Università degli Studi di Genova

Istituto Italiano di Tecnologia
Department of Pattern Analysis and Computer Vision
(PAVIS)

Investigating Social Interactions Using Multi-Modal Nonverbal Features

Nicolò Carissimi

Abstract

Every day, humans are involved in social situations and interplays, with the goal of sharing emotions and thoughts, establishing relationships with or acting on other human beings. These interactions are possible thanks to what is called social intelligence, which is the ability to express and recognize social signals produced during the interactions. These signals aid the information exchange and are expressed through verbal and non-verbal behavioral cues, such as facial expressions, gestures, body pose or prosody. Recently, many works have demonstrated that social signals can be captured and analyzed by automatic systems, giving birth to a relatively new research area called social signal processing, which aims at replicating human social intelligence with machines. In this thesis, we explore the use of behavioral cues and computational methods for modeling and understanding social interactions. Concretely, we focus on several behavioral cues in three specific contexts: first, we analyze the relationship between gaze and leadership in small group interactions. Second, we expand our analysis to face and head gestures in the context of deception detection in dyadic interactions. Finally, we analyze the whole body for group detection in mingling scenarios.

Acknowledgements

There are a number of people I would like to express my gratitude to.

To my supervisor, Prof. Vittorio Murino, for giving me this important chance.

To Cigdem and Paolo, for the continuous big support.

To my fellow PhD mates, Roy, Vaibhav, Shahab, Xiangping, Riccardo and Behzad, for sharing the hurdles.

Finally, but most importantly, to my beloved family. My mom, dad and my grandmas Eraclia (ciao nonna) and Tina, without whom this would have not been possible. My daughter, Noa, and her sister, Lizzie (my "first one"), for making me laugh. And angry (sometimes). And Lana, for the good and the bad.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Although humans are not the only social animals on earth, our interactions are among the most complex, nuanced and varied in the animal kingdom. Every day we are involved in social situations with other human beings, with the goal of sharing our emotions and thoughts, establishing relationships with and acting (both physically and mentally) on other human beings for different purposes. These interactions are made possible thanks to our bodies and their ability to express and perceive what are called as em social signals [Pantic et al., 2011, Mehu and Scherer, 2012, Brunet and Cowie, 2012].

In recent years, social signals have been described in different ways. One of the first working definitions [Pantic et al., 2011] defines them as "[...] communicative or informative signals that, either directly or indirectly, provide information about social facts, that is, about social interactions, social emotions, social evaluations, social attitudes, or social relations". In [Mehu and Scherer, 2012], social signals are defined as "acts or structures that influence the behaviour or internal state of other individuals". In [Brunet and Cowie, 2012], they are seen as "actions whose function is to bring about some reaction or to engage in some process". Even though existing definitions might appear different, they all share the common idea that social signals consist of observable behaviors people display while interacting with each other, and that they produce changes in individuals [Burgoon et al., 2017].

Concretely, social signals are expressed through the use of non-verbal behavioral cues, which can be described as temporal changes in neuromuscular and physiological activities, or as physical characteristics and configurations. Behavioral cues are usually grouped in five classes [Vinciarelli et al., 2009]: *physical appearance*, *gesture and posture*, *face and eyes behavior*, *vocal behavior* and *space and environment*. *Physical appearance* includes characteristics such as height, body shape, skin and hair color, clothes and ornaments, encoding social signals such as attractiveness and personality. Body *gestures and posture* are known indicators of emotions (such as happiness, sadness or fear) [Coulson, 2004], attitude towards social situations [Richmond et al., 1991, Scheflen, 1964] and social information (such as status and dominance [McArthur and Baron, 1983]). *Face and eyes* behavioral cues include facial expressions and gaze. They are, perhaps, the most expressive cues when emotions are conveyed, and several studies [Ambady and Rosenthal, 1992, Grahe and Bernieri, 1999] show that they play an important part when involved in social perception. *Vocal* nonverbal behaviors surround and influence the message that is being actually said. They include voice quality (e.g. pitch, tempo and energy), linguistic vocalizations (non-words such as "ehm" or "uhm"), non-linguistic vocalizations (e.g. laughing, crying, etc.), silence and turn taking. Finally, interpersonal space and spacial arrangements between people (*space and environment*) provide clues about mutual relationships and personalities.

Given the bidirectional nature of interactions, humans have the ability not only to produce social signals, but also to detect and understand those coming from other humans, exhibiting what social psychology calls *social intelligence* [Thorndike, 1920, Ambady and Rosenthal, 1992, Albrecht, 2006]. This ability has always been considered exclusive to human beings, but recent technological advancements have shown that social signals can be captured with sensors (such as cameras and microphones) and analyzed using computer vision and machine learning techniques. This technological leap led to the idea that social intelligence could be replicated with machines, giving birth to the research area called *socially-aware computing* [Pentland, 2004], or *social signal processing* (SSP) [Pentland, 2007]. The main problems SSP focuses on are three: the *modeling* of the laws that govern the use of social signals, the *analysis* of social signals and their *synthesis* through the use of virtual and physical means (e.g. virtual assistants and social robots). This thesis addresses the analysis problem and focuses on non-verbal behavior based on gaze, face and body.

1.1 Motivation

The rate at which data about humans - and the world we interact with - is being generated (in the form of images, text and audio) has quickly made unpractical its manual inspection and analysis. Although we are still far away from the goal of replicating social intelligence with machines, computers are faster than humans at processing huge amounts of information, and able to detect patterns that we might easily miss [Pfister et al., 2011]. In this context, SSP can be applied as a tool that can aid, not substitute, humans in the analysis and interpretation of this huge amount of data (as already seen in several domains, such as surveillance [Cristani et al., 2011, Ricci et al., 2015, Setti et al., 2015] or multimedia tagging [Pantic and Vinciarelli, 2009, Soleymani and Pantic, 2012]). Our research furthers the work done in this area.

More specifically, the applications we tackled, i.e. emergent leadership, deception and group detection, have all potential applications in real world scenarios. When a job candidate is applying for a managing role, recruiters might look for a leader-type personality. This trait is usually assessed via specific kinds of interviews and tests, all of which are executed manually by humans. Having systems that automatically provide additional information can support the hiring process and, ideally, reduce human bias. The same considerations hold true for deception detection, which has been subject of study in social psychology for many decades [DePaulo et al., 2003, Ekman and Friesen, 1969, Zuckerman et al., 1981]: automatic systems able to detect patterns and cues that even a trained observer would hardly notice, might be able to help social psychologists find correlations between specific human behaviors and deception. Automatic group detection is another useful tool that could benefit different activities, from surveillance, where automatic retrieval of video segments showing groups of people could speed up the video analysis, to the design and planning of public areas, where understanding where groups of people gather more frequently could help exploiting the available space in a better way.

1.2 Thesis Objective

The goal of our work is to advance research aimed at understanding the relationship between social signals, nonverbal behavioral cues and social interactions. We start by focusing on a single important cue, *gaze*; then, in a ”zoom out” approach, after shifting the focus to the cues generated by the *face*, we expand the analysis to the whole *body*. Specifically, we focus on nonverbal behavioral cues in three different scenarios: first, we analyze the relationship between emergent leadership and gaze in small groups of people. Second, we analyze the relationship between deception, face and head gestures in dyadic interactions. Finally, we exploit whole body poses to detect groups of interacting people in crowded mingling events.

1.3 Contributions

We first devised a novel method for detecting emergent leaders in meeting environments with small groups of people. Emergent leadership detection is not a novel problem and many existing works devised effective frameworks using multimodal (audio/video) [Sanchez-Cortes et al., 2012b, Sanchez-Cortes et al., 2012a, Sanchez-

Cortes, 2013] or only audio based features [Hung et al., 2011]. However, there might be scenarios where audio information is not available and the only way to analyze the social interaction is by using visual information. We, thus, present a novel detection method using features based on gaze (modeled in terms of visual focus of attention, VFOA). We also introduce a new dataset and present a comprehensive comparison of several VFOA methods involving imbalanced dataset classification methods (a problem never considered for leadership detection).

Second, we explored the use of several face-based features for deception detection. Existing works already used nonverbal features extracted from face cues, but they were extracted manually [Pérez-Rosas et al., 2015a] and based on handcrafted features [Mimansa et al., 2016] (such as facial action units). Our proposed work is the first method that automatically extracts learned features based on several deep neural networks architectures. Additionally, for the first time in deception detection research, we employ a feature fusion technique based on multi-view learning [Xu et al., 2013], which is more effective than simple feature concatenation used in existing works.

Finally, we extracted features for group detection in mingling events based on 2D and 3D body poses. Given the crowded nature of these scenarios, severe body occlusions are always present, leading to lower quality features and detection results. We devised a novel reconstruction algorithm that, given an incomplete 2D body pose, outputs a complete one. Results show increased detection performance.

1.4 Publications

The following papers have been published in literature:

- Beyan C., Carissimi N., Capozzi F., Vascon S., Bustreo M., Pierro A., Becchio C., Murino V., "Detecting emergent leader in a meeting environment using nonverbal visual features only". In Proceedings of the 18th ACM International Conference on Multimodal Interaction 2016 Oct 31 (pp. 317-324). ACM.

- Carissimi N., Beyan C., Murino V., "A Multi-View Learning Approach to Deception Detection". In Automatic Face & Gesture Recognition (FG 2018), 2018 13th IEEE International Conference on 2018 May 15 (pp. 599-606). IEEE.
- Carissimi N., Rota P., Beyan C. Murino V., Filling the Gaps: Predicting Missing Joints of Human Poses Using Denoising Autoencoders, HBUGEN Workshop, European Conference on Computer Vision (ECCV) 2018.

1.5 Thesis Outline

The rest of the thesis is organized as follows.

Chapter 2 reviews the related work. After an introduction on SSP, we present the related work on the specific topics of leadership detection, deception detection and group detection.

We start our analysis by focusing on gaze in Chapter 3, where we examine its relationship to emergent leadership in face-to-face small groups interactions. We first introduce our dataset and then describe features and methodology, concluding with results and discussions.

We expand to face and deception detection in Chapter 4. We first briefly describe the multimodal dataset we used and we explore the various features extracted from the face. We then describe the learning algorithms used in our experiments, and we report the obtained results and our conclusions.

In Chapter 5 we broaden our analysis to the whole body, describing the pipeline from feature extraction to the final group detection. Results and discussion conclude the chapter.

Finally, in Chapter 6 we summarize the contributions of this thesis, discuss about limitations and suggest possible future work.

Chapter 2

Related Work

In this Chapter we review the related work, focusing on the chosen application contexts of emergent leadership detection in Section 2.1, deception detection in Section 2.2 and group detection in Section 2.3.

2.1 Leadership

In a social context, a leader is a person who has authority and power over a group of people, and can exert dominance, influence and control over them [Sanchez-Cortes et al., 2012b]. Similarly, an emergent leader (EL) is a person who naturally shows these characteristics among a group [Sanchez-Cortes et al., 2012b].

Automatically detecting emergent leaders in a meeting environment is a recent problem among SSP topics. The existing studies can be grouped according to *i)* the modality nonverbal features are extracted from, i.e. audio [Feese et al., 2011, Hung et al., 2011], video (such as our work) or audio and video fusion [Sanchez-Cortes et al., 2012b, Aran and Gatica-Perez, 2010, Jayagopi et al., 2009], *ii)* the type of information extracted, e.g. head and body activity [Sanchez-Cortes et al., 2012b] or visual focus of attention (VFOA) from head pose [Hung et al., 2008, Sanchez-Cortes, 2013], *iii)* the leader evaluation method, i.e. detecting only the emergent leader (or the dominant person) [Sanchez-Cortes, 2013], or detecting the most and the least emergent leaders [Aran and Gatica-Perez, 2010, Jayagopi et al., 2009].

In [Sanchez-Cortes et al., 2012b], leadership detection in a meeting environment was investigated using nonverbal audio and video based features. The emergent leadership was measured using the concepts of dominance, influence and control. The main assumptions of that study [Sanchez-Cortes et al., 2012b] were that a socially dominant person receives more frequent and longer lasting glances by other people, looks at others while speaking, uses more gestures, is more talkative and has longer speech turns. Nonverbal audio features based on speaking turn duration and visual features extracted from head and body activities were defined accordingly. Unlike our study, head pose was not used. In [Feese et al., 2011] the leadership styles "authoritarian" and "individually considerate" were estimated using nonverbal features extracted from audio, such as average single speaking energy, total speaking length and change in speaking turn duration. The prediction of leadership style was performed using logistic regression using only the features obtained from leaders, meaning that not all participants were analyzed, which can be seen as a drawback. Another study, which only uses audio to detect the dominant person in a meeting environment, was presented in [Hung et al., 2011]. Speaker diarization was used and the results showed that dominance estimation is robust to diarization noise. In [Aran and Gatica-Perez, 2010] and [Jayagopi et al., 2009], dominance in group conversation was investigated using short meeting segments. Different from our work, [Aran and Gatica-Perez, 2010] used scenario meetings where a specific role was assigned to each participant. In both studies [Aran and Gatica-Perez, 2010, Jayagopi et al., 2009], the most dominant and the least dominant people were identified independently using different annotator agreements, i.e. full agreement and majority agreement (in contrast, we classified the most emergent and the least emergent person with a common model). Body motion was used to extract nonverbal features from visual activity. A ranking procedure based on Gaussian Mixture Model

(GMM) using ranked Support Vector Machine (SVM) scores was applied (whereas we utilize SVM and its variations). The results in [Aran and Gatica-Perez, 2010] showed that in general visual features were not successful to estimate the dominance, while their fusion with audio features usually performed better than audio only. In [Jayagopi et al., 2009], on the contrary, audio-visual fusion did not yield any better performance than audio-only features.

A study using head pose to obtain VFOA to find the visual attention for dominance estimation was presented in [Hung et al., 2008]. In that study [Hung et al., 2008], VFOA for a person was labeled manually (in contrast to our study, in which VFOA is estimated automatically) and also detected automatically using a Bayesian formulation. As nonverbal features, the total "received visual attention" and "looking while speaking" features were used. The results, using both manually and automatically extracted cues, showed that audio cues were more effective than visual cues, but the latter could still be useful in the absence of audio sensors.

The most similar study to ours is [Sanchez-Cortes, 2013], since it aimed to detect the emergent leaders in a meeting environment and used nonverbal video-based features (although combined with nonverbal audio based features as well) extracted from VFOA. In that study [Sanchez-Cortes, 2013], emergent leadership detection performance was evaluated in terms of the variables "leadership", "dominance", "competence" and "liking", while in our evaluation we used human annotations and social psychology questionnaires. Results showed low VFOA prediction accuracy (42%), which might be the reason for the poor performance (except dominance) of nonverbal visual features compared to nonverbal audio features.

In our study (Chapter 3), we use head pose to approximate gaze and predict the VFOA. The nonverbal visual features are defined using VFOA, which is estimated automatically with a supervised method. The most and the least emergent leaders in short meeting segments are estimated using SVM and its variants. Leadership annotations of human observers (based on full and majority agreement) are used to learn and evaluate the leadership model. The efficiency of proposed nonverbal visual features are evaluated based on human annotations and furthermore validated by the social psychology questionnaires.

2.2 Deception

Deception and its detection is an ongoing research subject, based on the belief that there exist specific cues that liars exhibit and cannot hide. As Freud wrote in one of his works, “He who has eyes to see and ears to hear can convince himself that no mortal can keep a secret. If his lips are silent, he chatters with his fingertips; betrayal oozes out from every pore.” [Freud, 1959].

Two main types of cues are explored for the detection process: behavioral and physiological cues. One of the most common devices used to capture physiological cues from a person is the polygraph, which measures respiration, skin conductance, blood volume and pulse rate. The oldest test used to assess deception through physiological responses (even though its unreliableness was shown in many studies, like [Lykken, 1985]) is the relevant-irrelevant test [Larson et al., 1932], in which the polygrapher asks a series of relevant (regarding the deception) and irrelevant questions to the suspect and his/her physiological responses to the different kinds of questions are compared. More sophisticated devices, like the electroencephalogram, were also used to measure brain activity during specific tests that should solicit specific memory activity [Gamer, 2014].

Recently, several studies focused on automated deception detection based on behavioral cues. One of the first works on multimodal deception detection is [Jensen et al., 2010], where the analyzed dataset consisted of audio and video recordings of face to face interviews in which people were instructed to lie about a fake crime (mock theft). Gesture and posture behavioral cues (e.g. head and hands pose, body posture) were extracted from video, vocal behavioral cues (e.g. voice pitch and energy, turn takings, speech time) were extracted from audio and verbal behavioral cues (e.g. pleasantness, emotiveness, etc.) were extracted from analysis of interviews transcription. The detected verbal and non-verbal cues were then used in two tests: in the first one, behavioral cues were used to predict meta-communicative meanings (i.e. involvement, dominance, tenseness and arousal), which were, in turn, used to infer the level of deception/truthfulness (these relationships were modelled by applying a modified Brunswikian lens model [Scherer, 1982]). In the second test, the behavioral cues were used to directly predict deception. The second test led to the most significant results. All the features, both verbal and nonverbal, were concatenated in a unique vector and used to build a model through discriminant analysis.

Recently, [Pérez-Rosas et al., 2015b] and [Pérez-Rosas et al., 2015a] presented a novel dataset containing video clips showing real life trial hearings and interviews collected from the web, where deception was, for the first time, not the result of an imposed role play, but, instead, a spontaneous act in a context with high stakes. Similarly to the work of [Jensen et al., 2010], they used verbal and nonverbal cues to detect deception. Verbal cues consisted in uni-grams and bi-grams obtained from the bag-of-words representation of the recordings transcriptions, psycholinguistic features based on the Linguistic Word Count (LIWC) lexicon [Pennebaker et al., 2001], and syntactic complexity features [Lu, 2010]. Nonverbal cues included facial displays and hand gestures manually annotated using the MUMIN coding scheme [Allwood et al., 2007]. Features were concatenated and used to build classifiers with decision trees and random forests. After executing feature ablation studies, they obtained the highest classification accuracy by using only facial displays and random forest classifiers.

Another form of deception that gained growing attention is the written deception. It recently became subject of study thanks to the increasing use of internet and the web as interaction and self-expression tools, and the consequent availability of huge amount of data that can be easily accessed and analyzed. Deception can be used in clickbait news, which are sensational and often false news with the only purpose of drawing more visitors to increase revenue from advertisements [Chen et al., 2015]. Another example are false reviews that can help promoting a hotel or damage a competitors restaurant [Ott et al., 2011]. It became ever so important to discern truthful from deceptive content. Even though data is unimodal (text), different kinds of linguistic features are used for learning and classifying.

In [Pérez-Rosas and Mihalcea, 2015], Perez-Rosas and Mihalcea focused on detection of deception in open domain, i.e. not related to any specific topic. They collected statements from volunteers, consisting of a balanced number of lies and truths, without enforcing any specific topic. After extracting numerous linguistic features (uni-grams, syntax, semantic, readability and syntactic complexity), they combined different sets by concatenating them in a unique vector and built a Support Vector Machine (SVM) classifier. The highest accuracy was achieved with syntactic features.

Perez-Rosas and Mihalcea in [Pérez-Rosas and Mihalcea, 2014] also explored deception in different cultures and languages. As in [Pérez-Rosas and Mihalcea, 2015], different statements were collected from volunteers, but this time they were selected from three specific different countries (USA, India and Mexico). Statements were written in English and Spanish, and specific topics were enforced. Only uni-grams and semantic features were extracted and used to build SVM, and several tests were executed, comparing classifiers accuracy within-culture and cross-culture.

In [Ott et al., 2011], deceptive opinion spam is analyzed. Specifically, truthful positive reviews for popular hotels were collected from a famous website, and deceptive positive reviews for the same hotels were created and gathered using Amazon Mechanical Turk. With a linear SVM and different combinations of n-grams, syntactic and psycholinguistic features, they obtained a classification accuracy of almost 90%.

Generally, all papers in literature employed handcrafted features. To the best of our knowledge, this is the first work that focuses on learned face-based features, showing improved classification performance. Additionally, papers in literature concatenated features extracted from different modalities into one single vector which was used to train models (e.g. SVM, random forests). However, this concatenation is not always statistically meaningful, because each group can have different properties [Kincaid et al., 1975]. We exploit this heterogeneity to build better models by employing multi-view learning, where each feature group corresponds to a view, each view is modelled by a function and all the functions are jointly optimized. We show that, by employing multi-view learning, we can improve the classification accuracy while getting also insights about the importance of the features.

2.3 Groups

Social gatherings are a common human activity which consist of individuals interacting with each other in a shared space. The phenomenon has received growing attention from the computer vision community in the last years, and its analysis has numerous potential applications in surveillance, human-robot interaction and video interpretation and retrieval. An example of social gathering are free standing conversational groups (FCGs), shown in Figure 2.1. FCGs happen in an unconstrained way and, contrary to other types of meetings (e.g. round-table ones), they are dynamic by nature, increasing and decreasing in size, moving and splitting. This makes them inherently difficult to analyze. One approach to detect FCGs is to use *f-formations* [Kendon, 1990], a concept coming from social psychology and based on proxemics, which provides FCGs with geometrical properties. Formally, f-formations are defined as socio-spatial organizations of people in what is called p-space, surrounding a shared convex area called o-pace, which everyone is oriented and has equal access to (Figure 2.2, top). Any other person staying outside the p-space (in the area called r-space) does not belong to the f-formation. Typical spacial arrangements of people are shown in Figure 2.2 (bottom).

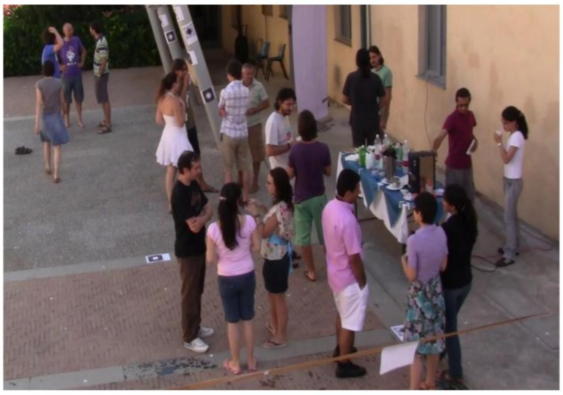

Figure 2.1: An example of free standing conversational groups. [Cristani et al., 2011].

