Deception Detection using Real-life Trial Data

Verónica Pérez-Rosas
Computer Science and Engineering
University of Michigan
Ann Arbor, MI 48109, USA
vrncapr@umich.edu

Mohamed Abouelenien
Computer Science and Engineering
University of Michigan
Ann Arbor, MI 48109, USA
zmohamed@umich.edu

Rada Mihalcea
Computer Science and Engineering
University of Michigan
Ann Arbor, MI 48109, USA
mihalcea@umich.edu

Mihai Burzo
Mechanical Engineering
University of Michigan, Flint
Flint, MI 48502, USA
mburzo@umich.edu

ABSTRACT

Hearings of witnesses and defendants play a crucial role when reaching court trial decisions. Given the high-stake nature of trial outcomes, implementing accurate and effective computational methods to evaluate the honesty of court testimonies can offer valuable support during the decision making process. In this paper, we address the identification of deception in real-life trial data. We introduce a novel dataset consisting of videos collected from public court trials. We explore the use of verbal and non-verbal modalities to build a multimodal deception detection system that aims to discriminate between truthful and deceptive statements provided by defendants and witnesses. We achieve classification accuracies in the range of 60-75% when using a model that extracts and fuses features from the linguistic and gesture modalities. In addition, we present a human deception detection study where we evaluate the human capability of detecting deception in trial hearings. The results show that our system outperforms the human capability of identifying deceit.

Categories and Subject Descriptors

I.2 [**Artificial Intelligence**]: Miscellaneous

Keywords

multimodal; verbal; non-verbal; real-life trial; deception detection

1. INTRODUCTION

With thousands of trials and verdicts occurring daily in courtrooms around the world, the chance of using deceptive statements and testimonies as evidence is growing. Given the high-stake nature of trial outcomes, implementing accurate and effective computational methods to evaluate the honesty of provided testimonies can offer valuable support during the decision making process.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from Permissions@acm.org.
ICMI 2015, November 9–13, 2015, Seattle, WA, USA.
© 2015 ACM. ISBN 978-1-4503-3912-4/15/11 ...$15.00.
.

The consequences of falsely accusing the innocents and freeing the guilty can be severe. For instance, in U.S. along there are tens of thousand of criminal cases filed every year. In 2013, there were 89,936 criminal cases filings in U.S. District Courts and in 2014 the number was 80,262. [1] Moreover, the average number of exonerations per year increased from 3.03 in 1973-1999 to 4.29 between 2000 and 2013. The National Registry of Exonerations reported on 873 exonerations from 1989 to 2012, with a tragedy behind each case [17]. Hence, the need arises for a reliable and efficient system to detect deceptive behavior and discriminate between liars and truth tellers.

In criminal settings, the polygraph test has been used as a standard method to identify deceptive behavior. This becomes impractical in some cases, as this method requires the use of skin-contact devices and human expertise. In addition, the final decisions are subject to error and bias [41, 15]. Furthermore, using proper countermeasures, offenders can deceive these devices as well as the human experts. Given the difficulties associated with the use of polygraph-like methods, learning-based approaches have been proposed to address the deception detection task using a number of modalities, including text [13] and speech [20, 29]. Unlike the polygraph method, learning-based methods for deception detection rely mainly on data collected from deceivers and truth-tellers. The data is usually elicited from human contributors, in a lab setting or via crowdsourcing [28, 33], for instance by asking subjects to narrate stories in deceptive and truthful manner [28], by performing one-on-one interviews, or by participating in “Mock crime" scenarios [33]. However, an important drawback identified in this data-driven research on deception detection is the lack of real data and the absence of true motivation while eliciting deceptive behavior. Because of the artificial setting, the subjects may not be emotionally aroused thus making it difficult to generalize findings to real-life scenarios.

In this paper, we describe what we believe is a first attempt at building a multimodal system that detects deception in real-life trial data using text and gestures modalities. While there is research work that has used court trial transcripts to identify deceptive statements [14], we are not aware of any previous work that took into consideration modalities other than text for deception detection on trial court data.

[1] www.uscourts.gov

We present a novel dataset consisting of 121 deceptive and truthful video clips, from real court trials. We use the transcription of these videos to extract several linguistic features, and we manually annotate the videos for the presence of several gestures that are used to extract non-verbal features. We then build a system that jointly uses the verbal and non-verbal modalities to automatically detect the presence of deception. Our experiments show that the multimodal system can identify deception with an accuracy in the range of 60-75%, which is significantly above the chance level. As deception detection research suggests that humans perform slightly above the chance level, we also place our results in context by performing a study where humans evaluate the presence of deception in court statements in single or multimodal data streams. Results show that our system outperforms humans on this task.

