mean that the algorithm cluster the most significantly different emotions (Anger / Happiness) and then, adjust parameters for other emotions. Additionally, it is also notable that Surprise has a significant difference among its confusions, in which the highest peak is Fear (0,171). The reason of this could lay in the similarities in the manifestation of the emotions when they are expressed.

Emotion	**Predictions**					
	Anger	**Disgust**	**Fear**	**Happiness**	**Sadness**	**Surprise**
Anger	**0.795**	0.045	0.022	0.090	0.0	0.045
Disgust	0.189	**0.432**	0.027	0.189	0.027	0.135
Fear	0 .076	0.025	**0.461**	0.102	0.102	0.230
Happiness	0.078	0.026	0.0	**0.842**	0.0	0.052
Sadness	0.054	0.081	0.162	0.027	**0.513**	0.162
Surprise	0.085	0.0	0.171	0.085	0.085	**0.571**

Table 4-1: Confusion Matrix for emotion classification using multimodal decision

4.3.2. Multimodal characteristic fusion strategy

Before the experimentation for the fusion of the data using the relational block, it was conducted an experiment to determine the amount of relational blocks to obtain best performance results. In this sense, there were evaluated, in terms of the amount of blocks shown in table

Number of Relational blocks	Recognition rate
2	0.44
3	0.64
4	0.58
5	0.47

Table 4-2: Performance comparison of multiple relational blocks

Emotion	Predictions					
	Anger	**Disgust**	**Fear**	**Happiness**	**Sadness**	**Surprise**
Anger	**0.795**	0.068	0.068	0.022	0.045	0.0
Disgust	0.027	**0.783**	0.054	0.081	0.027	0.027
Fear	0 .102	0.153	**0.564**	0.0	0.153	0.025
Happiness	0.078	0.157	0.026	**0.657**	0.0	0.078
Sadness	0.108	0.027	0.297	0.0	**0.540**	0.027
Surprise	0.085	0.114	0.142	0.028	0.2	**0.428**

Table 4-3: Confusion Matrix for emotion classification using multimodal characteristic fusion

, in which the model containing 3 relational blocks obtained best results. It is noteworthy to remark that no more experiments were performed due to hardware limitations (storage and processing). In table **4-3** are shown the results obtained for the model, achieving an overall accuracy of 0,64. In the case of this model, Anger (0,795) and Disgust (0,783) obtained highest performances. Besides, Happiness obtained a similar result (0,657) with low confusion rate except for Disgust, in which obtain a considerable rate (0,157). Other emotions such as Fear (0,564) and Sadness (0,540) obtained similar results, however, both confusion peaks correspond to the emotions Disgust and Fear respectively. On the other hand, the Surprise (0,428) obtained lowest results and high confusion rate with Disgust and Fear (0,11 and 0,14, respectively). This confusion is individual, since next lowest performance is Surprise (0,428) and there is a notable difference (0,112).

In table **4-4** are shown the comparison among the strategies proposed in this work. The results show that the combination of both modalities at characteristic fusion level outperforms other experiments to accomplish emotion recognition task. Additionally, table **4-5** compares this work with other state of the art strategies. The selection criteria for the comparison with previous works was the use of the same experimentation model for emotion recognition. For example, in [102] is presented a model for speech emotion recognition using a Deep Learning model. The work takes advantage of a pre-trained model to improve per-

Strategy	Recognition rate
Unimodal Audio emotion	0.54
Unimodal Video emotion	0.46
Multimodal decision fusion	0.62
Multimodal Characteristic fusion	0.64

Table 4-4: Performance comparison between Proposed strategies.

Strategy	Recognition rate
Datcu et al. [100]	0.563
Huang et al.[101]	0.611
Decision fusion (our)	0.62
Characteristic fusion (our)	0.64

Table 4-5: Performance comparison between Proposed strategies.

formance of the network taking the information of Mel spectrogram extracted from audio signals in a convolutional model to perform classification. However, they can not be compared with the proposed techniques since they only use speech data to validate their models; for the data augmentation strategy which increases significantly the amount of samples, the Gaussian weighting in subsamples in the data augmentation strategy and the LOSGO validation technique. Besides, they take advantage of transfer learning properties, which is proposed in the future works section below. On the other hand, the experimentation shows that the use of LOSGO experimentation is a more promising validation strategy to guarantee subject-independence in more realistic scenarios; compared with traditional cross validation strategies, in which authors have also obtained relevant results ([103, 104, 105]). However, the use of the implementation in this work have shown promising results facing implementation of these models in real life scenarios. Besides, other advantage lies in the extensible capabilities of the model to include more multimodal information for data fusion. Additionally, preliminary experimentations to improve performance have been realized during the development of this work, despite rigorous studies could outperform results obtained in this work.

