Multimodal Deception Detection

Mihai Burzo (mburzo@umich.edu)
Mohamed Abouelenien (zmohamed@umich.edu)
Veronica Perez-Rosas (vrncapr@umich.edu)
Rada Mihalcea (mihalcea@umich.edu)

October 2016

Abstract

Deception detection has received an increasing amount of attention due to the significant growth of digital media, as well as increased ethical and security concerns. Earlier approaches to deception detection were mainly focused on law enforcement applications and relied on polygraph tests, which had proven to falsely accuse the innocent and free the guilty in multiple cases. More recent work on deception has expanded to other applications, such as deception detection in social media, interviews, or deception in daily life. Moreover, recent research on deception detection has brought together scientists from fields as diverse as computational linguistics, speech processing, computer vision, psychology, and physiology, which makes this problem particularly appealing for multimodal processing.

This chapter will overview the state-of-the-art in multimodal deception detection, covering physiological (e.g., biosensors and thermal imaging), visual (e.g., facial expressions and gestures), speech (e.g., pitch and pause length), and linguistic modalities. We will describe the features that are typically extracted from each of these modalities, as well as means to combine these modalities into an overall system that can detect deception in multimodal content. We will cover methods that make use of lab recordings, as well as methods that rely on real-life data (e.g., recent work on multimodal deception detection from trial data). In general, a multimodal approach where features from different streams are integrated is found to lead to an improved performance as compared to the use of single modalities.

7.1 Introduction and Motivation

Deception is defined as an intentional attempt to mislead others [Depaulo et al. 2003]. Deceptive behavior ranges from simple harmless lies to major threats. Deception detection has been receiving increased attention from different research communities including computer vision, psychology, and language processing, as deception permeates on almost every human interaction and can have costly consequences. Additionally, there exists an international interest in detecting deceivers due to the alarming security incidents that occurred in the recent years. For example, airports are places where detecting deception is vital. Terrorists can deceive customs and borders interviewers and conceal essential information that could be life-threatening. Another example can be seen in the court of law. Thousands of trials occur daily where juries have to take serious decisions that can affect the lives of suspects and victims based not only on evidence, but on human judgment as well [Fornaciari and Poesio 2013a].

Applications such as security, business, and criminal investigation triggered research interest in different fields. Existing methodologies rely mainly on polygraph tests that extract physiological measurements such as heart rate, respiration rate, skin conductance, and skin temperature. This approach had proven to falsely accuse the innocent and free the guilty in multiple cases. Employing polygraph tests was shown to be unreliable in many cases as it requires decisions from human experts, which is subject to bias and error [Derksen 2012; Gannon et al. 2009]. Furthermore, with the appropriate training, suspects can easily fake innocence using specific countermeasures [Ganis et al. 2011], such as lying in the pre-test questions, muscle tensing or tongue biting.

As detecting deceit has expanded to other applications such as social media, interviews, online transactions, and deception in daily life, alternative approaches were proposed in order to improve the reliability of deception detection systems [Granhag and Hartwig 2008]. In particular, physiological, psychological, visual, linguistic, acoustic, and thermal modalities have been analyzed in order to detect discriminative features and clues to identify deceptive behavior [Feng et al. 2012; Hillman et al. 2012; Owayjan et al. 2012; Pfister and Pietikäinen 2012; Rajoub and Zwiggelaar 2014; Zhou et al. 2013].

Linguistic features were usually extracted from the language, words usage, and consistency of the statements made by a person[Howard and Kirchhübel 2011; Mihalcea and Strapparava 2009b; Vrij et al. 2010]. Visual clues of deception include facial emotions, expression intensity, hands and body movements, and microexpressions. These features were shown to be capable of discriminating between deceptive and truthful behavior [Ekman 2001; Owayjan et al. 2012]. The psychology of lying using non-verbal and verbal characteristics was analyzed to identify deception clues [Vrij 2001]. Deception was also detected by observing increased activity in the nervous system that were determined using physiological measurements, such as heart rate, blood pressure, skin conductance, and respiration rate. The physiological aspect of the human body was expanded in terms of the thermal variations that occurred in the faces and specifically in the periorbital areas as a person acted deceptively [Pavlidis et al. 2012; Shastri et al. 2012]. Acoustic features took into account the pitch and speaking rate among other measurements to specify whether certain features are associated with an act of deceit [Hirschberg et al. 2005].

Recently, multimodal analysis has gained a lot of attention due to their superior performance compared to the use of individual modalities [Abouelenien et al. 2015a, 2016a; Pérez-Rosas et al. 2015b].

Chapter V1-X presented an overview on multimodal approaches for affect recognition tasks. In the deception detection field, several multimodal approaches have been suggested to improve deception detection by integrating features from different modalities including thermal and visual data streams Abouelenien et al. [2014, 2015b, 2016b]. This integration created a more reliable system that is not susceptible to factors affecting sole modalities and

polygraph tests such as the fear of being caught in a lie, stress from daily responsibilities, and tiredness.

In order to be able to develop improved deception detection systems, deception data needs to be collected and evaluated. There are two ways to collect data, either using a lab-setting [Abouelenien et al. 2014] or using real-life data [Pérez-Rosas et al. 2015a,b]. Most of the work proposed for deception detection relies on crowdsourcing or artificial acted data. Only recently, techniques were proposed to detect deceit from real-life scenarios such as court trials and TV interviews.

These strategies have different strengths and weaknesses that need to be evaluated according to the research hypothesis. Observing deceptive behavior in natural settings allows for the collection of spontaneous and real-life responses, particularly during high stake scenarios. However, this type of data lacks the choice and availability of the modalities to be used and hence misses multiple features. On the other hand, simulated data allows for the use of multiple pre-determined modalities and scenarios, but instead has lower stakes and subjects are less motivated to elicit a deceptive response as compared to real-life situations.

In this chapter, we overview the state-of-the-art in multimodal deception detection and describe the features that are typically extracted from each of the aforementioned modalities, as well as means to combine these modalities into an overall system that can detect deception in multimodal content. Moreover, we cover methods that make use of lab recordings, as well as methods that rely on real-life data, and provide examples for both approaches. A terminology of the terms commonly used through the chapter can be found in Table 7.1.

7.2 Deception Detection with Individual Modalities

Multiple approaches have been explored targeting the identification of deceptive behavior. These approaches can be roughly divided into verbal and non-verbal [Henningsen et al. 2005] or into contact and non-contact approaches.

Earlier methodologies for detecting deceit, especially in law-enforcement, fell mostly under the contact-based approaches and were focused on polygraph tests, which use devices that measure responses from the nervous system [Vrij 2001]. In particular, techniques relying on the extraction of physiological and biological measurements such as skin conductance and heart rate fell under this category. With the limitations of the invasive contact-based methods, which included the need of physically attaching devices to the subject's body to measure a given response, and also require human interpretation, deception detection research shifted towards non-contact, non-invasive methods. Among others, non-contact approaches include the development of verbal and acoustic, psychological, and physiological, visual, and thermal techniques.

In the following sections we provide an overview of research work conducted from different research fields, using both verbal and non-verbal approaches, towards building automatic and reliable systems for deception detection.

> **Terminology**
>
> ***M*odality** A data stream that is generated from a single source. For example, facial features and gestures are extracted from the visual modality, while word counts and lexical features are extracted from the linguistic modality.
>
> ***M*ultimodal processing** A series of operations that are conducted on multiple modalities with the end goal of extracting meaningful features from each modality.
>
> ***M*ultimodal fusion** The process of integrating features from different modalities using different methodologies such as concatenating the features together or combining the results obtained from each modality into one.
>
> ***A*ccuracy** The degree to which a system is correct in predicting the labels of some instances, calculated as the total number of correctly classified instances divided by the total number of instances.
>
> ***L*eave-one-out cross validation** Cross validation is the process of dividing a dataset into batches where one batch is reserved for testing and all the other batches are used for training a system. Leave-one-out means each batch is formed of a single instance.
>
> ***B*iosensor** A device that uses a transducer and a biological element to collect physiological responses and convert them into an electrical signal.

Table 7.1   Terminology table with definitions for important terms used throughout this chapter.

7.2.1   Psychology

Initial explorations on deception detection were conducted in the psychology domain, where researchers examined lying and lie detection phenomena in search for behavioral cues to deception. These studies focused on the micro and macro analysis of verbal and non-verbal exchange between the deceiver and the lie detector [Zuckerman et al. 1981]. Psychology researchers posed questions related to deceiver's self-presentation such as, will their faces be prone to leakage by showing exaggerated or suppressed facial expressions?, will their voices be louder, slower or faster? Furthermore, which are the thoughts, feelings, or physiological processes that are more likely to occur when people are lying compared to when they are telling the truth? For instance, to what extent do liars show behaviors than indicate guilt and fear as compared to truth tellers? Are deceivers more fearful as the stakes becomes higher? The reader can find a more detailed discussion in [DePaulo 1992].

