MULTIMODAL INFORMATION FUSION FOR DECEPTION DETECTION IN VIDEOS

By
Rodrigo Rill García

B.Sc., ITESM

A Dissertation
Submitted to the Program in Computer Science,
Computer Science Department
in partial fulfillment of the requirements for the degree of

MASTER IN COMPUTER SCIENCE

at the

National Institute of Astrophysics, Optics and Electronics
August, 2019
Tonanzintla, Puebla

Advisors:

Ph.D. Hugo Jair Escalante Balderas
Ph.D. Luis Villaseñor Pineda
Principal Research Scientists
Computer Science Department
INAOE

MULTIMODAL INFORMATION FUSION FOR
DECEPTION DETECTION IN VIDEOS

by Rodrigo Rill García, M.Sc.
Advisors: Hugo Jair Escalante Balderas, Luis Villaseñor Pineda, Verónica Reyes Meza

Abstract

Deception (the action of deliberately causing someone to believe something that is not true) can have many different repercussions and it is inherent to our daily life. However, detecting lies is inherently complex for humans despite our continuous contact with them. Due to this, not only there is uncertainty on which features could or should be used as cues for (automatic) deception detection, but also labeled data is scarce. In this thesis, we explore features that can be automatically extracted from videos for affective computing and study their performance for the specific task of deception detection in videos. Additionally, we present a study on different multimodal fusion methods meant to improve the individual performance of the different feature sets extracted, including a novel set of methods based on boosting. For this study, high-level features are extracted using open automatic tools on the visual, acoustical and textual modalities, respectively. Experiments are conducted using a real-life trial dataset as well as a novel Mexican deception detection dataset using Spanish as the spoken language. Summarizing, in this thesis we study high-level features and perform a multimodal complementarity analysis between them to support the idea that multimodal fusion is a good approach for deception detection; with such evidence, we present one of the first works focused on multimodal deception detection methods further than early concatenation of features, including the first study (to the best of our knowledge) on automatic deception detection in clips from Mexican subjects speaking Spanish.

Resumen

El engaño (la acción de causar deliberadamente que alguien crea algo que no es cierto) puede tener varias diferentes repercusiones y es parte inherente de nuestra vida diaria. Sin embargo, detectar mentiras es inherentemente complejo para los humanos a pesar de nuestro continuo contacto con ellas. Debido a esto, no sólo hay incertidumbre en cuáles atributos podrían o deberían ser usados como pistas para detección (automática) de engaño, sino que también los datos etiquetados son escasos. En esta tesis, exploramos atributos que pueden ser automáticamente extraídos de videos para cómputo afectivo y estudiamos su desempeño para la tarea específica de detección de engaño en videos. Además, presentamos un estudio de diferentes métodos de fusión multimodal destinados a mejorar el desempeño individual de los diferentes conjuntos de atributos extraídos, incluyendo un nuevo conjunto de métodos basados en *boosting*. Para este estudio, atributos de alto nivel son extraídos usando herramientas automáticas de uso libre en las modalidades visual, acústica y textual, respectivamente. Los experimentos se llevan a cabo usando una base de datos de juicios de la vida real así como una nueva base de datos mexicana para detección de engaño usando el español como la lengua hablada. Resumiendo, en esta tesis estudiamos atributos de alto nivel y realizamos un análisis de complementariedad multimodal entre ellos para apoyar la idea de que la fusión multimodal es un buen acercamiento para la detección de engaño; con dicha evidencia, presentamos uno de los primeros trabajos enfocados en detección de engaño multimodal más allá de una concatenación temprana de atributos, incluyendo el primer estudio (hasta donde sabemos) en detección automática de engaño en videoclips de sujetos mexicanos hablando español.

ACKNOWLEDGMENT

I am grateful to the Consejo Nacional de Ciencia y Tecnología in Mexico (CONACYT) for the financial support that they gave me in order to finish my Master's thesis under the framework of project CB-2015-01-257383. I'm grateful too to Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE) and its Department of Computer Science, for the support provided in order to successfully finish my Master's program.

I also want to thank my Advisors, Hugo and Luis, for their kind help all along my thesis and during the academic courses they taught during my first Master's year. They supported my crazy ideas while still helping me to keep my feet on the ground in order to successfully conduct my research, providing me with valuable advises for both my thesis and my future career. I need to thank also to Verónica Reyes, who kindly accepted to be my external advisor and gave me a wider perspective of the research topic from the point of view of Psychology. Additionally, her support was invaluable in terms of the database collected thanks to her for deception detection in Mexican video clips.

I am also grateful to my colleagues from Tecnológico de Monterrey in Puebla, specially my immediate boss Juan de Dios Calderón, who supported me to get my Master's degree while keeping my job as adjunct professor in my alma mater. Specially, I wish to acknowledge Hugo González and Roberto Mora (former colleagues and great teachers during my college years) for their support during the process to get accepted into the Master's program and afterwards.

Finally, I wish to thank the friends who led me to the wonderful world of Computer Science, particularly Computer Vision and Machine Learning. This thesis is a consequence too of our nerdy talks.

Dedication

This thesis is dedicated to my family, who
have always supported me to
fulfill my goals.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter One

INTRODUCTION

According to the Oxford dictionary, deception is the action of deceiving someone, that is, “deliberately cause (someone) to believe something that is not true, especially for personal gain”. Deceptive behavior is part of our daily lives and, while there are many motives behind it, consequences go from innocuous cases to severe situations, specially when lies are told to escape unfavorable/undesirable situations (e.g. the statement of a witness in a judicial trial). The underlying problem comes when we analyze the human ability to detect lies; according to (Bond Jr and DePaulo, 2006), the average accuracy for this task -without special aids- is 54%; that is, just slightly better than random guessing.

1.1 Motivation

1.1.1 Non-automatic approaches for deception detection

To improve the natural human ability for deception detection, many strategies have been used. One well-known method is the Polygraph, which can be defined as a physiological method. However, it has many drawbacks; besides being impractical due to the need of skin-contact and a human expert, many counter-measures can be taken to fool those tests. Also in the category of phys-

iological methods, we can find the Magnetic Resonance Imaging; however, the cost of equipment for this method as well as its nature result in an impractical option (Farah et al., 2014).

When it comes to methods that take advantage of physical reactions without special equipment, another approach consists on the analysis of behavioral cues under the hypothesis that there are inherent unconscious behaviors associated to deception. Under this scope, the most relevant foundational work has been published by Paul Ekman. According to him (P. Ekman, 2009), facial micro-expressions reveal emotional information that subjects might wish to undercover. But then again, this approach has two major drawbacks: first at all, those hints (micro-expressions) are difficult to detect by untrained persons; furthermore, identifying the existence of such hidden emotions can be misleading of deception as stated by Ekman too (P. Ekman, 2003), since trying to suppress certain emotions is not necessarily due to a deceptive intention.

Despite those drawbacks, this approach gives evidence on the possibility to predict deception through cues obtained with non-invasive methods. If such cues could be obtained automatically, and expert human-knowledge about those cues could be replaced by computational systems, it would be possible to develop a framework for automatic deception detection.

1.1.2 Automatic approaches for deception detection

Aiming for such a framework, computer science has proposed alternative machine learning based approaches. A first idea that comes handy is to combine human expertise with classifiers: while humans are in charge of manual feature extraction, the data obtained from such human experts is used to draw a conclusion. Under this scheme, the most popular features to analyze are micro-expressions; according to (Wu et al., 2018), state of art methods for this task use these features.

However, visual attributes are not the only ones used for deception detection. Deceptive behavior is not limited to “live” lies, and this is particularly true nowadays because of the Internet and social media (this is basically a new context for human interaction, where lying about one’s identity is easier because of the lack of information). A closely related field of research is detection of unfaithful information via text extracted from the web. Many approaches have been used for this task; however, research have shown the effectiveness of features derived from text analysis to identify deceptive content from speech (Pérez-Rosas et al., 2015). Under this idea, a natural language processing approach can be used for videos by extracting transcripts from them; in other words, the underlying hypothesis is that deceptive speech can be identified with the help of semantic analysis using automatic methods.

However, this leads to a new idea: analyzing the non-semantic part of speech (e.g. extracting features that can’t be obtained from text like pauses, rhythm, intonation, tone, etc. to give some examples). Those are useful for humans to identify the emotional state of a speaker, and some works have used this as basis for automatic emotion recognition using acoustic features, like the one of (Espinosa and García, 2009). Also, the non-semantic part of speech is useful for humans to detect lies; in fact, some experiments by (Wu et al., 2018) show how humans are better to identify untruthful testimonies by hearing them rather than reading them. Furthermore, in their experiments, having access to the video itself rather than just audio does not really improve the person’s ability to detect false testimonies. When it comes to automatic deception detection, they find MFCC (Mel-frequency Cepstral Coefficients) to be useful for the task; MFCC have been used previously for many speech recognition tasks, including emotion recognition (Vogt and André, 2005).

Summarizing, there is a wide range of approaches for deception detection involving different sources of information, including methods meant for automatic deception detection using machine learning techniques. However, de-

spite the progress reached by these works, deception detection is still an open problem of interest for many areas.

1.2 Justification

So far, we have discussed different approaches based on single modalities to show the motivation behind automatic classification of deception. However, the particular interest of the current thesis is exploring the multimodal fusion of information. When it comes to videos, we are dealing with inherent multimodal data from which we can extract several feature sets. As we will explore in the next chapter, multimodal analysis is the current trend for deception detection in videos, where different types of features are combined to reach better results than those obtained by using them separately. Intuitively, a more informed decision is a better decision, so combining multiple information modalities should improve automatic classification.

However, it is well known that different representations (extracted features) of a single phenomenon can lead to better or worse results using the same type of classifier; additionally, long feature vectors are unlikely to provide satisfactory results when the number of training instances is small with respect to the length of the feature vectors. Therefore, the way in which features are extracted and how they are joined (fused) are two important research fields to take into account for multimodal classification, specially when available datasets contain few training instances (such as in the case of deception detection).

1.3 Problem Statement

Deception detection implies many problems and, as such, automatic detection of lies does too. The particular interest of this thesis resides on detection of

deceit without invasive methods, and that is why information extracted from videos is attractive.

When it comes to videos, there are two main temporal approaches for analysis: real time and forensic. While the first one is more useful for “in the wild” scenarios, the later one is better for controlled scenarios like criminal interrogation. Because of this and taking into account the related work, the focus for this thesis is a forensic analysis of videos from potential liars while speaking. Particularly, we aim to classify a video as deceptive if the speech as a whole intends to convince someone of something that is not true as a whole (e.g. even if the speaker is truthful during most of the video, the statement is deceptive if there are some parts intended to mislead the listener).

In order to do this, we want to build a predictive model from existing databases. Under this scheme, however, nowadays there are two great problems we need to address:

- Research on deception suggests there are many different cues to detect it on different domains and contexts
- Databases on this task are scarce and most of them were created under simulated conditions

To deal with the first problem, multimodal analysis seems useful since information is gotten from many perspectives: if different modalities provide different cues on deception, there may be a wider range of contexts that can be treated while using different combinations of multimodal features if they are chosen and combined properly. Common approaches involve simply concatenating the different features into a single vector for classification or using the decisions from each feature set as a new feature set for classification. However, these strategies might not be meaningful since such feature sets can be statistical not independent or have different statistical properties in general; therefore, a proper method of multimodal fusion goes beyond putting all data together.

Furthermore, the aiming is to create a system as general as possible within the limitations of available datasets. To do this, an important task is to learn cues for deception detection that are subject-independent (e.g. even if different persons lie differently, we should find cues that are shared among subjects), dealing then with the problem of scarce data to some extent.

1.3.1 Research questions

Under the scope of the aforementioned context, this thesis aims to explore the next questions:

- Is it possible to automatically extract high-level features that are useful for automatic deception detection?
- Given the temporal dimension of speech, how such features should be analyzed in order to deal with different length speeches?
- Given the multimodal nature of videos, is there complementarity between the features that can be extracted within and across modalities?
- Under the assumption that such complementarity exists, what is a proper fusion method to take advantage of the strengths of each feature set for deception detection?

1.3.2 Hypothesis

Despite the differences that exist between persons when it comes to lying, a multimodal computational system aimed to automatically detect deception in videos performs better when based on a non-trivial fusion rather than relying on single modalities or a simple early fusion of modalities.

1.3.3 Aims and goals

To provide a solution involving the two above mentioned problems, our general objective is *to develop methods for multimodal information fusion, inspired by current classifier ensemble methods, for automatic classification using high-level features in the task of binary deception detection in video recorded samples*.

Given the above mentioned aim, we need to complete some particular objectives:

- To define the possible modalities to extract from the available videos, as well as the different feature representations that can be gotten from them
- To evaluate such modalities separately according to their different extracted feature sets
- To develop a method for combining effectively the evaluated modalities
- To compare the results obtained from the method for deception detection with respect to other fusion strategies

1.4 Contribution of the thesis

In this thesis, we present the following contributions:

1. A study on high-level (interpretable by humans) feature sets that can be automatically extracted from videos for deception detection
2. An analysis of the complementarity between such features to provide evidence on the benefits that could be obtained from fusing them
3. A method based on LSTM networks and the k-means algorithm for multimodal encoding of variable length sequences into a fixed size vector

4. A study on multimodal fusion techniques inspired by classifier ensembles for deception detection in videos, including two novel methods based on boosting for multimodal fusion

5. A comparison between both single feature sets and fusions on two datasets

6. A novel dataset for deception detection in videos with Mexican subjects speaking Spanish (the first one of this kind to the best of our knowledge)

Additionally, based on the work done for this thesis, the following works were published:

1. *From Text to Speech: A Multimodal Cross-Domain Approach for Deception Detection*, at MIPPSNA @ ICPR 2018 (Rill-García, Rodrigo et al. (2018). "From Text to Speech: A Multimodal Cross-Domain Approach for Deception Detection". In: International Conference on Pattern Recognition. Springer, pp. 164–177.)

2. *High-level Features for Multimodal Deception Detection in Videos*, oral presentation at LatinX in AI Workshop @ ICML 2019

3. *High-Level Features for Multimodal Deception Detection in Videos*, at ChaLearn Looking at People series: Face Spoofing Attack Workshop and Competition @ CVPR 2019 (Rill-Garcia, Rodrigo et al. (2019). "High-Level Features for Multimodal Deception Detection in Videos". In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. –.)

1.5 General outline

The rest of this document is organized in the following way: Chapter 2 provides a general framework of the theoretical knowledge required to understand the

contents of this thesis; Chapter 3 provides an overview of the literature reviewed on state of the art works about deception detection in videos and multimodal information fusion in general; Chapter 4 aims to explain the different multimodal features extracted from videos for analysis in this thesis (the databases used for experiments are also described here); Chapter 5 presents an analysis of the performance of features from single modalities when used with a Machine Learning approach (as well as a study on the complementarity between this feature sets); Chapter 6 shows the results obtained when fusing the different modalities together using methods inspired by classifier ensembles, including two novel methods based on boosting described in this chapter; finally, Chapter 7 presents the conclusions derived from this work as well as its limitations and suggestions for future work.