Many works tried to tackle the problem of detecting f-formations, some of them focusing on feature extraction (i.e. distance, location and head/body orientation) and others focusing on the actual grouping of people. One of the pioneering works, [Cristani et al., 2011] proposed an unsupervised method which takes as input locations and head orientations and detects f-formations by employing a voting strategy based on a generalized Hough transform. [Setti et al., 2013b] extended [Cristani et al., 2011] by devising a multi-scale Hough transform, with a different voting strategy for different group sizes, modeling small and large aggregations. [Hung and Kröse, 2011] proposed a dominant set approach based on a learned affinity matrix which encodes relationships among persons. [Vascon et al., 2014] used game theory and probability distributions. Affinity between pairs of people is expressed as a distance between distributions over the most plausible oriented region of attention, while clustering is executed multi-payoff evolutionary game theory. [Setti et al., 2015] proposed an approach which iteratively assigns individuals to F-formations using a graph-cut based optimization, and updates the centres of the f-formations, pruning unsupported groups. [Ramírez et al., 2016] devised two algorithms based on a learn and forget strategy (Ebbinghaus forgetting curve [Ebbinghaus, 2013]). Group detection was performed by clustering graph's nodes representing people, where edges encoded relative distance, velocity, the learn/forget function value and the persons' field of view (FoV). An additional thresholding step based on inter-

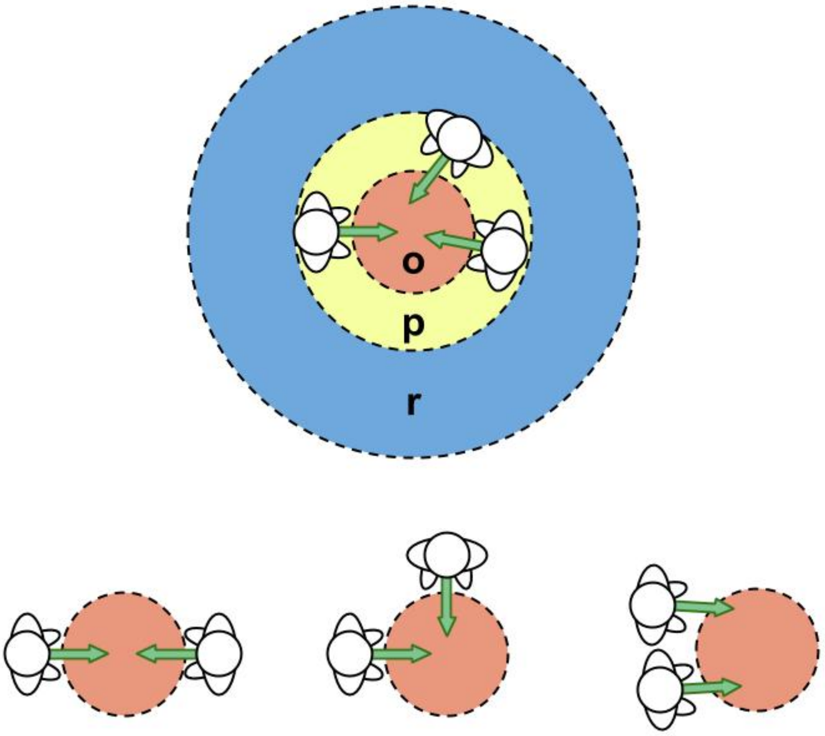

Figure 2.2: Top: example of circular f-formation with relative o-, p- and r-spaces. Bottom: other examples of f-formations configurations; vis-a-vis (left), l-shaped (center) and side-by-side (right).

personal synchrony was performed for the final group computation. All the parameters were empirically set, with no learning involved. Finally, [Zhang and Hung, 2016, Zhang and Hung, 2018] presented the first attempt to detect different levels of involvement in f-formations by using the concept of associate, defined by psychologists as a person who is attached to an f-formation, but is not a full member (i.e. he/she can be in either the p- or the r-space). Additionally, [Zhang and Hung, 2016, Zhang and Hung, 2018] learned the frustrum of attention that accounts for spatial context.

All the previous works used precomputed features (locations and orientations). Another family of methods focuses on feature extraction or joint feature extraction and group detection. [Subramanian et al., 2015] devised a semi-supervised method for jointly learning head and body pose regressors and f-formations centers and memberships, by using labelled and unlabelled training data. [Alameda-Pineda et al., 2015] dealt with the problem of estimating head and body pose in a scenario with several body occlusions in a robust way, by using a matrix completion approach on multimodal features (extracted from RGB cameras and wearable sensors). F-formations were found using existing state-of-the-art techniques [Setti et al., 2013b, Cristani et al., 2011, Setti et al., 2015]. [Alameda-Pineda et al., 2015] introduced also the novel task of finding *social attractors* (i.e. people inside groups who attract more attention) by counting how many times each person was looked at by the other group members. Similarly to [Subramanian et al., 2015], [Ricci et al., 2015] proposed another method for jointly learning head and body position and f-formations, by minimizing an error function which includes ground truth orientation and distance between the centers of the f-formations voted by each subject; groups were detected by finding f-formations implicitly, clustering voted o-space centers via an existing clustering algorithm.

Our work focuses on feature extraction, specifically body pose and orientation. “In the wild” free standing conversational groups scenarios present several complexities, such as body appearance variability, difficult lighting conditions and occlusions. Given these challenges, pose and orientation estimation tasks have always been simplified to a classification problem, where only “orientation classes”, not real-valued angles, are predicted. We propose a method for extracting fully articulated 3D body poses and fine grained orientations in degrees, not classes. Specifically, we devise a novel algorithm based on autoencoders for reconstructing incomplete 2D poses caused by partial body occlusions, improving the final 2D and 3D pose estimation.

Chapter 3

Emergent Leadership Detection by Analyzing Gaze and Visual Focus of Attention

Eyes are a fundamental component of human interaction. They are the primary mean for sensing nonverbal behaviors (with the only exception of vocal cues). They provide informative cues about our current focus of attention, whether it is a place, an object or a person [Subramanian et al., 2010]. They can also be used to control communication: [Jovanovic et al., 2006] showed that when a speaker is addressing someone, he/she gazes at that person the majority of time.

In this Chapter, we analyze gaze behavior in face-to-face small groups interactions. Specifically, we focus on the relationship between gaze and leadership and how it can be used to predict emergent leaders using their visual focus of attention (VFOA), which is defined as "what" or "where" a person is looking at at any specific moment. Given the relative difficulty of accurate gaze estimation in scenarios where the camera is not close to the face, we use the more reliable head orientation as a proxy. We start by presenting our dataset and the data collection method we performed, in Section 3.1. In Section 3.2 we describe how we model VFOA and the features based on it. Finally, results and conclusions are presented in Sections 3.3 and 3.4 respectively.

3.1 The Leadership Corpus

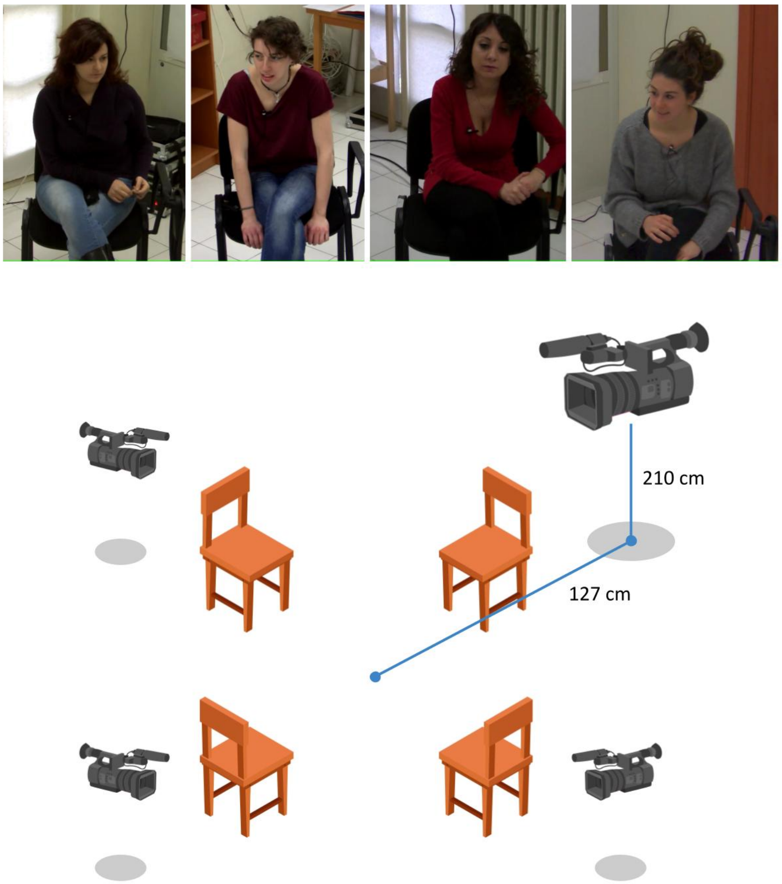

Figure 3.1: Example frames (top); seats and cameras setting (bottom).

The leadership dataset consists of 16 meeting sessions, where the longest meeting session lasts 30 minutes and the shortest lasts 12 minutes (for a total of 393 minutes). The sessions are composed of same gender, unacquainted four-person (in total 44 females and 20 males) with average age of 21.6 (2.24 standard deviation). The participants are seated as given in Figure 3.1. Videos were recorded using four frontal cameras with a resolution of 1280x1024 pixels and frame rate of 20 frame per second and a handy-cam which was located on a side of the room to capture the whole scene. Audio was recorded with four wireless lapel microphones, each one connected to person's corresponding frontal camera (audio sample rate=16 kHz). The participants performed one "survival task", randomly chosen between two tasks: "winter survival" and "desert survival" [Johnson and Johnson, 1991] which are most common tasks about small group decision making, dominance and leadership. In a typical survival task, participants are presented with a dramatic survival situation in a given geographical layout (e.g., a plane crash in the desert), with some details provided about the general conditions of the context (e.g., time of the day, nearest town distance, etc.). Participants are then given a list of objects that are left after the accident, and their task is to rank each of these items in the order of importance for the survival (from the most important to the least important). A single decision has to be taken by the group and follows a group discussion. In this study, instructions were given verbally, the use of pen and paper was not allowed, and the items to be ordered were 12.

3.1.1 Questionnaires

The SYstematic method for the Multiple Level Observation of Groups (SYMLOG) [Bales, 1980, Koenigs, 1999] is a comprehensive tool designed to evaluate individual dispositions along three bipolar dimensions: dominance versus submissiveness, acceptance versus non-acceptance of task orientation of established authority, and friendliness versus unfriendliness. The SYMLOG can be used both as a self-assessment instrument and as an instrument for external observation of a group interaction. Before the group task, volunteer participants were asked to complete the SYMLOG questionnaires and it was used to select the designated leaders of the task [Hare et al., 1998]. Specifically, subjects with scores of dominance and of task-orientation higher than the median of the sample were selected as designated leaders. The analysis regarding the designated leader is beyond the scope of this paper but it is worth to state that a designated leader may appear as an emergent leader.

Immediately after the group task, each participant was asked to rate the General Leader Impression Scale (GLIS) [Lord et al., 1984] questionnaire. The GLIS is an instrument designed to evaluate the leadership attitude that each member displays during a group interaction. It is a 5 item scale which asks participants to rate the other members of the group on their contribution to the group's overall effectiveness on the activity. GLIS were calculated for each individual by averaging the ratings given by the other group members.

Additionally, two independent judges observed the meetings of each group interactions and rated for each participant of each session both the GLIS (called as GLIS-Observers in this paper) ($InterClassCorrelation(ICC) = 0.771$; $p < 0.001$) and the SYMLOG (called as SYMLOG-Observers in this paper) (dominance $ICC = 0.866$, task-orientation $ICC = 0.569$, friendliness $ICC = 0.722$; $p < 0.001$). For SYMLOG-Observers only the dominance sub-scale of it, was used since the leadership impression obtained by GLIS-Observers and dominance tend to correlate with each other. The final scores for each participant were calculated as the average between their ratings.

3.1.2 Data Annotation

16 meeting sessions were divided into small segments, each lasting 4, 5 or 6 minutes, for ground truth annotation. In total, 75 meeting segments were used to analyze the proposed EL detection algorithm rather than using the original full meetings. The main reason for segmenting was to be able to have more data for training and testing, in a similar way to [Jayagopi et al., 2009]. This also resulted in more accurate ground truth annotations since people are more precise and more focused on annotation of videos when they were shorter, as mentioned in [Ambady et al., 2000].

Given that, psychology literature found that human observers can identify the emergent leaders [Sanchez-Cortes et al., 2012b], in total 50 human observers were used to

Emergent Leader	**Agreement Type**	**Average Agreement**	**Total # of Meetings/ Out of**
Most	Full Majority	1 0.73	26/75 49/75
Least	Full Majority	1 0.70	13/75 62/75

Table 3.1: Analysis of Leadership Annotations.

annotate each video segment. Each human observer annotated either 12 or 13 video segments. Each annotator judged no more than one segment per meeting session. During the annotation process, audio was not used in order to overcome any possible problem that might occur due to the level of understanding of the spoken language (similar to a recent study [Kindiroglu et al., 2014]). Annotators were requested to judge the four participants by ranking them from 1 to 4, where 1 corresponded to the person who exhibited the most leader behavior and 4 corresponded to the person who exhibited the least leader behaviour. In this paper, we used the annotations regarding the most EL and the least EL. The analysis about the annotations is given in Table 1. As can be seen from Table 3.1.2, annotating the least EL was more challenging than annotating the most emergent one.

3.2 Methodology

The proposed method (Figure 3.2) is divided into four parts. First, facial landmark detection and head pose estimation are applied. Then, visual focus of attention (VFOA) is modeled and estimated. Later, nonverbal visual features are extracted. As a result of facial landmark detection and head pose estimation, the pan, tilt and roll angles of a person for each video frames are obtained. Using the labeled VFOA of a person and a supervised learning algorithm, the entire VFOA of that person which gives the looking direction of the person is found. Finally, the nonverbal features are extracted from VFOA and they are used to detect the most and the least emergent leaders.

3.2.1 Facial Landmark Detection and Head Pose Estimation

Facial landmark detection and tracking are based on the Constrained Local Model (CLM) [Cristinacce and T.F.Cootes, 2006]. This method can be briefly summarized as follows: first, a model of faces is built from a training set by using shape (facial landmarks) and texture (patches around landmarks) information; then, the model is fit to a test image through an iterative algorithm, in which, at each iteration, the result of the correlation between the model's patches and the patches (called templates) sampled from the test image feature points is maximized and new feature points are chosen accordingly for the next iteration. When the algorithm converges, the resulting facial landmarks in 2D coordinates are converted to 3D coordinates and used to detect the head pose (pan, tilt, roll) and position in camera space [Baltrušaitis et al., 2013].

3.2.2 Modeling the Visual Focus of Attention

A person's VFOA can be defined as a person, object or, more generally, any position in the space the person is looking at [Stiefelhagen et al., 2002]. One way of inferring the VFOA is to use the person's eye gaze which is found by detecting and tracking the eyes. Current eye gaze tracking techniques are still constraining [Ba and Odobez, 2006] and challenging. For instance, they require the person to be close to the camera to track the eyes accurately [Hansen and Ji, 2010]. On the other hand, in many studies such as [Ba and Odobez, 2006, Stiefelhagen et al., 2002, Marin-Jimenez et al., 2011], it has been shown that the eye gaze can be estimated using the head pose representation.

In this study, we also use the head pose representation to find the VFOA. The pan and tilt angles are used to define the head pose which is in contrast to studies [Stiefelhagen et al., 2002, Carletta et al., 2005] that utilized only head pan angle while the roll angle is not used (similar to [Ba and Odobez, 2006]) since there is no effect of it to head direction. The VFOA of a person contains the other three persons who are on his/her right, left or front (shown as R, L and F, respectively, in Figure 2) and also no-one (shown as N in Figure 2) which refers to the time that the person is not looking to any participants but somewhere else such as ceiling, floor, door, etc. of the meeting room. It is important to highlight here that, in this VFOA definition, all the physical locations different than any other participant are considered as the same class.

In the literature, there are many supervised and unsupervised methods to estimate the VFOA in a meeting environment from head pose representation [Ba and Odobez, 2006, Stiefelhagen et al., 2002]. In this paper, SVM was used to learn and predict VFOA since it was significantly the best performing method among the compared state of the art methods (see Section 5 for details). Before applying SVM to find the VFOAs, we first interpolate the head pose representations using spline interpolation which is necessary since there are frame drops in different videos belonging to same meeting and the videos should be synchronized to be able to extract nonverbal features.

To train the SVM classifiers (one for each of 64 frontal videos) and also to evaluate its performance, the VFOA for a total of 25600 randomly selected frames (400 frames for each video which was determined by the confidence level=90% and margin error=4%) were annotated by two annotators. In total 23000 frames (in average 359.4 per video with standard deviation of 46.54) were used for evaluation which were obtained after removing differently labeled VFOAs. The VFOA annotation results show that we have highly imbalance VFOA classes when the least represented class is no-one which is 16% of the data. The labeled VFOA data were randomly divided into two folds (while having totally different but the same amount of instances from each classes) as training and validation sets and this process was repeated for 100 times to learn the individual SVM models. As SVM model, the radial basis kernel function (RBF) with varying kernel parameter was selected while hyper-planes were separated by sequential minimal optimization (SMO). As stated in [Beyan and

Fisher, 2015], SVM tended to be biased towards to well-presented class (majority class). To handle this class imbalance problem, the cost function [Fumera and Roli, 2002] (SVM-cost), the random under sampling [Yap et al., 2014] (SVM-RUS) and the SMOTE [Chawla et al., 2002] (SVM-SMOTE) methods were combined with SVM. To evaluate the performance of SVMs, the geometric mean of detection rates (see [Beyan and Fisher, 2015] for definition) were used. For each video, the method (SVM, SVM-cost, SVM-RUS or SVM-SMOTE), performing the highest geometric mean of the detection rates with corresponding parameters was selected to classify the whole unlabeled head pose. This results in VFOA per person for the entire video.

3.2.3 Nonverbal Visual Features Extraction

A fixation happens when a participant looks at another participant for a minimum amount of time. The number of frames that can be considered to start a fixation is called hysteresis. In our analysis, hysterisis was taken as 5 frames and all the VFOAs were smoothed with it as a post-processing step to denoise the VFOAs before extracting the visual non-verbal features. From the obtained VFOAs for each person the following nonverbal features are extracted:

- **totWatcher**: the total number of frames that a person is being watched by the other persons in the meeting
- **totME**: the total number of frames that a person is mutually looking at any other persons in the meeting (also called mutual engagement (ME))
- **totWatcherNoME**: the total number of frames that a person is being watched by any other persons in the meeting while there is no mutual engagement
- **totNoLook**: the total number of frames that are labeled as no-one in the VFOA vector meaning that a person is not looking at any other persons in the meeting

- **lookSomeOne**: the total number of frames that a person looked at other persons in the meeting

- **totInitiatorME**: the total number of frames to initiate the mutual engagements with any other persons in the meeting

- **stdInitiatorME**: the standard deviation of the total number of frames to initiate the mutual engagements with any other persons in the meeting

- **totInterCurrME**: the total number of frames intercurrent between the initiation of mutual engagement with any other persons in the meeting

- **stdtInterCurrME**: the standard deviation of the total number of frames intercurrent between the initiation of mutual engagement with any other persons in the meeting

- **totWatchNoME**: the total number of frames that a person is looking at any other persons in the meeting while there is no mutual engagement

- **maxTwoWatcherWME**: the maximum number of frames that a person is looked at by any other two persons while that person can have a mutual engagement with any of two persons

- **minTwoWatcherWME**: the minimum number of frames that a person is looked at by any other two persons while that person can have a mutual engagement with any of two persons

- **maxTwoWatcherNoME**: the maximum number of frames that a person is looked at by any other two persons while that person can have no mutual engagement with any of two persons

- **minTwoWatcherNoME**: the minimum number of frames that a person is looked at by any other two persons while that person can have no mutual engagement with any of two persons

- **ratioWatcherLookSOne**: the ratio between the *totWatcher* and *lookSomeOne*.

In total 15 features were extracted. All features (except *ratioWatcherLookSOne*) were divided by the total number of frames in a given meeting since the total number of frames per meeting is variable. The features *totWatcher*, *LookSomeOne* and *ratioWatcherLookSOne* were already used in [Sanchez-Cortes, 2013] by combining with nonverbal audio features for EL detection in a meeting environment and also in [Hung et al., 2008] to detect the dominant person in a meeting. To the best of our knowledge the rest of the features were never used in a SSP study, although they have been discussed in social psychology works related to dominance, leadership and nonverbal behavior. In addition to these features, the total number of frames that a person is looked by all other three persons in the meeting with/without a mutual engagement can also be extracted. However, for our dataset, we observed that, such a feature is not useful since there were no such a frame.

The motivation and the justification of the extracted features can be summarized as follows [Carney et al., 2005, Hall et al., 2005]: how many times and how long *i)* the EL is looked at by each person while there is no mutual engagement (ME) is a measure of the individual coordination to the leader, *ii)* the EL is looked at by the two or three members simultaneously when there is no ME is a measure of the group coordination of the leader, and it is expected that higher values of this index reflects the centrality of the leader, in other words, a person is looked at by another two or three persons simultaneously without ME reflect the group behavior towards an individual person and higher values of this feature could reflect the EL. *iii)* a peer is looked at by the leader without ME reflect's the leader's directiveness and correlate with the perceived efficacy of the leadership at the group level. *iv)* ME is a measure of the reciprocal engagement among the participants, higher values of this feature should reflect better leader-to-peer coordination. *v)* Being initiator of a ME can be seen as a measure of the ability to attract the attention of a person and it is expected that having high values of being initiator reflects the emergent leader's directive activity.

Using the extracted nonverbal visual features, the most and the least EL for each meeting segment were modeled and detected by the methods given in Section 3.3.

3.3 Results

In this section, we present *i)* the results corresponding to different VFOA detection algorithms, *ii)* emergent leader (EL) detection results by different algorithms and *iii)* the correlation analysis that was performed between each nonverbal feature and the questionnaires that were given in Section 3.1.1.

3.3.1 Results of VFOA Estimation

Different than SVM and its variations (SVM-cost [Fumera and Roli, 2002], SVM-RUS [Yap et al., 2014] and SVM-SMOTE [Chawla et al., 2002]), we applied methods based on OTSU [Otsu, 1979], k-means and Gaussian Mixture Model (GMM) [Stiefelhagen et al., 2002] to model and to estimate the VFOA. These methods are briefly summarized as follows:

OTSU [Otsu, 1979] based method. Pan and tilt angles per a frontal video (in other words per person) were first smoothed assuming that they can vary from -90 to 90 degrees. Then OTSU thresholding was applied to smoothed pan and tilt angles independently. This resulted in four thresholds (two for pan angles and two for tilt angles).