   To answer these questions, multiple approaches were explored focusing on four aspects: 1) Control, deceiver's attempted behavior control that might appear planned, rehearsed,

and lacking of spontaneity; 2) Arousal, indicators of deceiver's arousal responses such as pupil dilation, eye blinking, and speech disturbances; 3) Felt emotion, markers of deceiver's experience of negative or positive affect including grooming, scratching, anxiety, evasive responses, among others; and 4) Cognitive processing, indicators of cognitive load such as longer response latency, hesitation, and fewer illustrators. [Depaulo et al. 2003] presents an extensive analysis of psychology work conducted on deception detection exploring these factors and describe 158 cues to deception compiled from over 120 independent samples. Results show that there are indeed important differences among liars and true tellers. However, the extent of the conclusions was limited as most of the reviewed studies relied on behavioral observation from very specific and small groups.

Motivated by these findings and the increasing access to larger amounts of observational data, researchers from study fields such as computational linguistics, speech processing, computer vision, psychology, and physiology started exploring the identification of deceit from a data-driven perspective. Thus, allowing approaching the identification of deceit by automatic means.

For further reading, readers can refer to Chapter VII-1, which investigates multimodal behavioral and physiological signals as indicators of cognitive load. In particular, the chapter describes the integration of physiological features such as galvanic skin response and visual aspects such as eye-based features, which are shown to be robust measures of cognitive load.

7.2.2   Language

The identification of deceit in written content has been addressed in a large number of studies in the psychology and computational linguistics communities.

From the psychology perspective, several studies showed the relationship between people linguistic choices and deceptive behavior. Newman et al. [Newman et al. 2003] presented an examination of linguistic manifestations of falsehood in written stories. In this study, authors measured and tested several linguistic dimensions from a set of linguistic categories that were previously found correlated to deception, including self-references, negative emotion words, and markers of cognitive complexity [Depaulo et al. 2003]. Using a text analysis tool called Linguistic Inquiry and word count (LIWC) [Pennebaker and Francis 1999], a lexicon of words are grouped into semantic categories relevant to psychological processes, including thoughts, emotions, and motives, authors generated linguistic profiles of participants who were either lying or telling the truth about different topics in different contexts. Then, several regression models were built for each topic to test the discriminating power of the different linguistic categories over deceptive and truthful samples. Five scenarios were used in this study and each subject was asked to provide both a truthful and a deceptive response in four scenarios. They were equally divided into a deceptive and truthful groups for a fifth "Mock Crime" scenario, resulting in an overall balanced population with a baseline of 50%. Using this method authors were able to identify deception with an overall accuracy rate of 61%. Further analysis of word

usage provided evidence of linguistic differences between true-tellers and liars. In particular, liars were found to use fewer first person pronouns and more negative emotions words than true-tellers. On the other hand, liars seemed to use third-person references at higher rates.

Work on computational linguistics initially attempted to replicate the findings of psychological experimentation by applying computational approaches to distinguish between written samples of deceptive and truthful statements. [Mihalcea and Strapparava 2009a] proposed a data-driven method to build classifiers able to distinguish between deceptive and truthful essays covering three topics: opinions on abortion, opinions about death penalty, and feelings about a best friend. Data was collected via crowd sourcing and learning features consisted of counts of unique words (unigrams) present in each deceptive/truthful essay. Authors presented experiments using machine learning classifiers such as Support Vector Machines and Naive Bayes. Results showed a clear separation between truthful and deceptive texts regardless the topic being discussed. Further analysis identified salient words on deceptive text using the LIWC and Wordnet Affect [Strapparava and Valitutti 2004] dictionaries and reported similar findings to [Newman et al. 2003]. For instance, deceivers used more references to others i.e. third person pronouns, whereas true-tellers showed preference for words connected to the self, i.e. I, myself. A similar study is presented in [Feng et al. 2012], where authors focused on applying syntactic stylometry techniques to identify deception in text from essays and product reviews. Authors explore shallow and deep syntactic representations derived from Probabilistic Context Free Grammar (PCFG) parse trees, such as part of speech tags (POS), syntactic patterns encoded as production rules, as well as n-gram representations. Experimental results showed significant performance gain in deception detection when adding deep syntax information into the learning process.

Computational linguistic approaches have also covered the identification of deception on a variety of domains where computer mediated communication happens, including chats, forums, online dating websites, social networks e.g., Facebook, Twitter, as well as product reviews websites that are prone to have fake product reviews and spam content [Guadagno et al. 2012; Joinson and Dietz-Uhler 2002; Li et al. 2014; Ott et al. 2011; Toma and Hancock 2010; Warkentin et al. 2010].

In the product reviews domain, Ott et al. [Ott et al. 2011] addressed the identification of spam producers by analyzing linguistic patterns in deceptive reviews. Using a similar approach to [Mihalcea and Strapparava 2009a] i.e using n-grams and semantic features derived from the LIWC dictionary, authors built accurate machine learning classifiers that identified fake reviews with accuracies above the human baseline performance (which was found slightly better than chance). This study showed that the deception detection task can be accurately conducted on product reviews domain and that humans are generally poor deception detectors for this task. In addition, authors found that features derived from LIWC are not as effective for building deception detection models in the product review domain. In a following study, [Ott et al. 2013] presented an analysis of the sentiment associated to

deceitful reviews focusing particularly in those containing negative sentiment as it largely affects consumer purchase decisions.

Regarding studies that analyzed deception in online interaction, [Yu et al. 2015]analyzed the role of deception in online networks by detecting deceptive groups in a social elimination-game; [Toma and Hancock 2010] conducted linguistic analyses in online dating profiles and identified significant correlation between deceptive dating profiles, self-references, negations, and lower levels of words usage. Other works have targeted the identification of deceptive behavior during face-to-face interactions. A study focusing on deception aspects related to syntactic complexity in children speech is presented by [Yancheva and Rudzicz 2013], where authors examine the relation between speech syntactic complexity and children's age. Authors analyzed children's verbal responses in short interviews regarding an unambiguous minor transgression involving playing with a toy. Several linguistic features such as readability index of the verbal statements, sentence complexity based on T-unit analysis, and the use of passive constructions were evaluated to identify differences in the complexity of the language used by a child while either lying or telling the truth. Results showed a clear association between the complexity of deceptive speech and children's age.

There have been also a number of efforts on exploring the deception detection task in languages other than English. Almela et al. [Almela et al. 2012] approached the deception detection task in Spanish essays by using SVM classifiers and linguistic categories, obtained from the Spanish version of the LIWC dictionary. Fornaciari and Poesio [Fornaciari and Poesio 2013b] examined deception in Italian court cases. In this work, authors explore several strategies for identifying deceptive clues, such as utterance length, LIWC features, lemmas and part-of-speech patterns. [Pérez-Rosas and Mihalcea 2014] presented a study that examine cultural differences among deceptive and truthful essays written by English, Spanish, and Romanian speakers. Authors addressed the deception detection task by first building classifiers separately for each culture and then by conducting several experiments across cultures. Authors propose the use of automatic machine translation and the LIWC version in each language to build deception classifiers across-languages. Experimental results suggest important differences among cultures and also the feasibility of using semantic information as a cross-lingual bridge when deceptive data is not readily available for a given language. In addition, analyses on word usage showed interesting findings such as shared lying patterns among cultures including the use of negation, negative emotions, and references to others. Furthermore, true-tellers related patterns are also shared among cultures, where the most salient words were related to family, positive emotions, and positive feelings.

Overall, techniques used for deception detection frequently include n-grams, and word statistics such as sentence length, word type ratio and word diversity. The addition of syntactic information i.e. sentence grammatical structure, has also been found useful to identify linguistic patters associated to deception. Semantic information has been also a great source of

information about the deceiver's psychological processes. In this category, LIWC and Word-net Affect had been proved as valuable resources to analyze deceivers' word usage.

Finally, it is worth mentioning that learning resources for automatic deception detection are limited. Most of the research work on this area included a data collection step using either manual or crowd sourced means. However, there is an increasing number of research work that have directed their efforts to the construction of deception resources [Gokhman et al. 2012]. Some deception corpora publicly available include: a dataset on deceptive and truthful essays [Mihalcea and Strapparava 2009a], a fake hotel reviews dataset collected from trip advisor [Ott et al. 2011], a fake product review dataset collected using Mechanical Turk [Li et al. 2014]. In addition, there are a couple of deception datasets for languages other than English such as a German deception dataset of product reviews [Verhoeven and Daelemans 2014], a Spanish and Romanian essay datasets provided [Pérez-Rosas and Mihalcea 2014] covering opinions about different topics such as death penalty and abortion, and a Spanish essay dataset from [Almela et al. 2012] that includes topics such as homosexual adoption and bullfighting.

7.2.3 Vision

Vision is the most common way people can detect liars as deception occurs on daily basis in human interactions. Visual body language was explored in order to detect deceit. Spontaneous facial expressions and hand gestures were of special interest due to their usage to express people's emotions [Ekman 2001]. Using a machine learning approach, these features are used to train a classifier for automatic lie detection as well as multiple applications. More information on machine learning approaches can be found in Chapter IV-1.

