13 Multimodal Deception Detection

Mihai Burzo, Mohamed Abouelenien,
Veronica Perez-Rosas, Rada Mihalcea

13.1 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. The detection of such behaviors has been receiving increased attention from different research communities, including computer vision, psychology, and language processing, as deception permeates 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 on 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]. Reports dating back three decades indicated that polygraph results were false one third of the time [Lykken 1984].

Multiple factors can affect the reliability of polygraphs such as the fear of being perceived as deceptive and the anxiety about being tested [Council 2003]. 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.

An existing problem with evaluating polygraph testing is that the quality of the data available for evaluation is relatively low [Council 2003]. However, an evaluation attempt was conducted using 59 datasets collected during different decades from 57 studies (52 laboratory, 7 field) including 3,681 polygraph examinations. The study reported a wide range of accuracy index values starting from approximately 0.5 to more than 0.95 for the 52 laboratory studies and from approximately 0.7 to 1 for the field studies. As a result of the unreliability of polygraph testing, the U.S. Supreme court acted to restrict their use in legal proceedings in 1998.

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 [Owayjan et al. 2012, Pfister and Pietikäinen 2012, Hillman et al. 2012, Zhou et al. 2013, Rajoub and Zwiggelaar 2014, Feng et al. 2012].

Linguistic features were usually extracted from the language, words usage, and consistency of the statements made by a person [Howard and Kirchhübel 2011, Vrij et al. 2010, Mihalcea and Strapparava 2009b]. 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 [Shastri et al. 2012, Pavlidis et al. 2012]. Acoustic features took into account the pitch and speaking rate, among other measurements, to specify whether or not 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 [Pérez-Rosas et al. 2015b, Abouelenien et al. 2016a, Abouelenien et al. 2015a].

Chapter 6 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, Abouelenien et al. 2015b, Abouelenien et al. 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, Pérez-Rosas et al. 2015b]. While earlier work relied on polygraph tests and manual human efforts, most of the work proposed for automatic deception detection relies on crowdsourcing or artificial acted data. Only recently, automated 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.

This chapter will overview the state-of-the-art in multimodal deception detection, covering physiological (e.g., ***physiological sensors*** 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).

A terminology of the terms commonly used through the chapter can be found in the Glossary

Glossary

Feature-level multimodal fusion. The process of integrating features from different modalities using diverse methodologies such as concatenating the features together (early fusion) or combining the models obtained from each modality at decision level (late fusion).

Leave-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.

Physiological sensor. A device that uses a transducer and a biological element to collect physiological responses, such as heart rate and skin conductance, and convert them into an electrical signal. The measures obtained with such devices provide quantitative feedback about physiological changes or processes experienced by research subjects.

T-unit analysis. The analysis of terminable units of language (T-unit), which is the smallest group of words that could be considered as a grammatical sentence, regardless of how is punctuated. T-unit analysis is used extensively to measure the overall complexity of both speech and writing samples and consists mainly on measuring different aspects of their syntactic construction in text such as mean length of the t-units, and number of clauses present in each unit, among others.

13.2 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, toward building automatic and reliable systems for deception detection.

13.2.1 Psychology

Initial explorations on deception detection were conducted in the psychology domain, where researchers examined lying and lie detection phenomena in search of behavioral cues to deception. These studies focused on the micro and macro analysis of verbal and non-verbal exchanges 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.
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 truth tellers.

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 10, 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.

13.2.2 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's linguistic choices and deceptive behavior. 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 truth-tellers and liars. In particular, liars were found to use fewer first-person pronouns and more negative emotions words than truth-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 of 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 truth-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 and Twitter—as well as product review websites that are prone to have fake product reviews and spam content [Toma and Hancock 2010, Guadagno et al. 2012, Warkentin et al. 2010, Joinson and Dietz-Uhler 2002, Ott et al. 2011, Li et al. 2014].

In the product reviews domain, 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 automatic deception detection can be accurately conducted on the product reviews domain and that humans are generally poor deception detectors for this task with inter-annotator agreement scores in the range (0.00,0.20), which indicates "slight agreement" between annotators. Interestingly, this study also showed that annotators suffered from "truth bias," a psychological phenomenon in which humans judges tend to believe others thus making it more likely to classify information as truthful rather than deceptive. Furthermore, 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 word 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. [2012] approached the deception detection task in Spanish essays by using Support Vector Machine (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 examined cultural differences among deceptive and truthful essays written by English, Spanish, and Romanian speakers. The authors addressed the deception detection task by first building classifiers separately for each culture and then by conducting several experiments across cultures. The authors proposed 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, truth-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., a sentence's 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 Wordnet 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 in this area included a data collection step using either manual or crowd sourced means. However, there is an increasing number of research work that has 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],[1] and a fake hotel reviews dataset collected from trip advisor [Ott et al. 2011],[2] a fake product review dataset collected using Mechanical Turk [Li et al. 2014].[3] 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 dataset provided by 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.[4]

13.2.3 Vision

Vision is the most common way people can detect liars as deception occurs on a 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 1.

1. http://lit.eecs.umich.edu/~deceptiondetection/
2. http://myleott.com/op_spam/
3. http://www.cs.uic.edu/~liub/FBS/fake-reviews.html
4. Available from the authors upon request

Figure 13.1 An example of spontaneous expressions with a truthful response (left) and a deceptive response (right).

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 were studied to specify whether or not 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].[5] 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 13.1.

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.[6] Several attempts were made to

5. http://tomas.pfister.fi/

6. http://www.cs.cmu.edu/~face/facs.htm

code these gestures automatically for efficient detection of human behavior and emotions. For instance, a real-time automated system to recognize spontaneous facial expressions that 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[7]) [Littlewort et al. 2011].

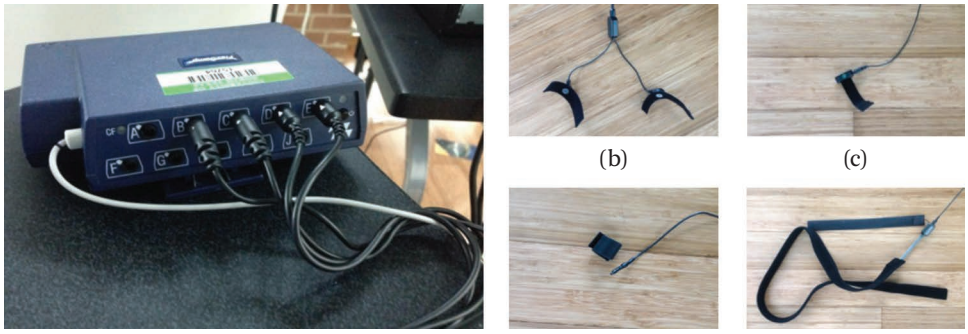
In addition to the action units, CERT provides 12 facial expressions such as yaw, pitch, roll, smile detector, anger, contempt, disgust, fear, joy, sad, surprise, and neutral. The software tool detects faces in each frame using Viola-Jones extension in a boosting framework followed by specifying the 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 an SVM-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.

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.

7. The CERT toolkit is no longer freely available. However, the CERT successor, The Facial Analysis Toolbox, is available as a commercial toolkit at http://imotions.com/



(a) (b) (c) (d) (e)

Figure 13.2 Physiological sensing system including (a) encoder, (b) skin conductance, (c) blood volume pulse, (d) skin temperature, and (e) abdominal respiration sensors.

2010]. Additionally, individuals acting truthfully produced more rhythmic pulsing gestures while those acting deceptively made more frequent speech prompting gestures [Hillman et al. 2012].

13.2.4 Physiology

Physiological signals play a crucial role in monitoring human health as well as detecting changes in human behavior. Chapter 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 13.2. Biological measurements, such as brain waves detected by MRI scanners, were also utilized as an indicator of deception [Kozel et al. 2004, Ganis et al. 2011]. 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.

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

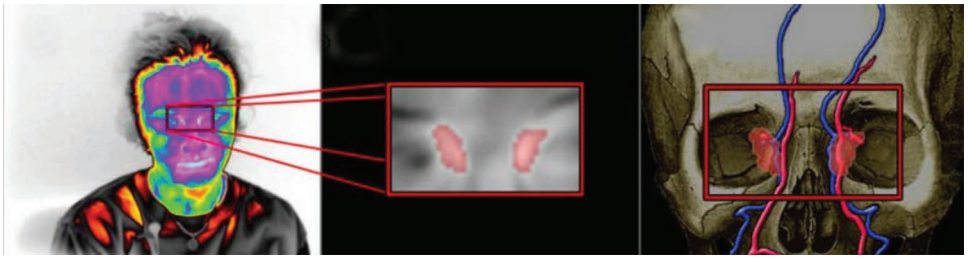
made to the style of questions directed to the subjects to avoid potential errors associated with polygraphs by using Guilty Knowledge Test (GKT) compared to the widely used Control Question Test (CQT) [Taylor et al. 2010]. GKT is a multiple choice form of questions that aimed at detecting concealed 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]. The study reported that bootstrapped analysis of reaction time method achieved 81.35% accuracy compared to 80% accuracy for event-related brain potentials approach.

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 toward the musculoskeletal tissue [Pavlidis and Levine 2001, Pavlidis and Levine 2002b]. 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% using two-class distinction; deceptive and non-deceptive. 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



(a) (b) (c)

Figure 13.3 (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]

thermal data from the periorbital area in the face to blood flow rates that had the potential of indicating deceit. Figure 13.3 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]. 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, Zhu et al. 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 13.4.

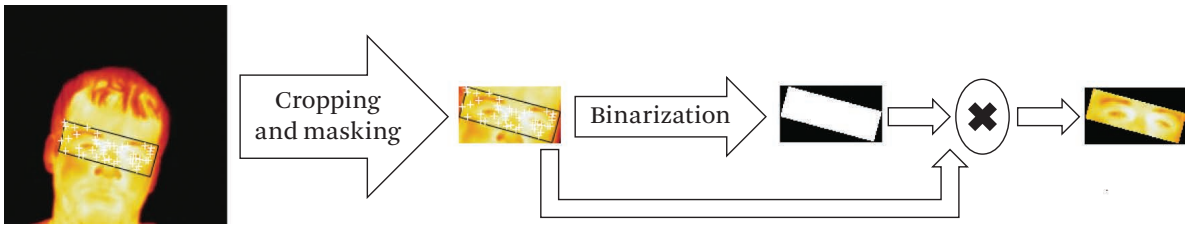


Figure 13.4 An overview of the region of interest segmentation process including determining the bounding box, cropping and masking, binarization, multiplication of the original and binarized images, and isolation of the region of interest.

13.3 Deception Detection with Multiple Modalities

In search of 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. [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.

13.3.1 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.

13.3.1.1 Dataset and Devices

Measurements were acquired in a lab setting using a thermal camera FLIR Thermovision A40 with a resolution of 340×240 and a frame rate of 60 frames per second, as well as 4 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.

13.3.1.2 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, representing the frequency of occurrence of words in the transcript of subjects responses, and features derived

from the frequency counts of word classes in the Linguistic Inquiry and Word Count (LIWC) lexicon.[8]

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.

13.3.1.3 Results

Feature-level multimodal 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 13.5 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.

Figures 13.6 and 13.7 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

8. The LIWC lexicon, available at http://liwc.wpengine.com/, is a resource developed for psycholinguistic analysis and contains about 70 word classes relevant to psychological processes (e.g., emotion, cognition).

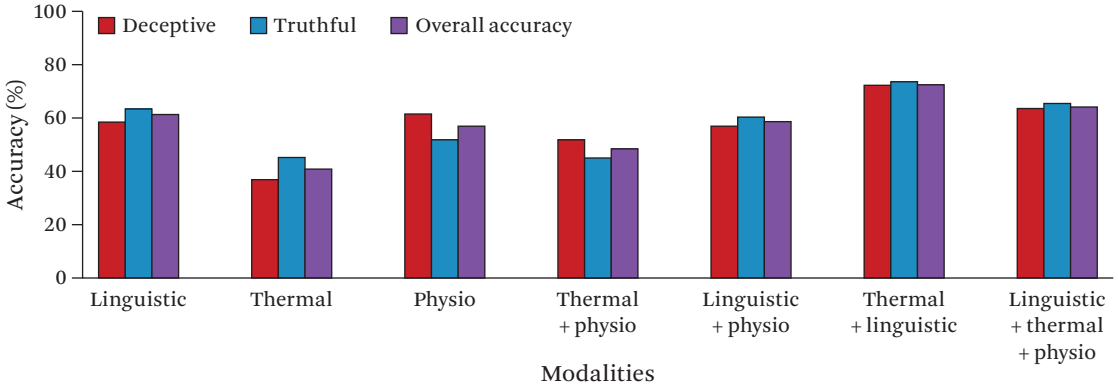


Figure 13.5 Deception, truthfulness, and overall accuracy percentages for individual and integrated modalities using features extracted from both the abortion and best friend topics.