K-means based method. The median and standard deviation of the tilt angles per frontal video were used to define the two thresholds which were obtained as median of tilt angles ± standard deviation of tilt angles. The pan angles per frontal video were clustered using k-means where the number of clusters was three. This resulted in three centers and the two thresholds were found by finding the middle point of the two consecutive cluster centers.

GMM [Stiefelhagen et al., 2002] based method. The thresholds from tilt angles per frontal video were obtained as given in k-means based method. Pan angles per frontal video were modeled using GMM with three components (representing left, right and front). The mean and covariance of components were initialized using k-means (having three clusters) where priors were set to uniform.

For all methods, VFOAs (right, left, front and no-one) from head poses per frame were classified as follows: A tilt value (which were obtained using the tilt angles from a frontal video) was classified as no-one if the tilt value was out of the thresholds. For OTSU and k-means, if the tilt value was between the thresholds, then the corresponding pan value was compared with the thresholds obtained using pan angles. If the pan value was smaller than the smallest pan threshold, the VFOA was classified as left; if the pan value was greater than the biggest pan threshold then the VFOA was classified as right; and finally if the pan value was between the pan thresholds then the VFOA was classified as front. For GMM, the maximum class probability was used to estimate the VFOA.

These methods were also combined with some pre-processing steps: 5% outlier removal and smoothing (by moving average filter) which were applied before calculating the thresholds that were obtained from pan and tilt angles (applying the outlier removal and smoothing always improved the results). All the results (average of right, left, front and no-one detection rates) regarding VFOA estimation are given in Table 3.3.1. As seen, SVM and its variations performed better than other methods especially to detect right, left and front. The detection rate of no-one by SVM and its variations was also better than the rest except OTSU which on the other hand performed very poorly to estimate the right, left and front. k-means and GMM were also performed almost as good as SVM and its variations (for detecting right,

Detection Rate Method	Right	Left	Front	No-One
OTSU	0.44	0.53	0.55	0.60
k-means	0.75	0.87	0.79	0.10
GMM	0.73	0.77	0.62	0.10
SVM	0.88	0.86	0.67	0.39
SVM-cost	0.85	0.85	0.72	0.52
SVM-RUS	0.83	0.82	0.70	0.56
SVM-SMOTE	0.87	0.86	0.70	0.51

Table 3.2: VFOA Estimation Results.

left, and front) however their no-one detection rate was very low. On the light of those results, as mentioned in Section 3.2.2, SVM and its variations were used to model and estimate the VFOAs.

3.3.2 Results of Emergent Leader Estimation

The variations of SVM (all with RBF with varying kernel parameters while hyperplanes were separated by SMO) using leave-one-out, leave-one-meeting-out and leave-one-meeting-segment-out approaches and rank-level fusion approach (RLFA) [Sanchez-Cortes, 2013, Aran and Gatica-Perez, 2010] using different feature groups were used to detect the most and the least emergent leaders using the proposed nonverbal visual features.

In Table 3.3.2, the best results for SVM (which is selected by the highest score of the geometric mean of the detection rates) and its variations and RLFA with different features were compared when the three classes (the most EL, the least EL and the other persons) were considered. As variations of SVM, SVM-cost [Fumera and Roli, 2002], SVM using the features after principal component analysis (PCA) was applied (SVM-afterPCA), SVM-cost [Fumera and Roli, 2002] using the PCA applied features (SVM-afterPCA-cost) and SVM which was applied using the features that were found correlated with the questionnaires (SVM-with-CorrFea, see Section 3.3.3 for more information) were used. For SVM, only the results with leave-one-meeting-out approach is given since all the results were similar to each other and also due

Method \ Detection Rate	Most EL	Least EL	Rest
SVM	0.71	0.59	0.75
SVM-cost	0.80	0.58	0.70
SVM-afterPCA	0.72	0.63	0.71
SVM-afterPCA-cost	0.79	0.63	0.64
SVM-with-CorrFea	0.67	0.62	0.72
RLFA	0.71	0.71	0.69
RLFA-with-CorrFea	0.72	0.67	0.68

Table 3.3: Emergent leader (EL) detection performances using nonverbal visual features.

to the space limitations. Assuming that the proposed nonverbal features can be correlated with each other which might affect the performance of SVM negatively, PCA was applied to the features as a dimensionality reduction technique. To obtain a useful set of components the smallest number of components that represent 90% of the sum of all eigenvectors was used. This left five features from the defined 15 features. On the other hand, the RLFA was applied using the whole nonverbal features and only with the features correlated with the questionnaires (RLFA-with-CorrFea).

As can be seen in Table 3.3.2, the best performing method for the most EL detection was SVM-cost while its least EL detection rate was the worst. Using PCA improved the detection rate of the least EL. Applying the cost function which penalize the misdetection of the most and the least emergent leaders more than the rest improved the detection rate of the most EL. The best performing method to detect the least EL was RLFA which in general performed as good as SVM and its variations although it is an unsupervised learning algorithm (similar to the results given in [Sanchez-Cortes et al., 2012b, Sanchez-Cortes, 2013, Jayagopi et al., 2009]). Overall, the best performing method can be considered as the method which performs well to detect the most EL while not performing poor in detecting the least EL and the rest as well. With such an assumption all methods performed almost the same with ±0.02 deviation.

Different from the results given here, SVM and its variations were also applied

Nonverbal Visual	**RLFA**		**SVM-cost**	
Features	**most**	**least**	**most**	**least**
totWatcher	0.71	0.68	0.74	0.55
totME	0.74	0.68	0.75	0.54
totWatcherNoME	0.68	0.68	0.76	0.54
totNoLook	0.24	0.15	0.38	0.26
lookSomeOne	0.26	0.23	0.38	0.26
totInitiatorME	0.46	0.46	0.50	0.20
stdInitiatorME	0.27	0.22	0.27	0.14
totInterCurrME	0.16	0.29	0.34	0.30
stdtInterCurrME	0.35	0.31	0.36	0.14
totWatchNoME	0.04	0.06	0.75	0.55
maxTwoWatcherWME	0.66	0.67	0.70	0.55
minTwoWatcherWME	0.60	0.60	0.59	0.50
maxTwoWatcherNoME	0.63	0.66	0.67	0.55
minTwoWatcherNoME	0.62	0.50	0.60	0.39
ratioWatcherLookSOne	0.72	0.67	0.72	0.57
Fea-[Sanchez-Cortes, 2013]	0.71	0.67	0.52	0.57

Table 3.4: Individual performance of nonverbal visual features for the most and the least emergent leaders

using binary classes as: i) the most EL versus the rest and ii) the least EL versus the rest. For the detection rate of the most and the least emergent leaders, the results were very similar to the results given in Table 3.3.2 while the detection rate of the rest were highly increased (in average 15%) no matter which cross validation approach (leave-one-out, leave-one-meeting-out and leave-one-meeting-segment-out) was applied.

To better investigate the performance of each nonverbal visual features for the most and the least emergent leaders detections, SVM and SVM-cost were applied using leave-one-meeting-out when the three classes were considered. Additionally, the features (*totWatcher*, sum of *totWatchNoME* and *totME* per frontal video, *ratioWatcherLookSOne*) used in [Sanchez-Cortes, 2013] (shown as Fea-[Sanchez-Cortes, 2013]) were also evaluated. The results are given in Table 3.3.2.

The results in Table 3.3.2 are the best results according to geometric mean of detection rates. These results show that the best features to detect the most EL accurately are: *totWatcher*, *totME*, *totWatcherNoME*, *totWatchNoME*, *maxTwoWatcherWME*, *maxTwoWatcherNoME*, and *ratioWatcherLookSOne*. The best features to detect the least EL more accurately are *totWatcher*, *totME*, *totWatcherNoME*, *maxTwoWatcherWME* and *ratioWatcherLookSOne*. Furthermore, using all features together (see Table 3.3.2) performed better for both classes in general. On the other hand, when the performance of the proposed features and the features presented in [Sanchez-Cortes, 2013] were compared, it has seen that the most EL detection performance of the proposed features was better no matter which classifier was applied while the least EL detection rates were similar.

3.3.3 Correlation Analysis

In Table 3.3.3, the correlation between variables derived from questionnaires and visual features are given when the meeting videos were evaluated as whole, rather than segmented, as defined in Section 3.1. As seen from Table 3.3.3, except *totNoLook*, *lookSomeOne*, *stdInitiatorME* all others nonverbal features found correlated (eight of them had high correlation, two of them had medium correlation and two of them had low correlation) with the results of SYMLOG-Observers. Similarly, except *totNoLook*, *lookSomeOne*, *stdInitiatorME* and *totInterCurrME* all other nonverbal features were correlated (seven of them had high correlation, three of them had medium correlation and one of them had low correlation) with the results of GLIS-Observer.

3.4 Conclusions

In this chapter we presented novel nonverbal visual features which are extracted from VFOA to detect the emergent leaders in a meeting environment. Different than many emergent leadership studies in the literature, we only used video cues although it was shown that audio cues were generally more effective. The proposed nonverbal

Nonverbal Visual Features	**SYMLOG-Observers**	**GLIS-Observers**
totWatcher	0.69	0.68
totME	0.61	0.59
totWatcherNoME	0.67	0.66
totNoLook	0.06	-0.08
lookSomeOne	-0.06	0.08
totInitiatorME	0.31	0.42
stdInitiatorME	0.005	0.08
totInterCurrME	-0.20	-0,06
stdtInterCurrME	0.23	-0.14
totWatchNoME	-0.61	-0.49
maxTwoWatcherWME	0.65	0.60
minTwoWatcherWME	0.51	0.52
maxTwoWatcherNoME	0.52	0.50
minTwoWatcherNoME	0.44	0.48
ratioWatcherLookSOne	0.65	0.59

Table 3.5: Correlation Coefficient Values Between Questionnaires and Nonverbal Visual Features

features performed well for detection of the most and the least emergent leaders (70% of detection rate in average) when the majority of the defined nonverbal features were highly correlated with the results of the social psychology questionnaires. The human annotations using the video segments showed very high overlap (94% overlap with SYMLOG-Observers for the most and the least leaders, and 88% overlap with GLIS-Observers for the most and the least leaders when the highest/lowest values of the questionnaires were used for EL inference) with the results of questionnaires which were filled by observers using the whole videos. In the 58 out of 75 video segments, the most EL annotated by the 50 human observers was also the designated leader. Similarly, in 12 out of 16 whole videos, the most EL inferred by GLIS-Observers was also the designated leaders. The applied supervised and unsupervised methods to detect the most and the least emergent leaders performed well, which can be a result of the accurate detection of VFOAs (72% detection rate in average) and the effectiveness of the used features.

Although gaze is an important nonverbal behavioral cue, conveying information about focus of attention and personal traits, its expressive power is limited. In the next Chapter, we extend our analysis to face-based behavioral cues.

Publication:

Beyan C., Carissimi N., Capozzi F., Vascon S., Bustreo M., Pierro A., Becchio C., Murino V., "Detecting emergent leader in a meeting environment using nonverbal visual features only". In Proceedings of the 18th ACM International Conference on Multimodal Interaction 2016 Oct 31 (pp. 317-324). ACM.

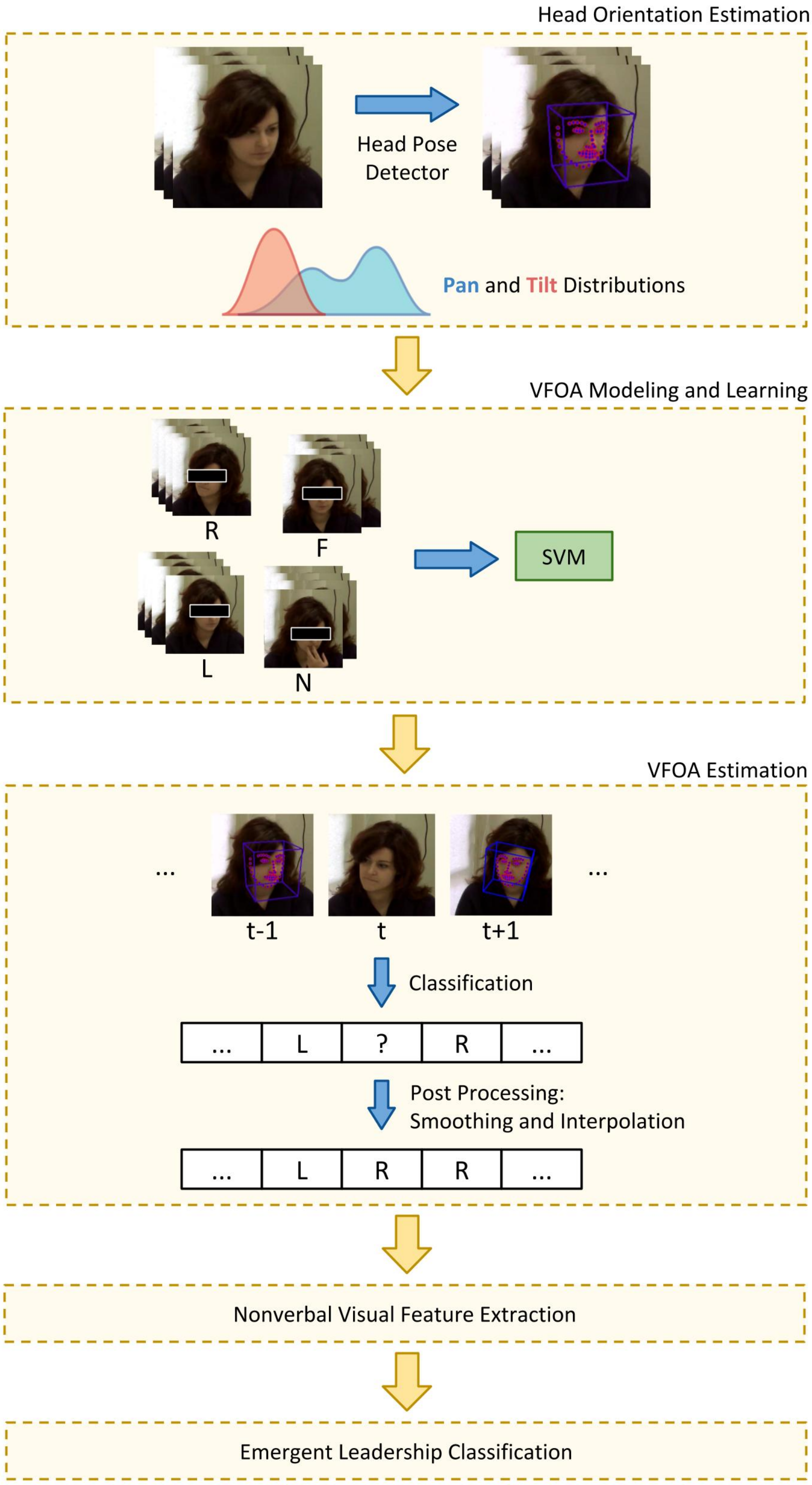
Head Orientation Estimation
Head Pose Detector
Pan and Tilt Distributions
VFOA Modeling and Learning
R
F
L
N
SVM
VFOA Estimation
...
t-1
t
t+1
...
Classification
... L ? R ...
Post Processing: Smoothing and Interpolation
... L R R ...
Nonverbal Visual Feature Extraction
Emergent Leadership Classification

Figure 3.2: Overview of the proposed emergent leadership detection pipeline.

Chapter 4

Face-Based Behavioral Cues and Deception Detection

The human face contains most of the apparatuses for *sensing* social signals, i.e. eyes, ears, mouth and nose. It also contains apparatuses for *producing* social signals, such as (again) eyes (gaze), mouth (vocal behavior) and facial muscles (face expressions). Among nonverbal behavioral cues, the ones related to face have been shown to play a major role in social interactions [Grahe and Bernieri, 1999, Ambady and Rosenthal, 1992]. It's no surprise, then, that the human face is the most essential part of our body for interpersonal interaction.

In this Chapter we examine the behavior of face in dyadic interactions. Specifically, we try to understand the relationship between cues expressed involuntarily (or "leaked") and the act of deception during conversations. In Section 4.1 we briefly introduce the dataset used in our analysis, while in Sections 4.2 and 4.3 we detail the extracted features. In Section 4.4 we discuss about feature fusion and the learning algorithm. Results and conclusions follow in Sections 4.6 and 4.7.

4.1 Real-Life Deception Dataset

To the best of our knowledge, the only public real-life (no role-play) dataset is [Pérez-Rosas et al., 2015a], which is a multi-modal dataset depicting deception in real-life court trials and TV interviews. It collects audio-video recordings from public multimedia sources, showing witnesses and defendants while testifying or being interviewed (Figure 4.1). A total of 121 recordings (61 deceptive and 60 truthful samples) are included, with an average length of 28 seconds. Text transcriptions and manually annotated nonverbal cues (based on the MUMIN coding scheme, Section 4.2.3) are also provided. Similarly to [Mimansa et al., 2016], three videos were discarded, since the face tracking algorithm (i.e. OpenFace [Baltrušaitis et al., 2016, Baltrušaitis et al., 2015]) was not able to detect the face because of severe occlusions or extreme face orientations (Fig. 4.2).

Figure 4.1: Example frames from the real-life deception dataset [Pérez-Rosas et al., 2015a].

4.2 Face-Based Features

In this section, we introduce all verbal and nonverbal features used in our analysis. The proposed face-based deep features are described in Section 4.2.1, while the

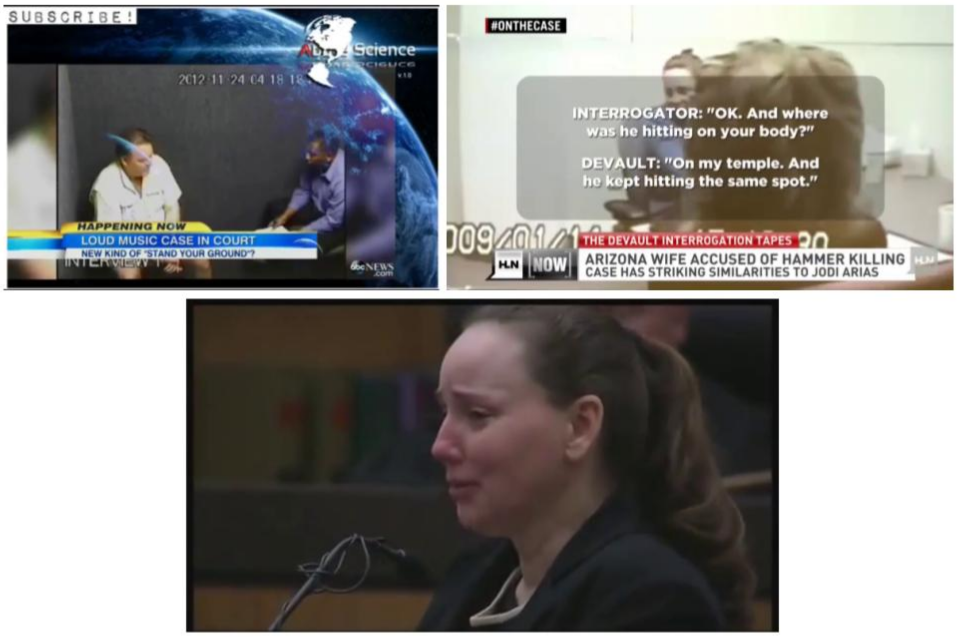
SUBSCRIBE!
Science
2012 11 24 04 18 18
HAPPENING NOW
LOUD MUSIC CASE IN COURT
NEW KIND OF 'STAND YOUR GROUND'?
abcNEWS .com
#ONTHECASE
INTERROGATOR: "OK. And where was he hitting on your body?"
DEVAULT: "On my temple. And he kept hitting the same spot."
THE DEVAULT INTERROGATION TAPES
HLN NOW
ARIZONA WIFE ACCUSED OF HAMMER KILLING
CASE HAS STRIKING SIMILARITIES TO JODI ARIAS

Figure 4.2: Example of discarded frames from the real-life deception dataset [Pérez-Rosas et al., 2015a]: faces partially occluded (top), extreme face orientation (bottom).

manually annotated nonverbal features based on the MUMIN coding scheme, as used in [Pérez-Rosas et al., 2015a], are described in Section 4.2.3. Finally, Section 4.2.2 introduces features based on facial Action Units [Mimansa et al., 2016].

4.2.1 Deep Neural Network Features

Deep learning based features (named as deep face features in this study) are extracted frame-wise, using several pre-trained DNNs (and applying fine-tuning in one case). We first use the OpenFace tool [Baltrušaitis et al., 2016, Baltrušaitis et al., 2015] to detect and crop faces from each frame, setting to zero the pixels corresponding to the background. We, then, apply the pre-trained DNNs on the resulting images and use the activation values of the layers before the final ones as feature vectors. Specifically, we use VGG-Face [Parkhi et al., 2015] with values from layers *fc6* and *fc7*, AlexNet [Krizhevsky et al., 2012] with values from layer *fc7*, GoogLeNet [Szegedy et al., 2015] with values from layer *pool5*, and ResNet50 [He et al., 2016] with values from layer *avgpool*.

The length of the feature vectors varies between 2048 and 4096.

Since AlexNet-based features lead to the lowest classification accuracy when used with MVL (see Table 4.6), we try to improve the results by fine-tuning the network on our dataset. The process is described in Section 4.6.

4.2.2 Facial Action Units

Facial Action Units (AUs) are defined as a contraction or relaxation of one or more facial muscles. They are a part of the Facial Action Coding System (FACS) [Ekman and Friesen, 1978, Ekman and Friesen, 1976], which is a taxonomy of human facial movements. AUs can be described by their presence (0 or 1, if AU is not visible or visible, respectively) or by their intensity (how intense is the AU, on a 0 to 5 point scale).

The OpenFace tool [Baltrušaitis et al., 2016, Baltrušaitis et al., 2015] is used to extract AUs. As in [Mimansa et al., 2016], nine different AUs {AU1, AU2, AU4, AU45, AU7, AU23, AU25, AU26 and AU28} are used for the analysis. The extraction of AUs is frame-wise, therefore they are combined into a single vector to represent the whole video using a threshold presence for each AU. Following [Mimansa et al., 2016], 3 is used as the threshold value. As a result of this feature extraction, each video clip is represented with an 18 dimensional feature vector (nine AUs with presence and intensity values for each of them).