Psychologists were interested in observing the expressions, movements, and emotions that occur spontaneously and the ones the subjects aim at hiding. Micro- and squelched-expressions where studied to specify whether they were associated with an act of deception [Ekman 2001]. Microexpressions are involuntary expressions that last for a short period of time while squelched-expressions last longer but are immediately changed into a different expression. The asymmetry, duration, and smoothness of these expressions were shown to vary as a person speaks deceptively [Ekman 2003]. A publicly available database of micro-expressions was published in [Pfister and Pietikäinen 2012]. A kernel-based method was integrated in a temporal interpolation framework in order to extract clues of lies from the microexpressions in the dataset. Furthermore, geometric-based dynamic templates were extracted from the video frames of the deception recordings to extract geometric measurements from microexpressions. Following this, multiple systems were developed to detect visual features, facial expressions, and emotions that could indicate deceptive behaviors [Bartlett et al. 2006; Pfister and Pietikäinen 2012]. An example of spontaneous expressions can be seen in Figure 7.1.

Figure 7.1   An example of spontaneous expressions with a truthful response (left) and a deceptive response (right).

In order to standardize the process, the Facial Action Coding System (FACS) [Ekman and Rosenberg 2005] was developed by psychologists and behavioral scientists. FACS provided taxonomy of facial features using muscle movements. Examples of these action units include inner brow raiser, nose wrinkle, lip raiser, cheek raiser, chin raiser, eye widen, and others[1]. Several attempts were made to code these gestures automatically for efficient detection of human behavior and emotions. For instance, a real-time automated system to recognize spontaneous facial expressions was introduced to detect attempts of deception using FACS can be found in [Ekman and Rosenberg 2005]. Another example and one of the most famous tools is the Computer Expression Recognition Toolbox (CERT) [Littlewort et al. 2011].

In addition to the action units, CERT provides twelve facial expressions such as yaw, pitch, roll, smile detector, anger, contempt, disgust, fear, joy, sad, surprise, and neutral. The software tool detects faces in each frame using Viola-Jones extension in a boosting framework followed by specifying the eyes corners, nose, and mouth corners and center. The algorithm determines the log-likelihood ratio of the presence of these regions in specific locations. Hence, the output of CERT consists of the distance to the hyperplane of a Support Vector Machine trained classifier for each action unit, which specifies the intensity of the facial actions. Using a combination of different action units, the global facial expressions are determined.

It was reported that automatically detecting these action units and expressions using CERT did not perform better than random guessing [Abouelenien et al. 2015b]. The performance was reported using a dataset that was collected in a lab-setting using several scenarios. However, using feature selection, it was reported that some of these features had potential of detecting deceit. The list consisted of eight action units and six expressions, which provided the highest accuracy of 63%. The list included brow lowering, chin raising, cheek raising, lip puckering, eye closure, distress brow, left turning AU 10, left AU 14, yaw, roll, contempt, disgust, sadness, and neutral. Additionally, with the integration with features from other modalities, the performance improved.

---

[1] http://www.cs.cmu.edu/ face/facs.htm

In order to detect visual features that more personalized to the subjects and their individual differences, templates from the subjects' video recordings were extracted to determine the neutral baseline. This is followed by comparing the deceptive and truthful responses to the neutral baseline to specify the differences, which achieved an accuracy exceeding 60% for measurements such as the blinking rates, head pose, and intensity of the facial expression [Tian et al. 2005].

Furthermore, using the visual modality, correlation between specific hand gestures and deception were detected [Caso et al. 2006]. A noticeable decrease in the frequency of gestures was observed when subjects narrated stories in a deceptive manner compared to narrating the same stories truthfully [Cohen et al. 2010]. Additionally, individuals acting truthfully produced more rhythmic pulsing gestures while those acting deceptively made more frequent speech prompting gestures [Hillman et al. 2012].

7.2.4 Physiology

Physiological signals play a crucial role in monitoring human health as well as detecting changes in human behavior. Chapter IX-5 discusses the theoretical foundations of multimodal interfaces and systems in the health care domain, especially multimodal interaction, distributing multimodal processing, and multisensory-multimodal facilitation of health systems.

For lie detection, physiological measurements were traditionally collected from sensors that were placed on the human body such as blood volume pulse (BVP sensor), skin conductance (SC sensor), skin temperature (T sensor), and abdominal respiration (BR sensor). An example of these sensors can be seen in Figure 7.2. Biological measurements, such as brain waves detected by MRI scanners, were also utilized as an indicator of deception [Ganis et al. 2011; Kozel et al. 2004]. The idea was to observe the variations that occur in the measurements generated from these sensors as the subjects shifted from truthful to deceptive responses.

Figure 7.2   An example of the biosensors used to extract physiological measurements.

Relying on such techniques were shown to have several shortcomings such as falsely accusing innocent people of committing crimes and freeing guilty persons [Derksen 2012; Gannon et al. 2009; Maschke and Scalabrini 2005; Verschuere et al. 2009; Vrij 2001]. By using proper countermeasures the suspects could take control of their physiological signals

or manipulate the results. Improvements were made to the style of questions directed to the subjects to avoid potential errors associated with polygraphs by using Guilty Knowledge Test (GKT) compared to the widely used Control Question Test (CQT) [Taylor et al. 2010]. GKT is a multiple choice form of questions that aimed at detecting concealed knowledge that a suspect might be hiding. However, GKT still ran across multiple challenges that could manipulate its performance [Carmel et al. 2003]. In an attempt to develop more accurate methods to detect deceit, reaction time analysis in combination with event-related brain potentials using electroencephalogram were used to identify liars from a pool of 62 participants [Mohammadian et al. 2008].

Alternative methods to improve deception detection rates were explored using biological measurements, such as the functional magnetic resonance imaging (fMRI) technology [Kozel et al. 2004]. Using fMRI, specific brain activity such as an increased activity in the right anterior frontal cortices of the brain was observed in the case of well-rehearsed lies [Ganis et al. 2003]. However, the employment of such methodology in large-scale applications was unfeasible.

The physiological aspect of the human body was expanded in terms of the responses of the nervous system and the changes in the blood distribution, which could be detected using thermal imaging. The new approach targeted exploring alternatives to the limitations and invasiveness of the polygraph tests. Pavlidis et al. [Pavlidis et al. 2002] developed a high definition thermal imaging method to analyze facial thermal reactions associated with deceptive responses determined by the physiological signature of the faces. It was shown that as the nervous system reacted with an act of deceit, a peripheral change in the blood flow distribution was detected towards the musculoskeletal tissue [Pavlidis and Levine 2001, 2002]. Hence, bioheat transfer models that described the geometry and anatomy of large blood vessels in the facial area were developed to analyze their relation to deceit [Garbey et al. 2004].

Pavlidis and his collaborators noticed that the subjects exhibited elevated blood flow in the orbital muscle area resulting in elevated temperatures in certain local areas [Tsiamyrtzis et al. 2007]. They reported an overall accuracy exceeding 80%. The system was compared with the traditional polygraph test designed and implemented by the Department of Defense Polygraph Institute, and was found to achieve equivalent result.

With further analysis, distinct non-overlapping facial thermal patterns were detected with an increase in the blood flow around the eyes when subjects acted deceptively. Hence, thermodynamic modeling was applied to transform the raw thermal data from the periorbital area in the face to blood flow rates that had the potential of indicating deceit. Figure 7.3 demonstrates the regions of interest found to be most indicative of deceit.

Further experiments were conducted to improve the detection accuracy achieved using thermal imaging. Tandem tracking and noise suppression methods were used to extract thermal features from the periorbital area without applying restrictions on the face movements of the subjects in order to improve deception detection rates [Tsiamyrtzis et al. 2005].
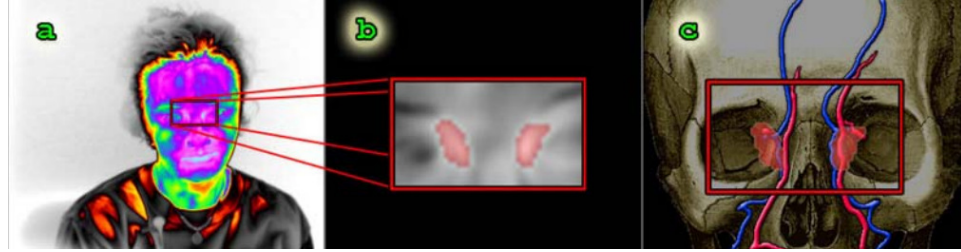
Figure 7.3 (a) a facial frame of the subject (b) the periorbital area found to be the most indicative of deceit with the 10% hottest pixels highlighted in pink and (c) the region of interest superimposed on the facial and ophthalmic arteriovenous complex. Image is provided by [Tsiamyrtzis et al. 2007]

Landmark detection systems were introduced to track landmarks on the regions of interest in the facial areas to track subjects as they lie [Jain et al. 2012].

Interestingly, a lie detection system was experimented in an airport using a set of 51 travelers by extracting thermal features such as the maximum, minimum, and average temperatures [Warmelink et al. 2011]. The system achieved accuracy above 64%. However, trained custom interviewers were able to detect liars with an accuracy exceeding 70%.