Despite the advantages shown previously, one of the drawback of the model lies in the amount of optimizing parameters, since it becomes a high dimensional parameter optimi-

zation to be considered to outperform results. This characteristic will require the use of complex computational problems to perform an effective search of parameters to increase results. Additionally, it will probably require a bias in the searching space to reduce the complexity of the problem for achieving implementation. Besides, other limitation lies in the merge layer included for each fusion model, previous to classification stage. The layer is based in a concatenation between the extracted audio and video representations, since it could bias the representation for the classification stage. However, in this case the drawback could be improved by implementing a similar relational block specially designed to apply a more effective merging. Despite these factors, the model could outperform results presented in this work to achieve higher robustness in comparison with previous strategies presented, by including an optimization stage or increasing the depth of the network. However, this will increase the computational complexity that should be considered.

On the other hand, a recent work have been found in the state of the art ([106]) in which an emotional model have been proposed. The proposed model have achieved a maximum accuracy in the multimodal fusion of 0,8597; however, the main difference of this work lies in the initialization of the network. They used a transfered learning for increase the performance the model; in which their networks are initialized using pre-trained models without fine tuning. This result is supported from the reported evaluation of the feature extraction using pre-trained models, in which they report 0,5435 and 0,7808 in audio and visual characterization respectively. The proposal for evaluation of pre-trained model is mentioned in section Considerations and Future works (5.2), despite it is currently in evaluation with our models.

5 Discussion, Conclusions and Future works

5.1. Discussion and conclusions

In this work, a Deep Learning model for emotion recognition using multimodal information fusion is proposed. Along the chapters of this work are proposed several strategies in which unimodal information is evaluated using audio and video data separately, and multimodal information fusion at decision level (after previous characterization) and characteristics level (combining video and audio features). The proposed approaches took from the well-known eNTERFACE'05 database a single sample and extracted video frames and audio; then a preprocessing stage is introduced in both cases. For audio samples, a first INTERSPEECH 2010 Paralinguistic challenge features extraction stage is proposed using the software OpenSMILE. On the other hand, for video frames, a Viola-Jones algorithm is used to extract facial regions of participants in the video along every frame. It is important to remark than the preprocessing stage in both cases was considered with the aim of reducing dimensionality of data to decrease computational resources. Additionally, for avoiding overfitting given the limitation of samples in the database, a data augmentation strategy is used by extracting several windows of data with an specific overlapping. The preprocessed audio and video data is then the input of the proposed strategies, which were evaluated using a 5-fold cross validation algorithm, extracting the overall accuracy from every video/audio samples by assigning a prediction to each window, and giving a Gaussian weight according to the position of the window in the sample.

However, there are several points which influence the performance of the algorithm, such as the data augmentation strategy. In our case, a 2 second window with a 0,9 overlapping was used, however, the window and the overlapping size are parameters with high impact in the learning model, since a very big window could significantly decrease the amount of samples, which may cause overfitting; other way to overlapping, which could generate loss in the temporary continuity of the data; nevertheless, a short overlapping could not guarantee the emotional content in all samples (silences, non-verbal expressions), generating confusion in the data for the model. Both parameters could be optimized to find best values to maximize results. The data augmentation technique is an important aspect as we have show in previous works [103, 105] in which we have proposed several models with and without data

augmentation that show increase in performance when it is included.

On the other hand, the evaluation metrics are also influencing the presented results in the way in which several authors make the cross validation without making any difference between amount of samples over subjects. However, the proposal in this work uses a cross validation per subjects in which it is separated the amount of samples per subjects instead of amount of individual samples. The metric is evaluated this way considering that voice corpora of participants change between subjects, which bias the data to the model. Additionally, this argument leads of thinking about a bias between gender (male - female) in which our proposed works, we demonstrate that the model could obtain better result if it is considered. However, in this work it is not evaluated due to the amount of samples available for female participants (only 19 % of the total subjects). On the other hand, it is important to remark that an evaluation for single subjects could increase the performance of the model, giving the adaptation taking into account the corpora of the participant. However, the evaluation has not been performed due to computational limitations to fit the model.

Regarding to emotion recognition models, four strategies were proposed in this work: The first and second concerning unimodal analysis using audio and video separately; the third and four using multimodal information fusion at characteristic and decision levels. The proposed multimodal models demonstrate that the combination of two data modalities outperform the unimodal models in both cases. At this point, several conclusions are derived from this work. First, the decision fusion modeling of temporal information using data windows take advantage of implicit correlations during fine-tuning stage to outperform the results, since the learning algorithm is based on parameter adapting according to each time step. However, the model based on decision fusion makes a separate characterization of the data corpus of video and audio, which intend to bias characterization, not in temporal dimension but in corpora singularities (such as noise peaks or intense facial movements). This problem is addressed in the fusion characteristic model, which adjust parameters according to a single kind of data, but including blocks of layers to share tuning parameters with after each characterization. The inclusion of these blocks increases performance, since the data corpus are share through combination blocks to efficiently combine patterns obtained in both modalities. The results of all experimentations shows that multimodal information fusion is more effective, compared to unimodal approaches to achieve the recognition taks. Besides, the characteristic fusion model outperform decision strategy in the same experimentation. Results in this work and other in the state of the art (such as [102]) using this experimentation are promising for the multimodal recognition problem; however, they can not be compared since relevant aspects such as the data augmentation strategy and the Gaussian weighting for subsamples are used in this work.. Nevertheless, this experimentation is considered to guarantee subject independency and more reliability and suitable to realistic scenarios.