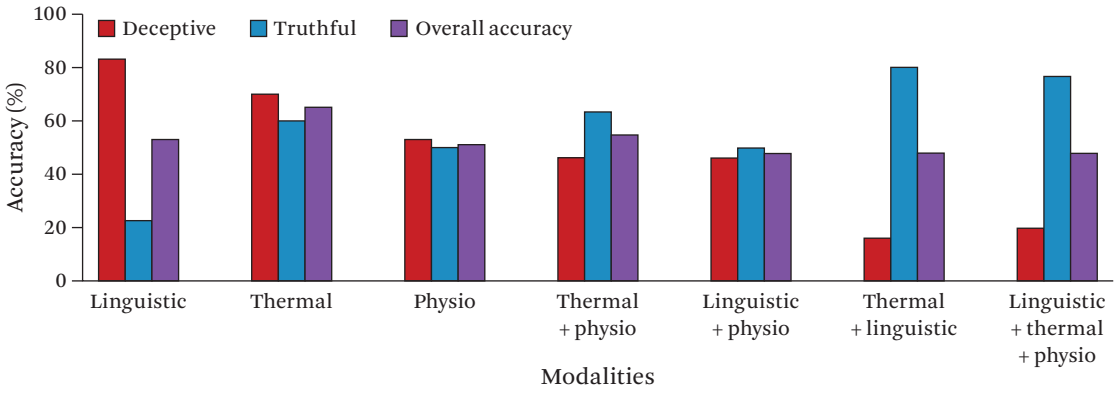


Figure 13.6 Deception, truthfulness, and overall accuracy percentages for individual and integrated modalities using across-topic learning. Best friend features are used for training and abortion features are used for testing.

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

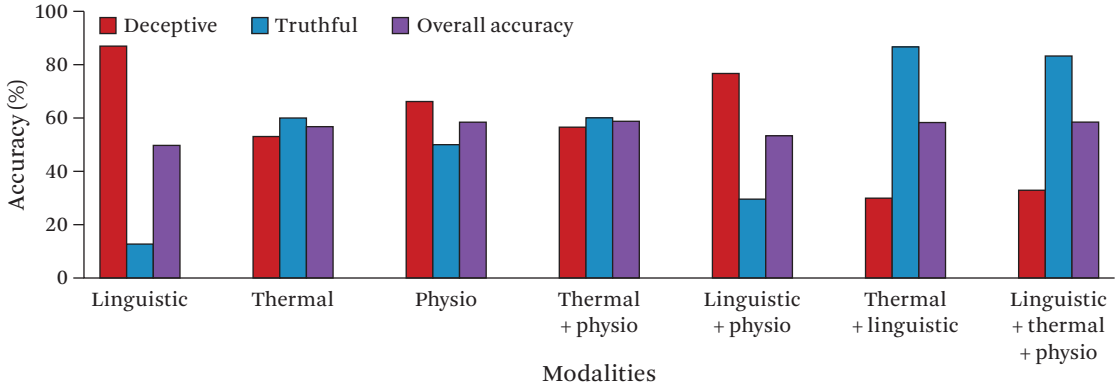


Figure 13.7 Deception, truthfulness, and overall accuracy percentages for individual and integrated modalities using across-topic learning. Abortion features are used for training and best friend features are used for testing.

modalities. The experiments showed that the quality of the 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.

13.3.2 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. The dataset consists of audio recorded interviews containing deceptive and truthful responses from 32 speakers and comprises approximately 7 hours of speech. In this work, authors built classifiers using the linguistic and speech modalities separately as well a combination of both modalities. Their experimental results showed noticeable improvement when combining the linguistic and acoustic modalities by reducing the baseline error by 6%. Overall, 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 use cepstral features to investigate speaker variability on the deception detection task. This 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. Following the same line of research, Enos et al. [2007] analyzed speaking segments, previously identified as emotionally charged and cognitively loaded, as a way to determine if a subject was 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 approach consisted of annotating the CSC corpus with critical segments and using lexical features, pauses, and vocal energy features to build models able to predict their occurrence. Their experimental results showed 20% relative improvement of performance over a random baseline while identifying deceptive speech.

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, and 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.

13.3.3 Vision and Language

More recently, the interest shifted towards detection of real-life deceptive behavior. A study used facial expressions, gestures, gaze, and conversational features in order to identify signals of trustworthiness between human negotiators [Lucas et al. 2016]. The study reported that multimodal approaches were better predictors of objective trustworthiness, whereas facial expression modality was more informative for perceived trustworthiness, suggesting that human mainly rely on facial expressions when judging trustworthiness.

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.

13.3.3.1 Dataset

The dataset consists of 121 videos including 61 deceptive and 60 truthful trial clips.[9] 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.

13.3.3.2 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.

9. http://deceptiondetection.eecs.umich.edu/

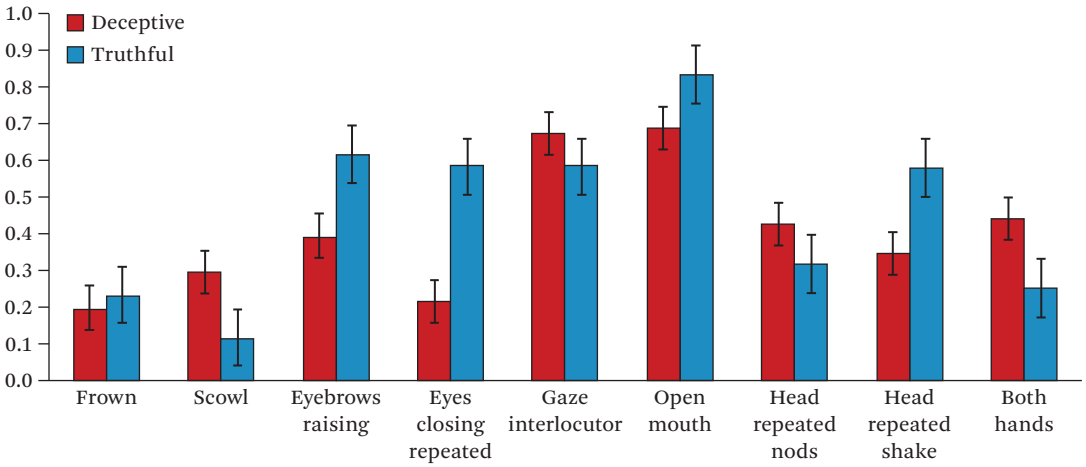


Figure 13.8 Distribution of non-verbal features for deceptive and truthful groups.

13.3.3.3 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 13.8 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.

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 13.1 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

Table 13.1 **Deception classifiers decision tree (DT) and random forest (RF) using individual and combined sets of verbal and non-verbal features**

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 13.2 **Feature ablation study**

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

(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 13.2 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 13.9 shows the five most predictive features of the presence of deception were the presence of frowning (Frowning), eyebrows movement (Eyebrows raising), lip gestures (Lip corners up, Lips protruded, Lips retracted), and head turns (Head side turn). These gestures were frequently portrayed by defendants and witnesses while being interrogated.

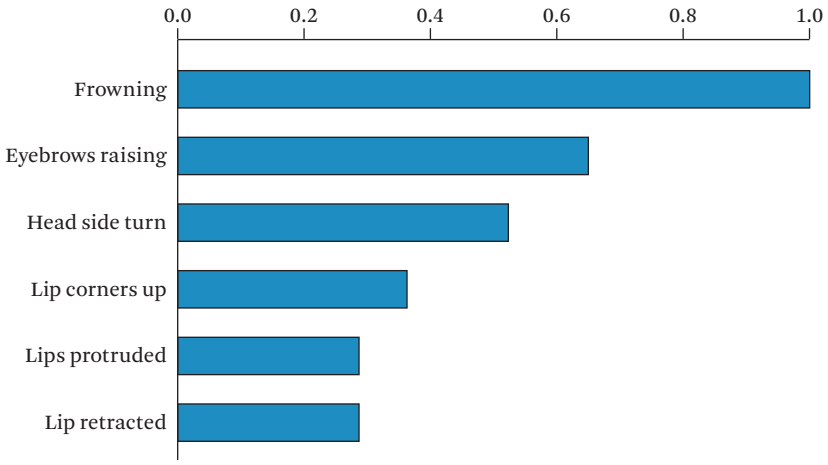


Figure 13.9 Weights of top non-verbal features in the multimodal deception classifier. The weights shown in this figure are normalized between 0 and 1 to easily observe the contribution of each feature.

Table 13.3 **Performance of three annotators (A1, A2, A3) and the developed automatic system (Sys) on the real-deception dataset over four modalities**

	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%

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 13.3, 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.

13.4 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 15–40 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 nonlinear transformations in order to interpret different levels of abstractions in the data, as can be seen in Chapter 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

This material is based in part upon work supported by National Science Foundation awards #1344257 and #1355633 and by DARPA-BAA-12-47 DEFT grant #12475008. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation or the Defense Advanced Research Projects Agency.

Focus Questions

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

13.2. What is meant by multimodal deception detection?

13.3. Which modalities can be used for deception detection?

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

13.5. How can the multimodal features be integrated?

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

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

13.8. How can deception detection be improved in the future? Design a high-performing deception detection system, and argue for its specific strengths.

13.9. What evidence is there that automatic multimodal-multisensor deception detection systems may outperform human judgment in the future? How can future systems be designed to further leverage these strengths of automated deception detection?

13.10. Discuss the problem of intentionally faking innocence on deception tests, which is a form of spoofing the system that creates potential security risks, and how systems can be designed to avoid it.

References

M. Abouelenien, V. Pérez-Rosas, R. Mihalcea, and M. Burzo. 2014. Deception detection using a multimodal approach. In *16th ACM International Conference on Multimodal Interaction, ICMI 2014*. DOI: 10.1145/2663204.2663229. 421, 433

M. Abouelenien, M. Burzo, and R. Mihalcea. 2015a. 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*, pp. 59:1–59:8. ACM, New York. DOI: 10.1145/2769493.2769505. 421

M. Abouelenien, R. Mihalcea, and M. Burzo. 2015b. Trimodal analysis of deceptive behavior. In *Proceedings of the 2015 ACM on Workshop on Multimodal Deception Detection, WMDD '15*, pp. 9–13. ACM, New York. DOI: 10.1145/2823465.2823470. 421, 429, 432, 437

M. Abouelenien, M. Burzo, and R. Mihalcea. 2016a. 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. DOI: 10.1145/2910674.2910705. 421

M. Abouelenien, R. Mihalcea, and M. Burzo. June 2016b. 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. DOI: 10.1145/2910674.2910682. 421

J. Allwood, L. Cerrato, K. Jokinen, C. Navarretta, and P. Paggio. 2007. The mumin coding scheme for the annotation of feedback, turn management and sequencing phenomena. *Language Resources and Evaluation*, 41(3-4): 273–287. DOI: 10.1007/s10579-007-9061-5. 440

A. Almela, R. Valencia-García, and P. Cantos. 2012. Seeing through deception: A computational approach to deceit detection in written communication. In *Proceedings of the Workshop on Computational Approaches to Deception Detection*, pp. 15–22. Association for Computational Linguistics, Avignon, France. DOI: 10.5195/lesli.2013.5 . 426, 427

M. Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. Fasel, and J. Movellan. 2006. Automatic recognition of facial actions in spontaneous expressions. *Journal of Multimedia*, 1(6): 22–35. DOI: 10.4304/jmm.1.6.22-35. 428

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. March 2009. Detecting concealment of intent in transportation screening: A proof of concept. *IEEE Transactions on Intelligent Transportation Systems*, 10(1): 103–112. DOI: 10.1109/TITS.2008.2011700. 433

D. Carmel, E. Dayan, A. Naveh, O. Raveh, and G. Ben-Shakhar. 2003. 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. DOI: 10.1037/1076-898X.9.4.261. 431

L. Caso, F. Maricchiolo, M. Bonaiuto, A. Vrij, and S. Mann. 2006. The impact of deception and suspicion on different hand movements. *Journal of Nonverbal Behavior*, 30(1): 1–19. DOI: 10.1007/s10919-005-0001-z. 429

G. Chittaranjan and H. Hung. 2010. 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)*, pp. 5334–5337. DOI: 10.1109/ICASSP.2010.5494961. 439

D. Cohen, G. Beattie, and H. Shovelton. 2010. Nonverbal indicators of deception: How iconic gestures reveal thoughts that cannot be suppressed. *Semiotica*, 2010(182): 133–174. DOI: 10.1515/semi.2010.055. 429, 430

N. R. Council. 2003. *The Polygraph and Lie Detection*. The National Academies Press, Washington, DC. http://www.nap.edu/catalog/10420/the-polygraph-and-lie-detection. DOI: 10.17226/10420. 420

B. M. DePaulo. 1992. Nonverbal behavior and self-presentation. *Psychological Bulletin*, 111(2): 203. DOI: 10.1037//0033-2909.111.2.203. 423

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. 2003. Cues to deception. *Psychological Bulletin*, pp. 74–118. 419, 423, 424, 437

M. Derksen. 2012. Control and resistance in the psychology of lying. *Theory and Psychology*, 22(2): 196–212. DOI: 10.1177/0959354311427487. 419, 430

P. Ekman. 2001. *Telling Lies: Clues to Deceit in the Marketplace, Politics and Marriage*. Norton, W.W. and Company. 420, 427, 428

P. Ekman. 2003. 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. DOI: 10.1196/annals.1280.010. 428

P. Ekman and E. Rosenberg. 2005. *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. 428, 429

F. Enos, E. Shriberg, M. Graciarena, J. Hirschberg, and A. Stolcke. 2007. Detecting deception using critical segments. In *INTERSPEECH*, pp. 2281–2284. 438

S. Feng, R. Banerjee, and Y. Choi. 2012. Syntactic stylometry for deception detection. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, ACL '12, pp. 171–175. Association for Computational Linguistics, Stroudsburg, PA. 420, 425