Other facial areas were additionally investigated in order to determine their capability of indicating deceit. A system for automatic blush detection was developed while focusing on areas such as the cheeks to identify changes in the skin temperature [Harmer et al. 2010]. A potential importance of the forehead region in detecting lies was suggested due to the presence of multiple blood vessels in this particular area [Zhu et al. 2007, 2008]. A comparison between different thermal facial regions in the face illustrated that the forehead area provided features that achieved improved performance compared to other regions [Abouelenien et al. 2015b]. An example of segmenting the region of interest can be seen in Figure 7.4.
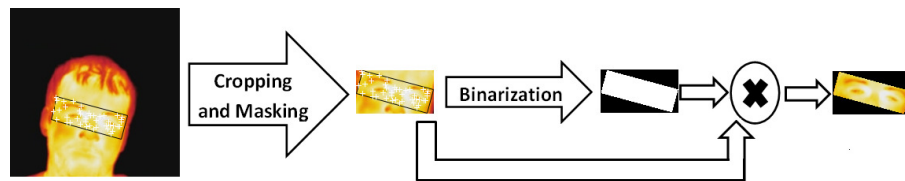

Figure 7.4 An overview of the region of interest segmentation process.

7.3 Deception Detection with Multiple Modalities

In search for more sophisticated lie detection systems, researchers explored multimodal approaches where features from several modalities are integrated. These approaches aim

to avoid the uncertainty associated with the use of single modalities, as well as the human efforts required for the analysis and decision-making processes used in earlier approaches. Additionally, the integration of features from different modalities enriches the dataset with information that is not available when these modalities are used separately, which can be reflected in the overall performance and the confidence level of the classifier.

For example, Henningsen et al. [Henningsen et al. 2005] examined the classification of deception cues into verbal and nonverbal, and how these cues influenced the perception of deception. [Burgoon et al. 2009] combined verbal and nonverbal features such as speech act profiling, feature extraction, and kinetic analysis for improved deception detection rates. [Jensen et al. 2010] extracted features from acoustic, verbal, and visual modalities following a multimodal approach. [Nunamaker et al. 2012] provided a review of approaches for evaluating human credibility using physiological, visual, acoustic, and linguistic features.

In the following section we provide an overview of research integrating multiple modalities in order to detect deceit. We also present some of the used datasets, extracted features, and evaluation results.

7.3.1   Thermal Imaging, Physiological Sensors, and Language Analysis

Recent work analyzed the combination of linguistic and thermal features [Nunamaker et al. 2012]. A novel approach that integrated features from the thermal, linguistic, and physiological modalities was presented in [Abouelenien et al. 2014] using data collected in a lab-setting environment. This research made two important contributions. First, a new dataset was collected with the participation of 30 subjects. The subjects were asked to discuss two different topics in both truthful and deceptive manners, while they were recorded using a microphone, a thermal camera, and several physiological sensors. Second, a multimodal system that integrated features extracted from three different modalities was developed in order to automate and improve the detection of deceptive behavior, avoid human efforts and the limitations associated with individual methods, and increase the efficiency of the decision making process. The research hypothesized that as a person acts/speaks deceptively, there will be subtle changes in his or her physiological and behavioral response, which can be detected using discriminant feature extraction.

7.3.1.1   Dataset and Devices

Measurements were acquired in a lab setting using a thermal camera FLIR Thermovision A40 with a resolution of 340x240 and a frame rate of 60 frames per second, as well as four biosensors including: blood volume pulse, skin conductance, skin temperature, and abdominal respiration sensors. Audiovisual recordings were also obtained using a Logitech web camera.

The scenarios that were used to elicit deceptive and truthful responses are as follows.

Abortion The subjects provided two separate statements, including a description of the subject's truthful opinion on abortion, and a deceptive description of the opposite opinion on abortion presented as if it was the subject's true opinion

Best Friend The subjects provided two separate statements including a true description of the subject's best friend, as well as a deceptive description about a person that the subject cannot stand described as if s/he were a best friend.

7.3.1.2 Multimodal Feature Extraction

The physiological features included assessments for temperature, heart rate, blood volume pulse, skin conductance, and respiration rate. Moreover, the features included a set of statistical descriptors of the raw measurements such as the maximum and minimum values, means, power means, standard deviations, and mean amplitudes (epochs).

The linguistic features included unigram counts which represented the frequency of occurrence of words in the transcripts of the responses in addition features derived from to the frequency counts of word classes in the Linguistic Inquiry and Word Count (LIWC) lexicon, which is a resource developed for psycholinguistic analysis and contains about 70 word classes relevant to psychological processes (e.g., emotion, cognition).

The thermal features were extracted by isolating the thermal facial areas in the video frames by employing image binarization techniques in addition to using relative measurements to locate the neck area and eliminate the back ground. Once the thermal faces were located in each frame, a thermal map was created by extracting the maximum, minimum, average, and standard deviation of the temperatures in addition to a histogram representing the temperature distribution in the faces.

7.3.1.3 Results

Feature-level fusion was used to integrate the features from individual modalities in order to train a decision tree classifier. A leave-one-out cross validation scheme was used and the average overall and per class accuracies were reported. This data distribution resulted in a baseline performance of 51.01% and 48.99% for the deceptive and truthful classes, respectively. Additionally, across-topic learning scheme was used, where the classifier was trained with features extracted from one topic while tested on the other.

Figure 7.5 illustrates the performance of the features extracted from both topics together for all modalities. The use of multimodal features further enhanced the classification accuracy. In particular, the integration of all three modalities together in addition to the integration of the thermal and linguistic features obtained higher accuracy in comparison to all other combinations as well as all individual modalities. Although the best performing single modalities were linguistic and physiological, the combination of thermal and linguistic modalities exceeded 70% for both classes and for the overall accuracy.
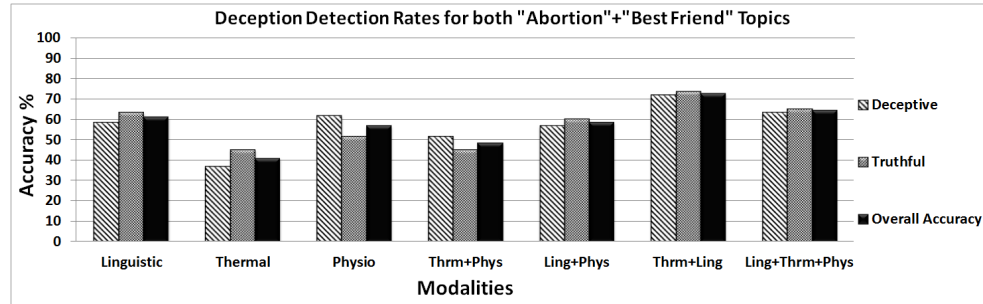
Figure 7.5   Deception, truthfulness, and overall accuracy percentages for individual and integrated modalities using features extracted from both the abortion and best friend topics.
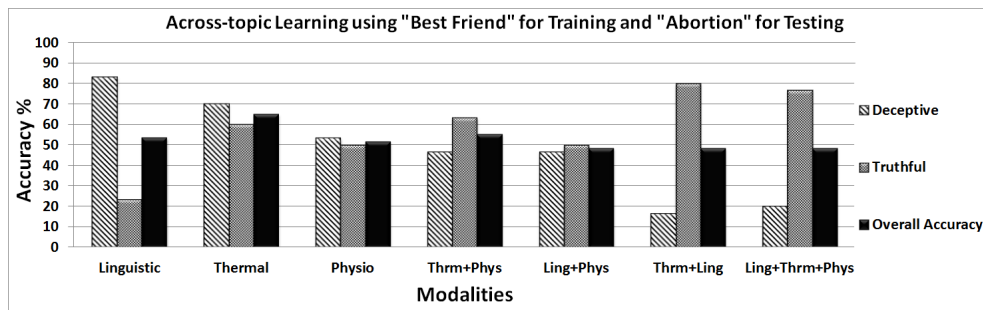

Figure 7.6   Deception, truthfulness, and overall accuracy percentages for individual and integrated modalities using across-topic learning. Best friend features are used for training and abortion features are used for testing.

Figures 7.6 and 7.7 illustrate the deceptive and truthful detection rates and the overall accuracy for the across-topic learning process using individual and combined modalities. In this learning scheme, the classifier was trained using features from one topic and then tested on the other topic. In both cases, it can be noticed that the linguistic modality created a large imbalance between the detection rate of the deception and truthfulness classes, which indicates the failure of the learning process. The disposition of the results can be explained with the dependency of the linguistic features on the corresponding topic.

Experimental results suggested that features extracted from linguistic and thermal modalities can potentially be good indicators of deceptive behaviors, which paves the way towards a completely automated, non-invasive deception detection process. Moreover, creating a multimodal classifier by integrating features from different modalities proved to be superior compared to learning from individual modalities. The experiments showed that the quality of the
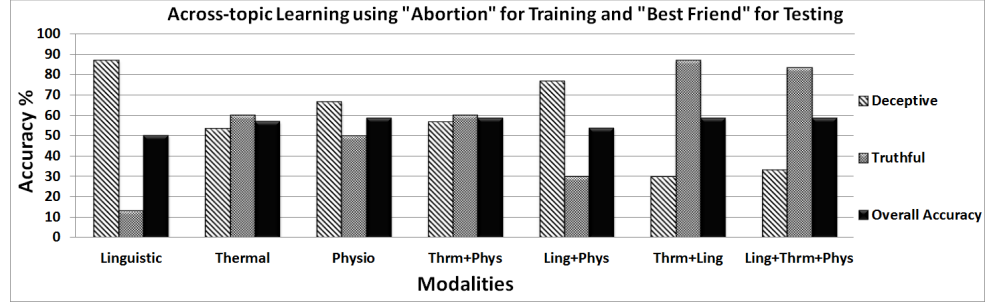
18

Figure 7.7   Deception, truthfulness, and overall accuracy percentages for individual and integrated modalities using across-topic learning. Abortion features are used for training and best friend features are used for testing.

extracted features is topic-dependent as the physiological and thermal features were topic-independent while the linguistic features were not.