Finally, the experiments proposed in this work demonstrate the effectiveness of the multi-

modal fusion to accomplish the emotion recognition task, using Deep Learning strategies at different levels (characteristics and fusion) as it will be described below: Regarding to the first specific objective .[Es]tablish a baseline of state-of-the art techniques for emotion recognition based on unimodal analysis, using Deep Learning techniques", a characterization strategy for audio and video data, based on previous works in the state of the art was established. The results of characterization research were the baseline of preprocessing stage (audio feature extraction, face extraction and windowing strategy), before evaluate the models. According to second specific objective "Proposing a feature extraction strategy based on Deep Learning approaches for multimodal data representation in a similar dimensional space", two unimodal strategies based on Deep Learning for emotion recognition were developed. In this sense, each model included a classification stage to evaluate characterization according to state of the art performance metrics. The results of these characterization are described in chapter 3. Regarding to third specific objective "Proposing a Deep Learning strategy for emotion recognition from multimodal information, using representation stage developed in specific objective 2"; two multimodal information fusion models at decision and characteristics levels were proposed. The model at decision level included two separate characterizations (for audio and video) for later merging the outputs to perform classification task. On the other hand, the model at characteristic fusion included combination blocks to fusion parameter tuning during training stage, with the aim of combine both corpora data patterns obtained in both modalities. The results of the multimodal information fusion models are described in chapter 4.
According to fourth specific objective .[Ev]aluate performance of emotion recognition proposed strategy using public and available emotional databases with multimodal information", all the experiments and metrics were obtained using the eNTERFACE'05 database, and authors ensure they were equivalent in all aspects (same set of data, time steps, weighting function and labeling, among others). The metrics and experimental setup is described in section 3.3 Finally, the explanation of accomplishment of each specific object described above demonstrate the fulfillment of the hypothesis described in section 1.3. Besides, other experimentations (such as different publications [103, 105, 104, 13]), parallel work (such as undergraduate student projects) and related works ([107, 108]) has been developed based on the main objective of this proposal (see Appendix section). These experiments and attached works guarantee the fulfillment of the proposed objectives and hypothesis in this document. However, several considerations, sub-experiments and hypothesis for future work proposals have been derived from the development of this work, and will be described in the subsection below 5.2.

5.2. Considerations and Future works

As several conclusions of this work have been arisen, similar future works are proposed. The first is addressed to the evaluation of a different emotional model. It is important to remark that this work evaluate the Ekman emotional model, which is based in the idea of discre-

te emotions with significantly different construct; since the main purpose of this work was focused in the development of a multimodal fusion information model taking advantage of Deep Learning approaches. However, the evaluation of a different emotional model (such as Plutchik Emotional model) could let us perform a comparative analysis between both models for computational algorithms. Additionally, a different emotional model (such as Russell circumflex model, positive activation-negative activation (PANA) or Pleasure, Arousal and Dominance (PAD) model), nevertheless, the learning task would completely change since these models are based on multidimensional continuous variables, transforming the classification task into a regression problem.

On the other hand, several parameters for the augmentation data strategy are critical for the performance of the model. An optimization study will be performed to find best parameters for the window and the overlapping size, using different algorithms (such as particle swarm optimization (PSO), Bayesian optimization (BO) or Random search). Gradient - based or stochastic algorithms could be considered as well, nevertheless, the computational cost for gradient computation and momentum estimation would have to be evaluated in order to avoid increasing the order of growth in algorithms.

Additionally, several works in the state of the art suggest that initialization of weights in the network is a important criteria to avoid divergence and increase generalization of the model. The main problem of initialization lays in the amount of data, the computational cost and time required to train a specific model. A shallow alternative could be fin in the usage of a pre-trained model (such as GoogleNet or AlexNet), however, the amount of parameters in the network would have to match, relieving flexibility for the model. The two proposal concerning model parameter optimization lay on the usage of a pre-trained model which include more randomly initialized parameters, and the usage of a large database to pre-train our proposed model before performing emotion recognition task.

Another aspect to take into account, which had been already mentioned previously is the gender dependence of subjects. Several works in the state of the art and previous works [103, 104] show that the gender consideration for emotion recognition is a relevant aspect to increase performance of the problem, since there are significantly differences between males and females, concerning facial expressions and voice corpora. This problem was not considered in this work due to the amount of samples from female participants (19 %); however, the usage of a larger database or the combination of various would let us perform an study of the classification task for comparative purposes (gender - dependent vs gender - independent). Additionally, another proposal consist in a single training model for one subject for emotion recognition would be analyzed, with the aim of performing a study on the creation of adaptive models for immersive environments to improve human - machine interactions.