2. RELATED WORK

2.1 Verbal Deception Detection

To date, several research publications on verbal-based deception detection have explored the identification of deceptive content in a variety of domains, including online dating websites [39, 18], forums [42, 23], social networks [21], and consumer report websites [31, 24]. Research findings have shown the effectiveness of features derived from text analysis, which frequently includes basic linguistic representations such as n-grams and sentence count statistics [28], and also more complex linguistic features derived from syntactic CFG trees and part of speech tags [13, 43]. Some studies have also incorporated the analysis of psycholinguistics aspects related to the deception process. Research work has relied on the Linguistic Inquiry and Word Count (LIWC) lexicon [34] to build deception models using machine learning approaches [28, 4] and showed that the use of psycholinguistic information was helpful for the automatic identification of deceit. Following the hypothesis that deceivers might create less complex sentences in an effort to conceal the truth and being able to recall their lies more easily, several researchers have also studied the relation between text syntactic complexity and deception [44].

While most of the data used in related research was collected under controlled settings, only few works have explored the used of data from real-life scenarios. This can be partially attributed to the difficulty of collecting such data, as well as the challenges associated with verifying the deceptive or truthful nature of real-world data. To our knowledge, there is very little work focusing on real-life high-stake data. The work closest to ours is presented by Fornaciari and Poesio [14], which targets the identification of deception in statements issued by witnesses and defendants using a corpus collected from hearings in Italian courts. Following this line of work, we present a study on deception detection using real-life trial data and explore the use of multiple modalities for this task.

2.2 Non-verbal Deception Detection

Earlier approaches on non-verbal deception detection relied on polygraph tests to detect deceptive behavior. These tests are mainly based on physiological features such as heart rate, respiration rate, and skin temperature. Several studies [41, 15, 10] indicated that relying solely on such physiological measurements can be biased and misleading. Chittaranjan et al. [7] created an audio-visual recordings of the "Are you a Werewolf?" game in order to detect deceptive behaviour using non-verbal audio cues and to predict the subjects' decisions in the game. In order to improve lie detection in criminal-suspect interrogations, Sumriddetchkajorn and Somboonkaew [37] developed an infrared system to detect lies by using thermal variations in the periorbital area and by deducing the respiration rate from the thermal nostril areas. Granhag and Hartwig [16] proposed a methodology using psychologically informed mind-reading to evaluate statements from suspects, witnesses, and innocents.

For hand gestures, blob analysis was used to detect deceit by tracking the hand movements of subjects [25, 40], or using geometric features related to the hand and head motion [27]. Caso et al. [6] identified several hand gestures that can be related to the act of deception using data from simulated interviews. Cohen et al. [8] found that fewer iconic hand gestures were a sign of a deceptive narration, and Hillman et al. [19] determined that increased speech prompting gestures were associated with deception while increased rhythmic pulsing gestures were associated with truthful behavior. Also related is the taxonomy of hand gestures developed by Maricchiolo et al. [26] for applications such as deception detection and social behavior.

Facial expressions also play a critical role in the identification of deception. Ekman [11] defined micro-expressions as relatively short involuntary expressions, which can be indicative of deceptive behavior. Moreover, these expressions were analyzed using smoothness and asymmetry measurements to further relate them to an act of deceit [12]. Tian et al. [38] considered features such as face orientation and facial expression intensity. Owayjan et al. [32] extracted geometric-based features from facial expressions, and Pfister and Pietikainen [36] developed a micro-expression dataset to identify expressions that are clues for deception.

Recently, features from different modalities were integrated in order to find a combination of multimodal features with superior performance [5, 22]. A multimodal deception dataset consisting of linguistic, thermal, and physiological features was introduced in [35], which was then used to develop a multimodal deception detection system [2]. An extensive review of approaches for evaluating human credibility using physiological, visual, acoustic, and linguistic features is available in [30].

To our knowledge, no work to date has considered the problem of deception detection on multimodal real-life trial data, which is the task that we are addressing in this paper.

3. DATASET

Our goal is to build a multimodal collection of occurrences of real deception during court trials, which will allow us to analyze both verbal and non-verbal behaviors in relation to deception.

3.1 Data Collection

To collect the dataset, we start by identifying public multimedia sources where trial hearing recordings were available, and deceptive and truthful behavior could be fairly observed and verified.