This work was extended later in [Abouelenien et al. 2015b], where a "Mock Crime" scenario was added and different thermal regions in the face were tracked. In this scenario, a $20 bill was hidden in an envelope and the subjects were supposed to steal the money and deny it. This work reported that the forehead thermal features outperformed other facial features in its ability in detecting deceit.

7.3.2   Language and Acoustics

Psychology literature have found a significant correlation between deceptive behavior and speech attributes such as pitch, pitch accent, intonation, rhythm, and loudness [Depaulo et al. 2003; Zuckerman et al. 1981]. The speech community have addressed the identification of deceptive speech using machine learning approaches mainly by combining prosodic and cepstral speech features. Speech feature extraction is usually conducted at small intervals, also called audio frames, or globally by calculating representative statistics of the whole utterances. Most researchers use descriptive statistics such as mean, medians, standard deviations, and ranges of prosody features. Among them, fundamental frequency, pitch, energy, pauses, and formants are the most commonly used features. While initial efforts explored the use of lexical features derived from speech transcriptions or acoustic features extracted from the raw speech signals separately, more recent studies have addressed the relation between language and acoustics on the identification of deceptive behaviors.

Hirschberg et al. [Hirschberg et al. 2005] presented one of the first studies to explore the potential of combining prosodic and lexical cues on the identification of deceptive speech. Their experiments were conducted on a self-acquired dataset, the Columbia-SRI-Colorado corpus CSC, which consists of audio recorded interviews containing deceptive and truthful

responses from 32 speakers and comprises approximately 7 hours of speech. Authors built classifiers using the linguistic and speech modalities separately as well a combination of both modalities; the linguistic modality consisted of manual transcripts of spoken interviews. Experiments using individual modalities showed accuracies below the majority class baseline. However, results revealed accuracy improvement when combining the linguistic and acoustic modalities. This study showed that identifying deception in speech content is a very challenging task as speech shows a high degree of variation among individuals making difficult to develop speaker independent models. [Graciarena et al. 2006] reported additional experiments on the CSC corpus where authors explore the use of cepstral features to address aspects related to speaker variability on the deception detection task. In addition, their study also evaluated the use of automatic speech recognition as alternative to manual transcription. Results showed a reasonable trade-off in quality of deception classifiers build from transcripts obtained with noisy speech to text recognition (ASR). Following the same line of research, [Enos et al. 2007] explored the automatic identification specific interview segments, called critical segments, that are emotionally charged and cognitively loaded as a way to determine if a subject is telling the truth or lying. These events, also termed as hot spots by the psychology community, are particularly useful in the identification of lies as they indicate salient topics of the speaker's deception that are highly associated to deceptive statements. Authors annotated the CSC corpus with critical segments and then presented an automatic approach to identify their occurrence using lexical features, pauses, and vocal energy features. Results indicate that the inclusion of critical segments improves deception detection accuracy.

In addition, acoustics and language analysis has been also applied to explore cultural differences in deceptive behavior. A study on examining cultural differences in deceptive behavior among American and Chinese native speakers – all speaking English – is presented in [Levitan et al. 2015]. This study also introduces a deception dataset that includes personality, gender, and ethnicity information as well as confidence ratings on subjects' ability to deceive and to detect deception. Deceptive and truthful responses were elicited using the "fake resume" paradigm, where subjects provided true and false biographical information in a game setting in which they played the role of interviewer or interviewee. This dataset contains information from 139 subject pairs and comprised about 100 hours of speech. The ground truth was provided by the participants during each interview using key presses to indicate truth or lie labels. In this work, authors sought to distinguish between deceptive and non-deceptive behavior using features derived from the individual's speech, speaker's gender, ethnicity and personality factors. Acoustic features included F0, pitch, voice quality, speaking rate among others while personally factors included measurements derived from the Neuroticism Extraversion Openness Five Factor Inventory. Several machine learning experiments were conducted to evaluate these features on the identification of deceptive utterances. Research findings indicate that information about speaker's gender and their native language improves

the performance of acoustic models for deception detection and further suggests cultural differences during deceptive behavior.

Deception detection on audio content has also been addressed in competitive role-playing games (RPGs). [Chittaranjan and Hung 2010] created an audio-visual recordings of the "Are you a Werewolf?" game in order to detect deceptive behavior using non-verbal audio cues and to predict the subjects' decisions in the game. Authors were able to identity suspicious behavior based on players interactions measured through several game features such as speaking statistics, speaker's turns information, player interruption activity, and pitch analysis.

Overall, the inclusion of the acoustic channel into deception detection models is a promising research direction. However, current technologies for speech processing make challenging to process noisy data coming from natural scenarios, particularly those where the speech signal suffer from significant quality loss such as data coming from phone calls or multi-party conversations. Other challenges include noise introduced due to speech recognition errors. In addition, speaker's individual variability including gender, age, accent, voice tone, and cultural background requires building specific models that incorporate these dimensions into the analysis.

7.3.3 Vision and Language

More recently, the interest shifted towards detection of real-life deceptive behavior. The first reported multimodal deception detection approach in high stakes real-life data was presented in [Pérez-Rosas et al. 2015b]. This work introduced a novel dataset consisting of 121 deceptive and truthful video clips, from real court trials. The transcription of these videos was used to extract several linguistic features, and the videos were manually annotated for the presence of multiple gestures that were used to extract non-verbal features. Moreover, a system that jointly used the verbal and non-verbal modalities was developed to automatically detect the presence of deception. The performance of the system was compared to that of human annotators.

7.3.3.1 Dataset

The dataset consists of 121 videos including 61 deceptive and 60 truthful trial clips [2]. The average length of the videos in the dataset is 28.0 seconds. The data consists of 21 unique female and 35 unique male speakers, with their ages approximately ranging between 16 and 60 years. The video clips were labeled as deceptive or truthful based on guilty verdict, non-guilty verdict, and exoneration. Examples of famous trials included in the dataset are the trials of Jodi Arias, Donna Scrivo, Jamie Hood, Andrea Sneiderman, Mitchelle Blair, Amanda Hayes, Crystal Mangum, Marissa Devault, Carlos Miller, Michael Dunn, Bessman Okafor, Jonathan Santillan, among other trials.

---

[2] http://deceptiondetection.eecs.umich.edu/

7.3.3.2 Multimodal Feature Extraction

All the video clips were transcribed via crowd sourcing using Amazon Mechanical Turk. The final set of transcriptions consisted of 8,055 words, with an average of 66 words per transcript. The verbal features consisted of unigrams and bigrams derived from the bag-of-words representation of the video transcripts.

The gesture annotation was performed using the MUMIN coding scheme, which is a standard multimodal annotation scheme for interpersonal interactions [Allwood et al. 2007]. In the MUMIN scheme, facial displays include several different facial expressions associated with overall facial expressions, eyebrows, eyes and mouth movements, gaze direction, as well as head movements. In addition, the scheme includes a separate category for general face displays, which codes four facial expressions: smile, laughter, scowl, and other. Hand movements are also labeled in terms of handedness and trajectory. Using this coding scheme, binary feature vectors were created from annotations that indicate the presence or absence of each gesture in the video clips.

7.3.3.3 Results

The results were reported as statistical measurements and frequency counts of the gestures associated with both classes in addition to a machine learning approach to learn from both modalities.
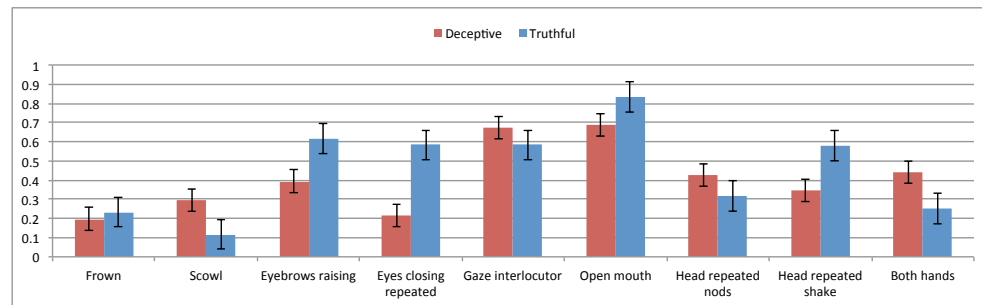

Figure 7.8   Distribution of non-verbal features for deceptive and truthful groups

Figure 7.8 shows the non-verbal features for which noticeable differences were observed in the two classes. Each bar pair shows the percentage distribution of the given gesture occurring during the deceptive and truthful conditions. For instance, it can be seen that eyebrow and eye gestures differentiated between the deceptive and truthful conditions as the non-overlapping error bars suggest statistically significant difference (P ¡ 0.05). In this figure, we can also observe that truth-tellers raised their eyebrows (Eyebrows raising), shook their head (Head repeated shake), and blinked (Eyes closing repeated) more frequently than deceivers. Interestingly, deceivers seemed to blink and shake their head less frequently than truth-tellers.