Appendix 1: Speech Emotion Recognition Based on a Recurrent Neural Network Classification Model

Rubén D. Fonnegra and Gloria M. Díaz
In: Cheok A., Inami M., Romão T. (eds) Advances in Computer Entertainment Technology
Lecture Notes in Computer Science, Springer, Cham. 2018

Appendix 2: Deep Learning Based Video Spatio-Temporal Modeling for Emotion Recognition

Rubén D. Fonnegra and Gloria M. Díaz
In: Masaaki Kurosu (ed) Human-Computer Interaction: Theories, Methods and Human Issues (Part I)
Lecture Notes in Computer Science, Springer, Cham. 2018

Appendix 3: Speech Emotion Recognition Integrating Paralinguistic Features and Auto-encoders in a Deep Learning Model

Rubén D. Fonnegra and Gloria M. Díaz
In: Masaaki Kurosu (ed) Human-Computer Interaction: Theories, Methods and Human Issues (Part I)
Lecture Notes in Computer Science, Springer, Cham. 2018

Appendix 4:Performance comparison of deep learning frameworks in image classification problems using convolutional and recurrent networks

Rubén D. Fonnegra and Bryan Blair and Gloria M. Díaz
In: 2017 IEEE Colombian Conference on Communications and Computing (COLCOM)
IEEE Xplorer, 2017

Appendix 5: MSpecFace: A Dataset for Facial Recognition in the Visible, Ultra Violet and Infrared Spectra

Rubén D. Fonnegra and Alexander Molina and Andrés F. Pérez-Zapata and Gloria M. Díaz
In: Botto-Tobar M., Esparza-Cruz N., León-Acurio J., Crespo-Torres N., Beltrán-Mora M. (eds) Technology Trends.
Communications in Computer and Information Science, Springer, Cham. 2017.

Appendix 6: Automatic Face Recognition in Thermal Images Using Deep Convolutional Neural Networks

Rubén D. Fonnegra and Andrés F. Cardona-Escobar and Andrés F. Pérez-Zapata and Gloria M. Díaz
In: XVII Latin American Conference on Automatic Control CLCA 2016. Universidad EAFIT. 2016.

Bibliography

[1] R. W. Picard *et al.*, "Affective computing," 1995.

[2] X. Zhou and W. Shen, "Research on interactive device ergonomics designed for elderly users in the human-computer interaction," *International Journal of Smart Home*, vol. 10, no. 2, pp. 49–62, 2016.

[3] F. Balducci, C. Grana, and R. Cucchiara, "Affective level design for a role-playing videogame evaluated by a brain–computer interface and machine learning methods," *The Visual Computer*, vol. 33, no. 4, pp. 413–427, 2017.

[4] A. Bartsch and T. Hartmann, "The role of cognitive and affective challenge in entertainment experience," *Communication Research*, vol. 44, no. 1, pp. 29–53, 2017.

[5] C. A. Corneanu, M. O. Simón, J. F. Cohn, and S. E. Guerrero, "Survey on rgb, 3d, thermal, and multimodal approaches for facial expression recognition: History, trends, and affect-related applications," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 8, pp. 1548–1568, 2016.

[6] G. Bernal and P. Maes, "Emotional beasts: Visually expressing emotions through avatars in vr," in *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, pp. 2395–2402, ACM, 2017.

[7] I. Mavridou, J. T. McGhee, M. Hamedi, M. Fatoorechi, A. Cleal, E. Ballaguer-Balester, E. Seiss, G. Cox, and C. Nduka, "Faceteq interface demo for emotion expression in vr," in *Virtual Reality (VR), 2017 IEEE*, pp. 441–442, IEEE, 2017.

[8] P. Ekman, W. V. Freisen, and S. Ancoli, "Facial signs of emotional experience.," *Journal of personality and social psychology*, vol. 39, no. 6, p. 1125, 1980.

[9] Y. Wu, E. Y. Chang, K. C.-C. Chang, and J. R. Smith, "Optimal multimodal fusion for multimedia data analysis," in *Proceedings of the 12th annual ACM international conference on Multimedia*, pp. 572–579, ACM, 2004.

[10] T. Liu, L. Han, L. Ma, and D. Guo, "Audio-based deep music emotion recognition," in *AIP Conference Proceedings*, vol. 1967, p. 040021, AIP Publishing, 2018.

[11] D. Torres-Boza, M. C. Oveneke, F. Wang, D. Jiang, W. Verhelst, and H. Sahli, "Hierarchical sparse coding framework for speech emotion recognition," *Speech Communication*, vol. 99, pp. 80 – 89, 2018.

[12] X. Xia, J. Liu, T. Yang, D. Jiang, W. Han, and H. Sahli, "Video emotion recognition using hand-crafted and deep learning features," in *2018 First Asian Conference on Affective Computing and Intelligent Interaction (ACII Asia)*, pp. 1–6, May 2018.