We specifically target trial recordings on which some of the constraints imposed by current data processing technologies could be enforced: the defendant or witness in the video should be clearly identified; her or his face should be visible enough during most of the clip duration; visual quality should be clear enough to identify the facial expressions; and finally, audio quality should be clear enough to hear the voices and understand what the person is saying.

We considered three different trial outcomes that helped us to correctly label a certain trial video clip as deceptive or truthful: guilty verdict, non-guilty verdict, and exoneration. Thus, for guilty verdicts, deceptive clips are collected from a defendant in a trial and truthful videos are collected from witnesses in the same trial. In some cases, deceptive videos are collected from a suspect denying a crime he committed and truthful clips are taken from the same suspect when answering questions concerning some facts that were verified by the police as truthful. For the witnesses, testimonies that were verified by police investigations are labeled as truthful

Figure 1: Sample screenshots showing facial displays and hand gestures from real-life trial clips. Starting at the top left-hand corner: deceptive trial with forward head movement (*Move forward*), deceptive trial with both hands movement (*Both hands*), deceptive trial with one hand movement (*Single hand*), truthful trial with raised eyebrows (*Eyebrows raising*), deceptive trial with scowl face (*Scowl*), and truthful trial with an up gaze (*Gaze up*).

Truthful	Deceptive
We proceeded to step back into the living room in front of the fireplace while William was sitting in the love seat. And he was still sitting there in shock and so they to repeatedly tell him to get down on the ground. And so now all three of us are face down on the wood floor and they just tell us "don't look, don't look" And then they started rummaging through the house to find stuff...	No, no. I did not and I had absolutely nothing to do with her disappearance. And I'm glad that she did. I did. I did. Um and then when Laci disappeared, um, I called her immediately. It wasn't immediately, it was a couple of days after Laci's disappearance that I telephoned her and told her the truth. That I was married, that Laci's disappeared, she didn't know about it at that point.

Table 1: Sample transcripts for deceptive and truthful clips in the dataset.

whereas testimonies in favor of a guilty suspect are labeled as deceptive. Exoneration testimonies are collected as truthful statements.

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. Clips containing exonerees testimonies are obtained from "The Innocence Project" website.[2]

Given our goals and constraints, data collection ended up being a lengthy and laborious process consisting of several iterations of Web mining, data processing and analysis, and content validation.

The final dataset consists of 121 videos including 61 deceptive and 60 truthful trial clips. The average length of the videos in the dataset is 28.0 seconds. The average video length is 27.7 seconds and 28.3 seconds for the deceptive and truthful clips, respectively. The data consists of 21 unique female and 35 unique male speakers, with their ages approximately ranging between 16 and 60 years.

3.2 Transcriptions

Our goal is to analyze both verbal and non-verbal behavior to understand their relation to deception.

Gesture Category	Agreement	Kappa
General Facial Expressions	66.07%	0.328
Eyebrows	80.03%	0.670
Eyes	64.28%	0.465
Gaze	55.35%	0.253
Mouth Openness	78.57%	0.512
Mouth Lips	85.71%	0.690
Head Movements	69.64%	0.569
Hand Movements	94.64%	0.917
Hand Trajectory	82.14%	0.738
Average	75.16%	0.571

Table 2: Gesture annotation agreement

All the video clips are transcribed via crowd sourcing using Amazon Mechanical Turk. We specifically asked transcribers to include word repetitions and fillers such as *um*, *ah*, and *uh*, as well as indicate intentional silence using ellipsis. Obtained transcriptions were manually verified to avoid spam and ensure their quality. The final set of transcriptions consists of 8,055 words, with an average of 66 words per transcript. Table 1 shows transcriptions of sample deceptive and truthful statements.