Feature Set	DT	RF
Unigrams	60.33%	56.19%
Bigrams	53.71%	51.20%
Facial displays	70.24%	76.03%
Hand gestures	61.98%	62.80%
Uni+Facial displays	66.94%	57.02%
All verbal	60.33%	50.41%
All non-verbal	68.59%	73.55%
All features	75.20%	50.41%

Table 7.2 Deception classifiers Decision Tree (DT) and Random Forest (RF) using individual and combined sets of verbal and non-verbal features.

Feature Set	DT
All	**75.20%**
– Hand gestures	71.90%
– Facial displays	59.50%
– Bigrams	66.94%
– Unigrams	61.98%

Table 7.3 Feature ablation study.

Deception classifiers were built using two classification algorithms: Decision Tree (DT) and Random Forest (RF) using leave-one-out cross-validation. The choice of these classifiers is based on their success and recommendation from previous work [Qin et al. 2004, 2005]. Moreover, a decision tree facilitates the visualization of the constructed tree model and determines the sequence and importance of the multimodal features at different tree levels. Table 7.2 shows the accuracy figures obtained by the two classifiers. As shown in this table, the combined classifier that learned from all the features (using Decision Tree) and the individual classifier that relied on the facial displays features (using Random Forest) achieved the best results. Comparing the integration of verbal features and visual features, the non-verbal features clearly outperformed the verbal features.

Table 7.3 shows the accuracies obtained when one feature group is removed and the deception classifier is built using the remaining features. Interestingly, the facial displays contributed the most to the classifier performance, followed by the unigram features.

Figure 7.9 shows the five most predictive features of the presence of deception were the presence of frowning (Frowning), eyebrows movement (Eyebrows raising), lip gestures (Lip
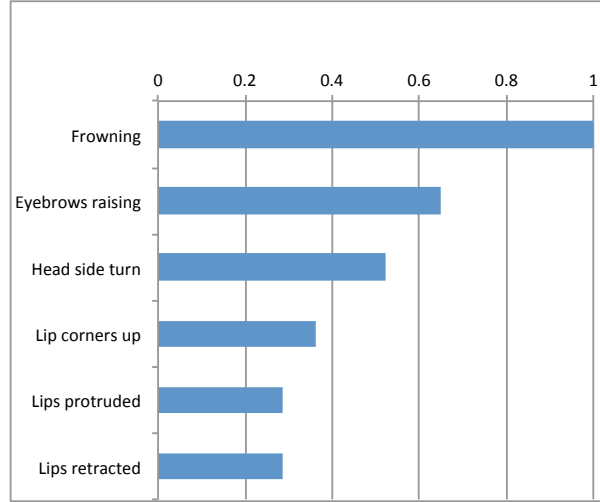
Figure 7.9 Weights of top non-verbal features in the multimodal deception classifier. The weights shown in this figure are normalized between zero and 1 to easily observe the contribution of each feature.

corners up, Lips protruded, Lips retracted), and head turns (Head side turn). These gestures were frequently portrayed by defendants and witnesses while being interrogated.

	Text	Audio	Silent video	Full video
A1	54.55%	51.24%	45.30%	56.20%
A2	47.93%	55.37%	46.28%	53.72%
A3	50.41%	59.50%	47.93%	59.50%
Sys	60.33%	NA	68.59%	75.20%

Table 7.4 Performance of three annotators (A1, A2, A3) and the developed automatic system (Sys) on the real-deception dataset over four modalities.

The proposed system was compared to the human ability to identify deceit on trial recordings when exposed to four different modalities: *Text*, consisting of the language transcripts; *Audio*, consisting of the audio track of the clip; *Silent video*, consisting of only the video with muted audio; and *Full video*, where audio and video are played simultaneously. The results, shown in Table 7.4, support the argument that human judges have difficulty performing the deception detection task [Ott et al. 2011]. Human detection of deception on silent video was more challenging than the rest of the modalities due to the lesser amount of deception cues available to the raters.

In summary, the analysis of non-verbal behaviors occurring in deceptive and truthful videos brought insight into the gestures that play a role in deception. Additional analysis showed the role played by the various feature sets used in the experiments. The proposed system achieved accuracies in the range of 60-75% and outperformed humans using different modalities with a relative percentage improvement of up to 51%. This showed that multimodal deception detection can provide valuable support for the trials decision making process.

7.4   The Way Forward

Based on the success of multimodal approaches in detecting deceit, improvements can be made to further achieve higher detection rates. For instance, improvements could be made in the multimodal data acquisition process, including the design of deceptive scenarios and data collection; in the selection of modalities to be extracted and their representation; or in the implementation of more efficient multimodal data fusion techniques.

Most of the developed deception datasets were in the range of 15 to 40 subjects. Larger datasets need to be collected in order to be able to detect reliable clues of deception as well as be able to generalize well to different real-life deception situations.

In a lab-setting environment where stakes are low or subjects are not motivated enough, the challenge is to develop creative scenarios other than the famous "Mock Crime" scenario in order to surprise the subjects and observe their initial reactions. This can be achieved by hiding the actual scenarios from the subjects before the recordings and surprising them with unexpected questions during the interviews. In real-life scenarios, there is a limit on the number of modalities used but no restrictions on the number of subjects. Efforts need to be exerted in order to collect larger datasets for deception detection. For instance, by taking advantage of publicly available data such as trials, 911 calls, police interrogations, political speeches, TV shows, and interviews.

For both lab-setting and real-life data, the cultural differences must be considered. Several cultural norms in a certain country could be easily considered suspicious behavior in another country. Hence, cross-cultural studies need to be conducted in order to identify such differences and develop a system that avoids bias and takes those differences into consideration.

The number of modalities used for feature extraction can further increase, which can result in a more reliable deception detection system. For instance, an integration of psychological, visual, physiological, linguistic, acoustic, and thermal modalities can reach the desired performance, especially in a lab-setting environment.

Finally, different techniques can be explored in order to enhance the quality of the extracted features. Temporal fusion for example can be used for this purpose. This type of fusion accounts for the temporal relationships between the modalities in the input datastream. One important research question when modeling the multimodal latent structure is the granularity of the input. Treating the deception data as a time series can also be used to determine the

relationships and dependencies between different features as well as modalities and specify the variations that occur within a certain window right before an act of deceit.

Furthermore, different classifiers and deep learning approaches can be used to detect deceit. For instance, deep learning uses multiple layers of linear and non-linear transformations in order to interpret different levels of abstractions in the data as can be seen in Chapter IV-4. In particular, Deep Neural Networks have shown success in detecting visual concepts in computer vision, which could add to the reliability of a multimodal deception detection system.

Acknowledgments

7.5 Supplementary Materials
Focus Questions

1. What is deception detection and why is it important?

2. What is meant by multimodal deception detection?

3. Which modalities can be used for deception detection?

4. What are the typical features that can be extracted from each modality to benefit the process of detecting deceit?

5. How can the multimodal features be integrated?

6. What are the advantages of using multimodal features compared to features from a single modality?

7. What is the difference between deception data collected in a lab and real-life deception data?

8. What are the advantages and limitations of processing multimodal lab-setting data and real-life deception data?

9. How can deception detection be improved in the future?

Bibliography

M. Abouelenien, V. Pérez-Rosas, R. Mihalcea, and M. Burzo. Deception detection using a multimodal approach. In *16th ACM International Conference on Multimodal Interaction, ICMI 2014*, November 2014.

M. Abouelenien, M. Burzo, and R. Mihalcea. Cascaded multimodal analysis of alertness related features for drivers safety applications. In *Proceedings of the 8th ACM International Conference on PErvasive Technologies Related to Assistive Environments*, PETRA '15, pages 59:1–59:8, New York, NY, USA, 2015a. ACM. ISBN 978-1-4503-3452-5. doi: 10.1145/2769493.2769505.

M. Abouelenien, R. Mihalcea, and M. Burzo. Trimodal analysis of deceptive behavior. In *Proceedings of the 2015 ACM on Workshop on Multimodal Deception Detection*, WMDD '15, pages 9–13, New York, NY, USA, 2015b. ACM. ISBN 978-1-4503-3987-2. doi: 10.1145/2823465.2823470.

M. Abouelenien, M. Burzo, and R. Mihalcea. Human acute stress detection via integration of physiological signals and thermal imaging. In *The 9th ACM International Conference on PErvasive Technologies Related to Assistive Environments, PETRA 2016*. ACM, June 2016a.

M. Abouelenien, R. Mihalcea, and M. Burzo. Analyzing thermal and visual clues of deception for a non-contact deception detection approach. In *The 9th ACM International Conference on PErvasive Technologies Related to Assistive Environments, PETRA 2016*. ACM, June 2016b.