[13] R. D. Fonnegra, B. Blair, and G. M. Díaz, "Performance comparison of deep learning frameworks in image classification problems using convolutional and recurrent networks," in *Communications and Computing (COLCOM), 2017 IEEE Colombian Conference on*, pp. 1–6, IEEE, 2017.

[14] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, pp. 1097–1105, 2012.

[15] J. Yan, W. Zheng, Z. Cui, C. Tang, T. Zhang, and Y. Zong, "Multi-cue fusion for emotion recognition in the wild," *Neurocomputing*, 2018.

[16] B. Sun, S. Cao, J. He, and L. Yu, "Affect recognition from facial movements and body gestures by hierarchical deep spatio-temporal features and fusion strategy," *Neural Networks*, vol. 105, pp. 36–51, 2018.

[17] L. F. Barrett, "Solving the emotion paradox: Categorization and the experience of emotion," *Personality and Social Psychology Review*, vol. 10, no. 1, pp. 20–46, 2006. PMID: 16430327.

[18] R. Plutchik, "Emotions: A general psychoevolutionary theory," *Approaches to emotion*, vol. 1984, pp. 197–219, 1984.

[19] P. Ekman, "Cross-cultural studies of facial expression," *Darwin and facial expression: A century of research in review*, vol. 169222, p. 1, 1973.

[20] P. Ekman, "An argument for basic emotions," *Cognition & emotion*, vol. 6, no. 3-4, pp. 169–200, 1992.

[21] P. Ekman and W. V. Friesen, "Constants across cultures in the face and emotion.," *Journal of personality and social psychology*, vol. 17, no. 2, p. 124, 1971.

[22] P. Ekman and W. V. Friesen, "Unmasking the face: A guide to recognizing emotions from facial clues.," 1975.

[23] P. Ekman and W. V. Friesen, "Measuring facial movement," *Environmental psychology and nonverbal behavior*, vol. 1, no. 1, pp. 56–75, 1976.

[24] R. Plutchik, *The circumplex as a general model of the structure of emotions and personality.* American Psychological Association, 1997.

[25] A. Mehrabian, "Framework for a comprehensive description and measurement of emotional states.," *Genetic, social, and general psychology monographs*, 1995.

[26] D. Watson and A. Tellegen, "Toward a consensual structure of mood.," *Psychological bulletin*, vol. 98, no. 2, p. 219, 1985.

[27] D. Watson, L. A. Clark, and A. Tellegen, "Development and validation of brief measures of positive and negative affect: the panas scales.," *Journal of personality and social psychology*, vol. 54, no. 6, p. 1063, 1988.

[28] B. Lance and S. Marsella, "Glances, glares, and glowering: how should a virtual human express emotion through gaze?," *Autonomous Agents and Multi-Agent Systems*, vol. 20, no. 1, p. 50, 2010.

[29] R. Plutchik, "The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice," *American scientist*, vol. 89, no. 4, pp. 344–350, 2001.

[30] N. Fragopanagos and J. G. Taylor, "Emotion recognition in human–computer interaction," *Neural Networks*, vol. 18, no. 4, pp. 389–405, 2005.

[31] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. G. Taylor, "Emotion recognition in human-computer interaction," *IEEE Signal processing magazine*, vol. 18, no. 1, pp. 32–80, 2001.

[32] R. L. Hazlett and S. Y. Hazlett, "Emotional response to television commercials: Facial emg vs. self-report," *Journal of Advertising Research*, vol. 39, pp. 7–24, 1999.

[33] L.-C. Yu, J.-L. Wu, P.-C. Chang, and H.-S. Chu, "Using a contextual entropy model to expand emotion words and their intensity for the sentiment classification of stock market news," *Knowledge-Based Systems*, vol. 41, pp. 89–97, 2013.

[34] O. Golan, E. Ashwin, Y. Granader, S. McClintock, K. Day, V. Leggett, and S. Baron-Cohen, "Enhancing emotion recognition in children with autism spectrum conditions: An intervention using animated vehicles with real emotional faces," *Journal of autism and developmental disorders*, vol. 40, no. 3, pp. 269–279, 2010.

[35] M. B. Harms, A. Martin, and G. L. Wallace, "Facial emotion recognition in autism spectrum disorders: a review of behavioral and neuroimaging studies," *Neuropsychology review*, vol. 20, no. 3, pp. 290–322, 2010.

[36] M. N. Pavuluri, M. M. O'Connor, E. Harral, and J. A. Sweeney, "Affective neural circuitry during facial emotion processing in pediatric bipolar disorder," *Biological psychiatry*, vol. 62, no. 2, pp. 158–167, 2007.

[37] M. L. Phillips, C. D. Ladouceur, and W. C. Drevets, "A neural model of voluntary and automatic emotion regulation: implications for understanding the pathophysiology and neurodevelopment of bipolar disorder," 2008.

[38] M. M. Bundele and R. Banerjee, "Detection of fatigue of vehicular driver using skin conductance and oximetry pulse: a neural network approach," in *Proceedings of the 11th International Conference on Information Integration and web-based applications & services*, pp. 739–744, ACM, 2009.