[2]http://www.innocenceproject.org/

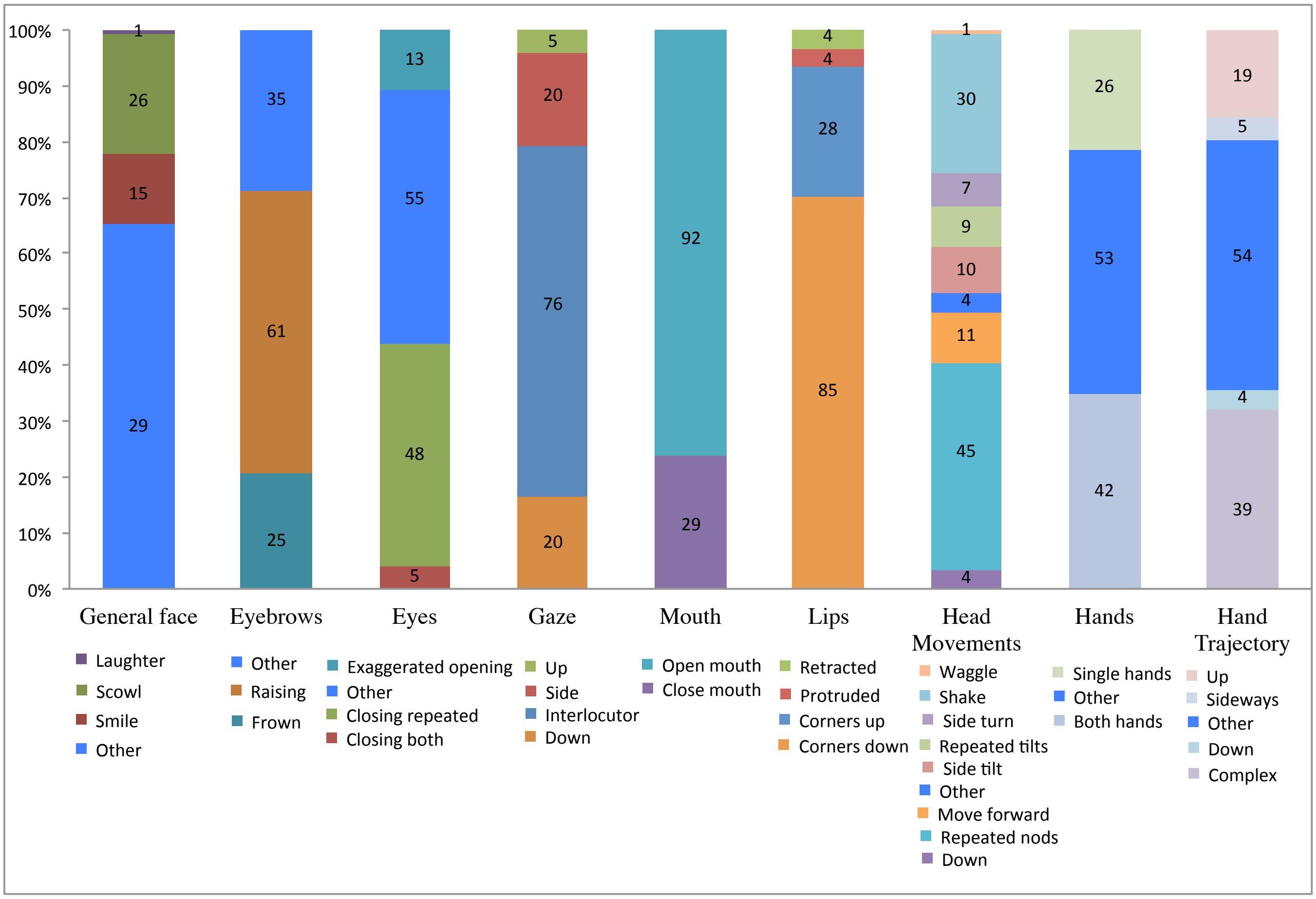

Figure 2: Distribution for nine facial displays and hand gestures

3.3 Non-verbal Behavior Annotations

We annotate the gestures[3] observed during the interactions in the video clips. Since our data occurs in interview-based scenarios where deceivers and truth-tellers are interacting with the interviewers, we decided to annotate the gesture behavior using a scheme that has been specifically designed for interpersonal communication.

We specifically focus on the annotation of facial displays and hand movements, as they have been previously found to correlate with deceptive behavior [9]. The gesture annotation is performed using the MUMIN coding scheme [3], which is a standard multimodal annotation scheme for interpersonal interactions.

In the MUMIN scheme, facial displays consist of 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. Figure 2 shows the nine gesture categories considered during the annotation.

The multimodal annotation was performed by two annotators using the Elan software. During the annotation process, annotators were allowed to watch each video clip as many times as they needed. They were asked to identify the facial displays and hand gestures that were most frequently observed or dominating during the entire clip duration. For each video clip, the annotators had to choose one label for each of the nine gestures listed in Figure 2. Annotations were performed at video level in accordance with the overall judgment of truthfulness and deceitfulness, which is based on the whole video content. During the annotation process, annotators chose only one label per gesture for every video clip.