T. Fornaciari and M. Poesio. 2013a. Automatic deception detection in italian court cases. *Artificial Intelligence and Law*, 21(3): 303–340. DOI: 10.1007/s10506-013-9140-4. 419

T. Fornaciari and M. Poesio. 2013b. Automatic deception detection in italian court cases. *Artificial Intelligence and Law*, 21(3): 303–340. DOI: 10.1007/s10506-013-9140-4. 426

G. Ganis, S. M. Kosslyn, S. Stose, W. L. Thompson, and D. A. Yurgelun-Todd. August 2003. Neural correlates of different types of deception: An fmri investigation. *Cerebral Cortex*, 13(8): 830–836. DOI: 10.1093/cercor/13.8.830. 431

G. Ganis, J. P. Rosenfeld, J. Meixner, R. A. Kievit, and H. E. Schendan. 2011. Lying in the scanner: Covert countermeasures disrupt deception detection by functional magnetic resonance imaging. *NeuroImage*, 55(1): 312–319. DOI: 10.1016/j.neuroimage.2010.11.025. 420, 430

T. A. Gannon, A. R. Beech, and T. Ward. 2009. *Risk Assessment and the Polygraph*, pp. 129–154. John Wiley and Sons Ltd. DOI: 10.1002/9780470743232.ch8 . 419, 430

M. Garbey, A. Merla, and I. Pavlidis. 2004. 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*, vol. 1, pp. I–356–I–363. DOI: 10.1109/CVPR.2004.1315054. 431

S. Gokhman, J. Hancock, P. Prabhu, M. Ott, and C. Cardie. 2012. In search of a gold standard in studies of deception. In *Proceedings of the Workshop on Computational Approaches to Deception Detection*, pp. 23–30. Association for Computational Linguistics. 427

M. Graciarena, E. Shriberg, A. Stolcke, F. Enos, J. Hirschberg, and S. Kajarekar. 2006. 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.*, vol. 1, pp. I–I. IEEE. DOI: 10.1109/ICASSP.2006.1660200. 438

P. A. Granhag and M. Hartwig. 2008. A new theoretical perspective on deception detection: On the psychology of instrumental mind-reading. *Psychology, Crime & Law*, 14(3): 189–200. DOI: 10.1080/10683160701645181. 420

R. Guadagno, B. Okdie, and S. Kruse. Mar. 2012. Dating deception: Gender, online dating, and exaggerated self-presentation. *Computers in Human Behavior*, 28(2): 642–647. DOI: 10.1016/j.chb.2011.11.010. 425

K. Harmer, S. Yue, K. Guo, K. Adams, and A. Hunter. December 2010. Automatic blush detection in "concealed information" test using visual stimuli. In *2010 International Conference of Soft Computing and Pattern Recognition (SoCPaR)*, pp. 259–264. DOI: 10.1109/SOCPAR.2010.5686076. 432

D. D. Henningsen, K. Valde, and E. Davies. 2005. Exploring the effect of verbal and nonverbal cues on perceptions of deception. *Communication Quarterly*, 53(3): 359–375. DOI: 10.1080/01463370500101329. 422, 433

J. Hillman, A. Vrij, and S. Mann. 2012. Um . . . they were wearing . . . : The effect of deception on specific hand gestures. *Legal and Criminological Psychology*, 17(2): 336–345. DOI: 10.1111/j.2044-8333.2011.02014.x. 420, 430

J. Hirschberg, S. Benus, J. Brenier, F. Enos, S. Friedman, S. Gilman, C. Gir, G. Graciarena, A. Kathol, and L. Michaelis. 2005. Distinguishing deceptive from non-deceptive speech. In *In Proceedings of Interspeech 2005 - Eurospeech*, pp. 1833–1836. 420, 438

D. Howard and C. Kirchhübel. 2011. Acoustic correlates of deceptive speech: an exploratory study. In *Proceedings of the 9th International Conference on Engineering psychology and Cognitive Ergonomics*, EPCE'11, pp. 28–37. Springer-Verlag, Berlin, Heidelberg. DOI: 10.1007/978-3-642-21741-8_4. 420

U. Jain, B. Tan, and Q. Li. March 2012. Concealed knowledge identification using facial thermal imaging. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1677–1680. DOI: 10.1109/ICASSP.2012.6288219. 432

M. L. Jensen, T. O. Meservy, J. K. Burgoon, and J. Nunamaker. 2010. Automatic, multimodal evaluation of human interaction. *Group Decision and Negotiation*, 19(4): 367–389. DOI: 10.1007/s10726-009-9171-0. 433

A. N. Joinson and B. Dietz-Uhler. 2002. Explanations for the perpetration of and reactions to deception in a virtual community. *Social Science Computer Review*, 20(3): 275–289. DOI: 10.1177/089443930202000305. 425

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. 2004. 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. DOI: 10.1176/appi.neuropsych.16.3.295. 430, 431

S. I. Levitan, G. An, M. Wang, G. Mendels, J. Hirschberg, M. Levine, and A. Rosenberg. 2015. Cross-cultural production and detection of deception from speech. In *Proceedings of the 2015 ACM on Workshop on Multimodal Deception Detection*, pp. 1–8. ACM. DOI: 10.1145/2823465.2823468. 438

J. Li, M. Ott, C. Cardie, and E. Hovy. June 2014. Towards a general rule for identifying deceptive opinion spam. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*. Baltimore, MD. DOI: 10.3115/v1/P14-1147. 425, 427

G. Littlewort, J. Whitehill, T. Wu, I. Fasel, M. Frank, J. Movellan, and M. Bartlett. March 2011. The computer expression recognition toolbox (cert). In *2011 IEEE International Conference on Automatic Face Gesture Recognition and Workshops (FG 2011)*, pp. 298–305. DOI: 10.1109/AFGR.2008.4813406. 429

G. Lucas, G. Stratou, S. Lieblich, and J. Gratch. 2016. Trust me: Multimodal signals of trustworthiness. In *Proceedings of the 18th ACM International Conference on*

Multimodal Interaction, ICMI 2016, pp. 5–12. ACM, New York. http://doi.acm.org/10.1145/2993148.2993178. DOI: 10.1145/2993148.2993178. 439

D. T. Lykken. 1984. Polygraphic interrogation. *Nature*, 307(5953): 681–684. DOI: 10.1038/307681a0. 419

G. Maschke and G. Scalabrini. 2005. *The Lie Behind the Lie Detector*. http://antipolygraph.org. 430

R. Mihalcea and C. Strapparava. 2009a. The lie detector: Explorations in the automatic recognition of deceptive language. In *Proceedings of the Association for Computational Linguistics, ACL 2009*. Singapore. 424, 425, 427

R. Mihalcea and C. Strapparava. 2009b. The lie detector: Explorations in the automatic recognition of deceptive language. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pp. 309–312. Association for Computational Linguistics, Suntec, Singapore. 420

A. Mohammadian, V. Abootalebi, M. Moradi, and M. Khalilzadeh. 2008. Multimodal detection of deception using fusion of reaction time and p300 component. In *Biomedical Engineering Conference, 2008, CIBEC 2008, Cairo International*, pp. 1–4. DOI: 10.1109/CIBEC.2008.4786064. 431

M. Newman, J. Pennebaker, D. Berry, and J. Richards. 2003. Lying words: Predicting deception from linguistic styles. *Personality and Social Psychology Bulletin*, 29. DOI: 10.1177/0146167203029005010. 424, 425

J. Nunamaker, J.F., J. Burgoon, N. Twyman, J. Proudfoot, R. Schuetzler, and J. Giboney. June 2012. Establishing a foundation for automated human credibility screening. In *2012 IEEE International Conference on Intelligence and Security Informatics (ISI)*, pp. 202–211. DOI: 10.1109/ISI.2012.6284309. 433

M. Ott, Y. Choi, C. Cardie, and J. Hancock. 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, HLT '11*, pp. 309–319. Association for Computational Linguistics, Stroudsburg, PA. 425, 427, 443

M. Ott, C. Cardie, and J. T. Hancock. 2013. 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*. Association for Computational Linguistics, Atlanta, GA. 426

M. Owayjan, A. Kashour, N. Al Haddad, M. Fadel, and G. Al Souki. Dec 2012. 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)*, pp. 33–38. DOI: 10.1109/ICTEA.2012.6462897. 420

I. Pavlidis and J. Levine. 2001. 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*, vol. 3, pp. 2826–2829. DOI: 10.1109/IEMBS.2001.1017374. 431

I. Pavlidis and J. Levine. 2002b. Thermal image analysis for polygraph testing. *IEEE Engineering in Medicine and Biology Magazine*, 21(6): 56–64. DOI: 10.1109/MEMB.2002.1175139. 431

I. Pavlidis, N. L. Eberhardt, and J. A. Levine. 2002. Human behavior: Seeing through the face of deception. *Nature*, 415(6867): 35–35. DOI: 10.1038/415035a. 431

I. Pavlidis, P. Tsiamyrtzis, D. Shastri, A. Wesley, Y. Zhou, P. Lindner, P. Buddharaju, R. Joseph, A. Mandapati, B. Dunkin. 2012. Fast by nature-how stress patterns define human experience and performance in dexterous tasks. *Scientific Reports*, 2. DOI: 10.1038/srep00305. 420

J. Pennebaker and M. Francis, 1999. Linguistic Inquiry and Word Count: LIWC. Erlbaum Publishers. 424

V. Pérez-Rosas and R. Mihalcea. 2014. Cross-cultural deception detection. In *Proceedings of the Association for Computational Linguistics*. DOI: 10.1002/9781118510001.ch8. 426, 427

V. Pérez-Rosas, M. Abouelenien, R. Mihalcea, Y. Xiao, C. Linton, and M. Burzo. 2015a. Verbal and nonverbal clues for real-life deception detection. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 2336–2346. Association for Computational Linguistics, Lisbon, Portugal. DOI: 10.1145/2818346.2820758. 421

V. Pérez-Rosas, M. Abouelenien, R. Mihalcea, Y. Xiao, C. Linton, and M. Burzo. 2015b. Deception detection using real-life trial data. In *17th ACM International Conference on Multimodal Interaction, ICMI 2015*. 421, 440

T. Pfister and M. Pietikäinen. 2012. Electronic imaging & signal processing automatic identification of facial clues to lies. *SPIE Newsroom*. http://tomas.pfister.fi/. 420, 428

T. Qin, J. Burgoon, and J. Nunamaker. 2004. 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*, pp. 23–32. DOI: 10.1109/HICSS.2004.1265083. 441

T. Qin, J. K. Burgoon, J. P. Blair, and J. F. Nunamaker. 2005. Modality effects in deception detection and applications in automatic deception-detection. In *Proceedings of the 38th Hawaii International Conference on System Sciences*. DOI: 10.1109/HICSS.2005.436. 441

B. Rajoub and R. Zwiggelaar. 2014. Thermal facial analysis for deception detection. *IEEE Transactions on Information Forensics and Security*, PP(99): 1–1. DOI: 10.1109/TIFS.2014.2317309. 420

D. Shastri, M. Papadakis, P. Tsiamyrtzis, B. Bass, and I. Pavlidis. July 2012. Perinasal imaging of physiological stress and its affective potential. *IEEE Transactions on Affective Computing*, 3(3): 366–378. DOI: 10.1109/T-A1FFC.2012.13. 420

C. Strapparava and A. Valitutti. 2004. Wordnet-affect: an affective extension of wordnet. In *Proceedings of the 4th International Conference on Language Resources and Evaluation*. Lisbon. 425

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. 2010. A comparison of approaches to detect deception. Technical Report ADA537848, Naval Aerospace Medical Research Laboratory. 431

Y.-L. Tian, T. Kanade, and J. Cohn. 2005. Facial expression analysis. In *Handbook of Face Recognition*, pp. 247–275. New York; Springer. 429

C. Toma and J. Hancock. 2010. 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, pp. 5–8. ACM, New York. DOI: 10.1145/1718918.1718921. 425, 426

P. Tsiamyrtzis, J. Dowdall, D. Shastri, I. Pavlidis, M. Frank, and P. Eckman. 2005. 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*, pp. 555–566. 432

P. Tsiamyrtzis, J. Dowdall, D. Shastri, I. T. Pavlidis, M. G. Frank, and P. Ekman. 2007. Imaging facial physiology for the detection of deceit. *International Journal of Computer Vision*, 71(2): 197–214. ISSN 0920-5691. DOI: 10.1007/s11263-006-6106-y. 431, 432

B. Verhoeven and W. Daelemans. 2014. 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, eds., *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. European Language Resources Association (ELRA), Reykjavik, Iceland. 427

B. Verschuere, V. Prati, and J. De Houwer. 2009. Cheating the lie-detector: Faking the autobiographical iat. *Psychological Science*, 20: 410–413. DOI: 10.1111/j.1467-9280.2009.02308.x. 430

A. Vrij. 2001. *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. 420, 422, 430

A. Vrij, P. Granhag, and S. Porter. dec 2010. Pitfalls and opportunities in nonverbal and verbal lie detection. *Psychological Science in the Public Interest*, 11(3): 89–121. DOI: 10.1177/1529100610390861. 420

D. Warkentin, M. Woodworth, J. Hancock, and N. Cormier. 2010. Warrants and deception in computer mediated communication. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work*, pp. 9–12. ACM. DOI: 10.1145/1718918.1718922. 425