Chapter Two

THEORETICAL FRAMEWORK

2.1 Deception

When it comes to research on deception, the figure of Paul Ekman stands out as a pioneer in the area, particularly under the scope of facial expressions and the psychology of emotion. Beyond the consequences of lying under determined circumstances, deception is an inherent component of human interaction (or, at least, the possibility of it). As stated by Ekman, “examining how and when people lie and tell the truth can help in understanding many human relationships. [...] Lying is such a central characteristic of life that better understanding of it is relevant to almost all human affairs” (P. Ekman, 2009).

For Ekman, lying and deceit are interchangeably words, and a key component is that a liar can choose not to lie. This is a key statement because it implies that giving untrue information is not equivalent to lying (e.g. a person can be untruthful without being a liar). To give a simple example, think on the next scenario: a child who believes in Santa Claus will ensure you that he is real; even if the information is false, the child is communicating you something that is real for him.

Therefore, for deception to occur, the liar must intend to mislead the victim (whatever it might be the motivation for such misleading). Furthermore, during the lie, the liar must know all the time the difference between lying and be-

ing truthful in the given case; additionally, the liar must know that lying is the decision that (s)he is deliberately taking.

However, deception is a two-sided interaction: not only there is a liar (or liars) but there is a victim (or victims). For deception to occur, the victim must be unaware of the attempt of misleading. Therefore, even if an actor in a movie tries to mislead you to make you believe he is the character in the story, as there is an implicit agreement between the actor and the spectators this is not an example of deception.

Deception, being a common phenomenon in humans, involves a set of higher cognitive functions. Research on such phenomenon aims to understand its underlying cognitive framework, but the ultimate ambition relies on the ability to detect deceptive behavior. Therefore, identifying valid indicators of it is the main focus on deception research (Gamer and Ambach, 2014).

2.1.1 Computational cognition and affective computing

Being a cognitive process, research on deception can be done under the scope of computational cognition (also known as computational psychology). According to *A Dictionary of Computing* published by Oxford University Press, computational psychology is “A discipline lying on the border between artificial intelligence and psychology. It is concerned with building computer models of human cognitive processes and is based on an analogy between the human mind and computer programs. The brain and computer are viewed as general-purpose symbol-manipulation systems, capable of supporting software processes, but no analogy is drawn at a hardware level”.

Therefore, in the field of deception detection it would be desirable to build a computer model able to detect lies based on indicators of deceptive behavior. The given model, if intended to work analogically to the human mind, should use indicators understandable for humans. However, humans have a low per-

formance on this task (just above random guessing as shown by the work of (Bond Jr and DePaulo, 2006)).

Nevertheless, humans have the ability to understand involuntary behaviors, such as for emotion recognition. This task has been explored under the scope of computational psychology too, particularly by the field known as affective computing. According to (Tao and Tan, 2005), “Affective computing is trying to assign computers the human-like capabilities of observation, interpretation and generation of affect features.”; therefore, it “concerns multidisciplinary knowledge background such as psychology, cognitive, physiology and computer sciences”.

As stated by (R. Ekman, 1997), facial micro-expressions are useful to undercover emotional information from people even if they try to hide it. This kind of information is useful because it can reveal the internal state of a person even if they try to fake it (e.g. mislead people about how they feel). Scenarios like this show the relation that can exist between sentiment analysis and deception detection, and it is actually a topic of interest as seen by the *2017 Looking at People ICCV Challenge - Fake vs. true facial emotion recognition* (Wan et al., 2017).

2.2 Machine learning

Challenges like the one mentioned above show the interest of detecting deceptive behavior with models learned automatically (e.g. possible indicators or *features* are used for detection without hard-coding explicit instructions). This is a perfect example of Machine Learning (term coined by Arthur Samuel in 1959 (Samuel, 2000) and often referred simply as ML), a subarea of Artificial Intelligence that studies algorithms and statistical models used by computer systems to perform a specific task without explicit instructions. Machine learning algorithms build mathematical models based on sample or *training* data in order

to make predictions or decisions relying on patterns and inference instead of being explicitly programmed (Bishop, 2006).

2.2.1 Supervised methods

For a machine learning task, data from real world must be codified into vectors. Applications in which the training data comprises input vectors along with their corresponding target vectors are known as supervised learning problems. Cases such as deception detection, in which the aim is to assign each input vector to one of two discrete categories (truthful or deceptive), are called *classification* problems (Bishop, 2006). Furthermore, cases like this one are referred as *binary classification* problems since there are just two possible outcomes from the classification task.

Support vector machines

Lets suppose a two-class classification problem using a linear model of the form:

$$y(x) = w^T \phi(x) + b \tag{2.1}$$

where $\phi(x)$ denotes a fixed feature-space transformation and b is a bias parameter. Supposing a training data set comprised by N input vectors $x_1, ..., x_N$, with corresponding target values $y_1, ..., y_N$ where $y_n \epsilon \{-1, 1\}$, a new data point x is classified according to the sign of y(x). For this model to work, we shall assume that the training data set is linearly separable in the feature space (e.g. there exists at least one choice of the parameters w and b such that a function of the form (2.1) satisfies $y(x) > 0$ for points having y=+1 and $y(x) < 0$ for points having y=$-$1 for all x_n). Therefore, we have a decision boundary where $y(x) = 0$.

In Support Vector Machines (SVM) the decision boundary is chosen to be the one for which a margin is maximized; such margin is chosen to be the

smallest distance between the decision boundary and any of the samples (the samples used to define margins are known as support vectors).

However, in problems where different class vectors overlap, a linear function doesn't work properly as a decision boundary. To deal with non-linear classification, SVM use a strategy called the *kernel trick* or *kernel substitution*. The simplest example of a kernel function is the identity mapping for the feature space so that $\phi(x) = x$, in which case $k(x, z) = x^T z$ (which is known as the linear kernel).

The general idea is that, if we have an algorithm formulated in such a way that the input vector x enters only in the form of scalar products, then we can replace that scalar product with some other choice of kernel, transforming the input vector from a n-dimensional space to a representation in a higher-dimensional one where all the training data can be linearly separable (without explicitly computing the coordinates of the vector in the new feature space).

Think of the polynomial kernel $K(x, z) = (x^T z + c)^n =< \phi(x), \phi(z) >$, where <a,b> is the inner product between vectors a and b. With the polynomial kernel, we can evaluate the similarity of vectors a and b in the feature-space defined by ϕ without actually calculating $\phi(a)$ and $\phi(b)$. This is important because, both at training and evaluation, SVM evaluates the inner product between support vectors and new instances to be classified.

For a better understanding of this section and SVM in general, the reader is referred to Chapter 7 of (Bishop, 2006).

2.3 Multimodal analysis

Modality refers to the way in which something happens or is experienced (e.g. we see objects, hear sounds, feel textures, smell odors, and taste flavors). When a phenomenon is studied using multiple modalities, it is told to be a multimodal research problem. Multimodal machine learning, therefore, aims to

build models that can process and relate information from multiple modalities instead of focusing on specific single modal applications (Baltrušaitis, Ahuja, and Morency, 2019). As a multimedia message composed of synchronized single streams (namely image and audio), a video is a perfect example of a multimodal phenomenon that can be used for multimodal machine learning tasks. This is an increasing field that faces many challenges, namely: representation, translation, alignment, fusion, and co-learning.

For this work we will focus on the fusion challenge, which consists on joining information from two or more modalities to perform a prediction (e.g. for audio-visual speech recognition, the visual description of the lip motion is fused with the speech signal to predict spoken words). The interest in multimodal fusion arises from three main benefits: 1) having access to multiple modalities observing the same phenomenon may allow for more robust predictions; 2) having access to multiple modalities might allow us to capture complementary information (i.e. something that is not visible in individual modalities on their own); 3) a multimodal system can still operate when one of the modalities is missing (e.g. recognizing emotions from the visual signal when the person is not speaking) (Baltrušaitis, Ahuja, and Morency, 2019).

Within this scope, there are three typical approaches with respect to the fusion level: early, late and hybrid (or intermediate). From those, we will discuss the first two as the last one is, actually, a hybrid between both.

2.3.1 Early fusion

The early fusion approach (also known as feature level) consists on combining all the features extracted from input data and use this new representation as a single input for a single analysis unit that performs the analysis task. Here, features refer to some distinguishable properties from a media stream and may be numerous among the same modality (Atrey et al., 2010).

Typical methods for merging features into the above mentioned combined input are: concatenation (vectors from each feature set are put together to form a bigger vector), selection (a selection criterion is used to choose a subset of features among all the available ones), and extraction (all the features are projected to a new space) (Hu, 2008).

Some advantages of the feature level fusion are that it can utilize the correlation between multiple features from different modalities at an early stage which helps in better classification tasks; also, it requires only one learning phase (on the combined feature vector). However, in this approach it is hard to represent the time synchronization between the multimodal features because coupled modalities could be extracted at different times. Moreover, the features to be fused should be represented in the same format before fusion (Atrey et al., 2010). Even so, many researchers have adopted the early fusion approach for the multimedia analysis of deception as seen in Chapter 3.

2.3.2 Late fusion

The late fusion approach (also known as decision level) consists basically on two steps: first, n feature sets are fed to n analysis units to provide n local decisions; then, local decisions are combined using a decision fusion unit to make a fused decision vector that is analyzed further to obtain the final decision.

Unlike feature level fusion, where the features from different modalities may have different representations when fused into a single vector, the decisions at semantic level (i.e. per feature set) usually have the same representation, therefore simplifying the fusion of decisions. Moreover, the late fusion approach allows us to use the most suitable methods for analyzing each single modality, providing flexibility over early fusion. However, the late fusion approach fails to utilize the feature level correlation among modalities. Moreover, the learning process becomes expensive (on time complexity and resources management)

(Atrey et al., 2010). Not as popular as the feature level approach, decision level fusion has been used too for multimedia deception detection as discussed in the next chapter (Chapter 3), which is dedicated to explore the literature related with deception detection in videos and multimodal information fusion.

Chapter Three

STATE OF THE ART

Deception detection from videos is of particular interest as a non−invasive method, and as such it has drawn attention from the machine learning community. Typical sources of information extracted from video include RGB images, thermal imaging, audio recordings and speech transcripts. In this revision of the state of the art, we will focus first on research work that perform deception detection using multimodal features extracted from videos, followed by research work focused on multimodal fusion of data.

3.1 Multimodal deception detection in videos

(Abouelenien, Pérez-Rosas, Mihalcea, et al., 2017) presented a database for deception detection consisting of both physiological features (heart rate, blood volume pulse, respiration rate, skin conductance) and thermal videos, as well as transcriptions from videos. Thermal images are analyzed by face regions, while traditional linguistic features such as Part-Of-Speech (POS) tags, word unigrams, Linguistic Inquiry and Word Count (LIWC) embeddings, etc. are extracted from transcriptions. They tested multiple different modal combinations with an early fusion approach using decision trees, concluding that following a multimodal approach outperformed relying solely on single modalities. Furthermore, the fusion of features extracted from videos outperformed the results

obtained by physiological features, supporting the idea that non-invasive methods are not outperformed by invasive ones.

With a similar approach, the group of (Abouelenien, Pérez-Rosas, Zhao, et al., 2017) created a new database intended for multimodal deception detection based on gender. Extracting a similar set of multimodal features (linguistic, physiological and thermal), they found different gender-based patterns. Even though there were specific feature sets that performed well for both genders, they concluded that it is beneficial to consider gender differences in order to improve performance on deception detection.

The databases for both works were constructed within three different scenarios. In two of them ("Abortion" and "Best Friend"), participants were asked to speak freely on the given topic (first truthfully and then deceptively). For the third one, they would be interviewed about a mock crime. However, the above mentioned datasets were constructed by the cooperation of test subjects under controlled circumstances (e.g. participants may not really be motivated to lie).

For studying deception in a more real context, Pérez-Rosas et al., 2015 presented a novel dataset of real court trial videos (a scenario where one could assume speakers are really motivated to convince the listeners about their truthfulness even if they are actually lying). A multimodal approach is used again consisting on linguistical (unigrams and bigrams from transcriptions) and visual (manual annotation of facial displays and hand gestures) features. Results obtained from individual features are compared to those from early fusions, reaching the highest score when using all the modalities together. Experiments are performed using Decision Trees (DT) and Random Forests (RF) with leave-one-out cross validation; while DT achieve better results when it comes to linguistic features, the best results using visual features are achieved using RF.

Furthermore, the behavior of both classifiers differs when fusing features. While DT tend to improve results when joining features, RF show the opposite tendency. By fusing all the multimodal features they obtain an accuracy of

75.20% (which is greater than the best single feature set e.g. 70.24%) using DT; by using RF, they reach a 50.41% accuracy (which is lower than the worst single feature set e.g. 51.20%). This gives an insight: multimodal fusion can certainly improve deception detection in real-life scenarios, but is also able to worsen the performance. As far as the authors and we are concerned, that is the first work on deception detection using both verbal and non-verbal features from real trial recordings.

The dataset presented in that work is of particular interest for many reasons: first at all, it is a database built with real-life cases, which ideally should allow to detect deception cues when people is really motivated to lie; additionally, the collected videos come from many different contexts: not only the trials are over different crimes, but details as camera angle and distance, video quality, etc. are different from video to video. This allows analysis under a "wild" (unconstrained) scenario, which can eventually lead to a real-time deception detection system. That is why it has become a popular database for real-life deception detection, particularly in recent works for multimodal deception detection.

One of such works is the one presented by (Wu et al., 2018), which trains an automated deception detection system with the above mentioned dataset. The approach is multimodal again, but adding a new non-verbal modality; therefore, the system is trained using features extracted from transcriptions, audio stream and images. Facial gestures (treated as micro-expressions) are used again, extracted by a trained classifier rather than human annotation; additionally, video sequences are analyzed employing IDT (Improved Dense Trajectories). MFCC (Mel-frequency Cepstral Coefficients) are extracted and encoded from the audio modality. Finally, transcriptions are analyzed by using Glove (Global Vectors for Word Representation) to acquire verbal features. Given the feature extraction approaches, the number of features obtained for each video is different; to get a fixed-length vector, they use a Fisher Vector encoding.

As the aforementioned works, they performed experiments on single fea-

tures as well as their combinations using a late fusion approach, reaching the best results when combining all the modalities (among different classifiers such as SVM, Naive Bayes, DT, RT and Logistic Regression). The evaluation metric used in this case was AUC, reaching 0.8773 using a linear kernel SVM with all the automatically extracted modalities. This result is particularly interesting because of the experimental setup, which performs 10-fold cross validation using identities instead of videos (e.g. no person in the training set is contained in the test set); to our consideration, subject cross validation is the fairest way of evaluation in small databases since we are willing to detect deception independently of the subject (even though works like the one of (Abouelenien, Pérez-Rosas, Zhao, et al., 2017) suggest that subject-related features like gender can be useful to improve overall detection).

As part of her Doctoral thesis, (Morales, 2018) evaluates a set of automatically extracted multimodal high-level features using the same experimental setup as (Pérez-Rosas et al., 2015) on the above mentioned dataset. For this work, three modalities are used: video, audio and text (automatic transcriptions from audio). Features are extracted at frame, time window, and sentence level, respectively, using OpenMM (Morales, Scherer, and Levitan, 2017). To deal with different length videos, each feature is encoded into a 11-length vector consisting of statistical functionals. For this set of multimodal features, fusion reaches an accuracy of 76.03% using RF, which matches the result obtained by Pérez-Rosas et al; however, this result is slightly lower than the one of the best single modality feature set (e.g. 76.86% obtained with the acoustical modality).