Note that the "Other" category indicates cases where none of the other gestures was observed. For instance, in the case of gestures associated with hand movements, the "Other" label also accounted for those cases where the speaker's hands were not moving or were not visible.

Before all the video clips were annotated for gestures, we measured the inter-annotator agreement in a subset of 56 videos. Table 2 shows the observed annotation agreement between the two annotators, along with the Kappa statistic. The agreement measure represents the percentage of times the two annotators agreed on the same label for each gesture category. For instance, 80.03% of the time the annotators agreed on the labels assigned to the *Eyebrows* category. On average, the observed agreement was measured at 75.16%, with a Kappa of 0.57 (macro-averaged over the nine categories). Differences in annotation were reconciled through discussions. After this, the remaining videos were split between the two annotators, and were labeled by only one annotator. Figure 2 shows

[3]As done in the Human-Computer Interaction community, we use the term "gesture" to broadly refer to body movements, including facial expressions and hand gestures.

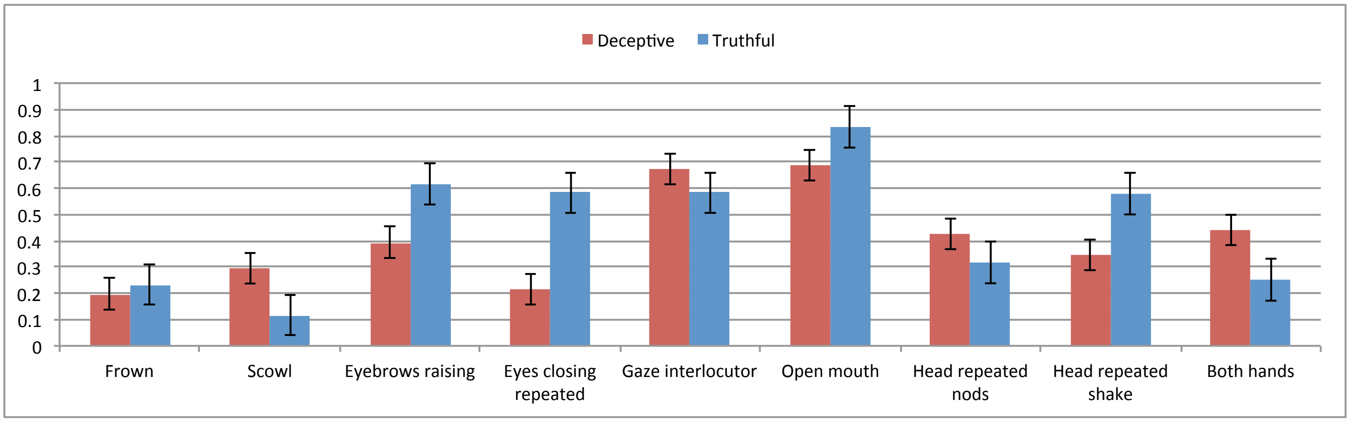

Figure 3: Distribution of non-verbal features for deceptive and truthful groups

the final frequency counts associated with each of the nine gestures considered during the annotation.

4. FEATURES OF VERBAL AND NON-VERBAL BEHAVIORS

Given the multimodal nature of our dataset, we focus on both the linguistic and gesture components of the recordings included in our collection. In this section, we describe the sets of features extracted from each modality, which will then be used to build classifiers of deception.

4.1 Verbal Features

The verbal features consist of unigrams and bigrams derived from the bag-of-words representation of the video transcripts. These features are encoded as word and word pairs frequencies and include all the words present in the transcripts with frequencies greater than 10. The frequency threshold cut was chosen using a small development set.

Previous work has also considered features derived from semantic lexicons, e.g., LIWC [34]. However, while these are great features to consider in order to gain insights into the semantic categories of words that represent useful clues for deception, their performance is often similar to that of the n-grams features [2]. Since in our current work we are not focusing on the insights that can be gained from linguistic analyses, we are not using these features in our current experiments.

4.2 Non-verbal Features

The non-verbal features are derived from the annotations performed using the MUMIN coding scheme as described in Section 3.3. We create a binary feature for each of the 40 available gesture labels. Each feature indicates the presence of a gesture only if it is observed during the majority of the interaction duration. The generated features represent nine different gesture categories covering facial displays and hand movements.

Facial Displays. These are facial expressions or head movements displayed by the speaker during the deceptive or truthful interaction. They include all the behaviors listed in Figure 2 under the General Facial Expressions, Eyebrows, Eyes, Mouth Openness, Mouth Lips, and Head Movements.