L. Warmelink, A. Vrij, S. Mann, S. Leal, D. Forrester, and R. P. Fisher. 2011. Thermal imaging as a lie detection tool at airports. *Law and Human Behavior*, 35(1): 40–48. ISSN 0147-7307. DOI: 10.1007/s10979-010-9251-3. 432

M. Yancheva and F. Rudzicz. August 2013. 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)*, pp. 944–953. Association for Computational Linguistics, Sofia, Bulgaria. 426

D. Yu, Y. Tyshchuk, H. Ji, and W. A. Wallace. 2015. 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, pp. 857–866. DOI: 10.3115/v1/P15-1083 . 426

Y. Zhou, P. Tsiamyrtzis, P. Lindner, I. Timofeyev, and I. Pavlidis. May 2013. Spatiotemporal smoothing as a basis for facial tissue tracking in thermal imaging. *IEEE Transactions on Biomedical Engineering*, 60(5): 1280–1289. DOI: 10.1109/TBME.2012.2232927. 420

Z. Zhu, P. Tsiamyrtzis, and I. Pavlidis. 2007. Forehead thermal signature extraction in lie detection. In *29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 2007. EMBS 2007*, pp. 243–246. DOI: 10.1109/IEMBS.2007.4352269. 432

Z. Zhu, P. Tsiamyrtzis, and I. Pavlidis. 2008. The segmentation of the supraorbital vessels in thermal imagery. In *IEEE Fifth International Conference on Advanced Video and Signal Based Surveillance, 2008. AVSS '08*, pp. 237–244. DOI: 10.1109/AVSS.2008.36. 432

M. Zuckerman, B. M. DePaulo, and R. Rosenthal. 1981. Verbal and nonverbal communication of deception. *Advances in Experimental Social Psychology*, 14(1): 59. 423, 437

Index

Biographies

Editors

Philip R. Cohen (Monash University) is Director of the Laboratory for Dialogue Research and Professor of Artificial Intelligence in the Faculty of Information Technology at Monash University. His research interests include multimodal interaction, human-computer dialogue, and multi-agent systems. He is a Fellow of the American Association for Artificial Intelligence, past President of the Association for Computational Linguistics, recipient (with Hector Levesque) of the Inaugural Influential Paper Award by the *International Foundation for Autonomous Agents and Multi-Agent Systems*, and recipient of the 2017 Sustained Achievement Award from the International Conference on Multimodal Interaction.

He was most recently Chief Scientist in Artificial Intelligence and Senior Vice President for Advanced Technology at Voicebox Technologies, where he led efforts on semantic parsing and dialogue. Cohen founded Adapx Inc. to commercialize multimodal interaction, deploying multimodal systems to civilian and government organizations. Prior to Adapx, he was Professor and Co-Director of the Center for Human-Computer Communication in the Computer Science Department at Oregon Health and Science University and Director of Natural Language in the Artificial Intelligence Center at SRI International. Cohen has published more than 150 articles and has 5 patents. He co-authored the book *The Paradigm Shift to Multimodality in Contemporary Computer Interfaces* (2015, Morgan & Claypool Publishers) with Dr. Sharon Oviatt. (Contact: philip.cohen@monash.edu)

Antonio Krüger (Saarland University and DFKIGmbH) is Professor of Computer Science and Director of the Media Informatics Program at Saarland University, as well as Scientific Director of the Innovative Retail Laboratory at the German Research Center for Artificial Intelligence (DFKI). His research areas focus on intelligent user interfaces, and mobile and ubiquitous context-aware systems. He has been General Chair of the Ubiquitous Computing Conference and Program Chair of MobileHCI,

IUI, and Pervasive Computing. He is also on the Steering Committee of the International Conference on Intelligent User Interfaces (IUI) and an Associate Editor of the journals *User Modeling and User-Adapted Interaction* and A*CM Interactive, Mobile, Wearable and Ubiquitous Technologies.* (Contact: krueger@dfki.de)

Sharon Oviatt (Monash University) is internationally known for her multidisciplinary work on multimodal and mobile interfaces, human-centered interfaces, educational interfaces, and learning analytics. She has been recipient of the inaugural ACM-ICMI Sustained Accomplishment Award, National Science Foundation Special Creativity Award, and ACM-SIGCHI CHI Academy award. She has published over 160 scientific articles in a wide range of venues and is an Associate Editor of the major journals and edited book collections in the field of human-centered interfaces. Her other books include *The Design of Future Educational Interfaces* (2013, Routledge) and *The Paradigm Shift to Multimodality in Contemporary Computer Interfaces* (2015, Morgan & Claypool Publishers). (Contact: sharon.oviatt@monash.edu)

Gerasimos Potamianos (University of Thessaly) is Associate Professor and Director of Graduate Studies in Electrical and Computer Engineering. His research spans multisensory and multimodal speech processing and scene analysis, with applications to human-computer interaction and ambient intelligence. He has authored over 120 articles and has 7 patents. He received a Diploma degree from the National Technical University of Athens, and a M.Sc. and Ph.D. from Johns Hopkins University, all in electrical and computer engineering. In addition to his academic experience, he has worked at AT&T Research Labs, IBM Thomas J. Watson Research Center (US), and at the FORTH and NCSR 'Demokritos' Research Centers in Greece. (Contact: gpotam@ieee.org)

Björn Schuller (University of Augsburg and Imperial College London) is currently ZD.B Chair of Embedded Intelligence for Health Care and Wellbeing at University of Augsburg and Reader in Machine Learning at Imperial College. He is best known for his work on multisensorial/multimodal intelligent signal processing for affective, behavioral, and human-centered computing. In 2015 and 2016, he was honored by the World Economic Forum as one of 40/50 extraordinary scientists under age 40. In 2018, he was elevated to Fellow of the IEEE and Senior Member of the ACM. He has published over 700 peer-reviewed scientific contributions across a range of disciplines and venues, and is Editor-in-Chief of *IEEE Transactions on Affective Computing*. His books include *Intelligent Audio Analysis* (2013, Springer) and *Computational Paralinguistics* (2013, Wiley). (Contact: bjoern.schuller@imperial.ec.uk)

Daniel Sonntag (German Research Center for Artificial Intelligence, DFKI) is a Principal Researcher and Research Fellow. His research interests include multimodal and mobile AI-based interfaces, common-sense modeling, and explainable machine learning methods for cognitive computing and improved usability. He has published over 130 scientific articles and was the recipient of the German High Tech Champion Award in 2011 and the AAAI Recognition and IAAI Deployed Application Award in 2013. He is the Editor-in-Chief of the *German Journal on Artificial Intelligence (KI)* and editor-in-chief of Springer's Cognitive Technologies book series. Currently, he leads both national and European projects from the Federal Ministry of Education and Research, the Federal Ministry for Economic Affairs and Energy, and Horizon 2020. (Contact: daniel.sonntag@dfki.de)

Authors and Challenge Discussants

Mohamed Abouelenien (University of Michigan-Dearborn) is an Assistant Professor in the Department of Computer and Information Science at the University of Michigan-Dearborn. He was a Postdoctoral Research Fellow in the Electrical Engineering and Computer Science Department at the University of Michigan, Ann Arbor from 2014–2017. In 2013, he received his Ph.D. in Computer Science and Engineering from the University of North Texas. His areas of interest broadly cover data science topics, including applied machine learning, computer vision, and natural language processing. He has worked on a number of projects in these areas, including affective computing, deception detection, ensemble learning, video and image processing, face and action recognition, and others. His recent research involves data analytics projects as well as modeling of human behavior for different applications. Abouelenien has published extensively in international journals and conferences in IEEE, ACM, Springer, and SPIE. He also served as the chair for the ACM Workshop on Multimodal Deception Detection, a reviewer for *IEEE Transactions* and Elsevier journals, and a program committee member for multiple international conferences.

Chaitanya Ahuja (Carnegie Mellon University) is a doctoral candidate at the Language Technologies Institute in the School of Computer Science at Carnegie Mellon University. His interests range in various topics in natural language, computer vision, computational music, and machine learning. Before starting graduate school, Chaitanya completed his Bachelor's degree at the Indian Institute of Technology, Kanpur, with a research focus on spatial audio.

Ethem Alpaydin (Bogazici University) received his degree of Docteur es Sciences from École Polytechnique Fédérale de Lausanne in 1990. Currently, he is a Professor in the Department of Computer Engineering of Bogazici University and a member of The Science Academy, Istanbul. As a visiting researcher, he worked in the Department of Brain and Cognitive Sciences at MIT in 1994, at the International Computer Science Institute at UC Berkeley in 1997, at the Idiap Research Institute in Switzerland in 1998, and at TU Delft in 2014. He was a Fulbright Senior Scholar in 1997/1998 and received the Research Excellence Award from the Bogazici University Foundation in 1998 (junior faculty) and 2008 (senior faculty), the Young Scientist Award from the Turkish Academy of Sciences in 2001, and the Scientific Encouragement Award from the Turkish Scientific and Technical Research Council in 2002. His book *Introduction to Machine Learning*, published by MIT Press, is now in its third edition and was translated into Chinese, German, and Turkish. *Machine Learning: The New AI* was also published by MIT Press as part of the Essential Knowledge Series in 2016, and has since been translated into Russian and Japanese. He is a senior Member of the IEEE and an Editorial Board Member of the Pattern Recognition journal, published by Elsevier.

Mehdi Ammi (University of Paris-Saclay) is an Associate Professor at the University of Paris-Saclay. He is also the head of the Pervasive and Ubiquitous Environments team and member of the Architecture and Models for Interaction group at the multidisciplinary LIMSI-CNRS lab. He earned his Ph.D. at the University of Orleans in 2005 with an emphasis on robotics and virtual reality.

Michel-Ange Amorim (Université Paris-Sud) is a Full Professor at the Université Paris-Sud (UPSUD), Université Paris-Saclay, Orsay, France. He received his Ph.D. in cognitive psychology from Université René Descartes, Paris, France, in 1997. At UPSUD, he leads CIAMS, a multidisciplinary laboratory investigating human movement, including motor control, psychology, biomechanics, physiology, and behavioral and cognitive neuroscience. His research interests are in embodied cognition combining psychophysics—the Information Integration Theory approach—with neuroimaging techniques in normals and patients, in order to decipher the neurocognitive processes underlying spatial and motoric embodiment of self, self-environment, and self-other relationships.

Elisabeth André (Augsburg University) is a Full Professor of Computer Science and Founding Chair of Human-Centered Multimedia at Augsburg University in Germany. She has a long track record in multimodal human-machine interaction, embodied conversational agents, social robotics, affective computing, and social

signal processing. Elisabeth André has served as a General and Program Co-Chair of major international conferences, including ACM International Conference on Intelligent User Interfaces (IUI), ACM International Conference on Multimodal Interfaces (ICMI), and International Conference on Autonomous Agents and Multiagent Systems (AAMAS). In 2010, Elisabeth André was elected a member of the prestigious Academy of Europe, the German Academy of Sciences Leopoldina, and AcademiaNet. To honor her achievements in bringing artificial intelligence techniques to HCI, she was awarded a EurAI (European Coordinating Committee for Artificial Intelligence) fellowship in 2013. Most recently, she was elected to the CHI Academy, an honorary group of leaders in the field of human-computer interaction.

Syed Z. Arshad (University of New South Wales) has a Ph.D. in computer science from the University of New South Wales, Australia. He is a member of IEEE, ACM, SIGCHI, and SIGKDD. His research interests include cognitive computing, intelligent user interfaces, machine learning, and data visualization techniques.

Tadas Baltrušaitis (Microsoft) is a scientist at Microsoft in Cambridge, UK. His primary research interests lie in the automatic understanding of non-verbal human behavior, computer vision, and multimodal machine learning. In particular, he is interested in the application of such technologies to healthcare settings, with a focus on mental health. Before joining Microsoft, he was a post-doctoral researcher at Carnegie Mellon University, working on multimodal machine learning and automatic facial behavior analysis. He received his Ph.D. at the University of Cambridge, where his work focused on automatic facial expression analysis in especially difficult real-world settings.

Samy Bengio (Google) has been a research scientist at Google since 2007. Before that, he had been a senior researcher in statistical machine learning at IDIAP Research Institute, where he supervised Ph.D. students and postdoctoral fellows. His research interests span many areas of machine learning such as deep architectures, representation learning, sequence processing, speech recognition, image understanding, support vector machines, mixture models, large-scale problems, multi-modal (face and voice) person authentication, brain–computer interfaces, and document retrieval. He was the program chair of NIPS 2017 and ICLR 2015 and 2016; was the general chair of BayLearn 2012–2015, the Workshops on Machine Learning for Multimodal Interactions (MLMI) 2004–2006, and the IEEE Workshop on Neural Networks for Signal Processing (NNSP) in 2002; and served on the program committees of several international conferences such as NIPS, ICML, ICLR, ECML and IJCAI.

Nigel Bosch (University of Illinois at Urbana-Champaign) is a postdoctoral researcher with the National Center for Supercomputing Applications. His research utilizes data mining and machine learning techniques to understand human experiences including emotion, cognition, and behavior. His current interests are focused on developing methods to model new forms of big multimodal data in online educational contexts, as well as studying the ethical implications of machine learning models trained in these contexts. He received his Ph.D. in computer science from the University of Notre Dame.