4.2.3 MUMIN Based Face Features

The MUMIN coding scheme [Allwood et al., 2007] is a multi-modal annotation scheme intended as a tool for studying hand gestures and facial displays in interpersonal interactions. As in [Pérez-Rosas et al., 2015a], six MUMIN groups, which contains 21 gestures, are considered. These gestures are listed as follows:

- **General Face**: Laughter, Scowl, Smile, Other.
- **Eyebrows**: Other, Raising, Frown.
- **Eyes**: Exaggerated Opening, Other, Closing Repeated, Closing Both.
- **Gaze**: Up, Side, Interlocutor, Down.
- **Mouth**: Open Mouth, Close Mouth.
- **Lips**: Retracted, Protruded, Corners Up, Corners Down.

Features are values that represent the presence (1) or absence (0) of a specific gesture in a video clip, which results in a 21 elements binary vector for each given recording.

4.3 Non Face-Based Features

We show the importance of face features compared to two other types of features: nonverbal features extracted from head and hands movements and verbal features based on n-grams [Pérez-Rosas et al., 2015a, Mimansa et al., 2016].

4.3.1 MUMIN Based Head and Hands Features

The MUMIN coding scheme [Allwood et al., 2007] includes three other groups of gestures related to head and hands movements, for a total of 17 gestures. These gestures are listed as follows:

- **Head Movements**: Waggle, Shake, Side Turn, Repeated Tilts, Side Tilt, Other, Move Forward, Repeated Nods, Down.
- **Hands**: Single Hand, Other, Both Hands.
- **Hand Trajectory**: Up, Sideways, Other, Down, Complex.

As described in Section 4.2.3, features are values that represent the presence (1) or absence (0) of a specific gesture in a video clip, which results in a 17 elements binary vector for each given recording.

4.3.2 N-Grams

N-grams are contiguous sequences of n elements (e.g. letters, syllables, words) in a given text. The most common n-grams used in natural language processing and text analysis are uni-grams ($n = 1$), bi-grams ($n = 2$) and tri-grams ($n = 3$), and they use words as elements. The verbal features are extracted by generating a bag-of-n-grams based on words from a text corpus obtained by aggregating all the transcriptions of the dataset, and computing vectors of n-grams frequencies for each text sample. As in [Pérez-Rosas et al., 2015a, Mimansa et al., 2016], we use uni-grams and bi-grams, which result in a feature vector with 1609 dimensions (124 for uni-grams and 1485 for bi-grams) for each video clip.

4.4 Combining Features

In this section, we describe the approach used for classification, which is based on multi-view learning (MVL) [Xu et al., 2013], a learning paradigm that gained popularity in the recent years.

It is very common, nowadays, to have ”heterogeneous” feature sets, i.e. sets containing features extracted from different modalities (e.g. audio, video) or features representing different properties of the data (e.g. color, texture data). Conventional machine learning algorithms, such as SVMs, discriminant analysis or spectral clustering, concatenate all the features into one single vector, or ”view”, which is then used to build a model. However, this concatenation does not take into account the specific statistical properties of the different feature types, and it can also lead to overfitting in case of small size training sets. In contrast, multi-view learning considers each feature type as a single view. Each view is modeled by a single function and all the functions are jointly optimized to improve the learning performance and obtain a better model.

MVL algorithms can be classified into three main groups: *co-training* [Blum and Mitchell, 1998], *multiple kernel learning* [Lanckriet et al., 2004] and *subspace learning* [Akaho, 2006, Chaudhuri et al., 2009]. Interested readers can refer to [Kincaid et al., 1975] for a detailed explanation, while the applied method falls under the co-training and multiple kernel learning techniques.

Although many MVL techniques exist, all of them use one of two principles to combine the different views, namely, the *consensus principle* and the *complementary principle* [Kincaid et al., 1975]. The goal of the consensus principle is to maximize the agreement of the output of different views. As demonstrated by [Dasgupta et al., 2002], by minimizing the disagreement of two views, the error rate of the output of the two views is minimized as well. The complementary principle states that each view may contain knowledge of the data that other views do not have, and that, by using multiple views, data can be described accurately and comprehensively.

In this work, we adopt the implementation of a MVL algorithm cast as a special case of a general vector-valued Reproducing Kernel Hilbert Spaces (RKHS) framework [Minh et al., 2013, Minh et al., 2016] which unifies manifold regularization and co-regularized MVL. The advantages of this method compared to other existing methods, such as [Sun, 2011, Luo et al., 2013], are *i)* the ability to specify an arbitrary number of views, and *ii)* the presence of powerful regularization terms that satisfy the consensus and complementary principles.

Manifold regularization tries to learn the geometry of the input space, assuming that data lie on a space with lower dimension than the input space, by using both labeled and unlabeled data.

In co-regularized MVL, the aim is to construct target functions based on different hypothesis spaces corresponding to different views of the input data. Output values from the different views are enforced to be consistent by using a regularization term and are combined in a principled way to give a final output. Co-regularized learning falls in the co-training family of MVL techniques and the regularization term implements the consensus principle.

The following equation shows the general minimization problem for the multi-view manifold regularization

$$f_{\mathbf{z},\gamma} = argmin_{f \in \mathcal{H}_K} \frac{1}{l} \sum_{i=1}^{l} V(y_i, Cf(x_i)) + \Gamma_A + \Gamma_I \tag{4.1}$$

where, given $\mathcal{X}$ as the input space and $\mathcal{Y}$ as the output space, $\mathbf{z} = \{(x_i, y_i)\}_{i=1}^{l}$ is a random training sample of l labeled examples, $\mathcal{W}$ is a separable Hilbert space, K is a positive definite kernel and $\mathcal{H}_K$ is its Reproducing Kernel Hilbert Space of $\mathcal{W}$-valued functions, f is a function belonging to $\mathcal{H}_K$, $C : \mathcal{W} \to \mathcal{Y}$ is a bounded linear operator, and Γ_A and Γ_I are regularization terms.

The first term in Eq. (4.1) measures the error between the final output $Cf(x_i)$ for given x_i and relative output y_i. In our case, we chose the least squares loss function for V, thus

$$V(y_i, Cf(x_i)) = \|y_i - Cf(x_i)\|_Y^2. \tag{4.2}$$

$Cf(x)$ is a linear combination of the output of all the views and has the form

$$Cf(x) = \sum_{i=1}^{m} c_i f^i(x) \in Y \tag{4.3}$$

where $c_i \in \mathbb{R}$ and $f^i(x) \in \mathcal{Y}$ is the output of a single view.

The second term, Γ_A, is the RKHS regularization term (the ambient regularizer)

$$\Gamma_A = \gamma_A \|f\|_{\mathcal{H}_K}^2. \tag{4.4}$$

The third and final term, Γ_I, is the multi-view manifold regularization term

$$\Gamma_I = \gamma_I \langle \mathbf{f}, M\mathbf{f} \rangle_{\mathcal{W}^{u+l}} \tag{4.5}$$

where $M : \mathcal{W}^l \to \mathcal{W}^l$ is a symmetric positive operator. If there is only one view, it simply consists of standard manifold regularization (the intrinsic regularizer); otherwise, if there are more views, it consists of manifold regularization for each view.

Eq. (4.5) is divided into two subterms

$$\gamma_I \langle \mathbf{f}, M\mathbf{f} \rangle_{\mathcal{W}^l} = \gamma_B \langle \mathbf{f}, M_B\mathbf{f} \rangle_{\mathcal{W}^l} + \gamma_W \langle \mathbf{f}, M_W\mathbf{f} \rangle_{\mathcal{W}^l} \tag{4.6}$$

where γ_B, $\gamma_W \geq 0$ and M_B, $M_W : \mathcal{W}^l \to \mathcal{W}^l$ are symmetric, positive operators.

The first term, $\gamma_B \langle f, M_B f \rangle_{\mathcal{W}^l}$, is called the *between-view regularization* term, which enforces consistency between the output of the different views $f^i(x)$ and, given $\mathcal{W} = \mathcal{Y}^m$, has the form

$$\gamma_B \langle \mathbf{f}, M_B\mathbf{f} \rangle_{\mathcal{Y}^{ml}} = \gamma_B \sum_{i=1}^{l} \sum_{j,k=1,j<k}^{m} \| f^j(x_i) - f^k(x_i) \|^2_{\mathcal{Y}}. \tag{4.7}$$

The second term, $\gamma_W \langle f, M_W f \rangle_{\mathcal{W}^l}$, is called the *within-view regularization* term, which enforces smoothness of the output for each view and has the form

$$\gamma_W \langle \mathbf{f}, M_W\mathbf{f} \rangle_{\mathcal{Y}^{ml}} = \gamma_W \sum_{i=1}^{m} \sum_{j,k=1,j<k}^{l} W^i_{jk} \| f^i(x_j) - f^i(x_k) \|^2_{\mathcal{Y}}. \tag{4.8}$$

The parameter C, represented by the vector $c \in \mathbb{R}^m$, is optimized together with $f_{\mathbf{z},\gamma}$. Let $S^{m-1}_\alpha = x \in \mathbb{R}^m : \|x\| = \alpha$ be the sphere at the origin in $\mathbb{R}^m$ with radius $\alpha > 0$. Thus Eq. (4.1) becomes

$$f_{\mathbf{z},\gamma} = argmin_{f \in \mathcal{H}_K, \mathbf{c} \in S^{m-1}_\alpha} \frac{1}{l} \sum_{i=1}^{l} V(y_i, Cf(x_i)) + \Gamma_A + \Gamma_I. \tag{4.9}$$

Eq. (4.9) is not convex and can be optimized via alternate minimization: first, $\mathbf{c} \in S_{\alpha}^{m-1}$ is fixed and the problem is solved for the optimal $f_{\mathbf{z},\gamma} \in \mathcal{H}_K$. Then, f is fixed and Eq. (4.9) becomes equivalent to

$$\min_{c \in S_{\alpha}^{m-1}} \frac{1}{l} \sum_{i=1}^{l} \|y_i - Cf(x_i)\|_{\mathcal{Y}}^2. \tag{4.10}$$

The combination parameter C represents the importance of the views: the larger is the absolute value of c_i, the greater is the importance of view f^i, and vice versa.

4.5 Baseline Methods

The performance of MVL is compared with single kernel SVM (the state of the art classifier when automatically extracted features were used for deception detection in [Mimansa et al., 2016]) and a popular multiple kernel learning (MKL) algorithm, Localized Multiple Kernel Learning (LMKL) [Gonen and Alpaydin, 2008, Gonen and Alpaydin, 2011] (which showed significantly better classification performance as compared to other MKL methods for analysis of different social interactions in [Beyan et al., 2016, Beyan et al., 2017]).

4.5.1 Support Vector Machines

SVM is applied using the proposed deep face features only, and also using a combination of deep face features and other state of the art nonverbal and verbal features. Radial basis function (RBF) and linear kernels are used. We perform a grid search on parameters, with kernel parameter set as 2^i, $i = -1, 1, 3...31$ and radial basis function's (RBF) γ set as 2^j, $j = -11, -9, -7...11$.

4.5.2 Localized Multiple Kernel Learning

Multiple Kernel Learning (MKL) methods use set of kernels in linear and non-linear way such that they find the optimal kernel combination for different features coming from multiple sources. For a comprehensive survey on different MKL methods and performance comparisons (particularly among LMKL and many other MKL methods), interested readers can refer to [Gonen and Alpaydin, 2011].

LMKL utilizes nonlinear combinations of kernel weights. Different kernel weights are assigned to different regions of the feature subsets. It includes two components: *i)* gating model and *ii)* locally combined kernel matrix. These two components are optimized jointly: first, the gating model selects the locally optimal kernel function by assigning kernel weights to a subset of data, while the optimization is performed using a fixed gating model. Later, the gating model is updated using the gradients calculated by the current solution (for more details see [Gonen and Alpaydin, 2008, Gonen and Alpaydin, 2011]). One advantage of LMKL, in addition to its better performance as compared to many MKL methods, is its ability to set the same type of kernel (e.g. linear) for different subsets of data.

In this study, we combine LMKL with SVM to perform fair comparisons with the state of the art [Mimansa et al., 2016] and also with the proposed method, i.e. MVL (Section 4.4). Several combinations of different numbers of linear kernels (from two to five) and gating models (sigmoid or softmax) are tried. Grid search is used to find the best kernel parameter (the trade-off between model simplicity and classification error), with values set as 2^i, $i = -1, 1, 3...31$.

4.6 Experimental Analysis

In this section, we discuss our experimental analysis and report the results we obtain by employing MVL and deep face features. We follow the same evaluation protocol of [Pérez-Rosas et al., 2015a, Mimansa et al., 2016], where accuracy was used as the evaluation metric (the dataset is class-balanced) and leave-one-out was used as the cross validation approach. Table 4.6 shows a comparison between state of the art and our results. "Face-Manual" refers to the manually annotated MUMIN-based face features (Section 4.2.3 [Pérez-Rosas et al., 2015a]), "Face-AU" refers to the automatically detected face action units (Section 4.2.2 [Mimansa et al., 2016]) and "Others" refers to the rest of the features, i.e. the MUMIN-based nonverbal features (i.e. gaze, head movement, hand gestures) and the verbal features based on uni-grams and bi-grams (Sections 4.3.1 and 4.3.2, respectively).

As mentioned before, in this study we focus on deception detection particularly using face-based nonverbal features; therefore the experimental analysis is applied accordingly. In detail, deep face features are compared with Face-Manual and Face-AU when fused with other nonverbal features and verbal features. Comparisons are made using three classifiers: SVM, LMKL and MVL, with the claims: *i)* MVL performs better than the other classifiers, *ii)* deep face features perform better than the other nonverbal features, which can be shown by investigating the contribution of each feature group using the multi-view regularization parameter and/or by using deep face features alone (without fusing with another feature). To analyze the effectiveness of the deep face features (particularly, the ones extracted from VGGFace [14], as they perform the best when they are combined with other features), we also use them alone when SVM is applied.

For MVL training, we perform a grid search over the following sets of regularization parameters $\gamma_A = \gamma_B = \gamma_W = \{2^{-19}, 2^{-18}, 2^{-16}, 2^{-12}, 2^{-8}, 2^{-4}, 2^{0}, 2^{4}\}$ and two types of kernels, i.e. linear and RBF.

Several different combinations of views are tested, but we find that the best results are generated by the following four: *a)* face view (MUMIN-based, facial action units-based or deep face features), *b)* MUMIN-based gaze and head movements view, *c)* MUMIN-based hands and hands trajectory view and *d)* bi-grams and uni-grams

Features	**Method**	**Accuracy**
Face-Manual & Others	SVM	0.67
Face-Manual & Others	LMKL	0.76
Face-Manual & Others	MVL	0.79
Face-AU & Others	SVM	0.63
Face-AU & Others	LMKL	0.67
Face-AU & Others	MVL	0.75
Face-VGGFace & Others	SVM	0.73
Face-VGGFace & Others	LMKL	0.76
Face-VGGFace & Others	MVL	**0.89**
Face-AlexNet & Others	SVM	0.80
Face-AlexNet & Others	LMKL	0.80
Face-AlexNet & Others	MVL	0.74
Face-ResNet & Others	SVM	0.71
Face-ResNet & Others	LMKL	0.73
Face-ResNet & Others	MVL	0.79
Face-GoogLeNet & Others	SVM	0.73
Face-GoogLeNet & Others	LMKL	0.74
Face-GoogLeNet & Others	MVL	0.78
Face-AlexNet-FT & Others	SVM	**0.99**
Face-AlexNet-FT & Others	LMKL	**0.99**
Face-AlexNet-FT & Others	MVL	**0.98**
Face-VGGFace	SVM	0.79
Face-AlexNet-FT	SVM	**0.99**

Table 4.1: The best results of each method with the features utilized. The best results are emphasized in bold-face. FT stands for fine-tuning.

view.

After applying the DNNs, we obtain a feature vector for each frame. We combine them in order to have a single vector for each video using two methods, i.e. by taking the average and the maximum of their values. The average leads to the best results for all cases.

In Table 4.6 we report the classification accuracy results. As can be seen, MVL outperforms LMKL and SVM in all cases, except when AlexNet features are used (Face-AlexNet). Additionally, the fusion of deep face and other features outperforms the state of the art features combinations (i.e. Face-Manual & Others and Face-AU & Others), especially when MVL is applied. Interestingly, SVM trained only on VGG-Face-based features performs as good as Face-Manual & Others and Face-

Face	C-Face	C-Gaze-Head	C-Hand	C-N-grams
VGGFace	**0.050**	0.007	0.004	0.016
AlexNet	**0.062**	4.542e-06	2.414e-05	0.018
ResNet	0.004	0.009	0.032	**0.034**
GoogLeNet	**0.059**	0.023	0.006	0.002
AlexNet-FT	**0.048**	4.123e-05	4.894e-05	0.032

Table 4.2: Importance (according to the values of C) of each feature group.

ResNet & Others with MVL, and better than all the other combinations (sometimes significantly, with p-value <0.05) expect for Face-VGGFace & Others with MVL, Face-AlexNet & Others with SVM and LMKL. This shows that the extracted deep face features are good enough to be used alone, particularly compared to the state of the art features.

Given the general better performance of the deep face features, we decide to apply fine-tuning only on the deep architecture that performs the worst when combined with MVL, namely AlexNet. We substitute the last classification layers with new ones for the classification of deceptive and truthful frames. Then we set a high learning rate for the new layers and a low one for the layers we keep from the pre-trained network. Finally, we train the network using stochastic gradient descent, cross validation (70/30 training-validation data split) and a low number of epochs (10). We refer to the resulting fine-tuned DNN as AlexNet-FT. The results shown in Table 4.6 suggest that, indeed, fine tuning is helpful.

Finally, in Table 4.6 we report the values of the optimized C parameter, which indicate the contribution of each view to the final classification results obtained with MVL. As can be seen, except for ResNet, the most important views are the ones corresponding to deep face features; this observation is also validated by the results obtained when these features are used alone to train the single-view classifier (SVM).

4.7 Conclusions

In this chapter we extracted and compared various features based on face, head and hands movements. As shown in previous studies, features extracted from face were the most important nonverbal features. This result was obtained when *a)* manually annotated face-based nonverbal features were used and when *b)* facial action-units based features representing the same manually annotated face-based nonverbal features were utilized. Motivated by the success of face-based nonverbal features for automatic deception detection, instead of using hand-crafted features, we utilized deep features obtained by applying various pre-trained Deep Neural Networks as feature extractors, with and without fine-tuning. Using deep face features resulted in improved deception detection accuracy.

We also showed improved accuracy by employing better learning techniques in the form of multi-view learning (MVL), leveraging the different statistical properties of each modality and each feature type. Additionally, the optimized values of the views combination parameter of MVL comfirmed the importance of face cues, showing that the feature group which contributed the most to the final classification result was the one based on deep face features (with the only exception of features extracted using ResNet).

Sometimes, faces cannot be captured with a high level of detail. Most social gatherings, for example, take place in wide areas, where the scene must be captured as a whole and zooming on single persons is not possible. Privacy is another constraint that could prevent faces to be captured in detail. In these scenarios, it might be useful to focus on whole-body based behavioral cues. In the next chapter we extend our analysis to body posture and orientation.

Publication:

Carissimi N., Beyan C., Murino V., "A Multi-View Learning Approach to Deception Detection". In Automatic Face & Gesture Recognition (FG 2018), 2018 13th IEEE

International Conference on 2018 May 15 (pp. 599-606). IEEE.

Chapter 5

Body Pose and Group Detection

Many studies showed that the majority of gestures produced consciously (and unconsciously) by the body are associated with speech. They represent social signals such as illustrators or regulators [McNeill, 1992] and, in general, they are used to regulate interactions [Morris, 2002]. Nonetheless, other works have studied other social signals communicated by the body, analyzing how posture and limbs movements express basic emotions, like happiness, surprise or anger [Coulson, 2004, Gross et al., 2007, Pollick et al., 2001]. Postures are also reliable cues about the attitude toward a social situation [Richmond et al., 1991]. Existing literature classifies them following three main criteria [Scheflen, 1964]: *inclusive* and *non inclusive* postures, accounting for how much someone takes into consideration the presence of someone else; *face-to-face* and *parallel* postures, accounting for the level of engagement in a conversation; *congruent* and *incongruent* postures, accounting for psychological involvement.

In this Chapter we analyze body posture and its relationship with free standing conversational groups and f-formations, focusing on fully articulated body pose estimation in highly complex social scenarios. In Section 5.1 a brief overview of the used pose estimation algorithms is given. In Section 5.2 our proposed approach for pose reconstruction in presence of occlusions is detailed and in Section 5.3 qualitative and quantitative results are presented. Finally, Section 5.4 evaluates the estimated orientation for the task of group detection and conclusions are presented in Section 5.5

5.1 Body Pose Estimation

Body pose estimation is a popular task in the computer vision community. Two main approaches to the problem exist. The first one deals with a complete and fully articulated estimation of the location of body parts and joints, and pose can be predicted in 2D [Toshev and Szegedy, 2014, Newell et al., 2016, Pishchulin et al., 2016, Insafutdinov et al., 2016, Cao et al., 2017] or 3D [Akhter and Black, 2015, Zhao et al., 2017, Tome et al., 2017, Rogez et al., 2017]. The second approach estimates only the average direction of a single body part, such as the head or upper body, expressed as the *yaw* angle [Chen and Odobez, 2012, Chamveha et al., 2013, Yan et al., 2016, Ricci et al., 2015, Alameda-Pineda et al., 2015, Subramanian et al., 2015, Tan and Hung, 2018]. Moreover, the estimation is cast as a classification problem, where a “directional class” (a quantized value of the angle) is predicted. Examples of both approaches can be seen in Figure 5.1.