Hand Gestures. The second broad category covers gestures made with the hands, and it includes the Hand Movements and Hand Trajectories listed in Figure 2.

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 3: Deception classifiers using individual and combined sets of verbal and non-verbal features.

5. EXPERIMENTS

We start our experiments with an analysis of the non-verbal behaviors occurring in deceptive and truthful videos. We compare the percentage of each behavior as observed in each class. For instance, there is a total of 61 videos in the dataset that include the Eyebrows raising feature (as shown in Figure 2), out of which 24 are part of the deceptive set of 61 videos, and 37 are part of the truthful set (60 videos). Hence, the percentages of existence of this feature are 39% in the deceptive class and 61% in the truthful class. Figure 3 shows the percentages of all the non-verbal features for which we observe noticeable differences for the deceptive and truthful groups. As the figure suggests, eyebrow and eye gestures help differentiate between the deceptive and truthful conditions. For instance, we can observe that truth-tellers appear to raise their eyebrows (Eyebrows raising), shake their head (Head repeated shake), and blink (Eyes closing repeated) more frequently than deceivers. Interestingly, deceivers seem to blink and shake their head less frequently than truth-tellers.

Motivated by these results, we proceed to conduct further experiments to evaluate the performance of the extracted features using a machine learning approach.

We run our learning experiments on the trial dataset introduced earlier. Given the distribution between deceptive and truthful clips, the baseline on this dataset is 50.4%. For each video clip, we create feature vectors formed by combinations of the verbal and non-verbal features described in the previous section.

We build deception classifiers using two classification algorithms: Decision Trees (DT) and Random Forest (RF).[4] We run several

[4]We use the implementation available in the Weka toolkit with the default parameters.

comparative experiments using leave-one-out cross-validation. Table 3 shows the accuracy figures obtained by the two classifiers on the major feature groups described in Section 4. As shown in this table, the combined classifier that uses all the features (using Decision Trees) and the individual classifier that relies on the facial displays features (using Random Forest) achieve the best results. We also evaluate classifiers that rely on combined sets of features, with the non-verbal features clearly outperforming the verbal features.

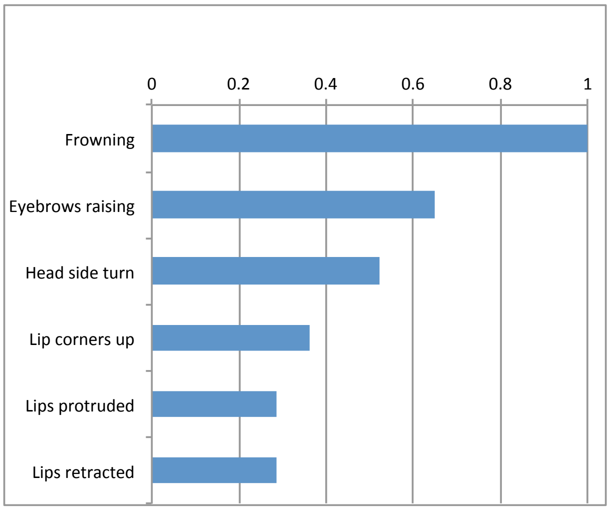

Figure 4: Weights of top non-verbal features

5.1 Analysis of Feature Contribution

To better understand the contribution of the different feature sets to the overall classifier performance, we conduct an ablation study where we remove one group of features at a time. Given that Decision Trees had the most consistent performance across different settings in our initial set of experiments, we run all our analysis experiments only using this classifier.

Table 4 shows the accuracies obtained when one feature group is removed and the deception classifier is built using the remaining features. From this table we made interesting findings for the combined model. Not surprisingly, the facial displays seem to contribute the most to the classifier performance, followed by the unigram features. This further suggests that a better feature fusion method may be beneficial, and we plan to explore this in future work.

For a closer look at the contribution of individual features included in the group of *Facial Displays*, we determine and compare the weights assigned by the learning algorithm to the features in this group, as shown in Figure 4.[5] The five most predictive features are 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.

6. HUMAN PERFORMANCE

An important remaining question is concerned with the human performance on the task of deception detection. An answer to this question can shed light on the difficulty of the task, and can also place our results in perspective.

[5]In the figure, the features are normalized with respect to the largest feature weight.

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

Table 4: Feature ablation study.

Modality	Agreement	Kappa
Text	59.80%	0.071
Audio	62.00%	0.196
Silent video	51.50%	0.014
Full video	57.60%	0.127

Table 5: Agreement among three human annotators on text, audio, silent video, and full video modalities.