Even though the two previous works perform automated feature extraction, this extraction is based on pre-trained models. Exploring an end-to-end framework, (Karimi, Tang, and Li, 2018) used a Deep Learning (DL) approach for automatic feature extraction from the video and audio channel, respectively. Unlike previous works, these are low-level features; however, to provide interpretability, an attention mechanism is used on the visual modality. To take

advantage of the sequential nature of videos, features are extracted at frame level and fed to Long-Short Term Memory (LSTM) networks to obtain a fixed-length representation for each modality. However, DL is meant to use large datasets for training; to deal with scarce data, the representations gotten from the LSTMs are manipulated with a variation of the Large Margin Nearest Neighbor (LMNN) method using triplets of videos to artificially increase the number of training instances. The final classification is performed with a simple k-nearest neighbors method.

As previous works, the analyzed modalities are used again individually and early fused in the court trial dataset. Their results show a improved performance using both modalities with respect to using them separately, reaching an average accuracy of 84.16% on ten randomly chosen test sets of 10 truthful and 10 deceptive videos. This work is of particular interest because it is able to show the frames which are given the most relevance according to the network, giving a step forward for spotting of deception in videos (which can eventually lead to real-time deception detection).

Another hybrid approach combining Deep Learning with traditional classifiers is presented by (Carissimi, Beyan, and Murino, 2018). According to them, and to the best of our knowledge too, it is the first time a multi-view learning (MVL) approach is used for deception detection instead of feature concatenation. Features are extracted from the video channel and transcriptions of videos from the court trial dataset, with special interest in faces. For face analysis, pre-trained DL networks are used for automatic feature extraction; this features are combined with automatic detection of facial Action Units, manual annotation of facial displays and hand gestures, uni-grams and bi-grams. Their MVL approach is compared with early concatenation fusion of features, showing a better performance with the non-trivial fusion strategy (reaching an accuracy of 89% using leave-one-out cross-validation); however, this time fused results are not compared with the performance of single views.

This work shows the recent rising interest on using more sophisticated fusion strategies for multimodal deception detection, with promising results. Therefore, it seems appropriate to explore fusion methods used for other tasks that could be further applied to deception detection.

3.2 Multimodal information fusion

Despite the variety of features extracted from different modalities in the aforementioned works, most of them share a common characteristic: a simple early fusion approach is used for final classification, that is, features are combined into a single vector before training a classifier; even for the one using late fusion, such fusion is done simply by combining the scores obtained from each modality. As far as we are concerned, there are almost no works exploring alternative multimodal fusion methods (either early or late approaches) for deception detection in videos; therefore, the related work explored in this section will focus on multimodal fusion for supervised classification independently from the classification task.

When talking about generic techniques for late fusion, (Barbu, Peng, and Seetharaman, 2010) present a multimodal fusion method inspired by classifier ensembles, where base classifiers are built independently from each feature set and combined with a boosting strategy. This is an iterative method, for which at each iteration a *weak learner* is trained for each feature set or *view*; at each step, the weak learner with the lower error rate is stored and given a weight based on its training error. At the next step, training instances are given a sampling weight according to if they were correctly or incorrectly classified in previous iterations (missclassified instances are given greater weights); as all the views share the same sampling distribution, this method was named Boosting With Shared Sampling Distribution (BSSD).

BSSD was tested on two datasets (FERET for face, gender and glasses-

presence classification, and CYGD for gene classification) and compared to majority vote, stacking, and semidefinite programming (SDP) using different kernels. For all the given tasks, fusion using BSSD showed an improvement with respect to individual views; additionally, BSSD outperformed the other methods tested.

However, BSSD is not able to deal naturally with the sequential nature of videos. When it comes to predictions with respect to input sequences, Recurrent Neural Networks have achieved good results. In the line of multimodal tasks using RNN to deal with temporal sequences, (Bouaziz et al., 2016) propose an architecture of parallel Long Short-Term Memory (LSTM) networks for multi-stream classification. LSTM are a special type of RNN that contain an internal cell state which is updated through input steps in a sequential data feed, allowing it to keep a *memory* of what the network has seen so far. Such memory is used as an additional parameter when evaluating new steps, therefore the network makes inferences including knowledge acquired from the past. Bidirectional LSTM (BLSTM) extend this principle by analyzing the input sequence backwards too (the very same sequence is analyzed independently from start to end and from end to start).

BLSTM are used in this case separately for independent input streams in the task of predicting the genre of the next TV show given a history of genres so far. Each input stream corresponds to the programming of a TV channel (using the EPG dataset); feature vectors obtained from each BLSTM are fed to independent fully connected layers and the outputs of this layers are summed element-wise to obtain the final decision derived from the multiple input streams, therefore using data from many streams (modalities) of sequential nature in a prediction task.

Extending the concept of Parallel Long Short-Term Memory (PLSTM) presented above, (Sawada, Masumura, and Nishizaki, 2017) present a new architecture with a hierarchical approach using attention mechanisms for multi-

class classification of conversations (from a Japanese call center). For this work, each speaker is considered as an input stream, following the next hierarchy: conversational document $\rightarrow$ speaker $\rightarrow$ utterance (sentence) $\rightarrow$ word. Sequences of words are analyzed at sentence level with a bidirectional RNN; hiden representations from BRNN are fed to an attention mechanism with a memory reader to obtain a representation of a sentence; sequences of sentences are fed too to a BRNN and a posterior attention mechanism with memory reader to obtain a representation of the conversation from the perspective of a single speaker; the final representation of a whole document for classification is the sum of the representation of each speaker (showing an improvement with respect to a single BRNN and parallel RNN).

This architecture is called a Parallel Hierarchical Attention Network (PHAN) because data is analyzed through parallel RNN using a hierarchy. The attention mechanism refers to learning a parameter that is able to give a normalized importance to each of the steps in a sequential input at output level (for example, in a network that prioritizes nouns, the attention derived from the input "A cat ran" could be the vector [0.2, 0.7, 0.1]); the final hidden representation of a sequence using a RNN and an attention mechanism is a weighted sum of the output of the RNN at each step given the attention provided by the mechanism (back at the previous example, supposing the outputs of the RNN were [[-2],[1],[-2]], the final representation would be 0.2*[-2] + 0.7*[1] +0.1*[-2] = [0.1]).

The memory reader is an additional trainable parameter used by the attention mechanism, and in the case of multistream text it is useful since it can be shared across streams (since all the streams are of a common type of data). This allows for parallel independent training of streams while still making emphasis on the common features between them. However, these works doesn't deal yet with a multimedia problem.

In the field of affective computing, (Gorbova et al., 2018) present a PLSTM

architecture for personality analysis from videos. In this case, videos are split into three modalities (audio signal, image sequence and transcription), features are extracted from each modality using external tools, and features per modality are fed into parallel LSTM in a similar way to the work of (Bouaziz et al., 2016). However, unlike the aforementioned architecture, the hidden representation obtained from the PLSTM is concatenated into a single vector.

This multimodal vector is then fed into a linear regressor, since the task to solve is the prediction of 5 personality traits in a scale [0,1] from the database collected for the 2017 ChaLearn LAP CVPR/IJCNN Competition. Although their results are slightly below the ones from the winners of the challenge, this particular approach have two main advantages of interest for us: the architecture is simple so easy to replicate, and more importantly, features extracted from each modality are high-level thus helping to provide interpretability from the constructed model.

This fusion, as well as the other ones mentioned in this section tend to the late fusion approach, where data is processed by intermediate analysis units and the representation gotten from those units is used together to take a final decision. One work focused on early fusion is the one presented by (Morales, 2018), which combines features from different modalities at feature level using temporal synchronization with “informed” methods. With this approach, one feature set is “informed” by other through time, being the first set analyzed under temporal circumstances defined by the second one. In the case of the “syntax informed method”, there are N time periods defined by the N different Part-Of-Speech (POS) tags present in the speech; each feature from the acoustic modality is analyzed in N different periods of time. Supposing two POS tags (Noun [N] and Verb [V]), and two acoustic features (Frequency 0 [F0] and Mel Frequency Cepstral Coefficient 0 [MFCC0]), the syntax informed vector would look like this: N x F0, N x MFCC0, V x F0, V x MFCC0; where N x F0 is the mean of F0 through all the time lapses when a Noun is told, N x MFCC0

is the mean of MFCC0 through all the time lapses when a Noun is told, and so on.

This informed approach is used for binary detection of depression in interviewed persons from videos using three modalities (audio, text and video). As the informed approach makes use of two modalities simultaneously, this method can just be evaluated by pairs of modalities; in that work, the combination of audio with text and audio with video were evaluated (being the audio informed by the other modalities), showing a better performance than a concatenation approach.

From this revision of the state of the art, we can conclude that not only the late fusion approach is barely explored when it comes to multimodal deception detection in videos (actually, there is no focus on fusion when it comes to the task), but the decision level approach is actually a tendency in multimodal analysis for many different tasks including video related ones.

For this thesis, we aim to pay special attention to the methods used for multimodal fusion (instead of simply concatenating the features extracted from each modality); as the tendency nowadays for multimodal information fusion is the late approach, we focus on late fusion approaches for deception detection. Additionally, unlike many of the works presented here, we split features not only by modality but also by type in order to explore them more congruently as high-level attributes (i.e. features that can be interpreted by humans as congruent sets of information). The next chapter is aimed to discuss the feature extraction strategy proposed in this thesis for the task of deception detection in videos.

Chapter Four

Multimodal characterization of variable-length videos

To analyze videos using a Machine Learning approach, the first step is to extract features from them. However, videos are an interesting type of raw data since they are inherently composed by more than one channel, and the channels that compose a video have a sequential nature. Therefore, to extract features from videos, we must define the different channels from which those features can be extracted. When it comes to videos, there are two main channels: images and audio. Additionally, from the audio channel we can extract text in the form of transcriptions. These three channels will be the *modalities* used for analysis in this work.

Once the different modalities have been determined, we need to separate and encode the channels into a set of features. Since the objective of this work involves the usage of high-level features, feature extraction is done using open tools developed for automatic extraction of feature sets typically used in affective computing. It is important to note that, from a single channel, different feature sets can be extracted according to the scope or phenomenon of interest (e.g. from a face picture, we can extract a set of features solely from the eyes and another one analyzing solely the mouth); each from these different feature sets that can be extracted from a modality will be called a *view* from the given

modality.

However, this views are constrained by the length of the analyzed video; furthermore, under a non-constrained context where videos are not expected to have a determined length, the number of features extracted per video will vary proportionally to its length. Unlike images, where input instances can be easily resized to meet the size conditions of a system, changing the duration of a video is not a trivial task. The next sections are aimed to describe the features extracted from each of the aforementioned modalities (Fig. 4.1 shows a summary of the different views extracted per modality), how they can or have been used for deception detection (or similar tasks), how they were extracted, and the strategies used in this work to deal with the variable length of videos from a database for their multimodal characterization.

	 *per frame*	 *per frame*	 *per video*
Modality	**Visual**	**Accoustic**	**Textual**
Views	*AU Int*	*Voice*	*Char 1-grams*
	AU Pres	*Glottal Flow*	*Char 2-grams*
	Eye LM	*MCEP*	*Char 3-grams*
	Facial LM	*HMPDM*	*Char 4-grams*
	Gaze	*HMPDD*	*POS 1-grams*
	Head		*POS 2-grams*
			POS 3-grams
			POS 1-grams
			BoW
			LIWC
			Syntax Info

Figure 4.1 The different views extracted for each of the 3 proposed modalities.

4.1 Visual

The visual modality is composed by the features that can be extracted from the image stream of a video. To get these feature sets, raw videos are fed to OpenFace (Baltrusaitis et al., 2018) 2.1.0, a facial behavior analysis toolkit. This toolkit analyses videos at frame level; from each image, a long vector is extracted composed by different types of features: facial landmarks, head pose, binary presence and estimated intensity of facial action units, eye landmarks and gaze direction (see Fig. 4.2 for a visual representation of this features). The given vector (composed of 465 features) is split according to the different types described above in order to get 6 different views.

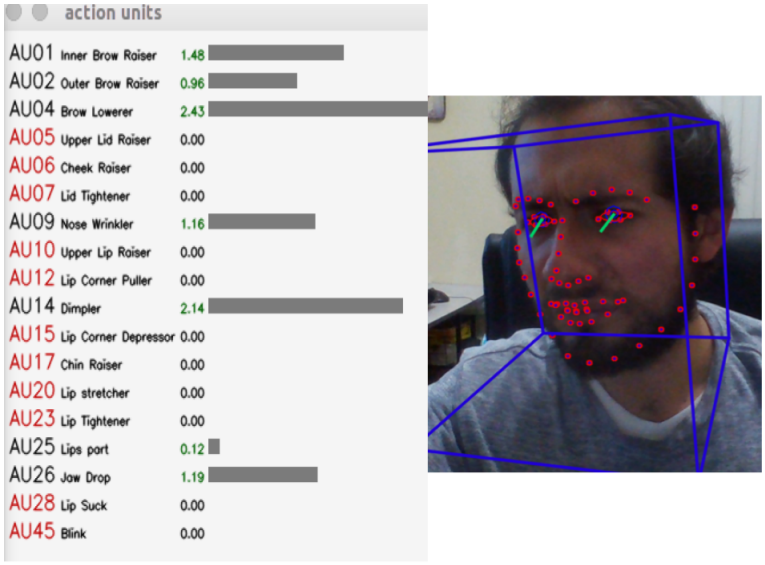

Figure 4.2 OpenFace working on a real-time video.

Facial analysis is of particular interest because of two main reasons: 1) face is usually the most visible body part when talking with someone, and 2) it reveals a vast amount of information about the internal state of speakers, including behavior that can be distinguished between liars and truth tellers as stated by (R. Ekman, 1997). Particularly, Action Units (AU) from the Facial

Action Coding System[1] (FACS) are useful to identify emotions that a person would want to keep in secret; however, identifying the existence of such hidden emotions can be misleading of deception as stated by (P. Ekman, 2003). Therefore, analyzing AU further than for emotion recognition can be useful for the deception detection task.

Systems meant to recognize AU automatically use facial landmarks; additionally, these landmarks are useful to describe faces and facial behavior in a more general way. Likewise, head pose estimation can be used as another descriptor of body language, giving an insight of involuntary movements beyond face.

When talking about pose, there is another feature of particular interest: eyes. Eyes can be described both in terms of gaze direction and opening (of eyelids and pupils). Gaze direction is a popular feature used to undercover the cognitive process going in the mind of the observed person (such as if they are reminding or inventing something); however, not only this prior knowledge can be used by a liar too in an attempt to improve their lies but there is not a general agreement about if there is a specific set of eye movements directly related with deception. As with AU, identifying certain movements can be misleading of detection; furthermore, there is evidence supporting than prior knowledge of the popular gaze direction pattern proposed by the Neuro-Linguistic Programming supporters does not make a significant difference when a person is trying to detect lies (Wiseman et al., 2012). However, there is in fact an agreement that there are involuntary eye movements associated with cognitive process; as an involuntary behavior, there may be still a correlation that can be captured by machine learning systems for deception detection. Similarly, opening of pupils is an involuntary behavior that could be exploited for deception detection; actually, there are some preliminary results that seem to support such

[1] For further information about the FACS and what facial action units look like, the reader is encouraged to visit https://www.noldus.com/facereader/facial-action-units

idea (Mitre-Hernandez et al., 2019).