[39] C. Li, C. Xu, and Z. Feng, "Analysis of physiological for emotion recognition with the irs model," *Neurocomputing*, vol. 178, pp. 103–111, 2016.

[40] K. Chang, K. Bowyer, and P. Flynn, "Face recognition using 2d and 3d facial data," in *ACM Workshop on Multimodal User Authentication*, pp. 25–32, Citeseer, 2003.

[41] A. De and A. Saha, "A comparative study on different approaches of real time human emotion recognition based on facial expression detection," in *Computer Engineering and Applications (ICACEA), 2015 International Conference on Advances in*, pp. 483–487, IEEE, 2015.

[42] A. Basu, A. Routray, S. Shit, and A. K. Deb, "Human emotion recognition from facial thermal image based on fused statistical feature and multi-class svm," in *India Conference (INDICON), 2015 Annual IEEE*, pp. 1–5, IEEE, 2015.

[43] M. Pantic and I. Patras, "Dynamics of facial expression: Recognition of facial actions and their temporal segments from face profile image sequences," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 36, no. 2, pp. 433–449, 2006.

[44] H. Gunes, C. Shan, S. Chen, and Y. Tian, "Bodily expression for automatic affect recognition," *Emotion Recognition: A Pattern Analysis Approach*, pp. 343–377, 2015.

[45] F. E. Pollick, H. M. Paterson, A. Bruderlin, and A. J. Sanford, "Perceiving affect from arm movement," *Cognition*, vol. 82, no. 2, pp. B51–B61, 2001.

[46] A. Camurri, I. Lagerlöf, and G. Volpe, "Recognizing emotion from dance movement: comparison of spectator recognition and automated techniques," *International journal of human-computer studies*, vol. 59, no. 1, pp. 213–225, 2003.

[47] L. Cui, S. Li, and T. Zhu, "Emotion detection from natural walking," in *International Conference on Human Centered Computing*, pp. 23–33, Springer, 2016.

[48] Z. Guendil, Z. Lachiri, C. Maaoui, and A. Pruski, "Emotion recognition from physiological signals using fusion of wavelet based features," in *Modelling, Identification and Control (ICMIC), 2015 7th International Conference on*, pp. 1–6, IEEE, 2015.

[49] J. Preethi, M. Sreeshakthy, and A. Dhilipan, "A survey on eeg based emotion analysis using various feature extraction techniques," *International Journal of Science, Engineering and Technology Research (IJSETR)*, vol. 3, no. 11, 2014.

[50] J. D. Echeverry and M. M. Pérez, "Reconocimiento de emociones en el habla," *Revista Tecno Lógicas*, no. 21, pp. 113–130, 2008.

[51] J. Deng, Z. Zhang, E. Marchi, and B. Schuller, "Sparse autoencoder-based feature transfer learning for speech emotion recognition," in *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on*, pp. 511–516, IEEE, 2013.

[52] S. Poria, E. Cambria, A. Hussain, and G.-B. Huang, "Towards an intelligent framework for multimodal affective data analysis," *Neural Networks*, vol. 63, pp. 104–116, 2015.

[53] M. S. Hossain, G. Muhammad, M. F. Alhamid, B. Song, and K. Al-Mutib, "Audio-visual emotion recognition using big data towards 5g," *Mobile Networks and Applications*, vol. 21, no. 5, pp. 753–763, 2016.

[54] S. Mirsamadi, E. Barsoum, and C. Zhang, "Automatic speech emotion recognition using recurrent neural networks with local attention," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*, pp. 2227–2231, IEEE, 2017.

[55] S.-H. Wang, P. Phillips, Z.-C. Dong, and Y.-D. Zhang, "Intelligent facial emotion recognition based on stationary wavelet entropy and jaya algorithm," *Neurocomputing*, 2017.

[56] H. Yan, "Collaborative discriminative multi-metric learning for facial expression recognition in video," *Pattern Recognition*, 2017.

[57] C. Mühl, B. Allison, A. Nijholt, and G. Chanel, "A survey of affective brain computer interfaces: principles, state-of-the-art, and challenges," *Brain-Computer Interfaces*, vol. 1, no. 2, pp. 66–84, 2014.

[58] S. Wang and Q. Ji, "Video affective content analysis: a survey of state-of-the-art methods," *IEEE Transactions on Affective Computing*, vol. 6, no. 4, pp. 410–430, 2015.

[59] S. Dobrišek, R. Gajšek, F. Mihelič, N. Pavešić, and V. Štruc, "Towards efficient multi-modal emotion recognition," *International Journal of Advanced Robotic Systems*, vol. 10, no. 1, p. 53, 2013.

[60] S. Zhalehpour, Z. Akhtar, and C. E. Erdem, "Multimodal emotion recognition based on peak frame selection from video," *Signal, Image and Video Processing*, vol. 10, no. 5, pp. 827–834, 2016.