We conduct a study where we evaluate 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.

We create an annotation interface that shows an annotator instances for each modality in random order, and ask him or her to select a label of either "Deception" or "Truth" according to his or her perception of truthfulness or falsehood. The annotators did not have access to any information that would reveal the true label of an instance. The only exception to this could have been the annotators' previous knowledge of some of the public trials in our dataset. A discussion with the annotators after the annotation took place indicated however that this was not the case.

To avoid annotation bias, we show the modalities in the following order: first we show either *Text* or *Silent video*, then we show *Audio*, followed by *Full video*. Note that apart from this constraint, which is enforced over the four modalities belonging to each video clip, the order in which instances are presented to an annotator is random.

Three annotators labeled all 121 video clips in our dataset. Since four modalities were extracted from each video, each annotator annotated a total of 484 instances. Annotators were not offered a monetary reward and we considered their judgments to be honest as they participated voluntarily in this experiment. Table 5 shows the observed agreement and Kappa statistics among the three annotators for each modality.[6] The agreement for most modalities is rather low and the Kappa scores show mostly slight agreement. As noted before [31], this low agreement can be interpreted as an indication that people are poor judges of deception.

We also determine each annotator performance for each modality. The results, shown in Table 6, additionally support the argument that human judges have difficulty performing the deception detection task. Not surprisingly, human detection of deception on silent video is more challenging than the rest of the modalities due to the lesser amount of deception cues available to the raters. An interesting, yet perhaps unsurprising observation is that the human performance increases with the availability of modalities. The

[6]Inter-rater agreement with multiple raters and variables. `https://mlnl.net/jg/software/ira/`

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

Table 6: Performance of three annotators (A1, A2, A3) and the developed automatic system (Sys) on the real-deception dataset over four modalities.

poorest accuracy is obtained in *Silent video*, followed by *Text*, *Audio*, and *Full video* where the judges have the highest performance.

Overall, our study indicates that detecting deception is indeed a difficult task for humans and further verifies previous findings where human ability to spot liars was found to be slightly better than chance [1]. Moreover, the performance of the human annotators appears to be significantly below that of our system.

7. CONCLUSIONS

In this paper we presented a study of multimodal deception detection using real-life high-stake occurrences of deceit. We introduced a novel dataset from public real trials, and used this dataset to perform both qualitative and quantitative experiments. Our analysis of non-verbal behaviors occurring in deceptive and truthful videos brought insight into the gestures that play a role in deception. We also built classifiers relying on individual or combined sets of verbal and non-verbal features, and showed that we can achieve accuracies in the range of 60-75%. Additional analyses showed the role played by the various feature sets used in the experiments.

We also performed a study of human ability to spot liars in single or multimodal data streams. The study revealed high disagreement and low deception detection accuracies among human annotators. Our automatic system outperformed humans using different modalities with a relative percentage improvement of up to 51%.

To our knowledge this is the first work to automatically detect instances of deceit using both verbal and non-verbal features extracted from real trial recordings. Future work will address the use of automatic gesture identification and automatic speech transcription, with the goal of taking steps towards a real-time deception detection system.

The dataset introduced in this paper is available upon request.

Acknowledgments

We are grateful to Hagar Mohamed, Aparna Garimella, Shibamouli Lahiri, and Charles Welch for assisting us with the data collection and the human evaluations. This material is based in part upon work supported by National Science Foundation awards #1344257 and #1355633, by grant #48503 from the John Templeton Foundation, 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, the John Templeton Foundation, or the Defense Advanced Research Projects Agency.

8. REFERENCES

[1] M. Aamodt and H. Custer. Who can best catch a liar? a meta-analysis of individual differences in detecting deception. *Forensic Examiner*, 15(1):6–11, 2006.

[2] M. Abouelenien, V. Pérez-Rosas, R. Mihalcea, and M. Burzo. Deception detection using a multimodal approach. In *Proceedings of the 16th International Conference on Multimodal Interaction*, ICMI '14, pages 58–65, Istanbul, Turkey, 2014. ACM.

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

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

[5] J. Burgoon, D. Twitchell, M. Jensen, T. Meservy, M. Adkins, J. Kruse, A. Deokar, G. Tsechpenakis, S. Lu, D. Metaxas, J. Nunamaker, and R. Younger. Detecting concealment of intent in transportation screening: A proof of concept. *IEEE Transactions on Intelligent Transportation Systems*, 10(1):103–112, March 2009.