4.2 Acoustical

The acoustical modality is formed using features that can be extracted from an audio stream. For each video, the open tool FFmpeg is used to extract a WAV audio file. Each of these files (audio stream) is fed to a MATLAB script from COVAREP (Degottex et al., 2014), an open-source repository of advanced speech processing algorithms.

Unlike the visual modality, analysis is done among time-windows (rather than frames). For this thesis, the windows size was chosen to match the frame rate of the image stream (i.e. there are 29.7 windows in a second since typical cams record 29.7 frames per second). For each time-window, 74 features are extracted to form a long vector; again, the given vector is split according to feature types in order to create the different views from the acoustical modality: glottal flow (NAQ, QOQ, H1-H2, HRF, PSP, MDQ, Peak Slope, Rd, Rd confidence, Creky Voice), voice (F0, V/UV), MCEP (MCEP 0-24), HMPDM (HMPDM 0-24) and HMPDD (HMPDD 0-12). For a deeper understanding of this features, the reader is encouraged to read the related paper of COVAREP (Degottex et al., 2014).

The glottis is the opening between the vocal folds; sound production that involves moving the vocal folds close together is called glottal. Glottal flow has an important contribution to the supra-segmental characteristics of speech and is known to significantly vary with changes in phonation type, so its parameterisation can be useful in many areas of speech research (Degottex et al., 2014).

Sounds generated by a human are also filtered by the shape of the vocal tract (tongue, teeth, etc.). This shape determines what sound comes out, and it manifests itself in the envelope of the short time power spectrum. The job

of Mel-frequency cepstral coefficients (MFCC) is to accurately represent this envelope. This is why MFCC have been widely used for different tasks such as Automatic Speech Recognition (Ganchev, Fakotakis, and Kokkinakis, 2005); however, COVAREP extracts an alternative set of MFCCs which are extracted from the "True Envelope" spectral representation (MCEP), which showed usefulness for emotion recognition (Degottex et al., 2014).

Sound (and consequently voice) is a periodic waveform propagated usually through air; the fundamental frequency (F0) is defined as the lowest frequency of such waveform. F0 and detection of voiced and unvoiced (V/UV) segments are used to study the pitch an rhythm of a person while lying or telling the truth. Harmonic model and phase distortion mean (HMPDM) and deviations (HMPDD) have been used before for depression detection in videos (Pampouchidou et al., 2016), showing its usefulness for an affective computing task hard for humans.

4.3 Textual

From the audio stream, transcriptions can be obtained to get a text file; by extracting different features from such text we get the textual modality. In order to have a system able to perform automatic analysis, video transcriptions are extracted automatically using Watson Speech to Text from IBM (using the model for English); since this task is done by an Automatic Speech Recognition (ASR) system, it lacks punctuation and does not identify the subject of interest (i.e. even if it is able to distinguish between speakers, it is not able to recognize who is the person being evaluated).

Based on the study presented by (Rill-García et al., 2018), the next views were extracted at video level using the Natural Language Toolkit from Python: bag of character n-grams (for $n = 1, 2, 3, 4$; where each value of n corresponds to a different view), bag of Part-Of-Speech n-grams (again for $n = 1, 2, 3, 4$) and

a LIWC dictionary encoding.

Other view is extracted using a typical Bag-Of-Words (BOW) representation as suggested by (Pérez-Rosas et al., 2015). To create a BOW, a vocabulary is obtained by extracting all the different words from a corpus of texts; afterwards, each text is represented as a vector counting the number of times each different word from the text is inside the text (the length of the vector is equal to the size of the vocabulary; a graphical example of this can be seen in Fig. 4.3). As transcription lengths can vary a lot from video to video, the resulting vector from each text is normalized; to reduce the size of the vocabulary, this is constructed using only words appearing in at least 10% of the texts. This representation is used because it is a common strategy for Natural Language Processing (NLP).

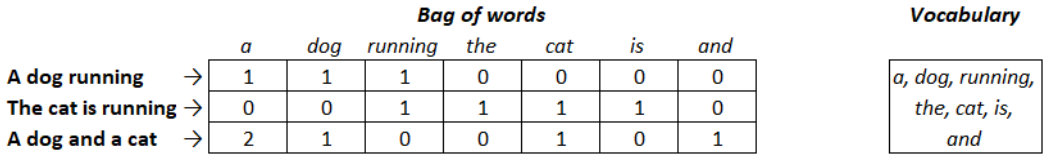

Figure 4.3 Example of a Bag-Of-Words extracted from a corpus of three sentences.

With respect to character n-grams, the term refers to sets of n consecutive characters (e.g. all the available character 3-grams in the sentence "*The cat*" are "*The*", "*he* ", "*e c*", "*ca*" and "*cat*"). To get a bag of character n-grams, a new corpus is obtained replacing each text with its corresponding set of ordered available character n-grams; once this is done, the same logic from a BOW is used, counting character n-grams instead of words (when something is used instead of words, the representation gets the generic name of Bag-Of-Terms). Character n-grams are often used to capture author style from texts, so they could be useful to capture style from liars/truth tellers.

Part-Of-Speech (POS) tags are used to classify words according to their grammatical properties. Although different POS tags exist for different lan-

guages, there is a set of universal POS tags: ADJ (adjective), ADP (adposition), ADV (adverb), CONJ (conjunction), DET (determiner, article), NOUN (noun), NUM (numeral), PRT (particle), PRON (pronoun), VERB (verb) and X (other). For each transcription, a new text is obtained by replacing each word with its corresponding POS using SyntaxNet. To get POS n-grams, we count POS tags instead of characters; consequently, a bag of POS n-grams is obtained by counting POS n-grams instead of words. As with character n-grams, the reasoning behind analyzing POS tags is that they are useful for capturing style.

The Linguistic Inquiry and Word Count (LIWC) dictionary defines one or more word categories or subdictionaries for different words/word stems. For example, the word "cried" is part of four word categories: *sadness, negative emotion, overall affect*, and a *past tense verb*. Similarly to a BOW representation, a LIWC enconding consists on representing a text with a fixed-size vector (the length of the vector is equal to the number of categories in the dictionary); unlike a BOW, each word can increase the count of many attributes at the same time (for example, with "cried", the positions for sadness, negative emotion, overall affect and past tense verb are increased by 1 at the same time). For this work, the 2007 English version of the LIWC dictionary[2] is used, normalizing each resulting vector.

Finally, a simple syntax analysis as done by OpenMM (Morales, Scherer, and Levitan, 2017) is used as another textual view; these "syntax features" are extracted taking advantage of the parse tree generated by SyntaxNet when POS tagging a text. These syntax features are considered because they have been used previously for sentiment and deception detection (Morales, 2018).

[2]For further information on this dictionary, the reader is encouraged to visit https://www.kovcomp.co.uk/wordstat/LIWC.html

4.4 Dealing with variable length videos

Despite the different text sizes obtained by transcribing variable length videos, the feature (views) extraction for the the textual modality results on fixed $1 \times N_i$-size vectors per video for each different view (with N_i equal to the number of different features corresponding to the view i). However, for the visual and acoustical modalities we extract a $M_j \times N_i$ matrix for each modality from each video (with M_j equal to the number of frames/time windows contained in the video j). As we expect videos with variable length, M_j will change from video to video; however, classic machine learning models are not able to deal with variable size inputs (not to say they typically expect a vector rather than a matrix).

Therefore, in order to use the extracted feature matrices, we need a strategy to code them into fixed-size vectors independent from the video length. The next subsections are aimed to explain the two approaches used for this in the present work.

4.4.1 Statistical functionals

A first intuition on how to solve this problem involves using mathematical functions that somehow describe the behavior of each feature in the whole video (that is, along all the frames). A typical approach to do this comes in the form of descriptive statistics; as done by OpenMM (Morales, Scherer, and Levitan, 2017), an open-source multimodal feature extraction tool, the final representation from a variable-length sequence of features is computed as 11 statistical functionals for each feature (a graphical representation of this process is shown in Fig. 4.4). The features extracted by OpenMM in the form of statistical functionals were aimed primarily for depression detection in videos (which is a similar task), but they were incidentally tested too for the deception detection

task showing promising results.

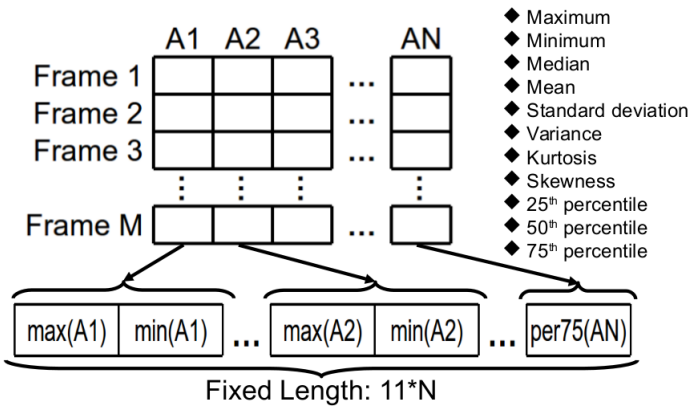

Figure 4.4 Creation of a $1 \times (11 * N)$ vector from a $M \times N$ matrix. The matrix is obtained by extracting N attributes AI for $I = 1, 2, ..., N$ from a video with M frames.

This approach is attractive since it is able to summarize hundreds (or thousands) of frames with an easy, low cost, implementation. However, it lacks the ability of capturing the inherent sequential nature of such frames; as these features are extracted from a video, the order of the frames is important to understand and analyze it. A strategy for coding variable-size matrices into fixed size vectors taking into account the order of the frames is presented in the next subsection.

4.4.2 Long-Short Term Memory

Long-Short Term Memory (LSTM) networks (Hochreiter and Schmidhuber, 1997) are a special type of Recurrent Neural Networks (RNNs) used in the field of Deep Learning. RNNs are well-suited to process time series data, while LSTMs are particularly useful to deal with long sequences in comparison with traditional RNNs (which comes handy when dealing with videos since a single second of recording involves $\sim$30 frames).

For this work, LSTMs are used as a feature extractor intended to encode a

sequence of high-level features into a fixed size vector. In order to do this, the resulting vector from a LSTM is connected to a fully connected layer intended to predict if a sequence of high-level features is associated with a truthful or deceptive video. Once the network is trained, the output of the LSTM layer when evaluating a input matrix is used as the fixed-size vector representing such matrix (as seen in Fig. 4.5).

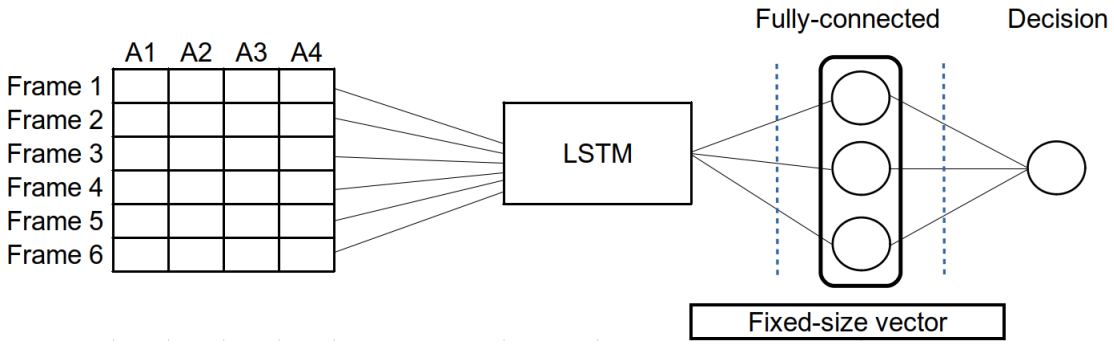

Figure 4.5 Creation of a 1×3 vector from a $6_{frames} \times 4_{attributes}$ matrix. The resulting vector is the output of the LSTM layer (which was used to calculate the final decision in the training phase).

However, training a network with very long input sequences is expensive in terms of time and memory. That is why training data fed to LSTM networks is usually padded in order to reduce the length of the training sequences (typically using a fixed length in terms of the average length of the training sequences). In the case of this work, a different approach for sequence reduction is used inspired by the work of (Gorbova et al., 2018) on automatic personality analysis from short video clips: instead of selecting a set of K consecutive frames, all the frames from a video are fed to a K-means algorithm; as this work aims to detect cues for deception, selecting representative frames from each video is an intuitive idea for focusing in the particular moments important for deception detection. To take advantage of the multimodal nature of videos, as well as the synchronized extraction of visual and acoustical views, all the multimodal views are concatenated into a multimodal matrix for each video. From this matrix, K

key frame numbers are selected as in Fig. 4.6; from this point, each view is used to train separate LSTMs as in Fig. 4.5 using just the frames obtained from the previous step.

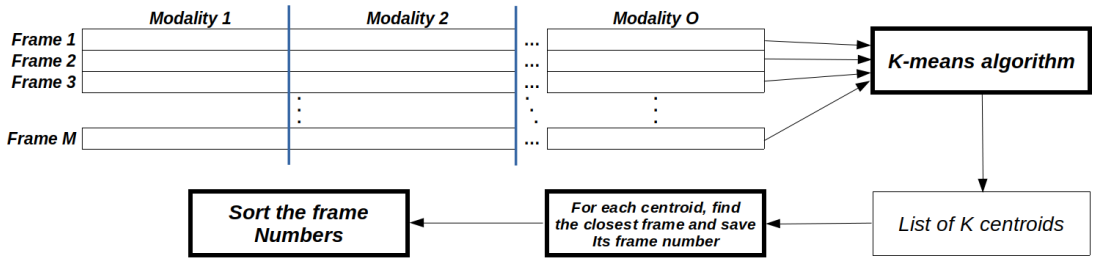

Figure 4.6 Selection of K ordered key frames using a K-means algorithm on the vectors resulting from concatenating the different views at frame level.

From a grid search, $K = 100$ was chosen. In a similar way, a grid search was performed for each view in order to find a proper size for the output of its corresponding LSTM (as well as other training parameters such as number of training epochs and batch size); however, as words are not synchronized with frames, the textual modality was excluded in this case.

The LSTM approach used here is an intermediate point between preserving full interpretability of the extracted features (which were meant to be high-level from the beginning) and exploiting the automatic feature extraction ability of Deep Learning models. More specifically, the high-level features extracted with the aforementioned tools substitute the convolutional layers from a typical DL architecture that are fed to a LSTM layer during training, thus creating a hybrid approach. This approach then uses high-level features that we, as humans, can understand; however the temporal encoding is learned automatically, making interpretability harder. Under the scope of training, this approach helps to deal with the small number of training instances, since typical DL models are trained using great amounts of data: by providing pre-extracted features, the network doesn't need to learn them itself, thus (intuitively) simplifying the training.

Once the strategy for feature extraction has been chosen, it is time to present the databases used for testing and validation.