Mihai Burzo (University of Michigan-Flint) is an Assistant Professor of Mechanical Engineering at the University of Michigan-Flint. Prior to joining University of Michigan in 2013, he was an Assistant Professor at University of North Texas. His research interests include heat transfer in microelectronics and nanostructures, thermal properties of thin films of new and existing materials, multimodal sensing of human behavior, and computational modeling of forced and natural heat convection. He has published over 50 articles in peer-reviewed journals and conference proceedings. He is the recipient of several awards, including the 2006 Harvey Rosten Award For Excellence for "outstanding work in the field of thermal analysis of electronic equipment," best paper award at the Semitherm conferences in 2013 and 2006, Young Engineer of the Year from the North Texas Section of ASME in 2006, and Leadership Award from SMU in 2002.

Fang Chen (DATA61, CSIRO) is a Senior Principal Research Scientist of Analytics in DATA61, CSIRO. She holds a Ph.D. in Signal and Information Processing, an M.Sc. and B.Sc. in Telecommunications and Electronic Systems, respectively, and an MBA. Her research interests are behavior analytics, machine learning, and pattern recognition in human and system performance prediction and evaluation. She has done extensive work on human-machine interaction and cognitive load modeling. She pioneered the theoretical framework of measuring cognitive load through multimodal human behavior and provided much of the empirical evidence on using human behavior signals and physiological responses to measure and monitor cognitive load.

Huili Chen (MIT) is a Ph.D. student at the MIT Media Lab. She obtained a Bachelor of Arts degree in Psychology and earned a Bachelor of Science degree in Computer Science from the University of Notre Dame in 2016. She is very interested in human-machine interaction and interactive artificial intelligence.

Lei Chen (Liulishuo Inc.) is a Principal Research Scientist at Liulishuo's AI Lab located in Silicon Valley who explores using AI technologies on improving language

education. Prior to Liulishuo, he worked at Educational Testing Service (ETS) from 2008–2017. At ETS, his research focused on the automated assessment of spoken language using speech recognition, natural language processing, and machine learning technologies. Since 2013, he has been working on multimodal signal processing technology for assessing video-based performance tests in areas such as public-speaking. In the 2009 International Conference of Multimodal Interface (ICMI), he won the Outstanding Paper Award sponsored by Google. He received a B.Eng. degree from Tianjin University in China, an M.Sc. degree from the Chinese Academy of Science (CAS), and a Ph.D. degree from Purdue University. All of his degrees are in electrical engineering.

Céline Clavel (Université Paris-Sud) received a Ph.D. in Cognitive Psychology from the Université Paris Ouest Nanterre La Défense in 2007. In September 2010, she became an Assistant Professor at Université Paris-Sud and teaches in the department of Accounting and Management in the Sceaux Institute of Applied Sciences and in the Ergonomic Master. Her research is focused on the emotional process in a virtual or real social interaction context and on the multi-user multimodal interactions in Collaborative Virtual Environments (CVEs). Her main research interest is to specify the psychology models to computer science applications and evaluate their contributions and/or study their impacts on behavior.

Jeffrey Cohn , Ph.D. (University of Pittsburg and Carnegie Mellon University), is Professor of Psychology and Psychiatry at the University of Pittsburgh and Adjunct Professor at the Robotics Institute, Carnegie Mellon University. He leads interdisciplinary and inter-institutional efforts to develop advanced methods of automatic analysis and synthesis of face and body movement and applies them to research in human emotion, communication, psychopathology, and biomedicine. His research has been supported by grants from the U.S. National Institutes of Health and U.S. National Science Foundation, among other sponsors. He chairs the Steering Committee of the IEEE International Conference on Automatic Face and Gesture Recognition (FG) and has served as General Chair of international conferences on automatic face and gesture recognition, affective computing, and multimodal interfaces.

Matthieu Courgeon (École Nationale d'Ingénieurs de Brest) defended his Ph.D. thesis in 2011 on affective computing and interactive autonomous virtual characters. He is the creator of the Multimodal Affective and Reactive Characters toolkit, used by several research teams around the world. His research area spans interactive expressive artificial humans and robots to collaborative immersive virtual reality. He

earned a Best Paper Award at Ubicomp 2013 for his work with expressive virtual humans with the Affective Computing team at the MIT Medialab.

Dr. Nicholas Cummins (University of Augsburg) is a habilitation candidate at the Chair of Embedded Intelligence for Health Care and Wellbeing at the University of Augsburg. He received his Ph.D. in Electrical Engineering from UNSW Australia in February 2016. He is currently involved in the Horizon 2020 projects DE-ENIGMA, RADAR-CNS, and TAPAS. His current research interests include multisenory signal analysis, affective computing, and computer audition with a particular focus on the understanding and analysis of different health states. He has (co)authored over 50 conference and journal papers (over 400 citations, h-index 12). Dr. Cummins is a reviewer for IEEE, ACM, and ISCA journals and conferences, as well as serving on their program and organizational committees. He is a member of ACM, ISCA, IEEE, and the IET.

Li Deng (Citadel) has been the Chief Artificial Intelligence Officer of Citadel since May 2017. Prior to Citadel, he was the Chief Scientist of AI, the founder of Deep Learning Technology Center, and Partner Research Manager at Microsoft (2000–2017). Prior to Microsoft, he was a tenured full professor at the University of Waterloo and held teaching and research positions at Massachusetts Institute of Technology (1992–1993), Advanced Telecommunications Research Institute (1997–1998), and HK University of Science and Technology (1995). He has been a Fellow of the IEEE since 2004, a Fellow of the Acoustical Society of America since 1993, and a Fellow of the ISCA since 2011. He has also been an Affiliate Professor at University of Washington, Seattle, since 2000. He was elected to the Board of Governors of the IEEE Signal Processing Society and served as editor-in-chief of the *IEEE Signal Processing Magazine* and IEEE/ACM *Transactions on Audio, Speech, and Language Processing* from 2008–2014, for which he received the IEEE SPS Meritorious Service Award. In recognition of his pioneering work on disrupting the speech recognition industry using large-scale deep learning, he received the 2015 IEEE SPS Technical Achievement Award for "Outstanding Contributions to Automatic Speech Recognition and Deep Learning." He also received numerous best paper and patent awards for contributions to artificial intelligence, machine learning, information retrieval, multimedia signal processing, speech processing and recognition, and human language technology. He is an author or co-author of six technical books on deep learning, speech processing, pattern recognition and machine learning, and natural language processing.

Sidney D'Mello (University of Colorado Boulder) is an Associate Professor in the Institute of Cognitive Science and Department of Computer Science at the University

of Colorado Boulder. He is interested in the dynamic interplay between cognition and emotion while individuals and groups engage in complex real-world tasks. He applies insights gleaned from this basic research program to develop intelligent technologies that help people achieve their fullest potential by coordinating what they think and feel with what they know and do. D'Mello has co-edited 6 books and published over 220 journal papers, book chapters, and conference proceedings (13 of these have received awards). His work has been funded by numerous grants and he serves or has served as associate editor for four journals and on the editorial boards for six others while also playing leadership roles in several professional organizations.

Julien Epps (UNSW Sydney and Data61, CSIRO) is an Associate Professor of Signal Processing with UNSW Sydney and a Contributed Principal Researcher at Data61, CSIRO. He has authored or co-authored more than 200 publications and 4 patents, mainly on topics related to emotion and mental state recognition of speech and behavioral signals. He is serving as an Associate Editor for *IEEE Transactions on Affective Computing* and *Frontiers in ICT* (the Human-Media Interaction and Psychology sections), and he recently served as a member of the Advisory Board of the ACM International Conference on Multimodal Interaction. He has delivered invited tutorials on topics related to this book for major conferences, including INTERSPEECH 2014 and 2015 and APSIPA 2010, and invited keynotes for the 4th International Workshop on Audio-Visual Emotion Challenge (part of ACM Multimedia 2014) and the 4th International Workshop on Context-Based Affect Recognition (part of AAAC/IEEE Affective Computing and Intelligent Interaction 2017).

Marc Ernst (Ulm University) heads the Department of Applied Cognitive Psychology at Ulm University. He studied physics in Heidelberg and Frankfurt/Main. In 2000, he received his Ph.D. from the Eberhard-Karls-University Tübingen for investigations into the human visuomotor behavior that he conducted at the Max Planck Institute for Biological Cybernetics. For this work, he was awarded the Attempto-Prize from the University of Tübingen and the Otto-Hahn-Medaille from the Max Planck Society. He was a research associate at the University of California, Berkeley working with Prof. Martin Banks on psychophysical experiments and computational models investigating the integration of visual-haptic information before returning to the MPI in Tübingen and becoming principle investigator of the Sensorimotor Lab in the Department of Professor Heinrich Bülthoff. In 2007, he became leader of the Max Planck Research Group on Human Multisensory Perception and Action, before joining the University of Bielefeld and the Cognitive Interaction Technology Center of Excellence (CITEC) in 2011.

Anna Esposito (Wright State University) received her Laurea degree *summa cum laude* in Information Technology and Computer Science from the Università di Salerno in 1989 with the thesis "The Behavior and Learning of a Deterministic Neural Net" (published in *Complex System*, 6(6), 507–517, 1992). She received her Ph.D. in Applied Mathematics and Computer Science from the Università di Napoli Federico II in 1995. Her Ph.D. thesis "Vowel Height and Consonantal Voicing Effects: Data from Italian" (published in *Phonetica*, 59(4),197–231, 2002) was developed at the MIT Research Laboratory of Electronics (RLE), under the supervision of professor Kenneth N. Stevens. She completed a postdoctoral program at the International Institute for Advanced Scientific Studies (IIASS), and was Assistant Professor in the Department of Physics at Università di Salerno, where she taught classes on cybernetics, neural networks, and speech processing (1996–2000). She was a Research Professor in the Department of Computer Science and Engineering at Wright State University (WSU) (2000–2002). She is currently a Research Affiliate at WSU and Associate Professor in Computer Science in the Department of Psychology at Università della Campania Luigi Vanvitelli. She has authored more than 170 peer-reviewed publications in international journals, books, and conference proceedings and edited or co-edited over 25 international books with Italian, EU, and overseas colleagues.

Yoren Gaffary (Insa Rennes) is an expert research engineer on haptic simulations in virtual environments at Insa Rennes, France. His research interests are in affective computing, haptics, and augmented reality. He received a Master's degree in Information, Learning, and Cognition at Université Paris-Sud. His Ph.D.thesis, also at Université Paris-Sud, concerned affective computing using mediated touch with robotic devices coupled with virtual humans.

Roland Goecke (University of Canberra) is Professor of Affective Computing in the School of Information Technology & Systems in the Faculty of Science & Technology at the University of Canberra. Professor Goecke holds a Master's degree in Computer Science (1998) from the University of Rostock, Germany, and a Ph.D. (2004) in Computer Science from the Australian National University, Canberra, Australia. Prior to joining the University of Canberra in 2008, he worked as a Senior Research Scientist with Seeing Machines, as a Researcher at the NICTA Canberra Research Lab, and as a Research Fellow at the Fraunhofer Institute for Computer Graphics, Germany. His research interests are in affective computing, computational behavior analysis, social signal processing, pattern recognition, computer vision, human-computer interaction, multimodal signal processing, and e-research.

Joseph F. Grafsgaard (University of Colorado Boulder) received a B.A. in Computer Science from the University of Minnesota Twin Cities in 2005, and M.Sc. and Ph.D. degrees in Computer Science from North Carolina State University in 2012 and 2014, respectively. He is currently a postdoctoral research associate at the Institute of Cognitive Science at the University of Colorado Boulder. His research interests include affective computing, advanced learning technologies, detection/modeling of nonverbal behavior and physiology, and multimodal affective interaction. He is a member of the Association for the Advancement of Affective Computing (AAAC), the ACM, the IAIED Society, the IEDM Society, and the IEEE.

Jyoti Joshi (University of Canberra) is a postdoctoral researcher at the University of Waterloo. She earned a Ph.D. at the Human-Centred Computing lab at the University of Canberra, Australia, and is supervised by Prof. Roland Goecke and Prof. Michael Wagner. Before starting her Ph.D., she worked as a research assistant with Vision and Sensing Group, University of Canberra. She also worked as a consultant at a leading EDA company Cadence Design Systems, India prior to coming to Australia. Her current research revolves around applications of pattern recognition, computer vision, and machine learning techniques with a focus on affect-based multimedia analysis.

Gil Keren (University of Passau) is a doctoral student at the University of Passau, Germany. Prior to that, he completed Bachelor's and Master's degrees in Psychology and Mathematics at Ben Gurion University, Israel. He conducts research and publishes academic papers on the topics of artificial neural networks, artificial intelligence, and models of human cognition.

Jean-Claude Martin (Université Paris-Sud) is first-class Full Professor of Computer Science at Université Paris-Sud. He is the head of the pluridisciplinary Cognition Perception Use research group at LIMSI-CNRS. He conducts research on the sensory-motor bases of social cognition in humans and in multimodal interfaces, such as expressive virtual agents and social robots. He considers several application areas related to social skills training: autism; job and medical interviews; virtual coaches; leadership and teamwork; stress management; and sports and e-Health. He is the Editor-in-Chief of the Springer *Journal on Multimodal User Interfaces* (JMUI). He has been involved in several projects about how we perceive (in)congruent blends of expressions of emotions in several modalities. He supervised and co-supervised 14 defended Ph.D. theses.