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

[7] G. Chittaranjan and H. Hung. Are you awerewolf? detecting deceptive roles and outcomes in a conversational role-playing game. In *2010 IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, pages 5334–5337, March 2010.

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

[9] B. Depaulo, B. Malone, J. Lindsay, L. Muhlenbruck, K. Charlton, and H. Cooper. Cues to deception. *Psychological Bulletin*, pages 74–118, 2003.

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

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

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

[13] S. Feng, R. Banerjee, and Y. Choi. Syntactic stylometry for deception detection. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers - Volume 2*, ACL '12, pages 171–175, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics.

[14] T. Fornaciari and M. Poesio. Automatic deception detection in Italian court cases. *Artificial Intelligence and Law*, 21(3):303–340, 2013.

[15] T. Gannon, A. Beech, and T. Ward. *Risk Assessment and the Polygraph*, pages 129–154. John Wiley and Sons Ltd, 2009.

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

[17] S. Gross and R. Warden. Exonerations in the united states, 1989 - 2012. Technical report, National Registry of Exonerations, 2012.

[18] R. Guadagno, B. Okdie, and S. Kruse. Dating deception: Gender, online dating, and exaggerated self-presentation. *Comput. Hum. Behav.*, 28(2):642–647, Mar. 2012.

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

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

[21] S. Ho and J. M. Hollister. Guess who? an empirical study of gender deception and detection in computer-mediated communication. *Proceedings of the American Society for Information Science and Technology*, 50(1):1–4, 2013.

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

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

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

[25] S. Lu, G. Tsechpenakis, D. Metaxas, M. Jensen, and J. Kruse. Blob analysis of the head and hands: A method for deception detection. In *Proceedings of the 38th Annual Hawaii International Conference on System Sciences (HICSS'05)*, HICSS '05, pages 20–29, Washington, DC, USA, 2005. IEEE Computer Society.

[26] F. Maricchiolo, A. Gnisci, and M. Bonaiuto. Coding hand gestures: A reliable taxonomy and a multi-media support. In A. Esposito, A. Esposito, A. Vinciarelli, R. Hoffmann, and V. Muller, editors, *Cognitive Behavioural Systems*, volume 7403 of *Lecture Notes in Computer Science*, pages 405–416. Springer Berlin Heidelberg, 2012.

[27] T. Meservy, M. Jensen, J. Kruse, D. Twitchell, G. Tsechpenakis, J. Burgoon, D. Metaxas, and J. Nunamaker. Deception detection through automatic, unobtrusive analysis of nonverbal behavior. *IEEE Intelligent Systems*, 20(5):36–43, September 2005.

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

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

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

[31] M. Ott, Y. Choi, C. Cardie, and J. Hancock. Finding deceptive opinion spam by any stretch of the imagination. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 309–319, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

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

[33] I. Pavlidis, N. Eberhardt, and J. Levine. Human behaviour: Seeing through the face of deception. *Nature*, 415(6867), 2002.

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

[35] V. Perez-Rosas, R. Mihalcea, A. Narvaez, and M. Burzo. A multimodal dataset for deception detection. In *Proceedings of the Conference on Language Resources and Evaluations (LREC 2014)*, Reykjavik, Iceland, May 2014.

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

[37] S. Sumriddetchkajorn and A. Somboonkaew. Thermal analyzer enables improved lie detection in criminal-suspect interrogations. In *SPIE Newsroom: Defense & Security*, 2011.

[38] Y. Tian, T. Kanade, and J. Cohn. Facial expression analysis. In *Handbook of Face Recognition*, pages 247–275. Springer New York, 2005.

[39] C. Toma and J. Hancock. Reading between the lines: linguistic cues to deception in online dating profiles. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work*, CSCW '10, pages 5–8, New York, NY, USA, 2010. ACM.

[40] G. Tsechpenakis, D. Metaxas, M. Adkins, J. Kruse, J. Burgoon, M. Jensen, T. Meservy, D. Twitchell, A. Deokar, and J. Nunamaker. Hmm-based deception recognition from visual cues. In *IEEE International Conference on Multimedia and Expo, 2005. ICME 2005*, pages 824–827, July 2005.

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

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

[43] Q. Xu and H. Zhao. Using deep linguistic features for finding deceptive opinion spam. In *Proceedings of COLING 2012: Posters*, pages 1341–1350, Mumbai, India, December 2012. The COLING 2012 Organizing Committee.

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