Even though existing methods produce high precision results, the problem of human pose estimation still remains challenging. For instance, real world images present several complexities, such as body appearance variability due to different clothing and lighting, uncommon body poses and crowded scenarios, which introduce occlusion problems. Specifically, when parts of the body are severely occluded, the resulting missing visual information might lead to the prediction of incomplete or wrong body poses (see examples in Figure 5.2). For this reason, prediction of the average direction has usually been favoured in challenging scenarios. However, this can lead to a less accurate orientation prediction and, in general, to less rich extracted information. We then propose an algorithm for the reconstruction of missing data

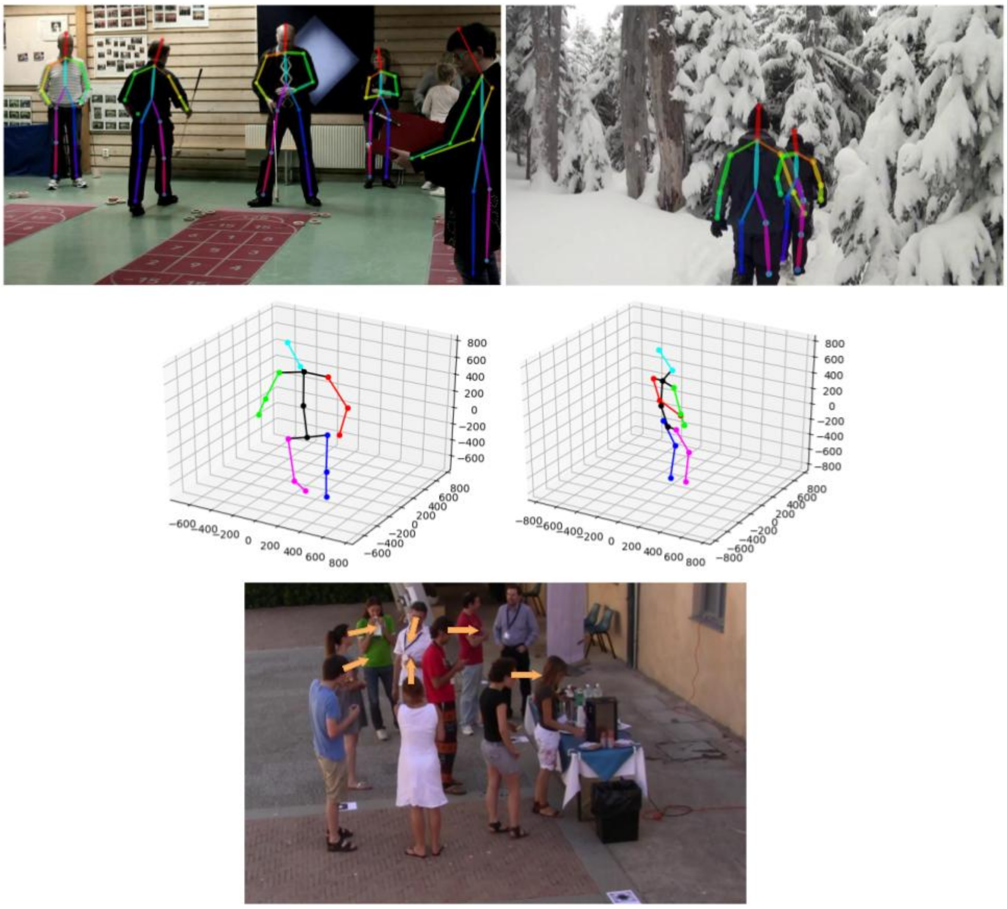

Figure 5.1: Example of body pose estimation.

that can aid the pose estimation process. Concretely, we propose a method for reconstructing complete articulated 2D poses from incomplete ones. The resulting poses can then be fed to a 3D pose estimation algorithm that generates full 3D human body representations even in the presence of occlusions. 3D poses are then, used, for detecting groups of people in RGB images.

5.1.1 3D Body Pose Estimation

We start by describing the algorithm used for 3D body pose estimation. We chose [Tome et al., 2017] for its robustness and computational speed. [Tome et al., 2017] performs a bottom-up estimation, where 2D poses are first estimated from the input image and then 3D poses are sampled from a learned model and fitted on the 2D joints. The process is iterative and, at each iteration, the resulting 3D pose is used to

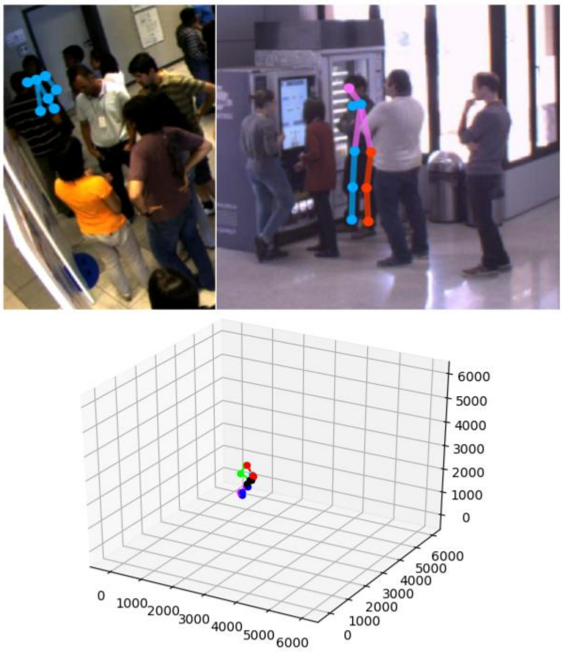

Figure 5.2: Example of wrong 2D (top) and 3D (bottom) body pose estimation.

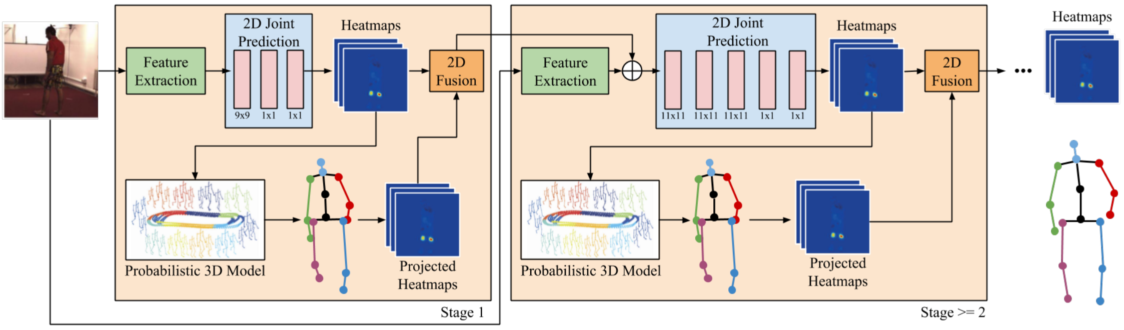

Figure 5.3: 3D pose estimation architecture [Tome et al., 2017].

refine the previous 2D prediction. Each stage (except the first one) receives as input a combination of outputs generated at previous stages, i.e. 2D and 3D heatmaps. The combination is a weighted sum, where the weights are learned during the end-to-end training. Figure 5.3 shows the architecture of the whole system.

5.1.2 2D Body Pose Estimation

We substituted the 2D pose estimator of [Tome et al., 2017] with the real-time multi-person pose estimator OpenPose [Cao et al., 2017, Wei et al., 2016]. Although OpenPose is not the best performing algorithm on pose estimation datasets, we found it to be the more robust when tested on real-life scenarios not strictly related to pose estimation (e.g. surveillance ones).

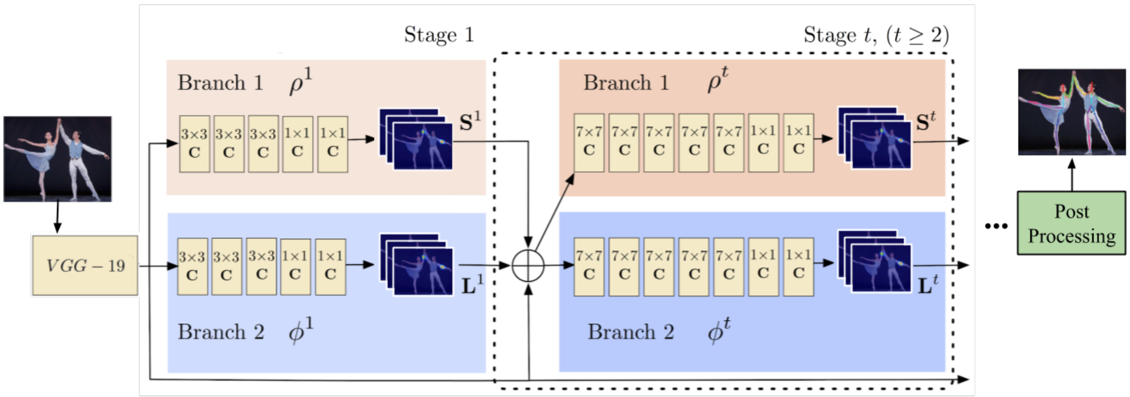

Figure 5.4: 2D pose estimation architecture [Cao et al., 2017].

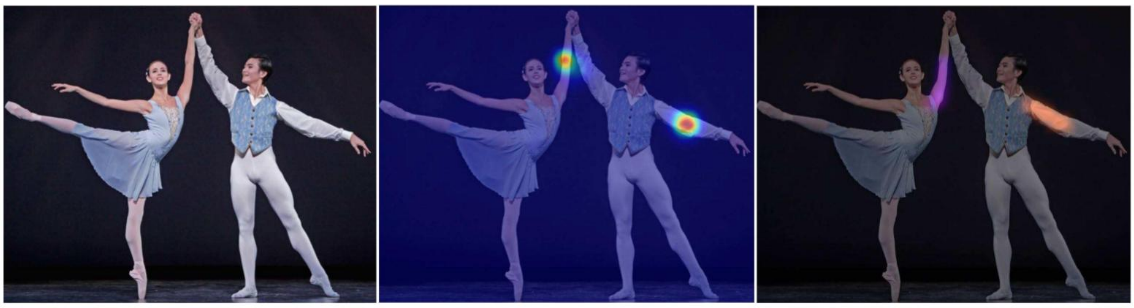

Figure 5.5: Example output of the network [Cao et al., 2017]: input image (left), left elbow heatmap (center), left upper arm affinity fields (right).

Figure 5.4 shows the overall architecture: a finetuned VGG network [Simonyan and Zisserman, 2014] extracts preliminary features from the input image. The extracted features are then fed to two branches of a multistage convolutional neural network (CNN); the upper branch (Branch 1) predicts body parts **S** in the form of confidence maps, one for each body part type; the lower branch (Branch 2) predicts part-to-part affinity fields (PAFs), one for each limb type. PAFs are 2D vector fields, where each pixel encodes location and orientation information across the region of support of each limb. As for [Tome et al., 2017], each stage receives as input a combination of outputs computed in previous stages, i.e. a concatenation of heatmaps, PAFs and features extracted by the VGG network. Examples of output confidence maps and affinity fields are shown in Figure 5.5. Discrete joints locations are obtained by performing non-maximum suppression on the heatmaps, while part association is computed by solving a bipartite matching problem on graphs where nodes are joints candidates and edges encode pair-wise association scores computed by summing the part affinity values along the line connecting the two joints (Figure 5.6).

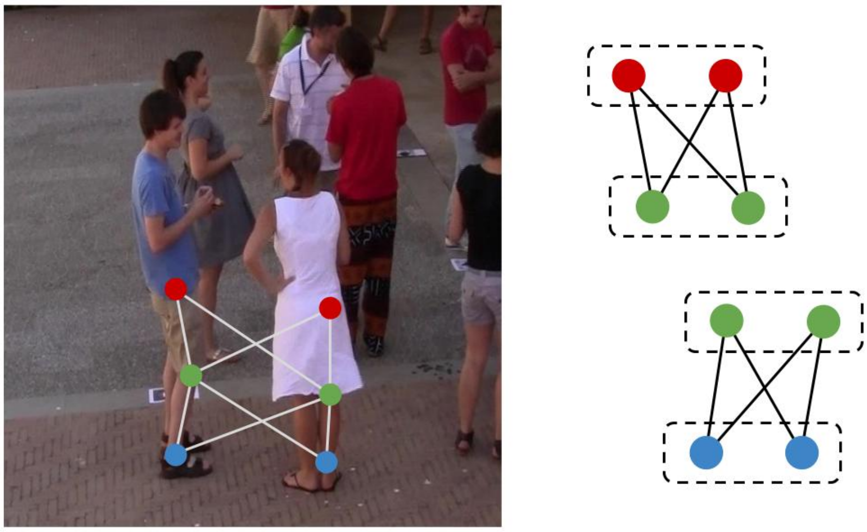

Figure 5.6: Body parts association. Left: joint locations for the joints right hip, right knee and right ankle. Right: the resulting graphs for the bipartite matching problem.

5.2 Pose Reconstruction

As previously mentioned, we propose a method for 2D pose reconstruction and not 2D pose prediction. We cast the task as a denoising problem, where the corrupted signal is represented by the partial human pose, and the resulting uncorrupted signal is the full reconstructed pose.

The choice of our model is motivated by two main reasons. In the first place, the model should be able to predict missing information and, second, has to deal with low dimensional data. Occlusions and degraded visual data might cause a pose detector to miss some types and number of joints in an unpredictable way. The resulting partial human pose can, thus, be seen as a noisy, stochastically corrupted version of the original data which is the complete human pose in our case. The model must, then, be able to learn a robust representation of the data even when parts of the data are missing. Unlike RGB images, which are composed by hundreds, or thousands of pixels, our domain data are small vectors of a few concatenated 2D coordinates (see Section 5.2.2), therefore we choose a model which is simple, yet powerful enough to learn a robust representation of this low dimensional domain data. Auto-encoders, as seen in previous works [Rumelhart et al., 1986, Vincent et al., 2008], are a powerful tool for learning representations of complex data distributions, and their denoising variant [Vincent et al., 2008] is specifically designed to deal with incomplete input data.

The next Section provides a short review of the theory behind auto-encoders, denoising auto-encoders and one of their most recent variants, variational auto-encoders.

5.2.1 Auto-Encoders

Auto-encoders have been introduced several years ago in [Rumelhart et al., 1986], they consist in an unsupervised learning model and have been used for different purposes such as dimensionality reduction, feature extraction [Vincent et al., 2008], pre-training of deep nets [Bengio et al., 2007, Vincent et al., 2010], data generation and reconstruction [Hou et al., 2017]. Concretely, an auto-encoder is a type of multi-layer neural network trained to map the input to a different representation of it, so that the input can be reconstructed from that representation. The simplest form of an auto-encoder has a single hidden layer which maps (*encodes*) an input $\mathbf{x}$ to its new representation $\mathbf{y}$

$$\mathbf{y} = s(\mathbf{W}\mathbf{x} + \mathbf{b}) \tag{5.1}$$

where s is a (usually non-linear) activation function, while W and b are, respectively, the weights and bias of the layer. The encoded input $\mathbf{y}$ is, then, mapped back (*decoded*) to a reconstruction $\mathbf{x_r}$ of the input

$$\mathbf{x_r} = s(\mathbf{W}'\mathbf{y} + \mathbf{b}'). \tag{5.2}$$

Training is performed by minimizing a loss function

$$L(\mathbf{x}, \mathbf{x_r}) \tag{5.3}$$

which, in our case, is the mean squared error MSE, calculated between the reconstructed input $\mathbf{x_r}$ and the target output, which is the input $\mathbf{x}$ itself. If the dimension of $\mathbf{y}$ is smaller than the dimension of $\mathbf{x}$, the auto-encoder is called undercomplete; on the other end, if the dimension of $\mathbf{y}$ is larger, the auto-encoder is called overcomplete.

If the dimension of the hidden units is larger than the original input, the auto-encoder might learn the identity function; however, there are different techniques to avoid this occurrence. One of these techniques introduces randomness during training: the network is fed with a stochastically corrupted version of the input $\mathbf{x_c}$, while the target output remains the original uncorrupted input $\mathbf{x}$. This training approach has been introduced by Vincent et al. in [Vincent et al., 2008] and the resulting models are called *denoising auto-encoders*. Their original purpose was to make the learned representation more robust to partial corruption of the input, but they present an additional useful property, i.e. the ability to reconstruct missing data from the input, which is well suited for our problem of missing joints prediction.

One downside of standard auto-encoders is that they tend to map similar input samples to latent vectors which might be very close to each other, resulting in almost identical reconstructions. This behavior is acceptable when input data represents classes (e.g. images of numbers or letters). On the contrary, preserving small input differences in the reconstruction is very important when dealing with human poses, which do not form a clustered space, but a continuous, smooth domain. Variational auto-encoders [Kingma and Welling, 2013] can learn such a continuous representation by design, making them more suited for our problem. Similarly to classic auto-encoders, they have the same encoder/decoder structure (where the encoder maps the input to a latent representation, and the decoder reconstructs the input from such representation). The main difference is that the latent variables are not a ”compressed representation” of the domain data itself, but they encode the parameters (i.e. the mean μ and standard deviation σ) of a *distribution* (typically an n-dimensional Gaussian one) modeling the input data. In order to force this, another term is added to the loss (see Eq. 5.3), i.e. the Kullback-Leibler divergence (D_{KL}) [Akaike, 1998], which measures the divergence between two probability distributions and has the form

$$D_{KL}(P\|Q) = -\sum_{i} P_i log(Q_i/P_i) \tag{5.4}$$

where P is the encoded n-dimensional Gaussian distribution $\mathcal{N}(\mu,\, \sigma^2)$ and Q is the target standard normal distribution.

5.2.2 Modeling the Human Pose

Since the chosen reconstruction loss needs a complete human pose as the target output, we need to select full human poses from the dataset as training data. Each pose is represented by a set of n 2D locations, where n is the number of joints (see Sections 5.3.1 and 5.3.2). The concatenation of these joints produces a vector of $n * 2$ elements (the x and y coordinates) which is the input of the network. Since the coordinates of the annotated joints are labeled in the image space, poses which are very similar to each other might appear in different parts of the image, resulting in input vectors with very different values. We, thus, normalize them using the following procedure: first we find the center of the torso (C_T) by averaging the coordinates of the neck and at least one of the shoulders and hip joints; the pose is then translated to the obtained 2D point and finally scaled by the distance between the neck and C_T. At testing time, this normalization technique requires an incomplete pose to have all the aforementioned joints, negatively affecting the number of poses that is processed by the network (see Section 5.3.4). Given that we are using a denoising auto-encoder, the training data must also be *corrupted.* We do this by adding noise to the previously normalized poses, randomly masking a small number of joints.

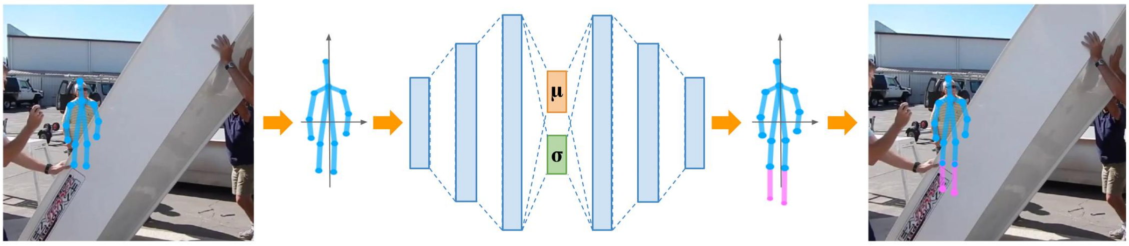

Figure 5.7: The pipeline of our method. Given an RGB image, a human pose prediction algorithm is used to generate one or more poses. The incomplete ones are, then, normalized and fed to the auto-encoder, which outputs the corresponding full human poses.

Fig. 5.7 shows the overall architecture: since the input vector is small compared to the space of the data we want to reconstruct, we choose to implement an over-complete auto-encoder. The encoder and decoder are composed by 2 hidden layers, each one gradually encoding (and decoding) more robust features. Layers μ and σ represent, respectively, the mean and standard deviation of the distribution we want to learn, while the final layer of the encoder represents a sample of it.

5.3 Pose Reconstruction Experiments

In this section, we report quantitative and qualitative results of our method, evaluated on two datasets, MPII Human Pose [Andriluka et al., 2014] and Microsoft's COCO Keypoint Detection [Lin et al., 2014], which are the most famous and widely datasets for multi-person pose estimation.

5.3.1 MPII Human Pose Dataset

The MPII Human Pose dataset [Andriluka et al., 2014] consists of around 25000 images and a total of 40000 annotated human poses. The training set is composed of 28000 of these poses, while the test set is composed of 11000 poses. Images contain people engaged in numerous activities and a variety of contexts, with a high variable of articulated poses and camera perspectives. People can be fully visible, severely occluded or partially out of the camera field of view. A full pose is composed of 16 landmarks, each one corresponding to the location, in image coordinates, of a body joint (head, neck, thorax and left and right shoulders, elbows, wrists, hips, knees and ankles).

5.3.2 COCO Keypoint Detection Dataset

The Microsoft's COCO Keypoint Detection dataset [Lin et al., 2014] is a subset of the whole COCO dataset, focused on the localization of person keypoints. The training and validation sets contain, respectively, around 260000 and 11000 annotated human poses. Unlike MPII, a full pose is composed of 17 joints, corresponding to nose and left and right shoulders, elbows, wrists, hips, knees, ankles, eyes and ears.

5.3.3 Experimental Settings

Our model takes a pose as the input and generates its reconstruction. If the pose is incomplete, i.e. with one or more missing joints, the output is a prediction of the corresponding full pose. As described in Section 5.2.2, the loss function needs a fully annotated pose; thus, we need to select a subset of the training data containing only complete poses. For the MPII dataset, this results in a total of, approximately, 20000 samples; we, then, use our own split (85%/15%) on the obtained data for training and validation purposes, and augment the remaining training data following a standard procedure for single pose estimation algorithms [Newell et al., 2016, Chen et al., 2017], obtaining a total of approximately 500000 training samples. In particular, we perform data augmentation by flipping and rotating the original poses (+/-30 degrees). We then normalize each pose, mask a random number of joints (from 0 up to 5, which roughly corresponds to 35% of the total number of joints) and feed the obtained data to the network.

The COCO dataset, on the contrary, has only a few thousands of complete poses, which, even after data augmentation, would not be enough for training purposes. Therefore, we decide to use the dataset only for testing. Since COCO and MPII have different annotated joint types, we feed the network (trained on MPII) with only the joints that are common between the two datasets (i.e. left and right shoulders, elbows, wrist, hips, knees and ankles) and set to zero the missing ones (head, neck and thorax).

The encoder is composed of two fully connected hidden layers, with 64 and 128 hidden units. Symmetrically, the decoder is composed of two hidden layers, with 128 and 64 hidden units, and an output layer with the same dimension as the input one. As in [Kingma and Welling, 2013, Feng et al., 2017], we use 20 latent dimensions. Every fully connected layer has ReLu non-linearities. The loss function is the sum of MSE (between the uncorrupted input and the reconstructed pose) and the D_{KL} (Section 5.2.1). The network is implemented using TensorFlow [Abadi et al., 2016] and trained with the Adam optimizer [Kingma and Ba, 2014] with a learning rate of 1e-3.