Rada Mihalcea (University of Michigan) is a Professor in the Computer Science and Engineering department at the University of Michigan. Her research interests

are in computational linguistics with a focus on lexical semantics, computational social sciences, and conversational interfaces. She serves or has served on the editorial boards of the *Journals of Computational Linguistics*, *Language Resources and Evaluations*, *Natural Language Engineering*, *Research in Language in Computation*, *IEEE Transactions on Affective Computing*, and *Transactions of the Association for Computational Linguistics*. She was a program co-chair for the Conferences of the Association for Computational Linguistics (2011) and the Empirical Methods in Natural Language Processing (2009), and general chair for the Conference of the North American Chapter of the Association for Computational Linguistics (2015). She is the recipient of a National Science Foundation CAREER award (2008) and a Presidential Early Career Award for Scientists and Engineers (2009).

Louis-Philippe Morency (Carnegie Mellon University) is Assistant Professor in the Language Technology Institute at Carnegie Mellon University, where he leads the Multimodal Communication and Machine Learning Laboratory (MultiComp Lab). He was formerly a research assistant professor in the Computer Sciences Department at the University of Southern California and a research scientist at the USC Institute for Creative Technologies. Professor Morency received his Ph.D. and Master's degrees from MIT Computer Science and Artificial Intelligence Laboratory. His research focuses on building the computational foundations to enable computers with the abilities to analyze, recognize, and predict subtle human communicative behaviors during social interactions. In particular, Professor Morency was lead co-investigator for the multi-institution effort that created SimSensei and MultiSense, two technologies to automatically assess nonverbal behavior indicators of psychological distress. He is currently chair of the advisory committee for ACM International Conference on Multimodal Interaction and associate editor at *IEEE Transactions on Affective Computing*.

Amr El-Desoky Mousa (Apple Inc.) received his Ph.D. in 2014 from the Chair of Human Language Technology and Pattern Recognition at RWTH Aachen University, Germany, where his research was focused on automatic speech recognition. In 2014–2015, he worked as a postdoctoral researcher at the Machine Intelligence and Signal Processing Group at the Technical University of Munich, Germany. In 2016–2017, he worked as a Research and Teaching Associate at the Chair of Complex and Intelligent Systems, University of Passau, Germany. In June 2017, he moved to Apple Inc. to work as a senior Machine Learning Engineer taking part in the research and development related to Siri, one of the most well-known speech-based Intelligent Assistants in the world. His main research interests include deep learn-

ing, large vocabulary continuous speech recognition, language modeling, acoustic modeling, and natural language processing.

Xavier Ochoa (Escuela Superior Politécnica del Litoral) is currently a Full Professor in the Faculty of Electrical and Computer Engineering at Escuela Superior Politécnica del Litoral (ESPOL) in Guayaquil, Ecuador. He directs the Information Technology Center (CTI) and the research group on Teaching and Learning Technologies (TEA) at ESPOL. He is currently the Vice President of the Society for the Research in Learning Analytics (SoLAR), member of the coordination team of the Latin American Community on Learning Technologies (LACLO), and president of the Latin American Open Textbooks Initiative (LATIn). He is editor of the *Journal of Learning Analytics* and member of the Editorial Board of the *IEEE Transactions on Learning Technologies*. He coordinates several regional and international projects in the field of learning technologies. His main research interests revolve around learning analytics, multimodal learning analytics, and data science.

Yannis Panagakis (Middlesex University and Imperial College London) is an assistant professor at Middlesex University London and research faculty at the Department of Computing, Imperial College London. His research interests lie in machine learning and its interface with signal processing, high-dimensional statistics, and computational optimization. Specifically, Yannis is working on models and algorithms for robust and efficient learning from high-dimensional data and signals representing audio, visual, affective, and social phenomena. He received his M.Sc. and Ph.D. from the Department of Informatics at Aristotle University of Thessaloniki and his B.Sc. in Informatics and Telecommunication from the University of Athens, Greece. Yannis has been awarded the prestigious Marie-Curie Fellowship, among various scholarships and awards for his studies and research.

Maja Pantić (Imperial College London) is a Professor of Affective and Behavioral Computing and leader of the i·BUG group at Imperial College London, working on machine analysis of human non-verbal behavior and its applications to human-computer, human-robot, and computer-mediated human-human interaction. Professor Pantić has published more than 250 technical papers in the areas of machine analysis of facial expressions, machine analysis of human body gestures, audiovisual analysis of emotions and social signals, and human-centered machine interfaces. She has served as the keynote speaker, chair and co-chair, and an organization or program committee member at numerous conferences in her areas of expertise. She received a B.Sc. from Delft University in 1995, followed by an M.Sc.

in Artificial Intelligence in 1997. Pantić earned a Ph.D. in 2001 at the Delft University of Technology, with a thesis on facial expression analysis by computational intelligence techniques.

Veronica Perez-Rosas (University of Michigan) is an Assistant Research Scientist of Computer Science and Engineering at the University of Michigan. Her main research interest areas are natural language processing (NLP), computational linguistics, applied machine learning, and multimodal interaction. Her work examines human communication to build computational models able to analyze, recognize, and predict affect-related behaviors during social interaction. She has developed methods that make use of multimodal information present during the social interaction (verbal and non-verbal) and combine data-driven approaches with linguistic and psycho-linguistic knowledge to build novel solutions to diverse NLP problems such as sentiment analysis and deception detection. She has publications in reputable journals and conferences, including IEEE, ACM, and ACL. She also served as a reviewer for *IEEE Transactions* and Elsevier journals and served as a program committee member for multiple international conferences in the NLP community.

Olivier Pietquin (Google) completed his Ph.D. under the Faculty of Engineering at l'Universiste de Mons, Belgium, and the University of Sheffield, UK. He then did a postdoctoral program with Philips in Germany before joining the Ecole Supérieure d'Electricité, France, in 2005, as an Associate Professor and, later, Professor. He headed the computer science program of the UMI GeorgiaTech-CNRS (joint lab in Metz) and also worked with the INSERM IADI team. In 2013, he moved to the University of Lille, France, as a Full Professor and joined the Inria SequeL (Sequential Learning) team. Since 2016, Olivier has been on leave with Google, first with DeepMind in London, and then with Brain in Paris. His current research is about direct and inverse reinforcement learning, learning from demonstrations, and applications to human-machine interaction.

Fabio Ramos (University of Sydney) is an Associate Professor in Machine Learning and Robotics at the School of Information Technologies and co-Director of the Centre for Translational Data Science at the University of Sydney. He received B.Sc. and M.Sc. degrees in Mechatronics Engineering at the University of Sao Paulo, Brazil, in 2001 and 2003, respectively, and a Ph.D. at the University of Sydney, Australia, in 2008. He was an ARC postdoctoral fellow from 2008–2010, and an ARC DECRA fellow from 2012–2014. He has authored over 130 peer-reviewed publications and received numerous awards. His research focuses on statistical machine learning

techniques for large-scale data fusion with applications in robotics, mining, environmental monitoring, and healthcare.

Arun Ross (Michigan State University) is a Professor in the Department of Computer Science and Engineering and the Director of the i-PRoBe Lab at Michigan State University (MSU). He was a faculty member at West Virginia University (WVU) from 2003–2012 and the Assistant Site Director of the NSF Center for Identification Technology and Research (CITeR) from 2010–2012. Arun received his B.E. (Hons.) in Computer Science from the Birla Institute of Technology and Science, Pilani, India, and his M.S. and Ph.D. in Computer Science and Engineering from MSU. He coauthored the textbook *Introduction to Biometrics* and the monograph *Handbook of Multibiometrics*, and coedited the *Handbook of Biometrics*. He received the IAPR JK Aggarwal Prize, the IAPR Young Biometrics Investigator Award, and the NSF CAREER Award and was an invited speaker at the Frontiers of Science Symposium organized by the National Academy of Sciences in 2006. He also received the 2005 Biennial Best Paper Award from the *Pattern Recognition Journal* and the Five-Year Highly Cited BTAS 2009 Paper Award. Arun served as a panelist at an event organized by the United Nations Counter-Terrorism Committee (CTC) at the UN Headquarters in 2013. He was an Associate Editor of *IEEE Transactions on Information Forensics and Security* (2009–2013) and *IEEE Transactions on Image Processing* (2008–2013). He currently serves as Associate Editor of *IEEE Transactions on Circuits and Systems for Video Technology*, Senior Area Editor of *IEEE Transactions on Image Processing*, Area Editor of the *Computer Vision and Image Understanding Journal*, Associate Editor of the *Image and Vision Computing Journal*, and Chair of the IAPR TC4 on Biometrics.

Ognjen (Oggi) Rudovic (MIT) is a Postdoctoral, Marie Curie, Fellow in the Affective Computing Group at MIT Media Lab, working on personalized machine learning for affective robots and analysis of human data. His background is in automatic control theory, computer vision, artificial intelligence, and machine learning. In 2014, he received a Ph.D. from Imperial College London, UK, where he worked on machine learning and computer vision models for automated analysis of human facial behavior. His current research focus is on developing machine-learning-driven assistive technologies for individuals on the autism spectrum.

Stefan Scherer (Embodied, Inc. and USC) is the CTO of Embodied, Inc. and a Research Assistant Professor in the Department of Computer Science at the University of Southern California (USC) and the USC Institute for Creative Technologies (ICT), where he leads research projects funded by National Science Foundation and the 765Army Research Laboratory. Stefan Scherer directs the lab for Behavior Analytics

and Machine Learning and is the Associate Director of Neural Information Processing at the ICT. He received the degree of Dr. rer. nat. from the faculty of Engineering and Computer Science at Ulm University in Germany with the grade *summa cum laude* (i.e. with distinction) in 2011. His research aims to automatically identify characterize, model, and synthesize individuals' multimodal nonverbal behavior within both human-machine as well as machine-mediated human-human interaction. His research was recently featured in the *Economist,* the *Atlantic,* and the *Guardian,* and was awarded a number of best paper awards in renowned international conferences. His work is focused on machine learning, multimodal signal processing, and affective computing. Stefan Scherer serves as an Associate Editor of the Journal *IEEE Transactions on Affective Computing* and is the Vice Chair of the IAPR Technical Committee 9 on Pattern Recognition in Human-Machine Interaction.

Mohamed Yacine Tsalamlal (LIMSI-CNRS Lab) received a Master's degree in Human-Machine Systems engineering from Université de Lorraine, France, and a Ph.D. degree in Computer Science from the Université Paris-Saclay. He is a postdoctoral researcher with the Architectures and Models for Interaction Group at the LIMSI-CNRS Lab. His research interests include the study of tactile interaction techniques for mediated social communication. He exploits multiple approaches (mechanics, robotics, experimental phycology) to help in the design of efficient multimodal affective interaction systems.

Alessandro Vinciarelli (University of Glasgow) is a Full Professor in the School of Computing Science at the University of Glasgow and an associated academic at the Institute of Neuroscience and Psychology. His main research interest is social signal processing, the domain aimed at modeling, analyzing, and synthesizing nonverbal behavior in human-human and human-machine interactions. He has published over 130 works and has been Principal Investigator or co-Principle Investigator of more than 15 national and international projects. He has chaired and co-chaired more than 30 international events and is the co-founder of Klewel, a knowledge management company recognized with several awards.

Johannes Wagner (University of Augsburg) graduated with a Master of Science in Informatics and Multimedia in 2007 and received a doctoral degree for his study "Online Systems for Multimodal Behaviour Analysis" in 2016. He is currently employed as a research assistant at the lab of Human Centered Multimedia (HCM) at he University of Augsburg and has been working on several European projects (Humaine, Callas, Ilhaire, CEEDs, Kristina, AriaValuspa). His main research focus

is the integration of Social Signal Processing (SSP) in real-life applications. He is the founder of the Social Signal Interpretation (SSI) framework, a general framework for the integration of multiple sensors into multimedia applications.

Yang Wang (DATA61, CSIRO) is a Principal Research Scientist of Analytics in DATA61, CSIRO. He received his Ph.D. in Computer Science from the National University of Singapore in 2004. His research interests include machine learning and information fusion techniques and their applications to intelligent infrastructure, cognitive, and emotive computing.

Kun Yu (DATA61, CSIRO) is a Research Scientist in DATA61, CSIRO. He received his Ph.D. in Electrical Engineering from University of New South Wales, Australia. His research interests include human-computer interaction, cognitive load examination, human trust calibration, and multimodal human-machine interfaces.

Stefanos Zafeiriou (Imperial College London) is a Reader in Machine Learning and Computer Vision with the Department of Computing, Imperial College London and a Distinguishing Research Fellow with the University of Oulu in the Finish Distinguishing Professor Program. In 2011, he was a recipient of the Prestigious Junior Research Fellowships from Imperial College London to start his own independent research group. He was the recipient of the President's Medal for Excellence in Research Supervision for 2016. He has coauthored more than 55 papers, mainly on novel statistical machine learning methodologies applied to computer vision problems, such as 2-D/3-D face analysis, deformable object fitting and tracking, shape from shading, and human behavior analysis, published in the most prestigious journals in his field of research and many top conferences, such as CVPR, ICCV, ECCV, and ICML.

Jianlong Zhou (DATA61, CSIRO) is a Senior Research Scientist of Analytics in DATA61, CSIRO. He earned a Ph.D. in Computer Science from the University of Sydney, Australia. His research interests include transparent machine learning, human-computer interaction, cognitive computing, visualization, spatial augmented reality, and related applications.