J. Allwood, L. Cerrato, K. Jokinen, C. Navarretta, and P. Paggio. The mumin coding scheme for the annotation of feedback, turn management and sequencing phenomena. *Language Resources and Evaluation*, 41(3-4):273–287, 2007. ISSN 1574-020X. doi: 10.1007/s10579-007-9061-5.

A. Almela, R. Valencia-García, and P. Cantos. Seeing through deception: A computational approach to deceit detection in written communication. In *Proceedings of the Workshop on Computational Approaches to Deception Detection*, pages 15–22, Avignon, France, April 2012. Association for Computational Linguistics.

M. Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. Fasel, and J. Movellan. Automatic recognition of facial actions in spontaneous expressions. *Journal of Multimedia*, 1(6):22–35, 2006.

J. K. Burgoon, D. P. Twitchell, M. L. Jensen, T. O. Meservy, M. Adkins, J. Kruse, A. Deokar, G. Tsechpenakis, S. Lu, D. N. Metaxas, J. Nunamaker, J. F., and R. E. Younger. Detecting concealment of intent in transportation screening: A proof of concept. *IEEE Transactions on Intelligent Transportation Systems*, 10(1):103–112, March 2009. ISSN 1524-9050.

D. Carmel, E. Dayan, A. Naveh, O. Raveh, and G. Ben-Shakhar. Estimating the validity of the guilty knowledge test from simulated experiments: The external validity of mock crime studies. *Journal of experimental psychology: Applied*, 9(4):261–269, 2003.

L. Caso, F. Maricchiolo, M. Bonaiuto, A. Vrij, and S. Mann. The impact of deception and suspicion on different hand movements. *Journal of Nonverbal Behavior*, 30(1):1–19, 2006. ISSN 0191-5886.

G. Chittaranjan and H. Hung. Are you awerewolf? detecting deceptive roles and outcomes in a conversational role-playing game. In *2010 IEEE International Conference on Acoustics Speech and*

Signal Processing (ICASSP), pages 5334–5337, March 2010. doi: 10.1109/ICASSP.2010.5494961.

D. Cohen, G. Beattie, and H. Shovelton. Nonverbal indicators of deception: How iconic gestures reveal thoughts that cannot be suppressed. *Semiotica*, 2010(182):133–174, 2010.

B. M. DePaulo. Nonverbal behavior and self-presentation. *Psychological bulletin*, 111(2):203, 1992.

B. M. Depaulo, B. E. Malone, J. J. Lindsay, L. Muhlenbruck, K. Charlton, H. Cooper, B. M. Depaulo, B. E. Malone, D. O. Psychology, J. J. Lindsay, L. Muhlenbruck, and K. Charlton. Cues to deception. *Psychological Bulletin*, pages 74–118, 2003.

M. Derksen. Control and resistance in the psychology of lying. *Theory and Psychology*, 22(2):196–212, 2012.

P. Ekman. *Telling Lies: Clues to Deceit in the Marketplace, Politics and Marriage*. Norton, W.W. and Company, 2001.

P. Ekman. Darwin, deception, and facial expression. *Annals of the New York Academy of Sciences*, 1000 (EMOTIONS INSIDE OUT: 130 Years after Darwin's The Expression of the Emotions in Man and Animals):205–221, 2003.

P. Ekman and E. Rosenberg. *What the Face Reveals: Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System (FACS)*. Series in Affective Science. Oxford University Press, 2005. ISBN 9780199792726.

F. Enos, E. Shriberg, M. Graciarena, J. Hirschberg, and A. Stolcke. Detecting deception using critical segments. In *INTERSPEECH*, pages 2281–2284, 2007.

S. Feng, R. Banerjee, and Y. Choi. Syntactic stylometry for deception detection. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, ACL '12, pages 171–175, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics.

T. Fornaciari and M. Poesio. Automatic deception detection in italian court cases. *Artificial Intelligence and Law*, 21(3):303–340, 2013a. ISSN 0924-8463. doi: 10.1007/s10506-013-9140-4.

T. Fornaciari and M. Poesio. Automatic deception detection in italian court cases. *Artificial Intelligence and Law*, 21(3):303–340, 2013b. ISSN 0924-8463. doi: 10.1007/s10506-013-9140-4.

G. Ganis, S. M. Kosslyn, S. Stose, W. L. Thompson, and D. A. Yurgelun-Todd. Neural correlates of different types of deception: An fmri investigation. *Cerebral Cortex*, 13(8):830–836, August 2003. ISSN 1460-2199.

G. Ganis, J. P. Rosenfeld, J. Meixner, R. A. Kievit, and H. E. Schendan. Lying in the scanner: Covert countermeasures disrupt deception detection by functional magnetic resonance imaging. *NeuroImage*, 55(1):312 – 319, 2011. ISSN 1053-8119.

T. A. Gannon, A. R. Beech, and T. Ward. *Risk Assessment and the Polygraph*, pages 129–154. John Wiley and Sons Ltd, 2009. ISBN 9780470743232.

M. Garbey, A. Merla, and I. Pavlidis. Estimation of blood flow speed and vessel location from thermal video. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004*, volume 1, pages I–356–I–363 Vol.1, June 2004.

S. Gokhman, J. Hancock, P. Prabhu, M. Ott, and C. Cardie. In search of a gold standard in studies of deception. In *Proceedings of the Workshop on Computational Approaches to Deception Detection*, pages 23–30. Association for Computational Linguistics, 2012.

M. Graciarena, E. Shriberg, A. Stolcke, F. Enos, J. Hirschberg, and S. Kajarekar. Combining prosodic lexical and cepstral systems for deceptive speech detection. In *2006 IEEE International Conference on Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings.*, volume 1, pages I–I. IEEE, 2006.

P. A. Granhag and M. Hartwig. A new theoretical perspective on deception detection: On the psychology of instrumental mind-reading. *Psychology, Crime & Law*, 14(3):189–200, 2008.

R. Guadagno, B. Okdie, and S. Kruse. Dating deception: Gender, online dating, and exaggerated self-presentation. *Comput. Hum. Behav.*, 28(2):642–647, Mar. 2012. ISSN 0747-5632. doi: 10.1016/j.chb.2011.11.010.

K. Harmer, S. Yue, K. Guo, K. Adams, and A. Hunter. Automatic blush detection in "concealed information" test using visual stimuli. In *2010 International Conference of Soft Computing and Pattern Recognition (SoCPaR)*, pages 259–264, December 2010.

D. D. Henningsen, K. Valde, and E. Davies. Exploring the effect of verbal and nonverbal cues on perceptions of deception. *Communication Quarterly*, 53(3):359–375, 2005.

J. Hillman, A. Vrij, and S. Mann. Um . . . they were wearing . . . : The effect of deception on specific hand gestures. *Legal and Criminological Psychology*, 17(2):336–345, 2012. ISSN 2044-8333.

J. Hirschberg, S. Benus, J. Brenier, F. Enos, S. Friedman, S. Gilman, C. Gir, G. Graciarena, A. Kathol, and L. Michaelis. Distinguishing deceptive from non-deceptive speech. In *In Proceedings of Interspeech 2005 - Eurospeech*, pages 1833–1836, 2005.

D. Howard and C. Kirchhübel. Acoustic correlates of deceptive speech: an exploratory study. In *Proceedings of the 9th international conference on Engineering psychology and cognitive ergonomics*, EPCE'11, pages 28–37, Berlin, Heidelberg, 2011. Springer-Verlag. ISBN 978-3-642-21740-1.

U. Jain, B. Tan, and Q. Li. Concealed knowledge identification using facial thermal imaging. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1677–1680, March 2012.

M. L. Jensen, T. O. Meservy, J. K. Burgoon, and J. Nunamaker, Jay F. Automatic, multimodal evaluation of human interaction. *Group Decision and Negotiation*, 19(4):367–389, 2010. ISSN 0926-2644.

A. N. Joinson and B. Dietz-Uhler. Explanations for the perpetration of and reactions to deception in a virtual community. *Social Science Computer Review*, 20(3):275–289, 2002.

F. A. Kozel, L. J. Revell, J. P. Lorberbaum, A. Shastri, J. D. Elhai, M. D. Horner, A. Smith, Z. Nahas, D. E. Bohning, and M. S. George. A pilot study of functional magnetic resonance imaging brain correlates of deception in healthy young men. *The Journal of Neuropsychiatry and Clinical Neurosciences*, 16(3):295–305, August 2004. ISSN 1460-2199.

S. I. Levitan, G. An, M. Wang, G. Mendels, J. Hirschberg, M. Levine, and A. Rosenberg. Cross-cultural production and detection of deception from speech. In *Proceedings of the 2015 ACM on Workshop on Multimodal Deception Detection*, pages 1–8. ACM, 2015.

J. Li, M. Ott, C. Cardie, and E. Hovy. Towards a general rule for identifying deceptive opinion spam. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, Baltimore, Maryland, June 2014.

G. Littlewort, J. Whitehill, T. Wu, I. Fasel, M. Frank, J. Movellan, and M. Bartlett. The computer expression recognition toolbox (cert). In *2011 IEEE International Conference on Automatic Face Gesture Recognition and Workshops (FG 2011)*, pages 298–305, March 2011.