5.3.4 Quantitative Analysis

In this section we show quantitative results of the proposed pipeline on the datasets described in Sections 5.3.1 and 5.3.2. For the generation of the input poses, we use the bottom-up multi-person pose estimator OpenPose [Cao et al., 2017, Wei et al., 2016] and its matlab CB: Matlab implementation, without modifying its preset parameters. Although OpenPose is not the best performing method on MPII and COCO anymore, and it's less precise in predicting complete human poses compared to other top-down approaches, we found it to be the more robust when tested on real-life datasets not strictly related to pose estimation and on which it wasn't trained on (such as Salsa [Alameda-Pineda et al., 2016]). Fig. 5.8 shows a comparison between poses generated by OpenPose and those generated by the state of the art

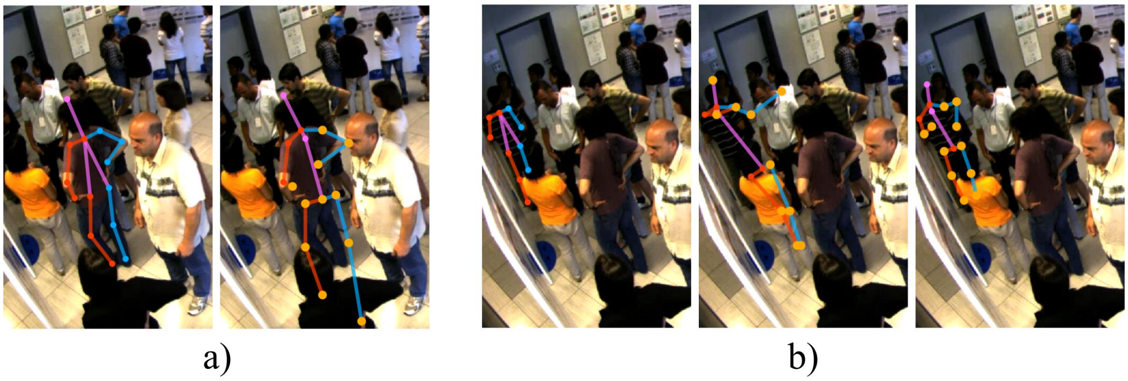

Figure 5.8: Comparison between OpenPose and Regional Multi-Person Pose Estimation (RMPE) [Fang et al., 2017]. a) and b) left show OpenPose predictions, while a) right and b) center, right show RMPE predictions. Orange joints have a confidence score below 0.2.

top-down approach called Regional Multi-Person Pose Estimation (RMPE) [Fang et al., 2017]. In a), OpenPose (left) produces a complete and better estimation of the pose, compared to RMPE (right). In b), OpenPose (left) cannot predict the head, the wrists and the right ankle, while RMPE (center, right) predicts all joints; however, RMPE generates two poses for the same person, due to redundant detections, and their quality is worse than the OpenPose one. Clearly, the underlying person detector is an important factor in the final performance of a top-down pose estimation algorithm. Also, top-down approaches learn not just local information (i.e. joints appearance) but also global information (i.e. joints relative location and appearance) and this information might be harder to generalize to unseen data.

We compare OpenPose's results with the results generated by our method using two metrics, the Miss Rate (MR) and the Percentage of Correct Keypoints (PCKh). MR is computed as

$$\#joints_{missed}/\#joints_{gt} \tag{5.5}$$

where $\#joints_{missed}$ is the number of missed (annotated) joints and $\#joints_{gt}$ is the number of all (annotated) joints. PCKh is a standard metric in pose estimation introduced in [Andriluka et al., 2014] for evaluation on the MPII dataset, where a keypoint is considered as correctly predicted if its distance from the ground truth

Method (All Joints)	**Ankle**	**Knee**	**Hip**	**Wrist**	**Elbow**	**Shoulder**	**Neck**	**Head**	**Average MR**
OpenPose [Cao et al., 2017]	0.072	0.040	0.021	0.066	0.037	0.019	0.011	0.019	0.039
Our Method	**0.020**	**0.015**	**0.016**	**0.012**	**0.011**	**0.010**	**0.011**	**0.011**	**0.014**

Table 5.1: Joints Missing Rate on the MPII dataset.

is less than a fixed threshold (specified as a fraction of the person's head size). The corresponding ground truth is assigned to each pose according to the highest PCKh.

While MR quantifies how many joints are failed to be predicted, PCKh quantifies the actual "quality" of the predictions.

We do not perform a comparison using the standard mean Average Precision (mAP) metric, which is commonly used in MPII for multi-person pose estimation, because it penalizes joints with no ground truth correspondence as false positives.

Table 5.3.4 shows that our method outperforms OpenPose in terms of number of missing joints. As can be seen, the highest missing rate differences correspond to joints which are body extremes (i.e. wrists and ankles) and thus more prone to be occluded. Even though our method is supposed to predict all missing joints, the missing rate is not 0 because it relies on the detection of the subjects by the baseline human pose estimator.

The quality of the predictions generated by our method can be seen in Table 5.3.4, where its PCKh is better than the OpenPose one, especially (as for the missing rate) for those joints which are frequently occluded. The highest difference in PCKh can be seen when computed over joints labeled as "occluded" only. Results on head and neck are omitted because they are never occluded.

One advantage of our method is that, by using 2D coordinates as input domain, it can be easily applied to different datasets it has not been trained on: Tables 5.3.4 and 5.3.4 show, respectively, the Missing Rate and the PCKh computed on COCO.

Finally, we report the computational time for training and testing. The analysis is performed on a laptop with 16GB of RAM and an NVIDIA GeForce GTX 960M with

Method	**Ankle**	**Knee**	**Hip**	**Wrist**	**Elbow**	**Shoulder**	**Neck**	**Head**	**Average**
(All Joints)									**PCKh**
OpenPose [Cao et al., 2017]	79.87	87.17	93.0	79.15	89.03	95.97	97.71	96.11	88.73
Our Method	**80.93**	**87.44**	**93.06**	**80.38**	**89.92**	**96.41**	**97.75**	**96.53**	**89.33**
Method	**Ankle**	**Knee**	**Hip**	**Wrist**	**Elbow**	**Shoulder**			**Average**
(Occluded Joints)									**PCKh**
OpenPose [Cao et al., 2017]	59.07	73.26	87.71	57.06	76.71	91.23			74.47
Our Method	**61.18**	**73.83**	**87.80**	**60.78**	**78.70**	**92.32**			**75.77**

Table 5.2: PCKh@0.5 on the MPII dataset, computed on all joints and only on joints labeled as occluded.

Method	**Shoulder**	**Elbow**	**Wrist**	**Hip**	**Knee**	**Ankle**	**Average**
(All Joints)							**MR**
OpenPose [Cao et al., 2017]	0.0021	0.0104	0.0214	0.0371	0.0752	0.0539	0.0333
Our Method	**0.0021**	**0.0032**	**0.0068**	**0.0150**	**0.0093**	**0.0052**	**0.0069**

Table 5.3: Joints Missing Rate on the COCO dataset.

4GB of RAM. Training requires only 3 hours, while reconstruction of a single pose requires, on average, 0.88 ms. This shows that our method can be easily combined with any existing pose estimation architecture without significantly affecting the overall computational time.

5.3.5 Qualitative Analysis

In this section we show qualitative results of our predictions. In Fig. 5.9 (images taken from MPII), the top row shows (in blue), predictions obtained from OpenPose, while the bottom row shows the corresponding complete poses generated by our model (the predicted missing joints are in magenta). In column a) and b) ankles are missing from sitting poses and our model is able to predict a plausible locations of them. In column c) a man is standing but both ankles are completely occluded by a

Method	**Shoulder**	**Elbow**	**Wrist**	**Hip**	**Knee**	**Ankle**	**Average**
(All Joints)							**PCKh**
OpenPose [Cao et al., 2017]	80.22	54.91	63.71	39.48	35.74	23.60	49.61
Our Method	**80.07**	**56.25**	**65.44**	**41.25**	**41.40**	**28.43**	**52.14**

Table 5.4: PCKh@0.5 trained on MPII and tested on COCO, computed on all joints.

foreground object and missing from the pose generated by OpenPose; however, our model is able to predict their position and produce a plausible complete standing pose. In column c) the right arm (elbow and wrist) is missing; our model generates the missing joints in a spatial configuration which is similar to the joints of the visible left arm. The last column shows an extreme case where the number of missing joints is very high, thus providing little context for the final prediction: a man is standing, with raised arms and head occluded by a foreground object. Although our model generates arms which are completely lowered, the resulting pose is still a plausible human pose.

Fig. 5.10 shows more examples of predictions obtained from OpenPose (top row) and the corresponding complete poses generated by our method (bottom row). In column a), not just an ankle but the entire left leg (knee and ankle) is missing; the predicted complete pose closely resembles the sitting person pictured in the image. In column b), the right wrist is not detected and both arms are raised, but our prediction is very close to the real wrist. Ankles (columns b), c) and d)), are outside the camera field of view; however our model is able to predict a full pose even when RGB information is missing.

Finally, Fig. 5.11 shows predictions on frames from Salsa (another dataset our model was not trained on), where it can be seen that our method can generate plausible human poses even when half of the body is missing (see columns c) and d), with completely occluded legs and arms).

5.4 F-formation Detection

As already described in Section 2.3, free standing conversational groups can be formally described using the concept of *f-formation*, which is a socio-spatial organization of people around multiple concentric spaces (o-, p- and r-space, Figure 2.2 top). More specifically, f-formations are defined by the location and orientation of

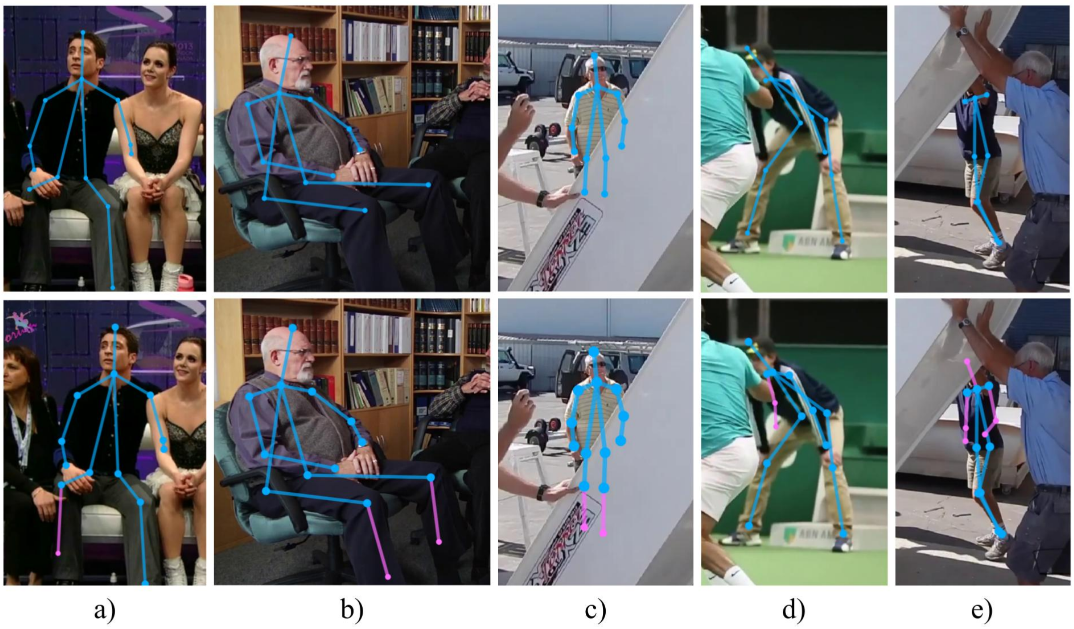

Figure 5.9: Examples of predictions obtained from OpenPose (top row, in blue) and the corresponding complete poses generated by our method (bottom row, in magenta) on MPII.

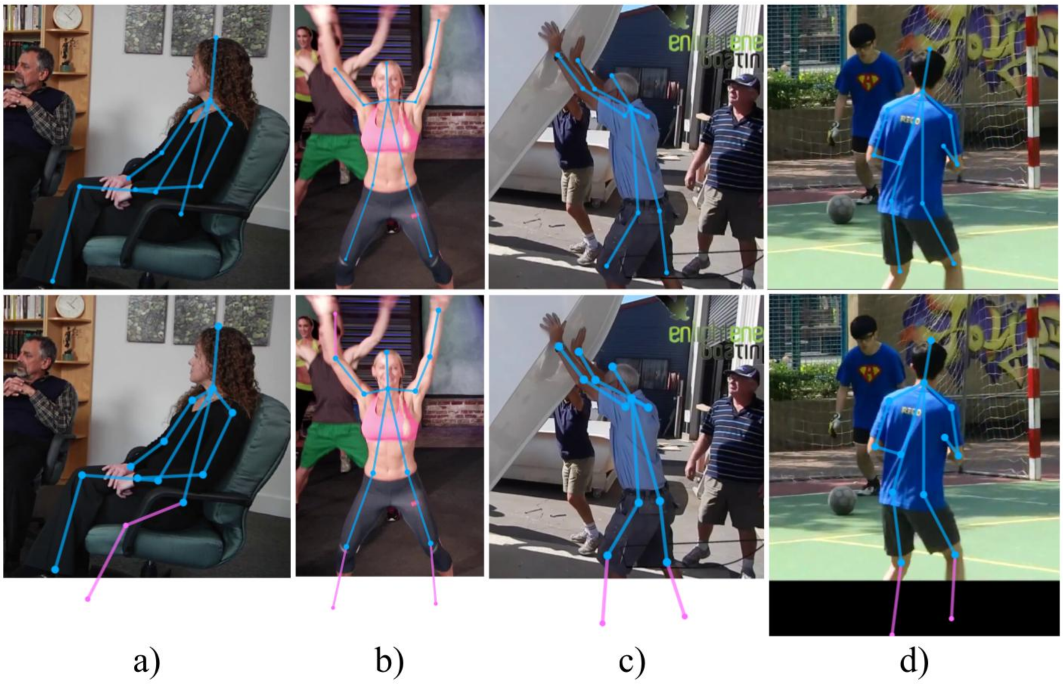

Figure 5.10: More examples of predictions obtained from OpenPose (top row, in blue) and the missing joints predicted by our method (bottom row, magenta) on MPII. As can be seen, the model is also capable of predicting joints which are outside of the camera field-of-view.

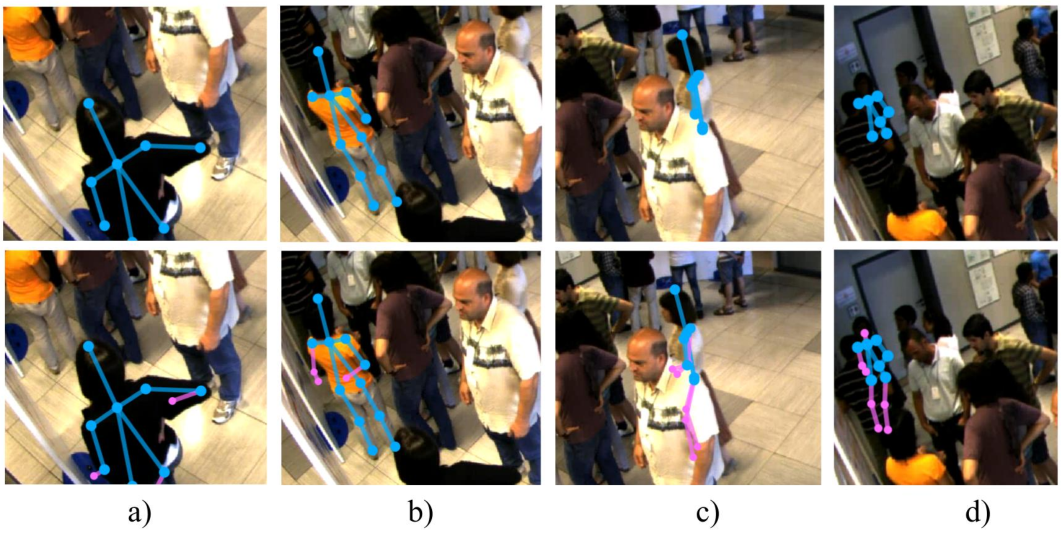

Figure 5.11: Examples of predictions on Salsa. Top row: OpenPose results (in blue). Bottom row: missing joints predicted by our method (in magenta).

people. We now describe the f-formation detection algorithms used in our experiments and the results obtained by using our method for the reconstruction of partial poses. We selected all methods with pubiclily available code.

5.4.1 F-formation Detection Algorithms

Single Scale Hough Voting [Cristani et al., 2011]. The algorithm in [Cristani et al., 2011] proposes a voting approach for the f-formations centers. For each individual $i \in L$ characterized by location (x_i, y_i) and orientation θ_i, a set of N votes $\{s_{i,n}\}$, $n = 1...N$, is generated by sampling from a uniform distribution $\mathcal{N}(\mu_i, \Sigma)$, where $\mu_i = (x_i, y_i, \theta_i)$ and Σ is a diagonal matrix with variances $\sigma_x^2, \sigma_y^2, \sigma_\theta^2$. Each sample votes for the center of an *o-space* with radius R along the subject's orientation and with a weight $w_{i,n} \propto \mathcal{N}(s_{i,n}; \mu_i, \Sigma)$. An accumulation space is defined by

$$\tilde{A}_I(x, y) = card(x, y) \cdot A_I(x, y), \forall (x, y) \in A_I(x, y) \tag{5.6}$$

where (x, y) is a voted center location, $card(\cdot, \cdot)$ is a function that returns the number of subjects voting for (x, y), and $A_I(x, y)$ accumulates the sum of weighted votes for center (x, y). The final f-formations centers are selected by evaluating in descending order the values of $\tilde{A}_I$ and checking if there are subjects inside the o-space. Radius R, variances $\sigma_x^2, \sigma_y^2, \sigma_\theta^2$ and number of samples N are free parameters, set according to sociological and empirical observations.

Multi-Scale Hough Voting [Setti et al., 2013b]. [Setti et al., 2013b] proposes an extension of [Cristani et al., 2011], where voting is performed several times, one for each f-formation cardinality value. Here, the radius is proportional to cardinality k and is set as $R_k \simeq \frac{s}{2\sin\frac{\pi}{k}}$, where s is an empirically set interpersonal space distance. The new equation for the accumulation space is

$$\tilde{A}_I(x, y) = card(x, y) \cdot \tilde{E}(x, y), \forall (x, y) \in A_I(x, y) \tag{5.7}$$

where $\tilde{E}(x, y)$ is the weighted entropy

$$\tilde{E}(x, y) = \sum_{i \in L} h_i(x, y) \cdot p_i(x, y) log_2 p_i(x, y), \tag{5.8}$$

with $h_i(x, y)$ as the normalized count of how many times subject i voted for center (x, y) and $p_i(x, y)$ as the sum of all weights $w_{i,n}$. An accumulation space $\tilde{A}_I(x, y)^{(k)}$ is built for each cardinality k and relative f-formations are obtained by following the same selection procedure from [Cristani et al., 2011]. The final multi-scale f-formations are obtained by merging f-formation results from the single cardinalities.

Game Theoretic Approach [Vascon et al., 2014]. [Vascon et al., 2014] develops a game-theoretic clustering approach which also models the uncertainty about location and orientation. First, for each person a socio-attentional view frustum is generated, modeled by a 2D histogram which represents a 2D gaussian distribution generated using the person location, orientation and a fixed view angle. Second, an affinity matrix for all the persons in a frame is computed, using the 2D histograms and an affinity measure based on the Kullback-Leibler or the Jensen-Shannon divergence. Finally, clusters of people are found using a non-cooperative clustering game

[Bulò and Pelillo, 2009].

Graph Cuts [Setti et al., 2015]. In [Setti et al., 2015], graph cuts are used to minimize a function based on people location and o-space center candidates. For each individual, a transactional segment TS is the area in front of the body that is easy to reach and where hearing and sight are most effective (which corresponds to the o-space); it is defined as $TS \sim N(\mu_i, \Sigma_i)$, where $\mu_i = [x_{\mu_i}, y_{\mu_i}]$ is the center of the area, defined by the location and orientation of person i and $\Sigma_i = \sigma \cdot \mathbf{I}$. Given O_G as the set of candidate o-space centers and $[u_{G_i}, v_{G_i}]$ the o-space center of an f-formation containing subject i, a cost function is defined as

$$J(O_G|TS) = \sum_{i \in [1,n]} (u_{G_i}, x_{\mu_i})^2 + (v_{G_i}, y_{\mu_i})^2 + \sigma^{-2}|O_G|, \tag{5.9}$$

where the last term is a minimum description length prior preventing f-formations with one single person. Starting from a set of candidate o-space centers, the algorithm iteratively minimizes Equation 5.9 with the graph cut based optimization [Ladický et al., 2013] and updates the o-centers with the mean of the centers voted by the members of the current f-formations.

5.4.2 F-formation Detection Experiments

CoffeeBreak Dataset [Cristani et al., 2011]. The CoffeeBreak dataset [Cristani et al., 2011] depicts a social scenario where people gather in an open space, forming free standing conversational groups. It is a surveillance-like setting, where the camera is placed above heads and is not in close range (Figure 5.12). People are free to move and bodies are often occluded. The dataset presents two sequences (seq1 and seq2) of consecutive frames, for a total of 120 frames. Each frame is annotated with location, head orientation and an id for each tracked person, and f-formations (lists of person ids).

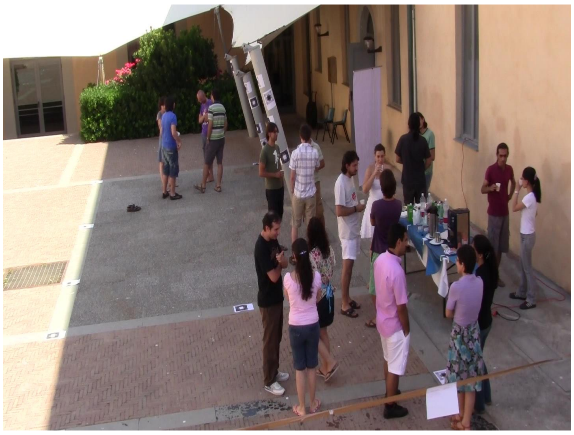

Figure 5.12: Example frame from [Cristani et al., 2011].

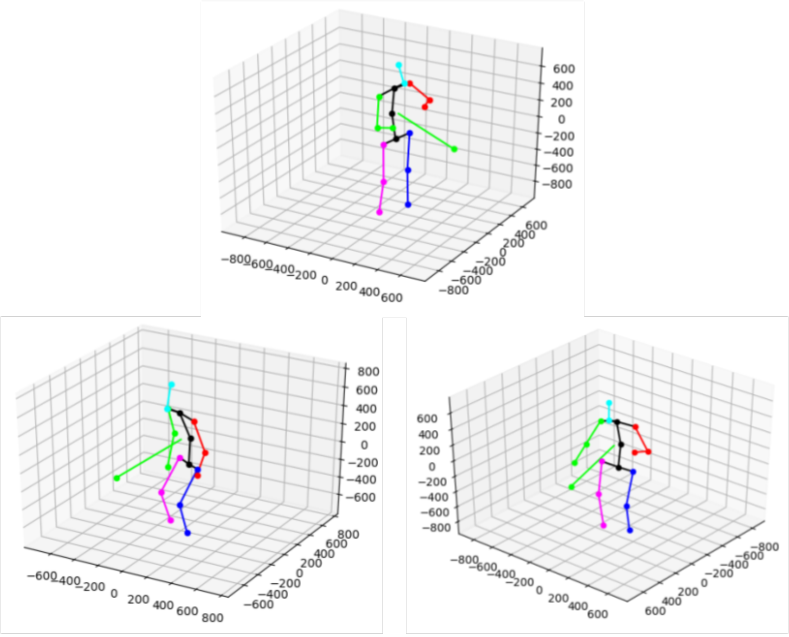

Figure 5.13: 3D poses and their corresponding orientation vectors (in green).