4.5 Datasets

As stated before, deception detection is a hard task for humans. As a consequence, it is hard to compile videos correctly labeled as a either truthful or deceptive. To deal with this problem, related works typically create artificial datasets by recording people who were explicitly asked to lie or tell the truth under controlled circumstances (actually, one of the contributions of this thesis is a database of this kind using Mexican subjects speaking Spanish); however, this databases are not usually publicly released due to Institutional Review Board restrictions. Furthermore, there is an underlying inconvenient, since people recorded for these databases are volunteers: there is no real motivation to lie, as there are not relevant rewards or punishments related to being trusted. Behavior is strongly influenced by context, and so is involuntary behavior too; therefore, by a context lacking of a real motivation to lie, the cues for deception found in videos like these may be misleading with respect to real-life scenarios.

Also, talking about behavior, it is known that it is conditioned by cultural background. As a consequence, we could think that the act of deceiving is influenced by the cultural background of the liar. While looking for universal cues of deception, it would be necessary to analyze people from different regions and countries; however, deception detection video datasets are not only scarce but they are usually composed from American people. Even in the cases where such databases are recorded with volunteers from different countries, they use English as the spoken language; these adds an additional factor to the problem, since speaking a foreign language already implies a different cognitive process. Actually, research suggests that either lying or detecting deception is

heavily influenced by the language spoken, in terms of speaking a native or a second language (Cheng and Broadhurst, 2005).

Concerning the points stated in this section, our experiments are performed with data extracted from two different databases developed for the deception detection task in videos, including a novel dataset composed of Mexicans speaking Spanish (to the best of our knowledge, this is the first database with these characteristics collected for the task). The next subsections are aimed to describe both datasets and how they were collected, as well as to discuss how both datasets help to deal with the issues described above.

4.5.1 Real-life trial database

Expecting to "to build a multimodal collection of occurrences of real deception during court trials, which will allow us to analyze both verbal and non-verbal behaviors in relation to deception", (Pérez-Rosas et al., 2015) introduce a novel dataset consisting of videos collected from real-life public court trials.

To collect such videos, they started by identifying public multimedia sources where trial hearing recordings were available; this is of particular interest since, videos being already public, the collection could be made publicly available without independent ethics committees restrictions. The second condition to choose a video was that deceptive and truthful behavior could be fairly observed and verified; this is highly relevant too, since it implies that there should be a mechanism able to fairly label videos from real-life scenarios (thus having a context where the subject is actually motivated to lie, i.e. a court trial).

Regarding data processing, some additional constrains were taken into account for video selection: the defendant or witness should be clearly identified in the video; their face should be visible enough during most of the clip; visual quality should be clear enough to identify facial expressions; audio quality should be clear enough to hear and understand what the person is saying.

With respect to video labeling, three different trial outcomes were taken into account to label a trial video clip as deceptive or truthful: guilty verdict, non-guilty verdict, and exoneration. For guilty verdicts, deceptive clips were collected from a defendant while truthful videos were collected from witnesses in the same trial; in some cases, deceptive videos were collected from a suspect denying a crime they committed while truthful ones were taken from the same suspect when talking about facts verified by the police. With respect to witnesses, testimonies verified by police investigations were labeled as truthful whereas testimonies in favor of a guilty subject were labeled as deceptive. In all cases, exoneration testimonies were labeled as truthful statements[3].

One should notice, however, that labeling is still subject to noise; without going farther, exonerees were first found guilty, so using a guilty verdict can be misleading of a correct label. But even so, it is a trade-off for the advantage of having a real-life scenario where labels can be assigned under certain factual evidence.

The final available dataset consists of 121 videos, including 61 deceptive and 60 truthful ones. The average length of the videos is 28.0 seconds (27.7 and 28.3 for deceptive and truthful clips, respectively). The dataset consists of 56 unique speakers including 21 female and 35 male (according to the authors of the set; the authors of this thesis counted out 58 different identities), from which each identity has an unbalanced set of deceptive and truthful videos (a person's videos are, usually, uniformly from a single class). As the videos were collected from many sources, they vary highly in camera position, movement, focus, scene change, background noise, volume, human editing, et cetera (some frame examples from different videos are shown in Fig. 4.7).

[3]Clips containing exonerees testimonies were obtained from "The Innocence Project" website (http://www.innocenceproject.org/)

Figure 4.7 Sample frames from 4 different videos of the court trial dataset.

4.5.2 Novel Mexican Spanish abortion/best friend database

As stated before, not only deception detection video datasets are scarce, but they are usually collected from American people; even if not, the language in those videos is English. Furthermore, most of them are not publicly available. Motivated by this facts, and looking to not only have more data for experiments but to study deception on not-American people speaking their native language, we worked in the development of a novel dataset composed by Mexican people speaking Spanish.

The above mentioned database was collected jointly with the Centro Tlaxcala de Biología de la Conducta (CTBC) from the Universidad Autónoma de Tlaxcala (UATx), by recording volunteer students from that institution. Similarly to related works, the recordings were done under a controlled environment; for this particular dataset, subjects were recorded facing the camera (they were recorded so that their heads and shoulders were fully visible) in front of a white wall in a silent room.

Regarding deception detection, subjects were asked to discuss two different topics: abortion, a controversial topic; and their best friend, a personal topic. This topics, as well as the protocol used for recording, was inspired by the ones used by (Abouelenien, Pérez-Rosas, Mihalcea, et al., 2017).

In the case of abortion, participants were asked to give their genuine position towards abortion; once they were done, they were asked to give a fake posture towards the same topic. For both cases, participants were instructed

to talk 2-3 minutes continuously (both postures are recorded as independent videos, and only their answers are recorded i.e. the interviewer's instructions are not recorded). When giving their genuine position, the video is labeled as truthful; when giving a fake posture, the clip is labeled as deceptive.

With respect to the best friend topic, the subjects were asked to describe their best friend; afterwards, they were instructed to think about a person they couldn't stand and describe them as if that person was their best friend. Similarly to the abortion topic, participants were asked to talk 2-3 uninterrupted minutes for each case. When talking about their best friend, the video is labeled as truthful; when describing the person they can't stand, the clip is labeled as deceptive.

Each participant was asked to give both postures about the two topics, but there were videos discarded when the subject wasn't able to fully follow the given instructions (in some cases, there are short utterances of the interviewer encouraging the subject to keep talking). The final collection of videos used for experiments consisted of 42 videos, including 21 deceptive and 21 truthful. As stated before, videos have a length between $\sim$2 and $\sim$3 minutes. The dataset consists of 11 unique speakers, from which each identity has a balanced set of deceptive and truthful videos. Unlike the court database, even if videos were recorded on different days, clips are homogeneous in terms of camera position, movement, focus, et cetera (some frame examples from different videos are shown in Fig. 4.8).

Figure 4.8 Sample frames from 4 different videos of the Mexican Spanish dataset.

Database	Court	Mexican Spanish
# Videos	121	42
Language	English	Spanish (Mexican)
Context	Court-trial, real life	Abortion/best friend, voluntary participation
# Subjects	58	11
Balanced? (Per subject)	No	Yes
# Deceptive clips	61	21
# Truthful clips	60	21

Table 4.1 Summary of the analyzed databases.

With both databases (summarized in Table 4.1), we contemplate the cases of both real-life and controlled lies, with both fixed and variable camera settings, under different contexts, on people using two different languages (including people from different ethnic origins). This diversity is useful to validate the features and methods used in this work under different constraints, thus helping in the objective of analyzing deception in a general context (i.e. independently from the conditions on which we want to detect deception).

Once the feature extraction strategy and the databases have been selected, it is time to evaluate the extracted features in the given datasets. A study of the performance of these features using a Machine Learning approach is presented in the next chapter.

Chapter Five

SINGLE-MODAL DECEPTION DETECTION

Before exploring multimodal fusion, we conducted preliminary experiments to evaluate each modality independently. Experiments were performed using scikit-learn 0.20.2 using SVC (a SVM classifier) as baseline (as it tended to show the best results in preliminary experiments without parameter tuning); a minor hyperparameter tuning was performed using grid search looking to optimize the average accuracy of all the views (i.e. all the views were trained separately using the same hyperparameters).

For all the experiments from now on, a 10-fold cross validation is used for evaluation. However, as we are exploring the multimodal analysis of deception cues, we want to avoid the classifiers to degenerate into identity detectors (e.g. it is not desirable to classify a person as a liar in the test set just because all their training examples were deceptive); to work around this, our 10 folds are identity based rather than instance based (i.e. no person in the test set was used in the training set as suggested by (Wu et al., 2018)). Additionally, as the labels per subject are unbalanced in the court dataset, AUC ROC is used as the evaluation metric. In the case of the Mexican Spanish dataset, as labels per subject are balanced, the AUC should be similar to accuracy, so AUC is conserved for convenience.

For each database, two experiments were performed: 1) using the features encoded with statistical functionals and 2) using features encoded with a LSTM. In both cases, modalities were evaluated as separate views; for statistical functionals, views are used as a single modality too (i.e. all the different view vectors from modality M were concatenated into a single long vector representing modality M -early fusion-). This is because we want to evaluate each view separately to gain insight of the performance of "intuitive" features separately, and then evaluate how all these views work together (when concatenated as a single feature set).

For the court database, the base classifier used a linear kernel; for the Mexican Spanish database, a polynomial kenrel was used with C = 0.01. For both cases, to ensure reproducibility, tolerance was set to 1e-7 and maximum number of iterations to 3000. The other hyperparameters were set to default.

5.1 Visual modality

For the visual modality, results are shown with blue bars in Figs. 5.1-5.4. The first two graphs refer to the court database, while the last two refer to the Mexican Spanish dataset; in both cases, the first graph shows results using statistical functionals while the second one using LSTM.

Focusing in the court dataset, we can see in both cases that gaze direction stands out among the different visual views, being the best feature set when using statistical functionals. Related with the eyes, Fig. 5.1 suggests that eye landmarks are useful too to recognize deception when using statistical functionals; in general, it seems that using statistical functionals outperforms the usage of LSTM to encode the temporal information of the chosen high-level features. This is particularly notorious too when analyzing the binary presence of action units; both AU presence and eye landmarks have a relevant better performance when using statistical functionals rather than LSTM. However, focusing on gaze

direction and head pose (both features referring to position and orientation of body parts), one should notice that LSTM seems to reach good results; actually, head pose achieves the best results from the LSTM encoding, reaching a result as good as analyzing gaze direction with statistical functionals.

A similar tendency can be seen when analyzing the Mexican Spanish dataset: overall, results using statistical functionals outperform the ones using LSTM. And again, gaze direction, eye landmarks and presence of AU seem to do well at detecting deception. However, we find that while using statistical functionals the overall performance of each view tends to improve, using LSTM has the opposite tendency. Particularly, we notice that the views that performed better in the court dataset have a considerable decrement in the Mexican Spanish database, while the intensity of action units saw a relevant improvement. Actually, the intensity of AU has a great improvement overall in the Mexican Spanish clips with respect to the court ones, as well as facial landmarks. This behavior could be explained by the camera distance/angle with respect to the speaker in both datasets, as in the Mexican Spanish dataset not only the camera is close to the person in all videos without changing position but there is no change of scene in the clips (unlike the court dataset), therefore simplifying the facial analysis.

To close this section, it is important to notice that concatenating all the visual modalities into a single vector (rightest blue column in Figs. 5.1 and 5.3) does not outperform the best view in the court dataset, but it does slightly improve the best individual result in the Mexican Spanish database.

5.2 Acoustical modality

For the visual modality, results are shown now with orange bars in Figs. 5.1-5.4. Again, the first two graphs refer to the court database, while the last two refer to the Mexican Spanish dataset; in both cases, the first graph shows results using

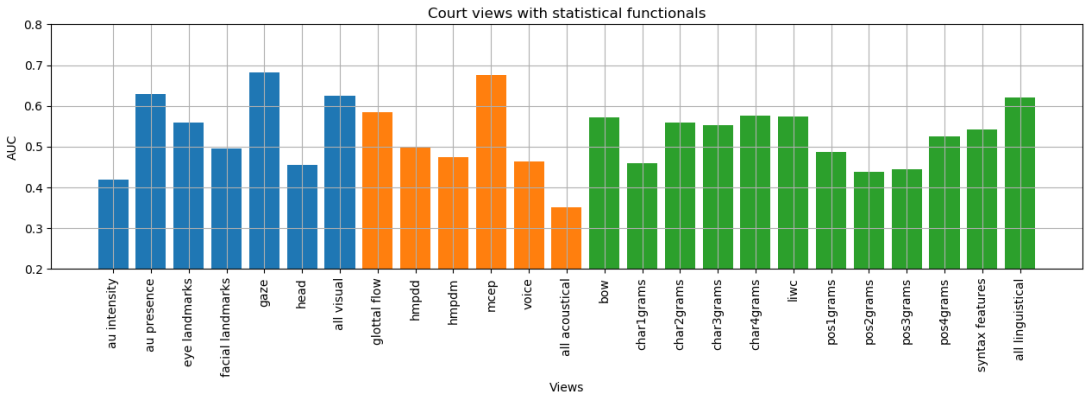

Figure 5.1 AUC achieved by the different views in the court-trial dataset when using statistical functionals. Views from the visual modality are in blue; the ones for the acoustical modality are in orange; green corresponds to the textual modality; for each case, the rightest column represents the unimododal concatenation.

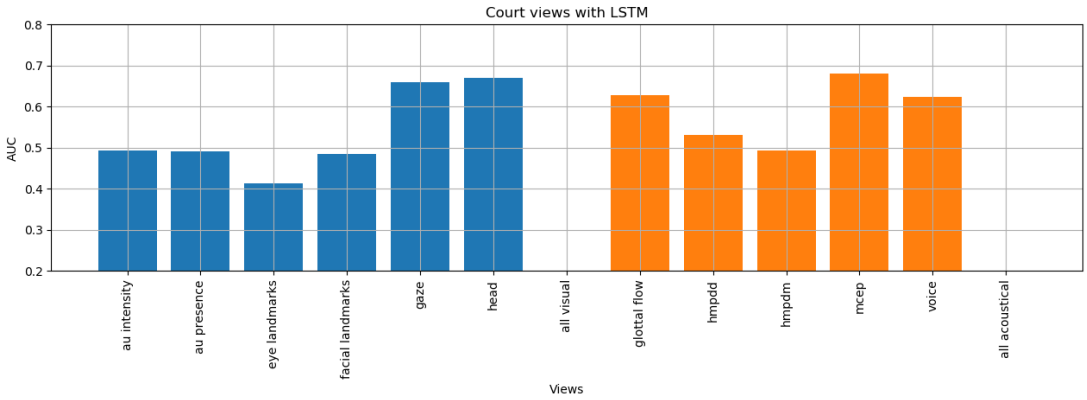

Figure 5.2 AUC achieved by the different views in the court-trial dataset when using LSTM. Views from the visual modality are in blue; the ones for the acoustical modality are in orange.

statistical functionals while the second one using LSTM.

In this case, for both cases in both datasets there is a view that outperforms all others: MCEP. In the case of the court clips, this view shows a behavior similar to gaze direction (visual modality); when analyzing the Mexican Spanish videos, using statistical functionals MCEP works much better than gaze

direction, but using LSTM gaze outperforms MCEP.