[61] J. Fu, Q. Mao, J. Tu, and Y. Zhan, "Multimodal shared features learning for emotion recognition by enhanced sparse local discriminative canonical correlation analysis," *Multimedia Systems*, pp. 1–11, 2017.

[62] M. Y. Alva, M. Nachamai, and J. Paulose, "A comprehensive survey on features and methods for speech emotion detection," in *Electrical, Computer and Communication Technologies (ICECCT), 2015 IEEE International Conference on*, pp. 1–6, IEEE, 2015.

[63] M. Swain, A. Routray, and P. Kabisatpathy, "Databases, features and classifiers for speech emotion recognition: a review," *International Journal of Speech Technology*, pp. 1–28, 2018.

[64] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org.

[65] Y. Bengio *et al.*, "Learning deep architectures for ai," *Foundations and trends® in Machine Learning*, vol. 2, no. 1, pp. 1–127, 2009.

[66] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[67] M. Raghu, B. Poole, J. Kleinberg, S. Ganguli, and J. Sohl-Dickstein, "On the expressive power of deep neural networks," *arXiv preprint arXiv:1606.05336*, 2016.

[68] J. G. H. Gutiérrez, C. A. Peña-Solórzano, C. L. Garzón-Castro, F. A. Prieto-Ortiz, and J. G. Ayala-Garzón, "Hacia el manejo de una herramienta por un robot nao usando programación por demostración," *Revista Tecno Lógicas*, vol. 17, no. 33, pp. 65–76, 2014.

[69] J. Dean, G. Corrado, R. Monga, K. Chen, M. Devin, M. Mao, A. Senior, P. Tucker, K. Yang, Q. V. Le, *et al.*, "Large scale distributed deep networks," in *Advances in neural information processing systems*, pp. 1223–1231, 2012.

[70] Z.-H. Ling, L. Deng, and D. Yu, "Modeling spectral envelopes using restricted boltzmann machines and deep belief networks for statistical parametric speech synthesis," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 10, pp. 2129–2139, 2013.

[71] T. N. Sainath, B. Kingsbury, B. Ramabhadran, P. Fousek, P. Novak, and A.-r. Mohamed, "Making deep belief networks effective for large vocabulary continuous speech recognition," in *Automatic Speech Recognition and Understanding (ASRU), 2011 IEEE Workshop on*, pp. 30–35, IEEE, 2011.

[72] H.-m. Shim and S. Lee, "Multi-channel electromyography pattern classification using deep belief networks for enhanced user experience," *Journal of Central South University*, vol. 22, no. 5, pp. 1801–1808, 2015.

[73] H.-I. Suk and D. Shen, "Deep learning-based feature representation for ad/mci classification," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 583–590, Springer, 2013.

[74] N. Nishida and H. Nakayama, "Multimodal gesture recognition using multi-stream recurrent neural network," in *Pacific-Rim Symposium on Image and Video Technology*, pp. 682–694, Springer, 2015.

[75] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural networks*, vol. 61, pp. 85–117, 2015.

[76] J. L. Elman, "Finding structure in time," *Cognitive science*, vol. 14, no. 2, pp. 179–211, 1990.

[77] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[78] A. Gunawan and H. Lau, "Fine-tuning algorithm parameters using the design of experiments approach," *Learning and Intelligent Optimization*, pp. 278–292, 2011.

[79] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.

[80] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[81] G. Hinton and T. Tieleman, "Lecture 6.5 - rmsprop," tech. rep., COURSERA: Neural Networks for Machine Learning, 2010.

[82] J. Duchi, E. Hazan, and Y. Singer, "Adaptive subgradient methods for online learning and stochastic optimization," *Journal of Machine Learning Research*, vol. 12, no. Jul, pp. 2121–2159, 2011.

[83] M. Wegrzyn, M. Vogt, B. Kireclioglu, J. Schneider, and J. Kissler, "Mapping the emotional face. how individual face parts contribute to successful emotion recognition," *PLOS ONE*, vol. 12, pp. 1–15, 05 2017.

[84] L. Abramson, I. Marom, R. Petranker, and H. Aviezer, "Is fear in your head? a comparison of instructed and real-life expressions of emotion in the face and body.," *Emotion*, vol. 17, no. 3, p. 557, 2017.

[85] M. Liu, D. Fan, X. Zhang, and X. Gong, "Human emotion recognition based on galvanic skin response signal feature selection and svm," in *2016 International Conference on Smart City and Systems Engineering (ICSCSE)*, pp. 157–160, Nov 2016.

[86] W. L. Zheng, J. Y. Zhu, and B. L. Lu, "Identifying stable patterns over time for emotion recognition from eeg," *IEEE Transactions on Affective Computing*, pp. 1–1, 2017.

[87] S. G. Koolagudi, R. Reddy, and K. S. Rao, "Emotion recognition from speech signal using epoch parameters," in *2010 International Conference on Signal Processing and Communications (SPCOM)*, pp. 1–5, July 2010.

[88] S. G. Koolagudi, N. Kumar, and K. S. Rao, "Speech emotion recognition using segmental level prosodic analysis," in *2011 International Conference on Devices and Communications (ICDeCom)*, pp. 1–5, Feb 2011.