F-formation Detection Pipeline. Figure 5.14 shows the pipeline for the f-formation detection. First, 2D poses are extracted from each frame (Section 5.1.2). The incomplete ones are then reconstructed (Section 5.2) and all the poses are fed to the 3D pose lifting algorithm (Section 5.1.1). Finally, each pose orientation and ids are computed and used as input for the f-formation detection algorithms (Section 5.4.1). The (3D) orientation of the body is computed by calculating the average of two cross products, one between hips and neck and one between shoulders and pelvis joints. Examples of the resulting vector can be seen in Figure 5.13. 2D pose tracking is performed by combining the multi-object tracker [Pirsiavash et al., 2011] with a graph matching problem: first, each joint is singularly tracked using [Pirsiavash et al., 2011]. Then, for each pair of consecutive frames a bipartite graph is created, where each subset of nodes correspond to poses predicted in one of the two frames, while the edges encode the number of common joints ids (assigned in the previous step by [Pirsiavash et al., 2011]) between poses. Solving the bipartite matching problem generates a correspondence between poses of consecutive frames. By chaining the obtained matches from the first to the last frame, we obtain the tracking of the poses.

Ids are, then, matched to the labelled ones and location is obtained from ground truth.

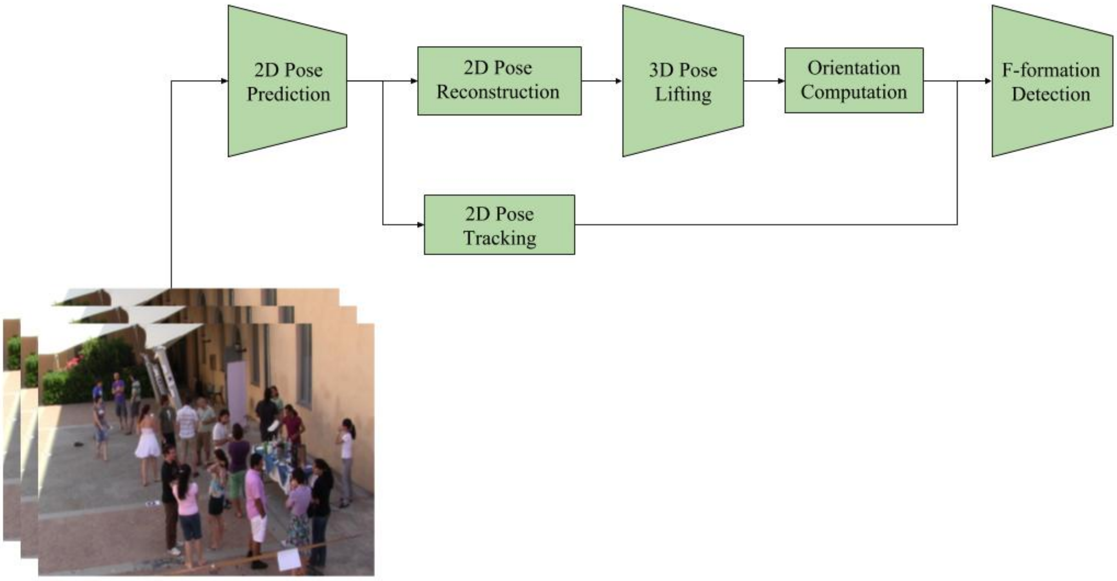

Figure 5.14: F-formation detection pipeline.

Results. Table 5.4.2 shows the comparison between the aforementioned methods for f-formation detection with and without our computed orientation, reporting precision, recall and F1 score. We also report the results of [Ricci et al., 2015] (last row), which jointly learns head/body orientation and f-formation detection. As can be seen, the orientation provided by our algorithm achieves results on par with the state-of-the-art, which uses annotated orientations (except for [Ricci et al., 2015]). Although our method is purely geometric, it is able to produce robust results and has the advantage of not being tied to the test dataset.

Method	**Precision**	**Recall**	**F1 Score**
HVFF lin [Cristani et al., 2011]	0.73	0.86	0.79
HVFF lin w 3D orientation	**0.74**	0.84	**0.79**
HVFF ent [Setti et al., 2013a]	0.81	0.78	0.79
HVFF ent w 3D orientation	0.75	**0.84**	**0.79**
HVFF ms [Setti et al., 2013b]	0.76	0.86	0.81
HVFF ms w 3D orientation	**0.76**	**0.86**	**0.81**
GTCG [Vascon et al., 2014]	0.83	0.89	0.86
GTCG w 3D orientation	**0.84**	0.88	**0.86**
GCFF [Setti et al., 2015]	0.85	0.91	0.88
GCFF w 3D orientation	0.83	**0.94**	**0.88**
Joint HBFF [Ricci et al., 2015]	0.84	0.88	0.86

Table 5.5: Comparison of state-of-the-art algorithms with and without body orientation computed with our method.

5.5 Conclusions

In this Chapter we focused on the body and used it for the analysis of free standing conversational groups.

Past works predicted "coarse" head and body orientation classes. However, we wanted to be able to predict and exploit full 3d articulated human poses. Although state of the art 2D and 3D pose estimators are able to cope with different challenges in in-the-wild social scenarios, occlusions are still a problem and might negatively affect the performance of predictions. In order to cope with this problem, we devised a method for the reconstruction of 2D poses that can be used in bottom-up 3D pose estimation. We approached the task as a denoising problem and showed that a simple model based on autoencoders leads to a satisfactory boost in performance. We reported quantitative and qualitative results on several pose estimation datasets and showed increased prediction performance over a well-known multi-person pose estimation algorithm and the ability to predict joints locations even when entire limbs are occluded.

We then estimated fine grained (real valued) body orientation from the obtained 3D poses and used it as a feature for f-formation detection. Analysis on a famous social interaction dataset showed group detection results on par with several state-of-the-art methods. Some of these methods use orientations which are either annotated ([Cristani et al., 2011, Setti et al., 2013b, Setti et al., 2015, Vascon et al., 2014]) or jointly learned from the dataset with f-formations ([Ricci et al., 2015]); some of them use temporal smoothing and information from the scene ([Ricci et al., 2015]). Our method is able to compute robust orientations which are not tied to the scene or the test dataset, making it more generalizable; additionally, it does not use any temporal information.

Publication:

Carissimi N., Rota P., Beyan C. Murino V., Filling the Gaps: Predicting Missing Joints of Human Poses Using Denoising Autoencoders, HBUGEN Workshop, European Conference on Computer Vision (ECCV) 2018.

Chapter 6

Conclusion

The goal of this thesis was to deepen the understanding of how nonverbal behavioral cues and social signals correlate to specific roles and behaviors, and how they can shape social structures. We pursued this goal by focusing on different parts of the body involved in the generation and perception of social signals, starting from the eyes and expanding to face and body. The resulting algorithms enrich the existing set of existing social signal processing tools for understanding human behavior.

6.1 Emergent Leadership Detection by Analyzing Gaze and Visual Focus of Attention

Eyes and gaze are fundamental tools for sensing our surroundings and the social signals coming from other beings. Gaze is also an informative cue about our focus of attention, revealing what or *who* we are currently interested in. In Chapter 3 we investigated the use of gaze as an indicator of emergent leaders in small groups. We approximated it by using head orientation and modelled each participant's visual focus of attention (VFOA). Different features based on VFOA were devised and then used to train a supervised classification model. Results showed that head orientation is an effective proxy for estimating gaze direction and the relatively high classification accuracy obtained for the most emergent leader class proved that VFOA is indeed a reliable indicator for this role. Intuitively, this can be explained by the fact that, as previously mentioned, leadership is about exerting dominance and control over other people. In face-to-face interactions, this translates into addressing and *looking* at someone. Symmetrically, the people addressed to by the leader have to look back, in order to give some sort of feedback. Our VFOA-based features model exactly these interactions.

Limitations and Future Work. The lack of direct gaze estimation is a limitation, since the actual VFOA does not always correspond to the head orientation of a person. At that time, gaze estimation algorithms were not able to produce reliable results on our dataset. Additionally, cameras could not be placed closer to the subjects, as it would have interfered with their interpersonal interactions and the VFOA. Current state-of-the-art algorithms could be evaluated in future work and using “real” gaze might further improve leadership classification accuracy.

Future work might also expand the dataset: participants were all young subjects, with an average age of 21.6 years (2.24 standard deviation) and all having the same occupation (i.e. psychology students). It would be interesting to see if the same VFOA dynamics hold true in sessions involving different age groups, mixed genders and different cultures.

6.2 Face-Based Behavioral Cues and Deception Detection

After the gaze, following the previously mentioned “zoom out” approach, we shifted our attention to face in Chapter 4. As the face is one of the most important parts of the body for interpersonal interactions and a mirror of our emotions and feelings, we decided to analyze its relationship with a complex human behavior: deception. In order to do that, we chose the only publicly available audio/video dataset [Pérez-Rosas et al., 2015a] depicting persons expressing "truthful" and "deceptive" statements in a high-stakes scenario (i.e. a courtroom), and compared several features based on face, head, hands and verbal content (i.e. n-grams). Specifically, for the face we used the manually annotated features (provided by the dataset) based on coarse facial movements (MUMIN coding scheme), automatically extracted handcrafted features based on fine facial movements (facial action units) and automatically extracted *learned* features based on deep neural networks encodings. For head and hands we used the provided manually annotated features (MUMIN coding scheme), while for verbal content we used n-grams. Different from the standard practice in deception detection of feature concatenation, we also employed a learning technique (multi-view learning) which combines together different types of features in a more effective way. Multi-view learning led to higher classification accuracy than state-of-the-art, proving that the method was able to better exploit the statistical characteristics of the different types of features. Most importantly, the computed features weights showed that, indeed, face-based features (especially the learned ones) were the most contributing to the final classification results, followed by n-grams. Together with other studies, this confirms the importance of face in detecting deception.

Limitations and Future Work. This importance, however, is relative to the chosen dataset. Even though its authors [Pérez-Rosas et al., 2015a] did a lengthy and laborious job, manually collecting, validating and annotating many video clips from the web, the dataset presents some drawbacks. The first one is the number of samples and subjects: there are only 121 videos in total, the number of different subjects is limited and not everyone expresses both a truthful and a deceptive statement, making it not statistically significant and unbalanced. Second, the quality and resolution of the images is often very low, limiting the visual information available and the features that can be extracted. The reason we chose this dataset is because, to the best of our knowledge, it was (and currently is) the only *publicly available* collection of *spontaneous* (not *faked*) deceptive statements recorded in a *high stakes*, *non controlled* scenario. Other datasets exist, but either they are not public [Radlak et al., 2015] or they present “acted” deceptive behavior in controlled environments, such as mock theft interviews [Derrick et al., 2010]. Future work could focus on creating a bigger and more diverse dataset, but the task might be difficult or even not feasible, if deception has to be real and not acted; privacy issues might also be a concern. With better video and audio recordings, audio analysis could also be performed.

From a methodological perspective, the most important limitation is the lack of behavior analysis using time. Classification is made on whole video clips, using features that encode and aggregate information from all frames, not single ones: MUMIN-based features encode the presence or absence of gestures in the entire video; action-units-based features are computed in a similar way and features based on deep neural networks encodings are an aggregation of features computed on all frames. This makes difficult to find correlations between deception, truth and specific features/behaviors. Additionally, in many single videos there are multiple statements and answers to different questions, which might contain both truths and lies; aggregating frame-wise features might, thus, lead to a wrong encoding of the information.

From an implementation point of view, searching for the best parameters is a time consuming process. The multi-view learning algorithm we chose becomes slow when dealing with big training sets and many features; thus, a single iteration of the grid search for the four major parameters (the kernel type and the three weights for the regularization terms) takes several hours to be completed. Future work could include an analysis of scalable (and novel) multi-view learning algorithms.

6.3 Body Pose and Group Detection

Finally, in Chapter 5 we moved to the whole body. We devised a method to extract the fully articulated body pose of a person immersed in a real-life social scenario, and used it to detect groups of people. Challenges included low-mid resolution images, difficult lighting conditions and body occlusions. We started by combining a robust multi-person 2D pose estimator [Cao et al., 2017] with a state-of-the-art “3D lifting” method [Tome et al., 2017]. The 2D pose estimator would produce incomplete poses due to the aforementioned challenges, leading to wrong 3D predictions. Thus, we devised a reconstruction algorithm (based on denoising autoencoders) that takes incomplete (or noisy) 2D poses (generated by [Cao et al., 2017]) as input and outputs reconstructed ones, showing improved estimation results. We then used the predicted 3D poses to compute the body orientation, which was then fed to f-formation detection algorithms, leading to results on par with the state-of-the-art. Contrary to existing algorithms which tackle the problem as a directional bin classification, our method is able to regress a real valued, fine grained orientation, which improves groups detection results. The pose reconstruction model is trained on the ground truth data of pose estimation datasets and takes as input only 2D joints locations, no RGB information. This makes it robust to domain changes and easy to *stack* on any newer and more robust 2D pose estimator.

Limitations and Future Work. The lack of visual data, however, represents also a drawback, especially in those cases where poses are severely incomplete. Although the autoencoder is able to output complete poses with a plausible (from a kinematic point of view) configuration, the reconstructed limbs might be in a very different position than the real one. Figure 6.1 shows an example: on the left, the actual pose has a raised left arm; the same arm is missing/occluded in the pose in the center; the pose on the right is the result of the reconstruction and has the left arm lowered. The predicted joints have a plausible configuration w.r.t. the rest of the body, but their position is far from the real one.

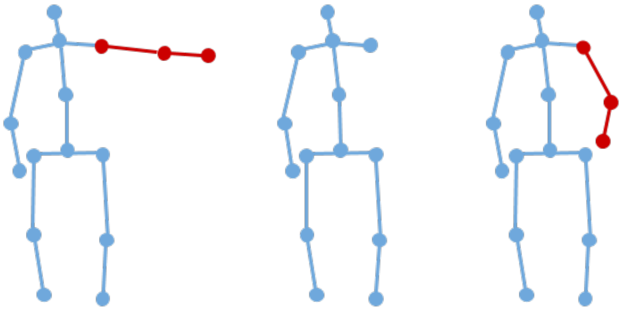

Figure 6.1: Example of reconstruction ambiguity. Left: actual pose with arm raised (in red). Center: pose with arm occluded/missing. Right: reconstructed pose with lowered arm (in red).

Poses with multiple missing limbs are very common in crowded environments and "wrong" predictions (such as the one in Figure 6.1) can lead to errors when classifying movements like gestures.

Having additional visual and contextual information might help the final reconstruction. Future work could include RGB data into the model. Richer input data would also aid the use of more complex architectures, e.g. CNNs, which might improve the final prediction over the simple fully connected autoencoders we used. The use of time (in the form of pose information coming from past frames) to strengthen the reconstruction results is another direction worth exploring.

Regarding the f-formation detection, results show that the fine grained predicted body orientation was effective. The 3D pose estimation algorithm proved to be able to generate plausible 3D poses with correct spatial orientation. In some cases, though, this orientation was wrong, with 3D poses facing opposite directions w.r.t. the people's ones. This is explained by how the algorithm works: the best transformation (scaling, rotation, translation, etc.) of a candidate 3D pose is found by minimizing the distance between its projection on the image plane and the input 2D pose; obviously, one 2D pose can correspond to the projections of multiple 3D poses, but if the 2D pose is asymmetrical, the 3D candidates are generally close to each other in terms of orientation and joints configuration. On the contrary, if the 2D pose is symmetrical (e.g. a person standing, facing the same direction of the camera), then 3D poses with different body orientations might generate the same projection. This problem is hard to solve, since it is intrinsic to bottom-up 3D pose estimation algorithms.

Finally, spatial location and orientation encode only geometrical properties of the group; this might not be enough for defining a group of interacting people in crowded and dynamic scenarios, where the r-, p- and o-spaces of the f-formation might be compressed, they might disappear or intersect with spaces belonging to other f-formations. Future work could enrich the current definitions of groups and model actual interactions between people as well as members roles, using features based on body movements and gestures.

Bibliography

[Abadi et al., 2016] Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., et al. (2016). Tensorflow: A system for large-scale machine learning.

[Akaho, 2006] Akaho, S. (2006). A kernel method for canonical correlation analysis. *arXiv preprint cs/0609071*.

[Akaike, 1998] Akaike, H. (1998). Information theory and an extension of the maximum likelihood principle. In *Selected Papers of Hirotugu Akaike*.

[Akhter and Black, 2015] Akhter, I. and Black, M. J. (2015). Pose-conditioned joint angle limits for 3d human pose reconstruction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1446–1455.

[Alameda-Pineda et al., 2016] Alameda-Pineda, X., Staiano, J., Subramanian, R., Batrinca, L., Ricci, E., Lepri, B., Lanz, O., and Sebe, N. (2016). Salsa: A novel dataset for multimodal group behavior analysis. *PAMI*.

[Alameda-Pineda et al., 2015] Alameda-Pineda, X., Yan, Y., Ricci, E., Lanz, O., and Sebe, N. (2015). Analyzing free-standing conversational groups: A multimodal approach. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 5–14. ACM.

[Albrecht, 2006] Albrecht, K. (2006). *Social intelligence: The new science of success.* John Wiley & Sons.

[Allwood et al., 2007] Allwood, J., Cerrato, L., Jokinen, K., Navarretta, C., and Paggio, P. (2007). The mumin coding scheme for the annotation of feedback, turn management and sequencing phenomena. *Language Resources and Evaluation*, 41(3-4):273–287.

[Ambady et al., 2000] Ambady, N., Bernieri, F., and Richeson, J. (2000). Toward a histology of social behavior: Judgmental accuracy from thin slices of the behavioral stream. *Advances in Experimental Social Psychology*, 32:201–257.

[Ambady and Rosenthal, 1992] Ambady, N. and Rosenthal, R. (1992). Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis. *Psychological bulletin*, 111(2):256.

[Andriluka et al., 2014] Andriluka, M., Pishchulin, L., Gehler, P., and Schiele, B. (2014). 2d human pose estimation: New benchmark and state of the art analysis. In *CVPR*.

[Aran and Gatica-Perez, 2010] Aran, O. and Gatica-Perez, D. (2010). Fusing audio-visual nonverbal cues to detect dominant people in small group conversations. In *ICPR*, pages 3687–3690.

[Ba and Odobez, 2006] Ba, S. O. and Odobez, J.-M. (2006). Recognizing people's focus of attention from head poses: a study. *IDIAP Research Report 06-42*, pages 1–27.

[Bales, 1980] Bales, R. (1980). *SYMLOG: case study kit with instructions for a group self study.* The Free Press, New York.

[Baltrušaitis et al., 2015] Baltrušaitis, T., Mahmoud, M., and Robinson, P. (2015). Cross-dataset learning and person-specific normalisation for automatic action unit detection. In *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on*, volume 6, pages 1–6. IEEE.

[Baltrušaitis et al., 2016] Baltrušaitis, T., Robinson, P., and Morency, L.-P. (2016). Openface: an open source facial behavior analysis toolkit. In *Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on*, pages 1–10. IEEE.

[Baltrušaitis et al., 2013] Baltrušaitis, T., Robinson, P., and Morency, L.-P. (2013). Constrained local neural fields for robust facial landmark detection in the wild. In *IEEE ICCVW 300 Faces in-the-Wild Challenge*, pages –.

[Bengio et al., 2007] Bengio, Y., Lamblin, P., Popovici, D., and Larochelle, H. (2007). Greedy layer-wise training of deep networks. In *NIPS*.

[Beyan et al., 2016] Beyan, C., Capozzi, F., Becchio, C., and Murino, V. (2016). Identification of emergent leaders in a meeting scenario using multiple kernel learning. pages 3–10. ACM ICMI-ASSP4MI.

[Beyan et al., 2017] Beyan, C., Capozzi, F., Becchio, C., and Murino, V. (2017). Prediction of the leadership style of an emergent leader using audio and visual nonverbal features. 20(2):441–456.

[Beyan and Fisher, 2015] Beyan, C. and Fisher, R. (2015). Classifying imbalanced data sets using similarity based hierarchical decomposition. *Pattern Recognition*, 48(5):1653–1672.

[Blum and Mitchell, 1998] Blum, A. and Mitchell, T. (1998). Combining labeled and unlabeled data with co-training. In *Proceedings of the eleventh annual conference on Computational learning theory*, pages 92–100. ACM.

[Brunet and Cowie, 2012] Brunet, P. M. and Cowie, R. (2012). Towards a conceptual framework of research on social signal processing. *Journal on Multimodal User Interfaces*, 6(3-4):101–115.

[Bulò and Pelillo, 2009] Bulò, S. R. and Pelillo, M. (2009). A game-theoretic approach to hypergraph clustering. In *Advances in neural information processing systems*, pages 1571–1579.

[Burgoon et al., 2017] Burgoon, J. K., Magnenat-Thalmann, N., Pantic, M., and Vinciarelli, A. (2017). *Social signal processing.* Cambridge University Press.

[Cao et al., 2017] Cao, Z., Simon, T., Wei, S.-E., and Sheikh, Y. (2017). Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*.

[Carletta et al., 2005] Carletta, J., Ashby, S., Bourban, S., Flynn, M., Guillemot, M., Hain, T., Kadlec, J., Karaiskos, V., Kraaij, W., and Kronenthal, M. (2005). The AMI meeting corpus: A pre-announcement. In *MLMI*, pages 28–39.

[Carney et al., 2005] Carney, D. R., Hall, J. A., and LeBeau, L. S. (2005). Beliefs about the nonverbal expression of social power. *Journal of Nonverbal Behavior*, 29(2):105–122.

[Chamveha et al., 2013] Chamveha, I., Sugano, Y., Sugimura, D., Siriteerakul, T., Okabe, T., Sato, Y., and Sugimoto, A. (2013). Head direction estimation from low resolution images with scene adaptation. *Computer Vision and Image Understanding*, 117(10):1502–1511.