Volume 2 Glossary

Accumulative GSR refers to the summation of GSR values over the task time. If the GSR is considered as a continuous signal over time, the accumulative GSR is the integration of GSR values over the task time.

Action units and action descriptors are the smallest visually discriminable facial movements. Action units are movements for which the anatomic basis is known [Cohn, Ambadar, & Ekman, 2007]. They are represented as either binary events (presence vs. absence) or with respect to five levels of ordinal intensity. Action units individually or in combinations can represent nearly all possible facial expressions.

Active Appearance Model (AAM). An AAM is a statistical model of shape and grey-level appearance that can generalize to almost any face [Edwards, Taylor, & Cootes, 1998; Matthews & Baker, 2004]. An AAM seeks to find the model parameters that can generate a synthetic image as close as possible to a target image [Cootes & Taylor, 2004]. AAMs are learned from hand-labelled training data.

Affect. Broad term encompassing constructs such as emotions, moods, and feelings. Is not the same as personality, motivation, and other related terms.

Affect and social signals can be described as temporal patterns of a multiplicity of non-verbal behavioral cues which last for a short time [Vinciarelli et al. 2009] and are expressed as changes in neuromuscular and physiological activity [Vinciarelli et al. 2008a]. Sometimes, we consciously draw on affect and social signals to alter the interpretation of a situation, e.g. by saying something in a sarcastic voice to signal that we actually mean the opposite. At another time, we use them without being aware of it, e.g., by showing sympathy towards our counterpart by mimicking his or her verbal and nonverbal expressions. See Chapter 7 of this volume for an overview of affective and social signals for which automated recognition approaches have been proposed.

Affect annotation. The process of assigning affective labels (e.g., bored, confused, aroused) or values (e.g., arousal = 5) to data (e.g., video, audio, text).

Affective computing. Computing techniques and applications involving emotion or affect.

Affective computing and social signal processing aims at perceiving and interpreting nonverbal behavior with a machine by detecting affect and social signals just as humans do among themselves. This will lead to a new generation of computers that is perceived as more natural, efficacious and trustworthy [Vinciarelli et al. 2008b, Vinciarelli et al. 2009], and it makes room for a more human-like and intuitive interaction [Pantic et al. 2007].

Affective experience-expression link. The relationship between experiencing an affective state (e.g., feeling confused) and expressing it (e.g., displaying a furrowed brow).

Affective ground truth. Objective reality involving the "true" affective state. Is a misleading term for psychological constructs like affect.

Alignment A third challenge is to identify the direct relations between (sub)-elements from two or more different modalities. For example, we may want to align the steps in a recipe to a video showing the dish being made. To tackle this challenge we need to measure similarity between different modalities and deal with possible long-range dependencies and ambiguities.

Classifier: in pattern recognition and machine learning, it is a function that maps an object of interest (represented through a set of physical measurements called features) into one of the classes or categories such an object can belong to (the number of classes or categories is finite).

Bag-of-words (BoW) is a data-driven algorithm for summarizing large volumes of features. It can be thought of as a histogram whose bins are determined by partitions (or clusters) of the feature space.

Canonical Correlation Analysis (CCA) is a tool to infer information based on cross-covariance matrices. It can be used to identify correlation across heterogeneous modalities or sensor signals. Let each modality be represented by a feature vector and let us assume there is correlation across these, such as is in audiovisual speech recognition. Then, CCA will identify linear combinations of the individual features with maximum correlation amongst each other.

Classifier Combination: in pattern recognition and machine learning, it is a body of methodologies aimed at jointly using multiple classifiers to achieve a collective

performance higher—to a statistically significant extent—than the individual performance of any individual classifier.

Classifier Diversity: in a set of classifiers that are being combined combined, the diversity is the tendency of different classifiers to have different performance in different regions of the input space.

Cognitive load is a multidimensional construct that refers to the momentary working memory load experienced by a person while performing a cognitive task. It can be increased by a wide variety of factors, including the difficulty of a task, the materials or tools used (e.g., computer interface), the situational context (e.g., distracting vs. quiet setting), the social context (e.g., working individually, vs. jointly with a group), a person's expertise in the domain, a person's physiological stress level, and so forth. Cognitive Load Theory describes cognitive load as having three components—intrinsic load, extraneous load, and germane load [Sweller et al. 2011]. For a detailed discussion of the dynamic and often non-linear interplay between cognitive load and domain expertise, including how cognitive load can either expand or minimize the performance gap between low- and high-performing students, see Oviatt [2013].

Cognitive load measurement refers to the methods to quantitatively discriminate the different levels of cognitive load experienced by the user. Usually the cognitive load is induced with varied task difficulty (i.e. the extraneous load is manipulated), and the methods to discriminate cognitive load include subjective methods, performance-based methods, physiological methods and behavioral methods.

Co-learning A fifth challenge is to transfer knowledge between modalities, their representation, and their predictive models. This is exemplified by algorithms of co-training, conceptual grounding, and zero shot learning. Co-learning explores how knowledge learning from one modality can help a computational model trained on a different modality. This challenge is particularly relevant when one of the modalities has limited resources (e.g., annotated data).

Communication: process between two or more agents aimed at the exchange of information or at the mutual modification of beliefs, shared or individual.

Confidence measure is the information on the assumed certainty of a decision made by a machine learning algorithm.

Congruent expressions of affects. A multimodal combination is said to involve congruent expressions of affects if each modality is conveying the same affect

in terms of the category or the dimension (e.g., the facial expressions express anger and the hand gestures also express anger).

Cooperative learning in machine learning is a combination of active learning and semi-supervised learning. In the semi-supervised learning part, the machine learning algorithm labels unlabeled data based on its own previously learned model. In the active learning part, it identifies which unlabeled data is most important to be labeled by humans. Usually, cooperative learning tries to minimize the amount of data to be labeled by humans while maximizing the gain in accuracy of a learning algorithm. This can be based on confidence measures such that the machine labels unlabeled data itself as long as it is sufficiently confident in its decisions. It asks humans for help only where its confidence is insufficient, but the data seem to be highly informative.

Construct. A conceptual variable that cannot be directly observed (e.g., intelligence, personality).

Continuous or discrete (i.e., categorical) representation refer to the modeling of a user state or trait. As an example, the age of a user can be modeled as continuum such as the age in years. As opposed to this, a discretized representation would be broader age classes such as "young," "adult's", and "elderly." In addition, the time can be discretized or continuous (in fact, it is always discretized in some respect—at least by the sample rate of the digitized sampling of the sensor signals). However, one would speak of continuous measurement if processing is delivering a continuous output stream on a (short) frame-by-frame basis rather than an asynchronous processing of (larger) segments or chunks of the signal such as per spoken word or per body gesture.

A **Convolutional Neural Network (CNN)** is a neural network that contains one or more convolutional layers. A *convolutional layer* is a layer that processes an image (or any other data comprised of points with a notion of distance between these points, such as an audio signal) by convolving it with a number of kernels.

A **correlation** is a single number that describes the degree of relationship between two variables (signals). It most often refers to how close two variables are to having a linear relationship with each other.

A **dense layer** is the basic type of layer in a neural network. The layer takes a one-dimensional vector as input and transforms it to another one-dimensional vector by multiplying it by a *weight matrix* and adding a *bias vector*.

Depression refers broadly to the persistence over an extended period of time of several of the following symptoms: lowered mood, interest, or pleasure; psychomo-

tor retardation; psychomotor agitation; diminished ability to think/concentrate; increased indecisiveness; fatigue or loss of energy; insomnia; hypersomnia; significant weight loss or weight gain; feelings of worthlessness or excessive guilt; and recurrent thoughts of death or recurrent suicidal ideation. It is important to note that there are multiple definitions of depression (see references in Section 12.2).

Domain adaptation refers to machine learning methods that learn from a source data distribution a well performing model on a different (but related) target data distribution.

Domain expertise refers to the level of working knowledge and problem-solving competence within a specific subject matter, such as algebra. It is a relatively stable state that influences how a person perceives and strategizes solving a problem. For the same task, a more expert student will group elements within it into higher-level patterns, or perceive the problem in a more integrated manner. A person's level of domain expertise influences a variety of problem-solving behaviors (e.g., fluency, effort expended, accuracy). A more expert person also will experience lower cognitive load when working on the same problem as a less expert person. Most people experience domain expertise in some subjects. This everyday experience of domain expertise is distinct from elite expert performance that occurs in a small minority of virtuosos or prodigies, which can take a decade or lifetime to achieve.

Dynamic Time Warping (DTW) is a machine learning algorithm to align two time series such as feature vectors extracted over time based on similarity measurement. This similarity is often measured by distance measures such as Euclidean distance or based on correlation such as when aligning heterogenous modalities. A classical application example is speech recognition, where words spoken at different speed are aligned in time to measure their similarity. DTW aims at a maximized match between the two observation sequences usually based on local and global alignment path search restrictions.

In **early combination**, the inputs from all the different modalities are concatenated and fed to a single model. In *late combination*, for each modality there is a separate model that makes a prediction based on its modality, and these model predictions are later fused by a combining model.

Early fusion models are models for processing multimodal or multisensorial data, in which a model is processing the concatenation of all the data representations from the different modalities. In *late fusion* models, there is a unimodal model

for each modality, and the outputs of all unimodal models are then combined to a final prediction based on all modalities.

Encoder-decoder architectures in deep learning start with an encoder neural network which—based on its input—usually outputs a feature map or vector. The second part—the decoder—is a further network that—based on the feature vector from the encoder—provides the closest match either to the input or an intended output. The decoder is in most cases employing the same network structure but in opposite orientation. Usually, the training is carried out unsupervised data, i.e., without labels. The target for learning is to minimize the reconstruction error, i.e., the delta between the input to the encoder and the output of the decoder. A typical application is to use encoder-decoder architectures for sequence-to-sequence mapping, such as in machine translation where the encoder is trained on sequences (phrases) in one language and the decoder is trained to map its representation to a sequence (phrase) in another language.

An **ensemble** is a set of models and we want the models in the set to differ in their predictions so that they make different errors. If we consider the space defined by the three dimensions that define a model as we defined above, the idea is to sample smartly from that space of learners. We want the individual models to be as accurate as possible individually, and at the same time, we want them to complement each other. How these two criteria affect the accuracy of the ensemble depends on the way we do the combination.

From another perspective, we can view each particular model as one noisy estimate to the real (unknown) underlying problem. For example, in a classification task, each base classifier, depending on its model, hyper-parameters, and input features, learns one noisy estimator to the real discriminant. In such a perspective, the ensemble approach corresponds to constructing a final estimator from these noisy estimators—for example, voting corresponds to averaging them.

When the different models use inputs in different modalities, there are three ways in which the predictions of models can be combined, namely, early, late, and intermediate combination/integration/fusion.

Extraneous load refers to the level of working memory load that a person experiences due to the properties of materials or computer interfaces they are using [Oviatt 2017].

FACS refers to the Facial Action Coding System [Ekman & Friesen, 1978; Ekman, Friesen, & Hager, 2002]. FACS describes facial activity in terms of anatomically

based action units (AUs). Depending on the version of FACS, there are 33 to 44 AUs and a large number of additional "action descriptors" and other movements.

Feature-level multimodal fusion. The process of integrating features from different modalities using diverse methodologies such as concatenating the features together (early fusion) or combining the models obtained from each modality at decision level (late fusion).

Fusion A fourth challenge is to join information from two or more modalities to perform a prediction. For example, for audio-visual speech recognition, the visual description of the lip motion is fused with the speech signal to predict spoken words. The information coming from different modalities may have varying predictive power and noise topology, with possibly missing data in at least one of the modalities.

Galvanic Skin Response (GSR) refers to galvanic skin response which is a measure of the conductivity of human skin, and can provide an indication of changes in the human sympathetic nervous system during the cognitive task time.

Gaussian mixture models (GMMs) are probability density functions comprising a weighted sum of individual Gaussian components, each with their own mean and covariance. They are commonly employed to compactly characterize arbitrary distributions (e.g. of features) that are not well-fitted by a single Gaussian.

Germane load refers to the level of a person's effort and activity compatible with mastering new domain content during learning. It pertains to the cognitive resources dedicated to constructing new schema in long-term memory.

Gross errors refer to non-Gaussian noise of large magnitude. Gross errors are often in abundance in audio-visual data due to incorrect localisation and tracking, presence of partial occlusions, enviromental noise etc.

Hand-crafted features refer to features developed to extract a specific type of information, usually as part of a hypothesis-driven research study. By contrast, data-driven features are those extracted automatically from raw signal data by algorithms (e.g., neural networks), whose physical interpretation often cannot easily be described.

Incongruent expressions of affects. A multimodal combination is said to involve incongruent expressions of affects if the combined modalities are conveying different affects in terms of the category or the dimension (e.g., the facial expressions express joy, while hand gestures express anger). Such combinations are also called **blends of emotions**. They might occur even in a single modality

such as the facial expressions (e.g., the upper part of the face may express a certain emotion, while the bottom part of the face conveys a different emotion).