[89] D. T. Robinson, J. Clay-Warner, C. D. Moore, T. Everett, A. Watts, T. N. Tucker, and C. Thai, *Toward an Unobtrusive Measure of Emotion During Interaction: Thermal Imaging Techniques*, pp. 225–266.

[90] J. Clay-Warner and D. T. Robinson, "Infrared thermography as a measure of emotion response," *Emotion Review*, vol. 7, no. 2, pp. 157–162, 2015.

[91] J. Yan, W. Zheng, Q. Xu, G. Lu, H. Li, and B. Wang, "Sparse kernel reduced-rank regression for bimodal emotion recognition from facial expression and speech," *IEEE Transactions on Multimedia*, vol. 18, no. 7, pp. 1319–1329, 2016.

[92] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM international conference on Multimedia*, pp. 1459–1462, ACM, 2010.

[93] B. W. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. A. Müller, S. S. Narayanan, *et al.*, "The interspeech 2010 paralinguistic challenge.," in *Interspeech*, vol. 2010, pp. 2795–2798, 2010.

[94] P. Viola and M. J. Jones, "Robust real-time face detection," *International journal of computer vision*, vol. 57, no. 2, pp. 137–154, 2004.

[95] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th international conference on machine learning (ICML-10)*, pp. 807–814, 2010.

[96] K. Jarrett, K. Kavukcuoglu, Y. LeCun, *et al.*, "What is the best multi-stage architecture for object recognition?," in *Computer Vision, 2009 IEEE 12th International Conference on*, pp. 2146–2153, IEEE, 2009.

[97] G. E. Dahl, T. N. Sainath, and G. E. Hinton, "Improving deep neural networks for lvcsr using rectified linear units and dropout," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pp. 8609–8613, IEEE, 2013.

[98] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *arXiv preprint arXiv:1412.3555*, 2014.

[99] O. Martin, I. Kotsia, B. Macq, and I. Pitas, "The enterface'05 audio-visual emotion database," in *Data Engineering Workshops, 2006. Proceedings. 22nd International Conference on*, pp. 8–8, IEEE, 2006.

[100] D. Datcu and L. J. Rothkrantz, "Emotion recognition using bimodal data fusion," in *Proceedings of the 12th International Conference on Computer Systems and Technologies*, pp. 122–128, ACM, 2011.

[101] K.-C. Huang, H.-Y. S. Lin, J.-C. Chan, and Y.-H. Kuo, "Learning collaborative decision-making parameters for multimodal emotion recognition," in *Multimedia and Expo (ICME), 2013 IEEE International Conference on*, pp. 1–6, IEEE, 2013.

[102] S. Zhang, S. Zhang, T. Huang, and W. Gao, "Speech emotion recognition using deep convolutional neural network and discriminant temporal pyramid matching," *IEEE Transactions on Multimedia*, vol. 20, pp. 1576–1590, June 2018.

[103] R. D. Fonnegra and G. M. Díaz, "Speech emotion recognition based on a recurrent neural network classification model," in *Cheok A., Inami M., Romão T. (eds) Advances in Computer Entertainment Technology*, vol. 10714, pp. 882–892, Lecture Notes in Computer Science, Springer, Cham, 2018.

[104] R. D. Fonnegra and G. M. Díaz, "Deep learning based video spatio-temporal modeling for emotion recognition (to appear)," in *Masaaki Kurosu (ed) Human-Computer Interaction: Theories, Methods and Human Issues (Part I)*, vol. 10901, Lecture Notes in Computer Science, Springer, Cham, 2018.

[105] R. D. Fonnegra and G. M. Díaz, "Speech emotion recognition integrating paralinguistic features and auto-encoders in a deep learning model (to appear)," in *Masaaki Kurosu (ed) Human-Computer Interaction: Theories, Methods and Human Issues (Part I)*, vol. 10901, Lecture Notes in Computer Science, Springer, Cham, 2018.

[106] S. Zhang, S. Zhang, T. Huang, W. Gao, and Q. Tian, "Learning affective features with a hybrid deep model for audio-visual emotion recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–1, 2017.

[107] R. D. Fonnegra, A. Molina, A. F. Pérez-Zapata, and G. M. Díaz, "Mspecface: A dataset for facial recognition in the visible, ultra violet and infrared spectra," in *Botto-Tobar M., Esparza-Cruz N., León-Acurio J., Crespo-Torres N., Beltrán-Mora M. (eds) Technology Trends. CITT 2017*, vol. 798, pp. 160–170, Communications in Computer and Information Science, Springer, Cham, 2017.

[108] R. D. Fonnegra, A. F. Cardona-Escobar, A. F. Pérez-Zapata, and G. M. Díaz, "Automatic face recognition in thermal images using deep convolutional neural networks," in *XVII Latin American Conference on Automatic Control CLCA 2016*, pp. 2–6, Universidad EAFIT, 2016.