[Chaudhuri et al., 2009] Chaudhuri, K., Kakade, S. M., Livescu, K., and Sridharan, K. (2009). Multi-view clustering via canonical correlation analysis. In *Proceedings of the 26th annual international conference on machine learning*, pages 129–136. ACM.

[Chawla et al., 2002] Chawla, V., Bowyer, K., Hall, L., and Kegelmeyer, W. (2002). Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357.

[Chen and Odobez, 2012] Chen, C. and Odobez, J.-M. (2012). We are not contortionists: Coupled adaptive learning for head and body orientation estimation in surveillance video. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 1544–1551. IEEE.

[Chen et al., 2015] Chen, Y., Conroy, N. J., and Rubin, V. L. (2015). Misleading online content: Recognizing clickbait as false news. In *Proceedings of the 2015 ACM on Workshop on Multimodal Deception Detection*, pages 15–19. ACM.

[Chen et al., 2017] Chen, Y., Shen, C., Wei, X.-S., Liu, L., and Yang, J. (2017). Adversarial posenet: A structure-aware convolutional network for human pose estimation. In *CVPR*.

[Coulson, 2004] Coulson, M. (2004). Attributing emotion to static body postures: Recognition accuracy, confusions, and viewpoint dependence. *Journal of nonverbal behavior*, 28(2):117–139.

[Cristani et al., 2011] Cristani, M., Bazzani, L., Paggetti, G., Fossati, A., Tosato, D., Del Bue, A., Menegaz, G., and Murino, V. (2011). Social interaction discovery by statistical analysis of f-formations.

[Cristinacce and T.F.Cootes, 2006] Cristinacce, D. and T.F.Cootes (2006). Feature detection and tracking with constrained local models. In *BMVC*, pages 929–938.

[Dasgupta et al., 2002] Dasgupta, S., Littman, M. L., and McAllester, D. A. (2002). Pac generalization bounds for co-training. In *Advances in neural information processing systems*, pages 375–382.

[DePaulo et al., 2003] DePaulo, B. M., Lindsay, J. J., Malone, B. E., Muhlenbruck, L., Charlton, K., and Cooper, H. (2003). Cues to deception. *Psychological bulletin*, 129(1):74.

[Derrick et al., 2010] Derrick, D. C., Elkins, A. C., Burgoon, J. K., Nunamaker Jr, J. F., and Zeng, D. D. (2010). Border security credibility assessments via heterogeneous sensor fusion. *IEEE Intelligent Systems*, (3):41–49.

[Ebbinghaus, 2013] Ebbinghaus, H. (2013). Memory: A contribution to experimental psychology. *Annals of neurosciences*, 20(4):155.

[Ekman and Friesen, 1978] Ekman, P. and Friesen, W. (1978). Facial action coding system: A technique for the measurement of facial movement. *Consulting Psychologists Press*.

[Ekman and Friesen, 1969] Ekman, P. and Friesen, W. V. (1969). Nonverbal leakage and clues to deception. *Psychiatry*, 32(1):88–106.

[Ekman and Friesen, 1976] Ekman, P. and Friesen, W. V. (1976). Measuring facial movement. *Environmental psychology and nonverbal behavior*, 1(1):56–75.

[Fang et al., 2017] Fang, H.-S., Xie, S., Tai, Y.-W., and Lu, C. (2017). Rmpe: Regional multi-person pose estimation. In *CVPR*.

[Feese et al., 2011] Feese, S., Muaremi, A., Arnrich, B., Troster, G., Meyer, B., and Jonas, K. (2011). Discriminating individually considerate and authoritarian leaders by speech activity cues. In *IEEE PASSAT, and IEEE SocialCom*, pages 1460–1465.

[Feng et al., 2017] Feng, W., Kannan, A., Gkioxari, G., and Zitnick, C. L. (2017). Learn2smile: Learning non-verbal interaction through observation. In *IROS*.

[Freud, 1959] Freud, S. (1959). Collected papers.(5 vols.).

[Fumera and Roli, 2002] Fumera, G. and Roli, F. (2002). Cost-sensitive learning in support vector machines. In *the Workshop Mach. Learn. Meth. Appl.*, pages –.

[Gamer, 2014] Gamer, M. (2014). Mind reading using neuroimaging: Is this the future of deception detection? *European Psychologist*, 19(3):172.

[Gonen and Alpaydin, 2008] Gonen, M. and Alpaydin, E. (2008). Localized multiple kernel learning. In *ICML*, pages 352–359.

[Gonen and Alpaydin, 2011] Gonen, M. and Alpaydin, E. (2011). Multiple kernel learning algorithms. *Journal of Machine Learning Research*, 12:2211–2268.

[Grahe and Bernieri, 1999] Grahe, J. E. and Bernieri, F. J. (1999). The importance of nonverbal cues in judging rapport. *Journal of Nonverbal behavior*, 23(4):253–269.

[Gross et al., 2007] Gross, M., Crane, E., and Fredrickson, B. (2007). Effect of felt and recognized emotions on body movements during walking. In *Proceedings of the International Conference on the Expression of Emotions in Health and Disease*, pages 615–625.

[Hall et al., 2005] Hall, J. A., LeBeau, L. S., and Coats, E. J. (2005). Nonverbal behavior and the vertical dimension of social relations: A meta-analysis. *Psychological Bulletin*, 131(6):898–924.

[Hansen and Ji, 2010] Hansen, D. and Ji, Q. (2010). In the eye of the beholder: a survey of models for eyes and gaze. *IEEE Trans. Pattern Anal. Mach. Intell.*, 32(3):478–500.

[Hare et al., 1998] Hare, A., Polley, R., and Stone, P. (1998). *The Symlog Practitioner: Applications of Small Group Research.* Praeger Press, New York.

[He et al., 2016] He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

[Hou et al., 2017] Hou, X., Shen, L., Sun, K., and Qiu, G. (2017). Deep feature consistent variational autoencoder. In *WACV*.

[Hung et al., 2011] Hung, H., Huang, Y., Friedland, G., and Gatica-Perez, D. (2011). Estimating dominance in multi-party meetings using speaker diarization. *IEEE Trans. Audio, Speech, Language Process*, 19(4):847–860.

[Hung et al., 2008] Hung, H., Jayagopi, D. B., Ba, S., Odobez, J.-M., and Gatica-Perez, D. (2008). Investigating automatic dominance estimation in groups from visual attention and speaking activity. In *ICMI*, pages 233–236.

[Hung and Kröse, 2011] Hung, H. and Kröse, B. (2011). Detecting f-formations as dominant sets. In *Proceedings of the 13th international conference on multimodal interfaces*, pages 231–238. ACM.

[Insafutdinov et al., 2016] Insafutdinov, E., Pishchulin, L., Andres, B., Andriluka, M., and Schiele, B. (2016). Deepercut: A deeper, stronger, and faster multi-person pose estimation model. In *ECCV*.

[Jayagopi et al., 2009] Jayagopi, D., Hung, H., Yeo, C., and Gatica-Perez, D. (2009). Modeling dominance in group conversations from nonverbal activity cues. *IEEE Trans. Audio, Speech, Language Process., Sp. Issue on Multimodal Processing for Speech-based Interactions*, 17(3):501–513.

[Jensen et al., 2010] Jensen, M. L., Meservy, T. O., Burgoon, J. K., and Nunamaker, J. F. (2010). Automatic, multimodal evaluation of human interaction. *Group Decision and Negotiation*, 19(4):367–389.

[Johnson and Johnson, 1991] Johnson, D. and Johnson, F. (1991). *Joining together: Group theory and group skills.* Prentice-Hall, Inc.

[Jovanovic et al., 2006] Jovanovic, N., op den Akker, R., and Nijholt, A. (2006). Addressee identification in face-to-face meetings. In *EACL*, pages 169–176.

[Kendon, 1990] Kendon, A. (1990). *Conducting interaction: Patterns of behavior in focused encounters*, volume 7. CUP Archive.

[Kincaid et al., 1975] Kincaid, J. P., Fishburne Jr, R. P., Rogers, R. L., and Chissom, B. S. (1975). Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical report, DTIC Document.

[Kindiroglu et al., 2014] Kindiroglu, A., Akarun, L., and Aran, O. (2014). Vision based personality analysis using transfer learning methods. In *IEEE SIU*, pages 2058–2061.

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

[Kingma and Welling, 2013] Kingma, D. P. and Welling, M. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

[Koenigs, 1999] Koenigs, R. (1999). *SYMLOG reliability and validity*. San Diego: SYMLOG Consulting Group.

[Krizhevsky et al., 2012] Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105.

[Ladickỳ et al., 2013] Ladickỳ, L., Russell, C., Kohli, P., and Torr, P. H. (2013). Inference methods for crfs with co-occurrence statistics. *International journal of computer vision*, 103(2):213–225.

[Lanckriet et al., 2004] Lanckriet, G. R., Cristianini, N., Bartlett, P., Ghaoui, L. E., and Jordan, M. I. (2004). Learning the kernel matrix with semidefinite programming. *Journal of Machine learning research*, 5(Jan):27–72.

[Larson et al., 1932] Larson, J. A., Haney, G. W., and Keeler, L. (1932). *Lying and its detection: A study of deception and deception tests*. University of Chicago Press Chicago, IL.

[Lin et al., 2014] Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In *ECCV*.

[Lord et al., 1984] Lord, R., Foti, R., and Vader, C. D. (1984). A test of leadership categorization theory: Internal structure, information processing, and leadership perceptions. *Organizational behavior and human performance*, 34(3):343–378.

[Lu, 2010] Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics*, 15(4):474–496.

[Luo et al., 2013] Luo, Y., Tao, D., Xu, C., Li, D., and Xu, C. (2013). Vector-valued multi-view semi-supervsed learning for multi-label image classification.

[Lykken, 1985] Lykken, D. T. (1985). The probity of the polygraph. *The psychology of evidence and trial procedure*, pages 95–123.

[Marin-Jimenez et al., 2011] Marin-Jimenez, M., Zisserman, A., and Ferrari, V. (2011). Here's looking at you, kid. detecting people looking at each other in videos. In *BMVC*, pages –.

[McArthur and Baron, 1983] McArthur, L. Z. and Baron, R. M. (1983). Toward an ecological theory of social perception. *Psychological review*, 90(3):215.

[McNeill, 1992] McNeill, D. (1992). *Hand and mind: What gestures reveal about thought*. University of Chicago press.

[Mehu and Scherer, 2012] Mehu, M. and Scherer, K. R. (2012). A psycho-ethological approach to social signal processing. *Cognitive processing*, 13(2):397–414.

[Mimansa et al., 2016] Mimansa, J., Tabibu, S., and Bajpai, R. (2016). The truth and nothing but the truth: Multimodal analysis for deception detection. In *IEEE 16th International Conference on Data Mining (ICDM) Workshops*, pages 938–943. IEEE.

[Minh et al., 2013] Minh, H. Q., Bazzani, L., and Murino, V. (2013). A unifying framework for vector-valued manifold regularization and multi-view learning. In *ICML (2)*, pages 100–108.

[Minh et al., 2016] Minh, H. Q., Bazzani, L., and Murino, V. (2016). A unifying framework in vector-valued reproducing kernel hilbert spaces for manifold regularization and co-regularized multi-view learning. *Journal of Machine Learning Research*, 17(25):1–72.

[Morris, 2002] Morris, D. (2002). *Peoplewatching*. Random House.

[Newell et al., 2016] Newell, A., Yang, K., and Deng, J. (2016). Stacked hourglass networks for human pose estimation. In *ECCV*.

[Otsu, 1979] Otsu, N. (1979). A threshold selection method from gray-level histograms. *IEEE Trans. Syst., Man, Cybern., Syst.*, 9(1):62–66.

[Ott et al., 2011] Ott, M., Choi, Y., Cardie, C., and Hancock, J. T. (2011). Finding deceptive opinion spam by any stretch of the imagination. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 309–319. Association for Computational Linguistics.

[Pantic et al., 2011] Pantic, M., Cowie, R., DErrico, F., Heylen, D., Mehu, M., Pelachaud, C., Poggi, I., Schroeder, M., and Vinciarelli, A. (2011). Social signal processing: the research agenda. In *Visual analysis of humans*, pages 511–538. Springer.

[Pantic and Vinciarelli, 2009] Pantic, M. and Vinciarelli, A. (2009). Implicit human-centered tagging [social sciences]. *IEEE Signal Processing Magazine*, 26(6).

[Parkhi et al., 2015] Parkhi, O. M., Vedaldi, A., and Zisserman, A. (2015). Deep face recognition. In *British Machine Vision Conference*.

[Pennebaker et al., 2001] Pennebaker, J. W., Francis, M. E., and Booth, R. J. (2001). Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, 71(2001):2001.

[Pentland, 2004] Pentland, A. (2004). Social dynamics: Signals and behavior. In *Proceedings of the third international conference on developmental learning (ICDL04). Salk Institute, San Diego. UCSD Institute for Neural Computation*, pages 263–267.

[Pentland, 2007] Pentland, A. (2007). Social signal processing [exploratory dsp]. *IEEE Signal Processing Magazine*, 24(4):108–111.

[Pérez-Rosas et al., 2015a] Pérez-Rosas, V., Abouelenien, M., Mihalcea, R., and Burzo, M. (2015a). Deception detection using real-life trial data. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, pages 59–66. ACM.

[Pérez-Rosas et al., 2015b] Pérez-Rosas, V., Abouelenien, M., Mihalcea, R., Xiao, Y., Linton, C., and Burzo, M. (2015b). Verbal and nonverbal clues for real-life deception detection. In *EMNLP*, pages 2336–2346.

[Pérez-Rosas and Mihalcea, 2014] Pérez-Rosas, V. and Mihalcea, R. (2014). Cross-cultural deception detection. In *ACL (2)*, pages 440–445.

[Pérez-Rosas and Mihalcea, 2015] Pérez-Rosas, V. and Mihalcea, R. (2015). Experiments in open domain deception detection. In *EMNLP*, pages 1120–1125.

[Pfister et al., 2011] Pfister, T., Li, X., Zhao, G., and Pietikäinen, M. (2011). Recognising spontaneous facial micro-expressions. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 1449–1456. IEEE.

[Pirsiavash et al., 2011] Pirsiavash, H., Ramanan, D., and Fowlkes, C. C. (2011). Globally-optimal greedy algorithms for tracking a variable number of objects. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1201–1208. IEEE.

[Pishchulin et al., 2016] Pishchulin, L., Insafutdinov, E., Tang, S., Andres, B., Andriluka, M., Gehler, P. V., and Schiele, B. (2016). Deepcut: Joint subset partition and labeling for multi person pose estimation. In *CVPR*.

[Pollick et al., 2001] Pollick, F. E., Paterson, H. M., Bruderlin, A., and Sanford, A. J. (2001). Perceiving affect from arm movement. *Cognition*, 82(2):B51–B61.

[Radlak et al., 2015] Radlak, K., Bozek, M., and Smolka, B. (2015). Silesian deception database: Presentation and analysis. In *Proceedings of the 2015 ACM on Workshop on Multimodal Deception Detection*, pages 29–35. ACM.

[Ramírez et al., 2016] Ramírez, O. A. I., Varni, G., Andries, M., Chetouani, M., and Chatila, R. (2016). Modeling the dynamics of individual behaviors for group detection in crowds using low-level features. In *Robot and Human Interactive Communication (RO-MAN), 2016 25th IEEE International Symposium on*, pages 1104–1111. IEEE.

[Ricci et al., 2015] Ricci, E., Varadarajan, J., Subramanian, R., Rota Bulo, S., Ahuja, N., and Lanz, O. (2015). Uncovering interactions and interactors: Joint estimation of head, body orientation and f-formations from surveillance videos. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4660–4668.

[Richmond et al., 1991] Richmond, V. P., McCroskey, J. C., and Payne, S. K. (1991). *Nonverbal behavior in interpersonal relations*. Prentice Hall Englewood Cliffs, NJ.

[Rogez et al., 2017] Rogez, G., Weinzaepfel, P., and Schmid, C. (2017). Lcr-net: Localization-classification-regression for human pose. In *CVPR 2017-IEEE Conference on Computer Vision & Pattern Recognition*.

[Rumelhart et al., 1986] Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning representations by back-propagating errors. *nature*, 323(6088):533.

[Sanchez-Cortes, 2013] Sanchez-Cortes, D. (2013). Computational methods for audio-visual analysis of emergent leadership. *PhD Thesis, EPFL, Lausanne*, pages –.

[Sanchez-Cortes et al., 2012a] Sanchez-Cortes, D., Aran, O., Jayagopi, D. B., Mast, M. S., and Gatica-Perez., D. (2012a). Emergent leaders through looking and speaking: from audio-visual data to multimodal recognition. *Journal on Multimodal User Interfaces*, 7(1–2):39–53.

[Sanchez-Cortes et al., 2012b] Sanchez-Cortes, D., Aran, O., Mast, M. S., and Gatica-Perez, D. (2012b). A nonverbal behavior approach to identify emergent leaders in small groups. *IEEE Trans. On Multimedia*, 14(3):816–832.

[Scheflen, 1964] Scheflen, A. E. (1964). The significance of posture in communication systems. *Psychiatry*, 27(4):316–331.

[Scherer, 1982] Scherer, K. R. (1982). Methods of research on vocal communication: Paradigms and parameters. *Handbook of methods in nonverbal behavior research*, pages 136–198.

[Setti et al., 2013a] Setti, F., Hung, H., and Cristani, M. (2013a). Group detection in still images by f-formation modeling: A comparative study. In *Image Analysis for Multimedia Interactive Services (WIAMIS), 2013 14th International Workshop on*, pages 1–4. IEEE.

[Setti et al., 2013b] Setti, F., Lanz, O., Ferrario, R., Murino, V., and Cristani, M. (2013b). Multi-scale f-formation discovery for group detection. In *Image Processing (ICIP), 2013 20th IEEE International Conference on*, pages 3547–3551. IEEE.

[Setti et al., 2015] Setti, F., Russell, C., Bassetti, C., and Cristani, M. (2015). F-formation detection: Individuating free-standing conversational groups in images. *PloS one*, 10(5):e0123783.

[Simonyan and Zisserman, 2014] Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

[Soleymani and Pantic, 2012] Soleymani, M. and Pantic, M. (2012). Human-centered implicit tagging: Overview and perspectives. In *Systems, Man, and Cybernetics (SMC), 2012 IEEE International Conference on*, pages 3304–3309. IEEE.

[Stiefelhagen et al., 2002] Stiefelhagen, R., Yang, J., and Waibel, A. (2002). Modeling focus of attention for meeting indexing based on multiple cues. *IEEE Trans. on Neural Networks*, 13(4):928–938.

[Subramanian et al., 2010] Subramanian, R., Staiano, J., Kalimeri, K., Sebe, N., and Pianesi, F. (2010). Putting the pieces together: multimodal analysis of social attention in meetings. In *ACM Multimedia*, pages 25–29.

[Subramanian et al., 2015] Subramanian, R., Varadarajan, J., Ricci, E., Lanz, O., and Winkler, S. (2015). Jointly estimating interactions and head, body pose of interactors from distant social scenes. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 835–838. ACM.

[Sun, 2011] Sun, S. (2011). Multi-view laplacian support vector machines. In *International Conference on Advanced Data Mining and Applications*, pages 209–222. Springer.

[Szegedy et al., 2015] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9.

[Tan and Hung, 2018] Tan, S. and Hung, H. (2018). Improving temporal interpolation of head and body pose using gaussian process regression in a matrix completion setting. *arXiv preprint arXiv:1808.01837*.

[Thorndike, 1920] Thorndike, E. L. (1920). Intelligence and its uses. *Harper's magazine*.

[Tome et al., 2017] Tome, D., Russell, C., and Agapito, L. (2017). Lifting from the deep: Convolutional 3d pose estimation from a single image. *CVPR 2017 Proceedings*, pages 2500–2509.

[Toshev and Szegedy, 2014] Toshev, A. and Szegedy, C. (2014). Deeppose: Human pose estimation via deep neural networks. In *CVPR*.

[Vascon et al., 2014] Vascon, S., Mequanint, E. Z., Cristani, M., Hung, H., Pelillo, M., and Murino, V. (2014). A game-theoretic probabilistic approach for detecting conversational groups. In *Asian conference on computer vision*, pages 658–675. Springer.

[Vincent et al., 2008] Vincent, P., Larochelle, H., Bengio, Y., and Manzagol, P.-A. (2008). Extracting and composing robust features with denoising autoencoders. In *ICML*.

[Vincent et al., 2010] Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., and Manzagol, P.-A. (2010). Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of Machine Learning Research*, 11(Dec):3371–3408.

[Vinciarelli et al., 2009] Vinciarelli, A., Pantic, M., and Bourlard, H. (2009). Social signal processing: Survey of an emerging domain. *Image and vision computing*, 27(12):1743–1759.

[Wei et al., 2016] Wei, S.-E., Ramakrishna, V., Kanade, T., and Sheikh, Y. (2016). Convolutional pose machines. In *CVPR*.

[Xu et al., 2013] Xu, C., Tao, D., and Xu, C. (2013). A survey on multi-view learning. *arXiv preprint arXiv:1304.5634*.

[Yan et al., 2016] Yan, Y., Ricci, E., Subramanian, R., Liu, G., Lanz, O., and Sebe, N. (2016). A multi-task learning framework for head pose estimation under target motion. *IEEE transactions on pattern analysis and machine intelligence*, 38(6):1070–1083.

[Yap et al., 2014] Yap, B., Rani, K., Rahman, H., Fong, S., Khairudin, Z., and Abdullah, N. (2014). An application of oversampling, undersampling, bagging, and boosting in handing imbalanced datasets. *In DaEng, Lecture Notes in Electrical Engineering*, 285:13–22.

[Zhang and Hung, 2016] Zhang, L. and Hung, H. (2016). Beyond f-formations: Determining social involvement in free standing conversing groups from static images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1086–1095.

[Zhang and Hung, 2018] Zhang, L. and Hung, H. (2018). On social involvement in mingling scenarios: Detecting associates of f-formations in still images. *IEEE Transactions on Affective Computing*.

[Zhao et al., 2017] Zhao, R., Wang, Y., and Martinez, A. M. (2017). A simple, fast and highly-accurate algorithm to recover 3d shape from 2d landmarks on a single image. *IEEE transactions on pattern analysis and machine intelligence*.

[Zuckerman et al., 1981] Zuckerman, M., DePaulo, B. M., and Rosenthal, R. (1981). Verbal and nonverbal communication of deception. *Advances in experimental social psychology*, 14:1–59.