However, the acoustical views tend to get much worse in general when using LSTM instead of statistical functionals in the Mexican Spanish database; we observe the opposite phenomenon in the court dataset, where using LSTM tends to improve performance with respect to statistical functionals. This behavior could be explained by the average length of the videos, since Mexican Spanish ones are approximately five times longer than the clips from court; thus, LSTMs are able to get the most from not-so-long videos in the court dataset while lacking information because of the sequence padding in the Mexican Spanish videos.

This is particular relevant for the voice view, since using a LSTM encoding highly improves the performance of this view in the court dataset with respect to using statistical functionals; the very opposite case occurs for the Mexican Spanish clips, where using statistical functionals is much better than using a LSTM encoding. However, it seems that properly analyzing the F0 and pauses (voice) along time can be useful to detect deception in different scenarios: either short participations with interruptions (like in the court database) or long (2-3 minutes) speeches without interruptions (like in the Mexican Spanish dataset).

We have another particular case with respect to glottal flow, that reach results $\gtrsim 0.60$ in 3 out of 4 cases (being the only exception using LSTM on the Mexican Spanish dataset, that was a case already discussed to have low performance in general for the acoustical modality). This suggests that acoustical cues for deception can be reached out at primitive levels of the phonetic process.

To close this section, it is important to notice that concatenating all the acoustical modalities into a single vector (rightest orange column in Figs. 5.1 and 5.3) severely hurts the score reached by the best performing view in both datasets. This observation is of particular interest, because even if the acousti-

cal modality has the best performing view, a trivial fusion of the acoustical views has a performance below the worst scored visual view: one could be misled to think that the acoustical modality is a bad source of information while, in reality, it contains many views useful for the task.

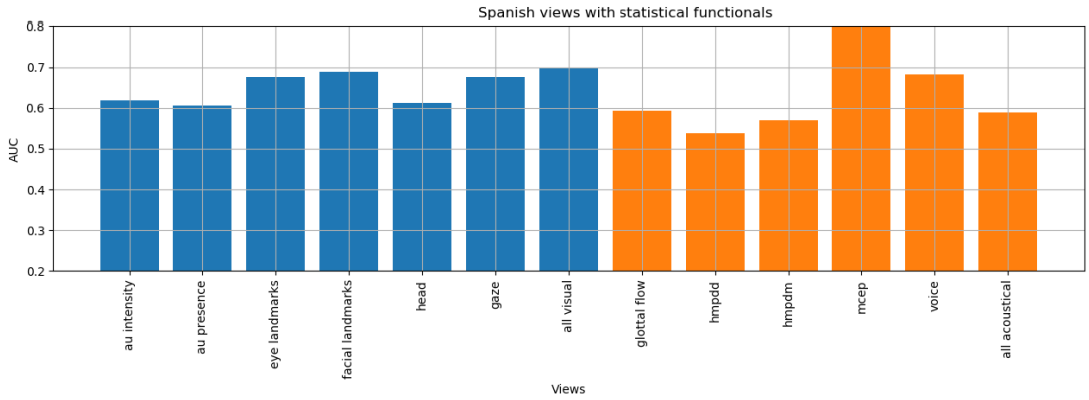

Figure 5.3 AUC achieved by the different views in the Mexican Spanish dataset when using statistical functionals. Views from the visual modality are in blue; the ones for the acoustical modality are in orange; for each case, the rightest column represents the unimododal concatenation.

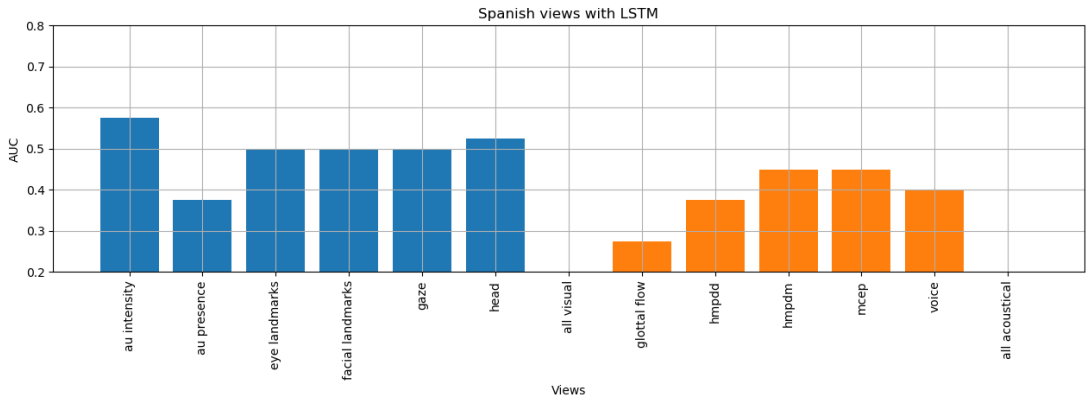

Figure 5.4 AUC achieved by the different views in the Mexican Spanish dataset when using LSTM. Views from the visual modality are in blue; the ones for the acoustical modality are in orange.

5.3 Textual modality

As mentioned before, the textual modality was extracted from text generated by an ASR system. However, Watson Speech to Text does not have a model trained for Mexican Spanish and the performance of transcriptions made with the model for Spain was undesirable. In consequence, transcriptions were only extracted for the court database; additionally, as text was not paired with images/sound at frame level, the LSTM encoding was not used; as the representations used to extract features already deal with variable length texts, there was no need of using statistical functionals.

Results for the textual modality are shown with green columns in Fig. 5.1. In this case, the best views are variations of Bag-Of-Terms representations, particularly using words and character n-grams as the terms; however, a LIWC encoding and extracting syntax features achieve similar results. Character n-grams and the LIWC encoding are congruent with the results presented by (Rill-García et al., 2018), but this is not the case for bags of POS n-grams. We find two possible explanations for the low performance of these views: 1) POS tagging is done based on context (for example, in the sentence “There was an orange in the orange bowl”, “orange” works both as a noun and an adjective); the automatic transcription done by Watson comprise every utterance in the audio stream, retrieving then a mixed text composed of sentences/words spoken by all the people in an audio file. With that conditions, not only transcription is harder (thus having a lower quality text) but the text itself is hard to understand without hearing the original conversation (therefore making automatic POS tagging harder). 2) They used a different tagger (provided by Stanford), which used different POS tags (while we use universal tags, they use a more specialized set of tags for English).

For this modality, concatenating all the views together (rightest green column in Fig. 5.1) improves the score reached by the best single view (LIWC

encoding, that is below the best single view results from the visual and acoustical modalities).

5.4 Multimodal complementarity

Some of the results presented in the previous section suggest empirically that it is useful to combine views to improve single-view performance. However, we want to find out if it is potentially useful to combine modalities in order to achieve better results. In order to do this, we analyze the predictions done at instance level when using statistical functionals (as not only this approach tends to achieve better results but it contemplates the three proposed modalities). This analysis is performed to see how complementary the predictions are at both views and modalities levels: even if each type of features (view or modality) has many mistakes, we want them to be wrong at different instances (so that if we combine them in a proper way we get better results).

To know the best possible result after fusion, we measure the Maximum Possible Accuracy: at instance level, if any of the studied views/modalities classified the instance correctly, the instance is considered as correctly classified; once this has been done for all the instances in the database, the accuracy is calculated. This is then an optimistic measure that considers a perfect fusion.

Also, we want a numeric measure to evaluate how diverse are the errors between the different views/modalities. For this purpose, we use the Coincident Failure Diversity (CFD) metric (Escalante, Montes, and Sucar, 2010), which ranges from 0 (in the case where all the studied views/modalities always make the same label predictions) to 1 (when misclassifications are unique to one view/modality).

As it can be seen in Figs. 5.5 (court) and 5.6 (Mexican Spanish), not only the CFD is far from 0 both at views and modality levels (meaning the different feature sets are far from mispredicting the same examples), but the MPA is

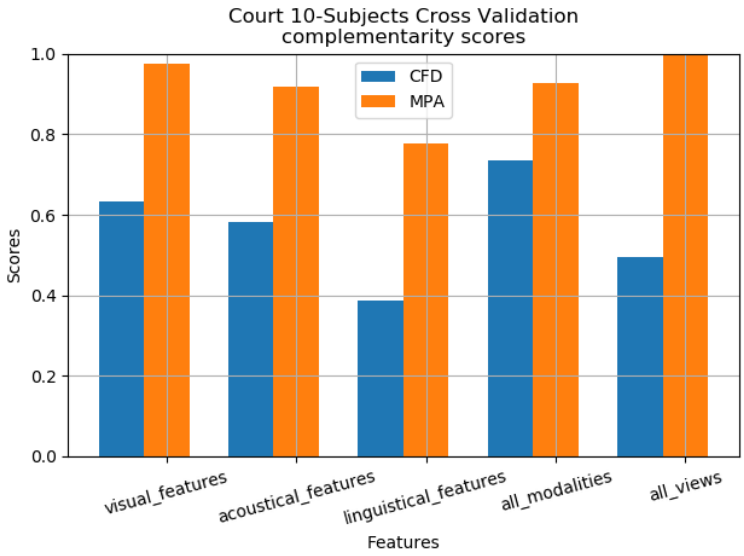

Figure 5.5 CFD between views and modalities from the court dataset, as well as their MPA. From left to right, these metrics are measured using: all the views from the visual modality independently, all the views from the acoustical modality independently, all the views from the textual modality independently, all the views grouped as three modality vectors, all the views independently.

greater at views level rather than at modalities level. This suggests there is, in fact, complementarity both at views and modalities level; also, it seems like there are complementarity reasons to split the different modalities into views (as a perfect fusion of all the proposed views achieves better results than a perfect fusion of three modality vectors).

At this point, we have evidence of the multimodal complementarity between the extracted feature sets; furthermore, there is evidence too on the potential improvement of performance that can be reached by performing multimodal fusion. However, there are results showing that a trivial fusion (early concatenation) is not really useful to bring out this performance improvement (this approach, actually, can hurt the overall scores obtained by the different views). Therefore, it is time to explore other multimodal approaches for deception detection. This is done inspired by ensemble techniques, as detailed in the next

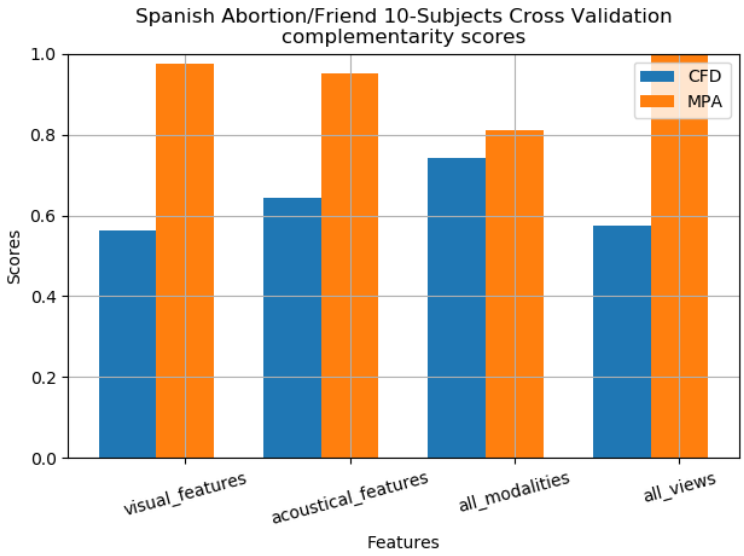

Figure 5.6 CFD between views and modalities from the Mexican Spanish dataset, as well as their MPA. From left to right, these metrics are measured using: all the views from the visual modality independently, all the views from the acoustical modality independently, all the views grouped as three modality vectors, all the views independently.

chapter.

Chapter Six

MULTIMODAL DECEPTION DETECTION

There is a maxim that says “two heads are better than one”. And, overall, this can be further extended to more “heads”; for example, when making important decisions, we often seek for a second, third or more opinions. As a human process, we seek for individual opinions of experts, weight them, and combine them to get a decision as informed as possible (presumably the best one).

In the terrain of machine learning, an ensemble consists of a set of individually trained classifiers whose predictions are combined whenever a new instance needs to be classified. This process is analogous to consulting several “experts” (classifiers) before making a final decision. In the case of multimodal fusion, instead of consulting different “experts” from the same area (different classifiers trained on the same modality) we consulte “experts” from different areas (classifiers trained on different types of data).

On this chapter, we describe some of these fusion methods as well as two methods proposed by us, and use them to fuse the multimodal features used in Chapter 5.

6.1 Traditional fusion methods

As aforementioned, we took an approach based on classifier ensembles; with this paradigm, we ensemble views rather than classifiers. Based on the complementarity study done in Chapter 5 and the work presented in (Rill-Garcia et al., 2019), here we don't present results of fusion done with modality vectors (i.e. vectors built by concatenating all the views from a single modality); this is because splitting data into views seems to be more effective than splitting it just into modalities. As baseline methods, we use a set of traditional ensemble techniques (Polikar, 2006).

Two of them are in the category of late fusion (fusion is done after independent classifications): majority votes and stacking. With majority votes, the final decision is the most frequent class predicted among all the classifiers. Stacking consists on putting together the predictions done by each classifier for an instance to form a new training vector; the vectors extracted from all the predicted instances are then used as a new database for a late classifier and the prediction from that classifier (which we call the "stacker") is the final decision. For this work, we used a SVC classifier with a linear kernel as the stacker (this decision was made after some preliminary experiments, where different linear classifiers were tested without hyperparameter tuning). A variation of both methods (majority votes and stacking) is tested too, where instead of using hard labels the probabilities of both classes are used.

Also, for a more complete evaluation, an additional approach is used as baseline too: early fusion. This consists simply on concatenating all the views together into a single vector before using a classifier. As the simplest form of fusion, this strategy works as the lower bound to surpass for the rest of fusion methods. Two of them are the novel fusion methods introduced in this work and described in the next section.

6.2 A novel fusion strategy

Among the classifier ensemble techniques we can find boosting; particularly, AdaBoost has been shown to improve the prediction accuracy of base learners through an iterative weighting process (Freund and Schapire, 1997). (Barbu, Peng, and Seetharaman, 2010) first presented a multiple view generalization of AdaBoost called Boosting With Shared Sampling Distributions (BSSD), where weak learners are built at “view" level at each iteration (our “view" definition was actually inspired by them). The weak learner with the lower error rate is chosen at each step, and its errors are used to calculate a new probability distribution of the training instances for the next iteration, giving greater weights to the wrongly classified instances. All views share the same sampling distribution, so each weak learner at each iteration gives greater importance to those examples that were “harder” to predict in the previous iterations (pseudocode for this can be seen at Algorithm 1).

Our first approach was extending BSSD with a hierarchical strategy (hierarchical BSSD), by using BSSD with the views from each modality independently and then using the label calculated for each modality as a new feature for late fusion (a diagram of this method is shown in Fig. 6.1). For consistency, the classifier used for late modality fusion is the same used as stacker in the stacking method, since hierarchical BSSD is a way to extend classical BSSD with a stacking approach at modality level.