Researchers exploring the perception of human (or computer-generated) multimodal expressions of affects usually consider the following attributes related to perception and affects that are impacted by or do impact multimodal perception:

- Recognition rate
- Reaction time
- Affect categories
- Affect dimensions
- Multimodal integration patterns
- Inter-individual differences and personality
- Timing: synchrony vs. sequential presentation of the signals in different modalities
- Modality dominance
- Context: environment and task related information (e.g., food or violent scenes, and associated applications for specific users with food disorders or PTSD)
- Task difficulty

In **intermediate combination**, each modality is first processed to get a more abstract representation and then all such representations from different modalities are fed together to a single model. This processing can be in the form of a **kernel function**, which is a measure of similarity, and such an approach is called **multiple kernel learning**. Or the intermediate processing may be done by one or more layers of a neural network, and such an approach corresponds to a **deep neural network**.

The level of combination depends on the level we expect to see a **dependency** between the inputs in different modalities. Early combination assumes a dependency at the lowest level of input features; intermediate combination assumes a dependency at a more abstract or semantic level that is extracted after some processing of the raw input; late combination assumes no dependency in the input but only at the level of decisions.

Intrinsic load is the inherent difficulty level and related working memory load associated with the material being processed during a user's primary task.

Learning analytics involves the collection, analysis, and reporting of data about learners, including their activities and contexts of learning, in order to understand and provide better support for effective learning. First-generation learning analytics focused exclusively on computer-mediated learning using keyboard-and-mouse computer interfaces. This limited analyses to click-stream activity patterns and linguistic content. Early learning analytic techniques mainly have been applied to managing educational systems (e.g., attendance tracking, tracking work completion), advising students based on their individual profiles (e.g., courses taken, grades received, time spent in learning activities), and improving educational technologies. Learning analytics data typically are summarized on dashboards for educators, such as teachers or administrators.

Leave-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.

The user **(long-term) traits** include biological trait primitives (e.g., age, gender, height, weight), cultural trait primitives in the sense of group/ethnicity membership (e.g,. culture, race, social class, or linguistic concepts such as dialect or first language), personality traits (e.g., the "OCEAN big five" dimensions openness, conscientiousness, extraversion, agreeableness, and neuroticism or likability), and traits that constitute subject idiosyncrasy, i.e., ID.

A **longer-term state** can subsume (partly self-induced) non-permanent, yet longer-term states (e.g., sleepiness, intoxication, mood such as depression (see also Chapter 12 the health state such as having a flu), structural (behavioral, interactional, social) signals (e.g., role in dyads and groups, friendship and identity, positive/negative attitude, intimacy, interest, politeness), and (non-verbal) social signals (see Chapters 7 and 8 and discrepant signals (e.g., deception (see also Chapter 13) irony, sarcasm, sincerity).

Longitudinal data refers to multiple recordings of the same type from the same individual at different points in time, between which it is likely that the individual's state (e.g., depression score) has changed.

$L()$ is the **loss function** that measures how far the prediction $g(x^t|\theta)$ is from the desired value r^t. The complexity of this optimization problem depends on the particular $g()$ and $L()$. Different learning algorithms in the machine learning literature differ either in the model they use, the loss function they employ, or the how the optimization problem is solved.

This step above optimizes the parameters given a model. Each model has an inductive bias that is, it comes with a set of assumptions about the data and the model is accurate if its assumptions match the characteristics of the data. This implies that we also need a process of *model selection* where we optimize the model structure. This model structure depends on dimensions such as (i) the learning algorithm, (ii) the hyper-parameters of the model (that define model complexity), and (iii) the input features and representation, or modality. Each model corresponds to one particular combination of these dimensions.

In **machine learning**, the **learner** is a model that takes an input x and learns to give out the correct output y. In **pattern recognition**, typically we have a classification task where y is a class code; for example in face recognition, x is the face image and y is the index of the person whose face it is we are classifying.

In building a learner, we start from a data set $\mathcal{X} = \{x^t, r^t\}, t = 1, \ldots, N$ that contains training pairs of instances x^t and the desired output values r^t (e.g., class labels) for them. We assume that there is a dependency between x and r but that it is unknown—If it were known, there would be no need to do any learning and we would just write down the code for the mapping.

Typically, x^t is not enough to uniquely identify r^t; we call x^t the observables and there may also be unobservables that affect r^t and we model their effect as noise. This implies that each training pair gives us only a limited amount of information. Another related problem is that in most applications, x has a very high dimensionality and our training set samples this high dimensional space very sparsely.

Our prediction is given by our predictor $g(x^t|\theta)$ where $g()$ is the model and θ is its set of parameters. Learning corresponds to finding that best θ^* that makes our predictions as close as possible to the desired values on the training set:

$$\theta^* = \arg\min_{\theta} \sum_{t=1}^{N} L(r^t, g(x^t|\theta))$$

Mel frequency cepstral coefficients (MFCCs) are features that compactly represent the short-term speech spectrum, including formant information, and are widely used to characterize both spoken content (for automatic speech recognition) and speaker-specific qualities (for automatic speaker verification). Briefly, a mel-scale frequency-domain filterbank is applied to the spectrum to obtain mel filterbank energies, the log of which is transformed to a lower-dimensional representation using the discrete cosine transform.

Metacognitive awareness involves higher-level self-regulatory behaviors that guide the learning process, such as an individual's awareness of what type of problem they are working on and how to approach solving it, the ability to diagnose error states, or understanding what tools are best suited for a particular task.

Moment of insight refers to one of several phases during the process of problem solving. It involves the interval of time immediately before and after a person consciously realizes the solution to a problem they've been working on. This idea represents what the person believes is the solution, although it may or may not be correct.

Multi-level multimodal learning analytics refers to the different levels of analysis enabled by multimodal data collection. For example, speech and handwriting can be analyzed at the signal, activity pattern, representational, or metacognitive levels. During research on learning, it frequently is valuable to analyze data across behavioral and physiological/neural levels for a deeper understanding of any learning effects. In this regard, multi-level multimodal learning analytics can support a more comprehensive systems-level view of the complex process of learning.

A **multimodal corpus** targets the recording and annotation of multiple communication modalities including speech, hand gesture, facial expression, body posture, etc. Today, most corpora that are multimodal consist of audio-visual data. Other modalities such as 3D body and gaze tracking, or physiological signals are hardly present, but are needed to provide a broader picture of human interaction. The collection of large databases rich of social behavior expressed through a variety of modalities is key to model the complexity of social interaction [Vinciarelli et al. 2012, Eerekoviae 2014].

Multimodal fusion is the process of combining information from multiple modalities, such as audio and video, into a homogenous and consistent representation. Combining affective and social cues across channels is important to resolve situations where social behavior is expressed in a complementary [Zeng et al. 2009] or even contradictory way [Douglas-Cowie et al. 2005]. This also involves a proper modeling of the complex temporal relationships that exist between the diverse channels. It can help to achieve higher precision such as in cognitive load measurement, or better reliability via overcoming the limitations of individual signal or interaction modalities. The fusion can be done at different stages: mid-fusion and late-fusion. Mid-fusion refers to the fusion of features extracted from multimodalities before classifications, while late-fusion is the fusion of classification scores from single modality decisions.

Multimodal learning analytics is an emerging area that analyzes students' natural communication patterns (e.g., speech, writing, gaze, non-verbal behavior), activity patterns (e.g., number of hours spent studying), and physiological and neural patterns (e.g., EEG, fNIRS) in order to predict learning-oriented behaviors during educational activities. These rich data can be analyzed at multiple levels, for example at the signal level (e.g., speech amplitude), activity level (e.g., frequency of speaking), representational level (e.g., linguistic content), and others. These second-generation learning analytic techniques are capable of predicting mental states during the process of learning, in some cases automatically and in real time.

Multimodal or multiview signals are sets of heterogeneous data, captured by different sensors, such as various types of cameras, microphones, and tactile sensors and in different contexts.

Online recognition means that a system is able to detect and analyze affective and social cues on-the-fly from the raw sensor input. Decisions based on the perceived user state need to be made fast enough to allow for a fluent interaction and it is not possible to look ahead in time. Setting up an online system is more complex than processing data offline.

Overfitting is a problem that occurs when the training or estimation of a machine learning method is performed on data with too few training examples relative to the number of parameters to be estimated. The resulting problem is that the method becomes too closely tuned to the training data, and generalizes poorly to unseen test data.

Physiological sensor. A device that uses a transducer and a biological element to collect physiological responses, such as heart rate and skin conductance, and convert them into an electrical signal. The measures obtained with such devices provide quantitative feedback about physiological changes or processes experienced by research subjects.

Prerequisites for learning are precursors for successful learning to occur, which can provide early markers. They include attention to the learning materials, emotional and motivational predisposition to learn, and active engagement with the learning activitites.

A **pseudo-multimodal** approach exploits a modality not only by itself, but in addition to estimate another modality's behavior to replace it. An example is estimating the heart rate from speech parameters and using it alongside (other) speech parameters.

A **Recurrent Neural Network (RNN)** is a neural network that contains one or more recurrent layers. A *recurrent layer* is a layer that takes a sequence x indexed by t and processes it element by a element, while maintaining a *hidden state* for each unit in the layer: $h_t = RNN(h_{t-1}, x_t)$, where h_t is the hidden state at step t, and RNN is a transition function to compute the next hidden state, that depends on the type of hidden layer.

Redundancy: tendency of multiple signals or communication channels to carry the same or widely overlapping information.

Representation A first fundamental challenge is learning how to represent and summarize multimodal data in a way that exploits the complementarity and redundancy of multiple modalities. The heterogeneity of multimodal data makes it challenging to construct such representations. For example, language is often symbolic while audio and visual modalities will be represented as signals.

A **sequence-to-sequence** model is a neural network that processes a sequence as its input and produces another sequence as its output. Example of such models include neural machine translation and end-to-end speech recognition models.

Shared hidden layer is a layer within a neural network which is shared within the topology. For example, different modalities, or different output classes, or even different databases could be trained within parts of the network mostly. In the shared hidden layer, however, they would share neurons by according connections. This can be an important approach to model diverse information types largely independently but provide mutual information exchange at some point in the topology of a neural network.

A **short-term state** includes the mode (e.g., speaking style and voice quality), emotions, and affects (e.g., confidence, stress, frustration, pain, uncertainty, see also Chapters 6 and 8.

Skilled performance involves the acquisition of skilled action patterns, for example when learning to become an expert athlete or musician. Skilled performance also can involve the acquisition of communication skills, as in developing expertise in writing compositions or giving oral presentations. The role of deliberate practice has been emphasized in acquiring skilled performance.

Social Signals: constellations of nonverbal behavioural cues aimed at conveying socially relevant information such as attitudes, personality, intentions, etc.

Social Signal Processing: computing domain aimed at modeling, analysis and synthesis of social signals in human-human and human-machine interactions.

A **softmax layer** is a dense layer followed by the softmax nonlinearity. The softmax nonlinearity takes a one-dimensional vector v of real numbers and normalizes it into a probability distribution, by applying $\text{softmax}(v)_j = \frac{e^{v_j}}{\sum_i e^{v_i}}$, where the sum is over all coordinates of the vector v.

Spatio-temporal features are those which have both a spatial and a time dimension. For example, the intensity of pixels in a video vary both in terms of their position within a given frame (spatial dimension) and in terms of the frame number for a given pixel coordinate (temporal dimension).

Support vector machine (SVM) is a widely used discriminative classification method, which defines a separating hyperplane between two classes of features, which is defined in terms of particular feature instances that are close to the class boundaries, called support vectors.

Support vector regression (SVR) is a commonly used method for multivariate regression, which concentrates on fitting a model by considering only training features that are not very close to the model prediction.

Temporal dynamics of facial expression: rather than being like a single snapshot, facial appearance changes as a function of time. Two main factors affecting temporal dynamics of facial expression is the speed with which they unfold and the changes of their intensity over time.

Transfer learning helps to reuse knowledge gained in one task in another task in machine learning. It can be executed on different levels, such as the feature or model level. For example, a neural network can be trained on a related task to the task of interest at first. Then, the actual task of interest is trained “on top" of this pre-training of the network. Likewise, rather than starting to train the target task of interest based on a random initialization of a network, related data could be used to provide a better starting point.

Transfer of learning refers to students’ ability to recognize parallells and make use of learned information in new contexts, for example to apply learned information outside of the classroom, in contexts not resembling the original framing of the problem, with different people present, and so forth. This requires generalizing the learned information beyond its original concrete context.

Translation A second challenge addresses how to translate (map) data from one modality to another. Not only is the data heterogeneous, but the relationship between modalities is often open-ended or subjective. For example, there exist

a number of *correct* ways to describe an image and one perfect translation may not exist.

T-unit analysis. The analysis of terminable units of language (T-unit), which is the smallest group of words that could be considered as a grammatical sentence, regardless of how is punctuated. T-unit analysis is used extensively to measure the overall complexity of both speech and writing samples and consists mainly on measuring different aspects of their syntactic construction in text such as mean length of the t-units, and number of clauses present in each unit, among others.

User-independent model A model that generalizes to a different set of users beyond those used to develop the model.

Voice quality refers to the type of phonation during voiced speech. Depending on the physical movement of the vocal folds during phonation, the perceived quality of speech can change, even for the same speech sound uttered at the same pitch. Descriptors such as "creaky" and "breathy" are applied to specific modes of vocal fold vibration.

Vowel space area is a term given to the two dimensional area enclosed by lines connecting pairs of vowels in the formant (F1/F2) space.

Zero-shot learning is a method in machine learning to learn a new task without any training examples for this task. An example could be recognizing a new type of object without any visual example but based on a semantic description such as specific features that describe the object.