G. Maschke and G. Scalabrini. *The Lie Behind the Lie Detector*. antipolygraph.org, 2005.

R. Mihalcea and C. Strapparava. The lie detector: Explorations in the automatic recognition of deceptive language. In *Proceedings of the Association for Computational Linguistics (ACL 2009)*, Singapore, 2009a.

R. Mihalcea and C. Strapparava. The lie detector: Explorations in the automatic recognition of deceptive language. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 309–312, Suntec, Singapore, August 2009b. Association for Computational Linguistics.

A. Mohammadian, V. Abootalebi, M. Moradi, and M. Khalilzadeh. Multimodal detection of deception using fusion of reaction time and p300 component. In *Biomedical Engineering Conference, 2008. CIBEC 2008. Cairo International*, pages 1–4, Dec 2008.

M. Newman, J. Pennebaker, D. Berry, and J. Richards. Lying words: Predicting deception from linguistic styles. *Personality and Social Psychology Bulletin*, 29, 2003.

J. Nunamaker, J.F., J. Burgoon, N. Twyman, J. Proudfoot, R. Schuetzler, and J. Giboney. Establishing a foundation for automated human credibility screening. In *2012 IEEE International Conference on Intelligence and Security Informatics (ISI)*, pages 202–211, June 2012.

M. Ott, Y. Choi, C. Cardie, and J. Hancock. 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*, HLT '11, pages 309–319, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics. ISBN 978-1-932432-87-9.

M. Ott, C. Cardie, and J. T. Hancock. Negative deceptive opinion spam. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Short Papers*, Atlanta, Georgia, USA, June 2013. Association for Computational Linguistics.

M. Owayjan, A. Kashour, N. Al Haddad, M. Fadel, and G. Al Souki. The design and development of a lie detection system using facial micro-expressions. In *2012 2nd International Conference on Advances in Computational Tools for Engineering Applications (ACTEA)*, pages 33–38, Dec 2012.

I. Pavlidis and J. Levine. Monitoring of periorbital blood flow rate through thermal image analysis and its application to polygraph testing. In *Proceedings of the 23rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society.*, volume 3, pages 2826–2829, 2001.

I. Pavlidis and J. Levine. Thermal image analysis for polygraph testing. *IEEE Engineering in Medicine and Biology Magazine*, 21(6):56–64, Nov 2002. ISSN 0739-5175.

I. Pavlidis, N. L. Eberhardt, and J. A. Levine. Human behavior: Seeing through the face of deception. *Nature*, 415(6867):35–35, 2002.

I. Pavlidis, P. Tsiamyrtzis, D. Shastri, A. Wesley, Y. Zhou, P. Lindner, P. Buddharaju, R. Joseph, A. Mandapati, B. Dunkin, et al. Fast by nature-how stress patterns define human experience and performance in dexterous tasks. *Scientific Reports*, 2, 2012.

J. Pennebaker and M. Francis. Linguistic inquiry and word count: LIWC, 1999. Erlbaum Publishers.

V. Pérez-Rosas and R. Mihalcea. Cross-cultural deception detection. In *Proceedings of the Association for Computational Linguistics*, 2014.

V. Pérez-Rosas, M. Abouelenien, R. Mihalcea, Y. Xiao, C. Linton, and M. Burzo. Verbal and nonverbal clues for real-life deception detection. In *Proceedings of the 2015 Conference on Empirical Methods*

in Natural Language Processing, pages 2336–2346, Lisbon, Portugal, September 2015a. Association for Computational Linguistics.

V. Pérez-Rosas, M. Abouelenien, R. Mihalcea, Y. Xiao, C. Linton, and M. Burzo. Deception detection using real-life trial data. In *17th ACM International Conference on Multimodal Interaction, ICMI 2015*, November 2015b.

T. Pfister and M. Pietikäinen. Electronic imaging & signal processing automatic identification of facial clues to lies. *SPIE Newsroom*, January 2012.

T. Qin, J. Burgoon, and J. Nunamaker, J. F. An exploratory study on promising cues in deception detection and application of decision tree. In *System Sciences, 2004. Proceedings of the 37th Annual Hawaii International Conference on*, pages 23–32, January 2004.

T. Qin, J. K. Burgoon, J. P. Blair, and J. F. Nunamaker. Modality effects in deception detection and applications in automatic deception-detection. In *Proceedings of the 38th Hawaii International Conference on System Sciences*, 2005.

B. Rajoub and R. Zwiggelaar. Thermal facial analysis for deception detection. *IEEE Transactions on Information Forensics and Security*, PP(99):1–1, 2014. ISSN 1556-6013.

D. Shastri, M. Papadakis, P. Tsiamyrtzis, B. Bass, and I. Pavlidis. Perinasal imaging of physiological stress and its affective potential. *IEEE Transactions on Affective Computing*, 3(3):366–378, July 2012. ISSN 1949-3045. doi: 10.1109/T-AFFC.2012.13.

C. Strapparava and A. Valitutti. Wordnet-affect: an affective extension of wordnet. In *Proceedings of the 4th International Conference on Language Resources and Evaluation*, Lisbon, 2004.

M. K. Taylor, D. S. Horning, J. F. Chandler, J. B. Phillips, J. Y. Khosravi, J. E. Bennett, H. Halbert, B. J. Fern, and H. Gao. A comparison of approaches to detect deception. Technical Report ADA537848, Naval Aerospace Medical Research Laboratory, December 2010.

Y.-L. Tian, T. Kanade, and J. Cohn. Facial expression analysis. In *Handbook of Face Recognition*, pages 247–275. Springer New York, 2005. ISBN 978-0-387-40595-7.

C. Toma and J. Hancock. Reading between the lines: linguistic cues to deception in online dating profiles. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work*, CSCW '10, pages 5–8, New York, NY, USA, 2010. ACM. ISBN 978-1-60558-795-0. doi: 10.1145/1718918.1718921.

P. Tsiamyrtzis, J. Dowdall, D. Shastri, I. Pavlidis, M. Frank, and P. Eckman. Lie detection - recovery of the periorbital signal through tandem tracking and noise suppression in thermal facial video. In *SPIE Conference on Sensors and Command Control Communications and Intelligence Technologies for Homeland Security and Homeland Defense IV*, pages 555–566, 2005.

P. Tsiamyrtzis, J. Dowdall, D. Shastri, I. T. Pavlidis, M. G. Frank, and P. Ekman. Imaging facial physiology for the detection of deceit. *International Journal of Computer Vision*, 71(2):197–214, 2007. ISSN 0920-5691.

B. Verhoeven and W. Daelemans. Clips stylometry investigation (csi) corpus: A dutch corpus for the detection of age, gender, personality, sentiment and deception in text. In N. C. C. Chair), K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, and S. Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, may 2014. European Language Resources Association (ELRA). ISBN 978-2-9517408-8-4.

B. Verschuere, V. Prati, and J. De Houwer. Cheating the lie-detector: Faking the autobiographical iat. *Psychological Science*, 20:410–413, 2009.

A. Vrij. *Detecting Lies and Deceit: The Psychology of Lying and the Implications for Professional Practice*. Wiley series in the psychology of crime, policing and law. Wiley, 2001.

A. Vrij, P. Granhag, and S. Porter. Pitfalls and opportunities in nonverbal and verbal lie detection. *Psychological Science in the Public Interest*, 11(3):89–121, dec 2010. ISSN 1529-1006. doi: 10.1177/1529100610390861.

D. Warkentin, M. Woodworth, J. Hancock, and N. Cormier. Warrants and deception in computer mediated communication. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work*, pages 9–12. ACM, 2010.

L. Warmelink, A. Vrij, S. Mann, S. Leal, D. Forrester, and R. P. Fisher. Thermal imaging as a lie detection tool at airports. *Law and Human Behavior*, 35(1):40–48, 2011. ISSN 0147-7307.

M. Yancheva and F. Rudzicz. Automatic detection of deception in child-produced speech using syntactic complexity features. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 944–953, Sofia, Bulgaria, August 2013. Association for Computational Linguistics.

D. Yu, Y. Tyshchuk, H. Ji, and W. A. Wallace. Detecting deceptive groups using conversations and network analysis. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 1: Long Papers*, pages 857–866, 2015.

Y. Zhou, P. Tsiamyrtzis, P. Lindner, I. Timofeyev, and I. Pavlidis. Spatiotemporal smoothing as a basis for facial tissue tracking in thermal imaging. *IEEE Transactions on Biomedical Engineering*, 60(5): 1280–1289, May 2013. ISSN 0018-9294.

Z. Zhu, P. Tsiamyrtzis, and I. Pavlidis. Forehead thermal signature extraction in lie detection. In *29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 2007. EMBS 2007*, pages 243–246, August 2007.

Z. Zhu, P. Tsiamyrtzis, and I. Pavlidis. The segmentation of the supraorbital vessels in thermal imagery. In *IEEE Fifth International Conference on Advanced Video and Signal Based Surveillance, 2008. AVSS '08*, pages 237–244, September 2008.

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