Based on the previous idea, we could extend classical BSSD with a stacking strategy at view level (stacking BSSD). Adaboost (and therefore BSSD) classifies an instance with a linear function of the labels predicted for such instance from each weak learner trained. The weights from this linear function are learned by the boosting algorithm as a function of the error rate of each weak learner; however, there might be a benefit from learning these weights outside the boosting algorithm. Hatami and Ebrahimpour Hatami and Ebrahim-

Algorithm 1 Boosting With Shared Sampling Distributions (BSSD) algorithm as presented by (Barbu, Peng, and Seetharaman, 2010).

Input: $z_0^j = \{x_i^j, y_i\}_{i=1}^n, j = 1, \ldots, M.$

Initialization: $W_1 = \{w_1(i) = \frac{1}{n}\}_{i=1}^n$

1. **for** $k = 1$ to k_{max} **do**
2. Sample z_k^j from z_0^j using the distribution W_k.
3. Compute hypothesis h_k^j from z_k^j for each view j.
4. Calculate error $\epsilon_k^j : \epsilon_k^j = P_{i \sim W_k}[h_k^j(x_i^j) \neq y_i]$
5. If for each view: $\{\epsilon_k^j\}_{j=1}^M \leq 0.5$, select h_k^* corresponding to $\epsilon_k^* = min_j\{\epsilon_k^j\}$.
6. Calculate $\alpha_k^* = \frac{1}{2}\ln(\frac{1-\epsilon_k^*}{\epsilon_k^*})$.
7. Update $w_{k+1}(i) = \frac{w_k(i)}{Z_k^*} \times e^{-h_k^*(x_i^*)y_i\alpha_k^*}$, where Z_k^* is a normalizing factor.
8. **end for**
9. **Output:** $F(x) = \sum_{k=1\ max}^{k} \alpha_k^* h_k^*(x^*)$.
10. **Final hypothesis:** $H(x) = sign(F(x))$.

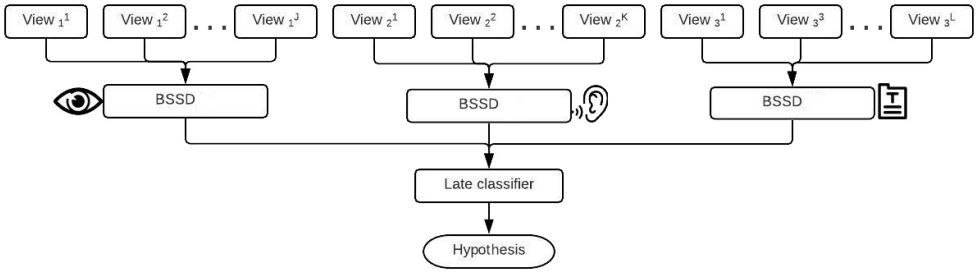

Figure 6.1 Block diagram of Hierarchical Boosting with Shared Sampling Distribution.

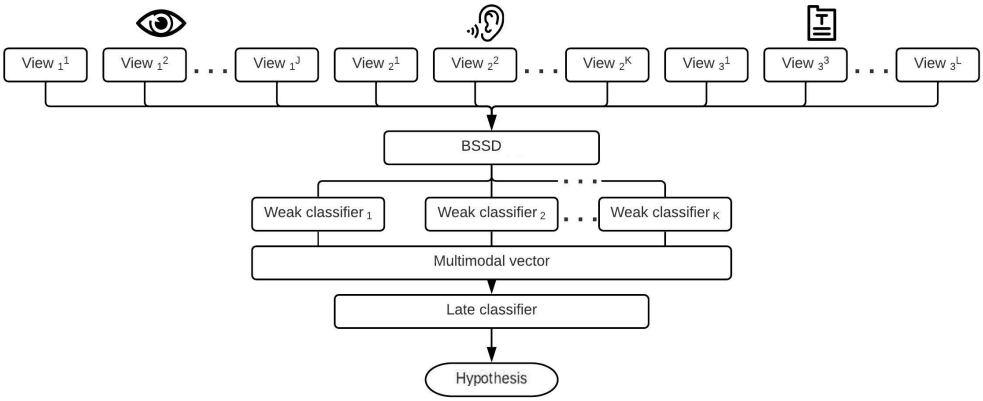

Figure 6.2 Block diagram of Stacking Boosting with Shared Sampling Distribution.

pour, 2007 try a similar approach, using the weak learners obtained by a boosting algorithm as base classifiers for a stacking method, achieving better results than using boosting alone. We use the same approach, using the weak learners generated by BSSD as base classifiers for stacking (a diagram of this method can be seen in Fig. 6.2).

6.3 Fusion results

Four different experiments were performed for each database. First, all the fusion strategies discussed in the previous sections were tested using all the

available views using statistical functionals; then, the same fusion strategies except for hierarchical BSSD are tested with all the available views using LSTM (these experiments are shown in Figs. 6.3-6.6 for both datasets). The next two experiments are basically the same, but using just the best two views per modality according to the results shown in Chapter 3 (these experiments are shown in Figs. 6.7-6.10 for both datasets).

First at all, when using all the available views, fusion methods using statistical functionals tend to work better than using LSTM (there was a similar case when analyzing single views in Chapter 5). However these fusions don't really improve the results obtained by the best single modalities for each case (these scores can bee seen in the captions of each graph). In the court dataset, the best fused results are around the best single view but without surpassing it. In the Mexican Spanish database, when using statistical functionals the best fusion method is far from the best single view; however, when using LSTM the best single view is outperformed by two fusion methods (stacking and majority vote). However, we can see that except for the case where the best single view greatly surpasses fusion methods a simple early fusion is outperformed by more sophisticated fusion strategies.

As with any Machine Learning problem, training can be harmed by the presence of noisy input data. This can be extended to ensembles, where bad classifiers can hurt the overall performance of the whole ensemble. From Figs. 5.1-5.4 we can notice many views with a performance < 0.5 (that is, potentially worse than random guessing). Thus, it could be deduced that the performance of the current ensembles is being decreased because of these "noisy" features. It is to look further into this hypothesis that the next experiments were done using the best views per modality (this is a process analogous to an empirical feature selection, dealing with whole feature sets instead of single features).

For the court database, the results support the given hypothesis since the best results using fusion methods are improved and get to outperform the best

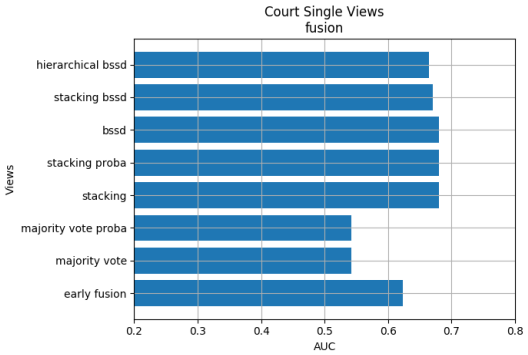

Figure 6.3 Fusion results obtained by using all the views from the court database using statistical functionals. The best single view was gaze (AUC=0.683).

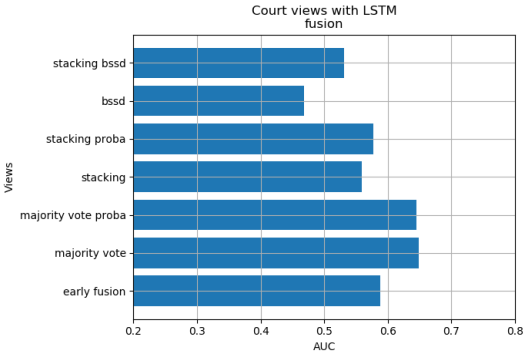

Figure 6.4 Fusion results obtained by using all the views from the court database using LSTM. The best single view was MCEP (AUC=0.680).

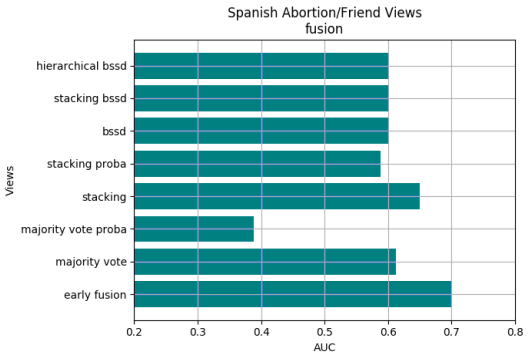

Figure 6.5 Fusion results obtained by using all the views from the Mexican Spanish database using statistical functionals. The best single view was MCEP (AUC=0.856).

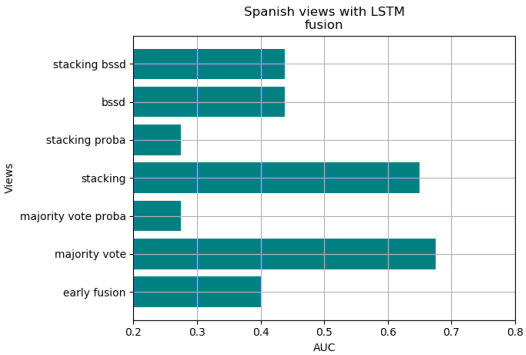

Figure 6.6 Fusion results obtained by using all the views from the Mexican Spanish database using LSTM. The best single view was AU intensity (AUC=0.575).

single view score. Interesting enough, the fusion of features encoded with LSTM improved to the point of surpassing the features encoded with statistical functionals (actually, 5 of the proposed fusion methods outperformed the best single view with statistical functionals despite the lower performance of LSTM encoding with respect to statistical functionals in single view tests).

With the Mexican Spanish dataset, however, we have a different case. With features using statistical functionals, the overall performance of the fusion methods tends to get worse; however, with LSTM encoded features the overall performance of fusion improves with respect to not doing feature set selection (nevertheless, the best fused score is lower than before i.e. using all the available views).

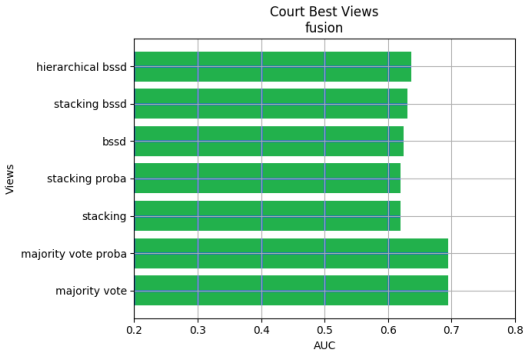

Figure 6.7 Fusion results obtained by using the best views from the court database using statistical functionals. The best single view was gaze (AUC=0.683).

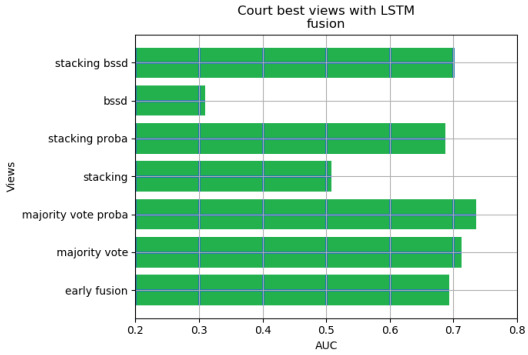

Figure 6.8 Fusion results obtained by using the best views from the court database using LSTM. The best single view was MCEP (AUC=0.680).

6.4 Comparison of fusion methods

Evidence so far suggests that reducing the number of views by choosing the ones with best individual performance can improve the overall performance of ensembles (as one would have expected). However, with such reduction of the number of available views, the performance of boosting based methods tends to get worse even if the chosen views are the ones with the best independent performance.

Boosting methods benefit themselves much from diversity unlike methods like majority vote (that tended to show the best performance along experi-

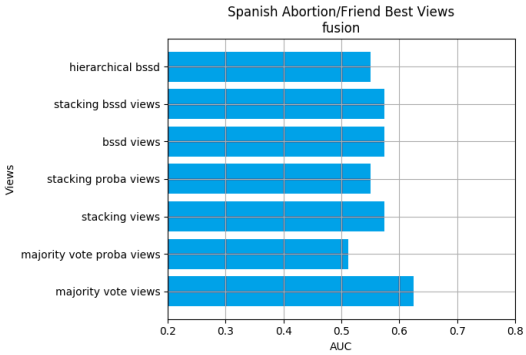

Figure 6.9 Fusion results obtained by using the best views from the Mexican Spanish database using statistical functionals. The best single view was MCEP (AUC=0.856).

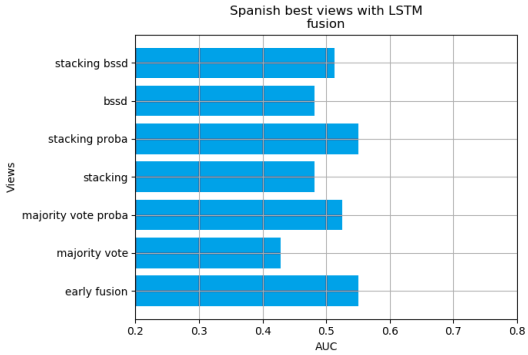

Figure 6.10 Fusion results obtained by using the best views from the Mexican Spanish database using LSTM. The best single view was AU intensity (AUC=0.575).

ments): while majority vote can be severely hurt by having many “poor” voters, boosting can use these poor voters when the best ones are not able to classify properly a hard instance. However, unlike majority vote, boosting methods are highly prone to overfitting; by looking at the algorithm of BSSD (Algorithm 1), one can notice that the method can attach itself to a single view if this is able to classify the training data with 100% accuracy (even if training accuracy is not representative of the general performance of a classifier).

This problem, indeed, was the motivation behind hierarchical BSSD: even if at modality level BSSD overfits to a single view, the method forces the system to take into account at least one view from each modality, thus ensuring that the final decision is in fact multimodal. This is something important to highlight about the boosting methods managed in this thesis, since it deals with the multimodal interpretability of the final decision made by the system.

Looking back at the motivation behind classifier ensembles (that is, searching for the opinions of different experts to take a final decision), it would be desirable to understand the weighting done by the system (i.e. understand-

ing the way in which deception cues are managed to take a decision about a sensible topic). Thanks to the linear nature of boosting, the final decision is a weighted sum of the decision made by the best view at each iteration of the BSSD algorithm: we can know perfectly which are the views taken into account and the weight assigned to such views for the final prediction. Additionally, by using a linear classifier (e.g. SVM) as the base learner, we can know the weight associated to each single feature from each view separately. Therefore, we can trace back the weights assigned to every single feature with a pipe-line of linear functions.

Overall, the problem with fusion for this task is the great amount of information sources available. Early fusion deals with the curse of dimensionality, since we have a huge number of attributes for a small number of training instances (due to the nature of the problem). Under this scope, distributing the data hierarchically into different feature sets seems like a good strategy to deal with the great amount of features obtained from each video; the results presented in this chapter seem to support this idea, since the majority vote, stacking and BSSD variant methods are based precisely on this principle of grouping data and often surpass the performance of early fusion.

However, there are concerns for each of these methods that impact in the final result of fusion. Taking back the idea of “experts”, a classifier trained with a single view is an expert in that area. Majority vote, stacking and BSSD methods take the decisions of these experts in different ways: democratically, by giving each expert the same importance; with preferences, by giving some experts more importance according to how useful they were for decision making during training; and discriminatorily, by only taking into account the opinions of experts who committed the lesser amount of errors during training.

6.4.1 Pros and cons of the different fusion approaches

The democratic approach is useful whenever all the voters are good individually, since even if a few percentage of them is wrong for a certain instance, this weakness is compensated by the rest of the experts; however, even if one of the voters is flawless, the ensemble will have a poor performance if all the rest of the experts have a bad performance (just as observed in our experiments).

The next intuition, then, is to give weights to each vote, trying to minimize the opinion of voters with low performance. To learn this weights the stacking strategy is utilized, using a late classifier to do this automatically; this method, however, is more prone to overfitting. When dealing with many experts, the stacking approach has the advantage of learning automatically which are the bad voters; this is the most probable explanation on why, when using all the available views, stacking surpasses majority vote. However, this happens only when using statistical functionals; this could be explained because using the LSTM encoding has an overall bad performance on single views, thus stacking has the hard task to assign weights to naturally bad voters.

When reducing the number of views to the best ones, however, majority vote tends to outperform stacking. This is probably explained by the different natures of videos in both datasets; e.g. even if stacking decided that visual views are the most important for detecting deception, this may be true just for certain subjects (the ones in the training set); consequently, this weighting may be ineffective for the general case, while the democratic approach of majority vote can work more properly since all the experts are fairly good to discriminate liars in general (confirmed by the cross-validation used to rank individual views).

In the case of the analyzed boosting methods, some voters are discriminated automatically instead of being discarded manually. As with stacking, these methods tend to perform better than majority vote when using all the available views, because these methods discard automatically a great amount

of bad voters. However this can be seen as a disadvantage too, since it can reduce the amount of data considered for the final decision; this tendency is shown when contrasted with stacking, as when using all the available views the performance of stacking tends to be slightly better than the performance of boosting methods. During training, glottal flow and MCEP tended to be the best views chosen at each iteration of the BSSD algorithm, then limiting the final decision to just these two views. However, when using just the best performing views, boosting methods tend to work slightly better than stacking; while stacking deals with weighting this views, boosting just selects the best one and complements it whenever it is necessary by taking the hypothesis from other good expert. However, again, this is prone to overfitting to a single view, thus having a final decision with less information than the one taken by majority vote.

Comparing BSSD with our proposed variants, we can note that stacking BSSD tends to work as well as or better than traditional BSSD; this is more notorious when using the LSTM encoding. Traditional BSSD assigns weights to each weak learner based on the mistakes made in the training instances, thus being prone to the same risks of validating a system by its results on training data; by using a late classifier to learn this weights, we have the same advantages of stacking allowing to ignore base learners that could have a high weight just due to overfitting on the training set while having a bad general performance.

By the other hand, hierarchical BSSD tends to work worse than the other two boosting methods. This behavior can be most likely explained by the same reason of the decreased performance of stacking when using just selected views instead of all the views, as we are taking again a stacking approach with a limited number of features (in this case, modality decisions instead of selected views).

Further conclusions about this and the thesis as a whole are presented in the next closing chapter.

Chapter Seven

CONCLUSIONS

In this thesis we explored high-level features extracted automatically from videos for the deception detection task. These features, analyzed hierarchically as "views" from three different "modalities" (visual, acoustical and textual), were studied with a machine learning approach.

We conducted preliminary experiments with these feature sets separately to gain an insight of their effectiveness on deception detection. Additionally, a study of complementarity between the different feature sets (at view and modality level) was performed to find evidence on the convenience of approaching the deception detection in videos as a multimodal problem.

Afterwards, work was done on multimodal fusion of features to improve the predictive power with respect to single views independently. This fusion was performed with methods based on classifier ensemble techniques, including two novel methods based on boosting first introduced in an article (Rill-Garcia et al., 2019) derived from this thesis.

Experiments were performed on two different datasets (summarized in Table 4.1). The first one is a real-life court trial database composed of public video clips collected from the web. The second one is a database collected for this thesis, composed by videos of Mexican people talking about a sensible and a personal topic in Spanish (to the best of our knowledge, this is the first dataset for deception detection in videos using Latin Spanish as native language).

Validation was done using 10-fold cross-validation based on subjects rather than instances (i.e. no person seen in the training set is part of the test set). This scheme was used because we are interested in subject-independent deception detection, so we want to avoid a classifier to label a video as deceptive because it depicts a subject who was always deceptive in the training set.

With the above mentioned work, we obtained the following answers for the research questions raised in Chapter 1:

There are high-level features that can be extracted automatically using open tools that are useful under different experimental settings for deception detection. Particularly, despite the cultural, language, context and topic-related differences, there were views that showed a tendency as good discriminators of deception in both datasets, namely: gaze direction, eye landmarks, AU (visual modality), MCEP, glottal flow and voice (acoustical modality).

In order to deal with variable length in videos, we found two approaches that can be useful under certain conditions. When it comes to large videos (i.e. videos with a high number of frames), statistical functionals are a good way to summarize video frames into a fixed-size vector; however, the functionals are not able to capture the sequential nature of videos. In order to take advantage of this nature, LSTM networks can be used to encode frame sequences into fixed-size vectors since this neural network architecture deals naturally with sequential data. However, large sequences are expensive in terms of memory and training time; therefore, data extracted from videos (many frames per second) needs to be padded somehow to deal with this problem. The method proposed in this thesis consisted in selecting K ordered key frames with a K-means algorithm based on multimodal vectors. This method is recommended for videos with an average length of $\sim$30 seconds (where data loss due to padding is not relevant); for longer videos, statistical functionals are recommended.

Measures on Coincident Failure Diversity show a complementarity between

the predictions done by different features sets, while the Maximum Possible Accuracy metric suggests a possible improvement by fusing such predictions. Particularly, the results suggest that different views can achieve a better result separately rather than concatenated as multimodal vectors; even if we have 22 different views (compared to 3 different modalities), CFD still shows a good complementarity between the errors committed by the different feature sets in both databases.

When it comes to multimodal fusion, methods based on boosting (BSSD) take advantage of the multimodal diversity of different features that can be extracted from videos; particularly, they have the advantage of allowing an easy interpretation of the decisions made by the system thanks to the linear nature of their predictions. However, the results achieved by other traditional ensemble methods can outperform the ones reached with the proposed boosting methods; this is particularly true when doing feature set selection: in this case, fusion methods are able to improve single-view results. Additionally, doing feature set selection helps to highly improve the results obtained when using features encoded using LSTM networks (in single-view experiments, LSTM encoding was outperformed by statistical functionals).

Overall, non-trivial fusion strategies can improve a simple early fusion approach, and multimodal fusion can improve the results obtained by single views. However, deciding the best fusion strategy for the task is not trivial, and there are still many areas of improvement. A brief discussion on future work for this thesis is presented in the next section.

7.1 Future work

With respect to fusion methods, the first step would be hyperparameter tuning for the classifiers trained with each view separately, in order to improve the performance of the base learners. It would be also useful to perform hyperpa-

rameter tuning for the classifiers used as stackers in the boosting methods.

With respect to feature selection, it would be convenient to perform analysis beyond empirical results to select the best feature sets (views) from the features extracted. Furthermore, it would be useful to perform feature selection at view level in order to reduce the dimensionality of training data.

With respect to deep learning approaches, it would be interesting to train LSTM networks using a greater number of frames (thus attempting to capture more information keeping into account the sequential nature of frames). Furthermore, the K-means algorithm for key multimodal frame selection could be improved by performing a synchronization of text with the other modalities. Although there are tools able to assign time stamps to text, the time used to pronounce a single word implies more than a single frame; a first idea to deal with this is using the statistical functions strategy, by using descriptive statistics on the elapsed frames during the pronunciation of a word (a word embedding would be needed for the textual modality).

Finally, with respect to the videos used for experiments, it is clear that more data is needed to build more robust systems (this is particularly true when exploring deep learning approaches). Automatic deception detection has gained much interest in recent years (not to say that deception detection in general has always been a topic of interest); however, as expressed in this thesis, databases available for the task are scarce and often not publicly available. Therefore, one of the more important tasks left is to increase the amount of data for training; with respect to us, we are still gathering more samples hoping to be able to distribute our database publicly.

REFERENCES

Abouelenien, Mohamed, Verónica Pérez-Rosas, Rada Mihalcea, et al. (2017). "Detecting deceptive behavior via integration of discriminative features from multiple modalities". In: *IEEE Transactions on Information Forensics and Security* 12.5, pp. 1042–1055.

Abouelenien, Mohamed, Verónica Pérez-Rosas, Bohan Zhao, et al. (2017). "Gender-based multimodal deception detection". In: *Proceedings of the Symposium on Applied Computing*. ACM, pp. 137–144.

Atrey, Pradeep K et al. (2010). "Multimodal fusion for multimedia analysis: a survey". In: *Multimedia systems* 16.6, pp. 345–379.

Baltrušaitis, T., C. Ahuja, and L. Morency (2019). "Multimodal Machine Learning: A Survey and Taxonomy". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41.2, pp. 423–443. ISSN: 0162-8828. DOI: 10.1109/TPAMI.2018.2798607.

Baltrusaitis, Tadas et al. (2018). "Openface 2.0: Facial behavior analysis toolkit". In: *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE, pp. 59–66.

Barbu, Costin, Jing Peng, and Guna Seetharaman (2010). "Boosting information fusion". In: *2010 13th International Conference on Information Fusion*. IEEE, pp. 1–8.

Bishop, Christopher M (2006). *Pattern recognition and machine learning*. springer.

Bond Jr, Charles F and Bella M DePaulo (2006). "Accuracy of deception judgments". In: *Personality and social psychology Review* 10.3, pp. 214–234.

Bouaziz, Mohamed et al. (2016). "Parallel long short-term memory for multi-stream classification". In: *2016 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, pp. 218–223.

Carissimi, Nicolo, Cigdem Beyan, and Vittorio Murino (2018). "A multi-view learning approach to deception detection". In: *2018 13th IEEE International*

Conference on Automatic Face & Gesture Recognition (FG 2018). IEEE, pp. 599–606.

Cheng, Keens Hiu Wan and Roderic Broadhurst (2005). “The detection of deception: The effects of first and second language on lie detection ability”. In: *Psychiatry, Psychology and Law* 12.1, pp. 107–118.

Degottex, Gilles et al. (2014). “COVAREP—A collaborative voice analysis repository for speech technologies”. In: *2014 ieee international conference on acoustics, speech and signal processing (icassp)*. IEEE, pp. 960–964.

Ekman, Paul (2003). “Darwin, deception, and facial expression”. In: *Annals of the New York Academy of Sciences* 1000.1, pp. 205–221.

— (2009). *Telling lies: Clues to deceit in the marketplace, politics, and marriage (revised edition)*. WW Norton & Company.

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

Escalante, Hugo Jair, Manuel Montes, and Enrique Sucar (2010). “Ensemble particle swarm model selection”. In: *The 2010 International Joint Conference on Neural Networks (IJCNN)*. IEEE, pp. 1–8.

Espinosa, Humberto Pérez and Carlos A Reyes García (2009). “Detection of negative emotional state in speech with ANFIS and genetic algorithms.” In: *MAVEBA*, pp. 25–28.

Farah, Martha J et al. (2014). “Functional MRI-based lie detection: scientific and societal challenges”. In: *Nature Reviews Neuroscience* 15.2, p. 123.

Freund, Yoav and Robert E Schapire (1997). “A decision-theoretic generalization of on-line learning and an application to boosting”. In: *Journal of computer and system sciences* 55.1, pp. 119–139.

Gamer, Matthias and Wolfgang Ambach (2014). “Deception research today”. In: *Frontiers in Psychology* 5.

Ganchev, Todor, Nikos Fakotakis, and George Kokkinakis (2005). “Comparative evaluation of various MFCC implementations on the speaker verification task”. In: *Proceedings of the SPECOM*. Vol. 1. 2005, pp. 191–194.

Gorbova, Jelena et al. (2018). “Integrating Vision and Language for First-Impression Personality Analysis”. In: *IEEE MultiMedia* 25.2, pp. 24–33.

Hatami, Nima and Reza Ebrahimpour (2007). “Combining multiple classifiers: diversify with boosting and combining by stacking”. In: *International Journal of Computer Science and Network Security* 7.1, pp. 127–131.

Hochreiter, Sepp and Jürgen Schmidhuber (1997). “Long short-term memory”. In: *Neural computation* 9.8, pp. 1735–1780.

Hu, Jiaqi (2008). “Data fusion: a first step in decision informatics”. PhD thesis. Rensselaer Polytechnic Institute.

Karimi, Hamid, Jiliang Tang, and Yanen Li (2018). “Toward End-to-End Deception Detection in Videos”. In: *2018 IEEE International Conference on Big Data (Big Data)*. IEEE, pp. 1278–1283.

Mitre-Hernandez, Hugo et al. (2019). “Assessing cognitive load using oculometrics to identify deceit during interviews”. In: *Applied Cognitive Psychology* 33.2, pp. 312–321.

Morales, Michelle Renee (2018). “Multimodal Depression Detection: An Investigation of Features and Fusion Techniques for Automated Systems”. In:

Morales, Michelle Renee, Stefan Scherer, and Rivka Levitan (2017). “OpenMM: An Open-Source Multimodal Feature Extraction Tool.” In: *INTERSPEECH*, pp. 3354–3358.

Pampouchidou, Anastasia et al. (2016). “Depression assessment by fusing high and low level features from audio, video, and text”. In: *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*. ACM, pp. 27–34.

Pérez-Rosas, Verónica et al. (2015). “Deception detection using real-life trial data”. In: *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*. ACM, pp. 59–66.

Polikar, Robi (2006). “Ensemble based systems in decision making”. In: *IEEE Circuits and systems magazine* 6.3, pp. 21–45.

Rill-Garcia, Rodrigo et al. (2019). “High-Level Features for Multimodal Deception Detection in Videos”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. –.

Rill-García, Rodrigo et al. (2018). “From Text to Speech: A Multimodal Cross-Domain Approach for Deception Detection”. In: *International Conference on Pattern Recognition*. Springer, pp. 164–177.

Samuel, A. L. (2000). “Some studies in machine learning using the game of checkers”. In: *IBM Journal of Research and Development* 44.1.2, pp. 206–226. ISSN: 0018-8646. DOI: 10.1147/rd.441.0206.

Sawada, Naoki, Ryo Masumura, and Hiromitsu Nishizaki (2017). “Parallel Hierarchical Attention Networks with Shared Memory Reader for Multi-Stream Conversational Document Classification.” In: *INTERSPEECH*, pp. 3311–3315.

Tao, Jianhua and Tieniu Tan (2005). “Affective computing: A review”. In: *International Conference on Affective computing and intelligent interaction*. Springer, pp. 981–995.

Vogt, Thurid and Elisabeth André (2005). “Comparing feature sets for acted and spontaneous speech in view of automatic emotion recognition”. In: *2005 IEEE International Conference on Multimedia and Expo*. IEEE, pp. 474–477.

Wan, Jun et al. (2017). “Results and analysis of chalearn lap multi-modal isolated and continuous gesture recognition, and real versus fake expressed emotions challenges”. In: *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3189–3197.

Wiseman, Richard et al. (2012). “The eyes don’t have it: lie detection and neuro-linguistic programming”. In: *PloS one* 7.7, e40259.

Wu, Zhe et al. (2018). “Deception detection in videos”. In: *Thirty-Second AAAI Conference on Artificial Intelligence*.