Verbal and Nonverbal Clues for Real-life Deception Detection

Verónica Pérez-Rosas, Mohamed Abouelenien, Rada Mihalcea,
Yao Xiao, CJ Linton, Mihai Burzo
University of Michigan
(vrncapr,zmohamed,mihalcea,xyaoinum,cjlint,mburzo)@umich.edu

Abstract

Deception detection has been receiving an increasing amount of attention from the computational linguistics, speech, and multimodal processing communities. One of the major challenges encountered in this task is the availability of data, and most of the research work to date has been conducted on acted or artificially collected data. The generated deception models are thus lacking real-world evidence. In this paper, we explore the use of multimodal real-life data for the task of deception detection. We develop a new deception dataset consisting of videos from real-life scenarios, and build deception tools relying on verbal and nonverbal features. We achieve classification accuracies in the range of 77-82% when using a model that extracts and fuses features from the linguistic and visual modalities. We show that these results outperform the human capability of identifying deceit.

1 Introduction

As deceptive behavior occurs on a daily basis in different areas of life (Meyer, 2010; Smith et al., 2014), the need arises for automated methodologies to detect deception in an efficient, yet reliable manner. There are many applications that can benefit from automatic deception identification, such as airport security screening, crime investigation and interrogation, interviews, advertisement, and others. In many of these settings, the polygraph test has been used as the main method to identify deceptive behavior. However, this method requires the use of skin-contact devices and human expertise, making it infeasible for large-scale applications. Moreover, polygraph tests were shown to be misleading in multiple cases (Vrij, 2001; Gannon et al., 2009), as human judgment is often biased.

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 (Feng et al., 2012) and speech (Hirschberg et al., 2005; Newman et al., 2003). Unlike the polygraph methods, 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. An important problem identified in this data-driven research is the lack of real data. Because of the artificial setting, the subjects may not be emotionally aroused, as they may not take the experiments seriously given the lack of motivation and/or penalty.

In this paper, we describe what we believe is a first attempt at building a multimodal system that detects deception in real-life settings. We collect a dataset consisting of 118 deceptive and truthful video clips, from real trials and live street interviews aired in television shows. 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 nonverbal features. We then build a system that jointly uses the verbal and nonverbal 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 77-82%, significantly improving over the baseline. In addition, we present a study on the human ability to detect deception in single or multimodal data streams, and show that our system outperforms humans on this task.

2 Dataset

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

Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 2336–2346,
Lisbon, Portugal, 17-21 September 2015.

Figure 1: Sample screenshots showing facial displays and hand gestures from real-life deception and truthful clips. Starting at the top left-hand corner: deceptive interview with up gaze (*Up*), deceptive interview with side gaze (*Side*), deceptive trial with both hands (*Both-H*), truthful trial with forward head (*Forward*), truthful interview with side turn (*Side-Turn*), and truthful interview with single hand (*Single-H*).

Truthful	Deceptive
I was sentenced to forty to sixty years in prison for this crime that I didn't commit. At the trial the judge had exceeded the sentence guidelines because he said I failed to show remorse. And I told him, you know, I felt terrible for what happen to this woman, shouldn't happen to anyone, but I can't show remorse for something I didn't do.	We had some drinks at the bar, maybe one ... two. um I got onto the dance floor myself as I explained, um I have been a trained dancer for some time, going to be able to dance freely is like a ... release. I'm very much in my own space when I do that and so I got up, and I was dancing alone on the dance floor.
It's difficult to pick just one but um I think Tender Mercies uh is ... really captured my imagination um when I was in junior high. Had a lot to do with Robert Duval's performance certainly and that got me excited about the possibility of um .... pulling off an acting career for myself.	Yeah, yeah he was convincing as a wolf. Ahhh actually you know ahhh this is like crazy I'm terrified from wolves, it's my worst fear even though they don't exist but thats my worst fear, sharks and stuff like that. Yeah its my worst fear, I am being honest with you.

Table 1: Sample transcripts for deceptive and truthful clips. The first row presents transcripts from the *Trials* domain while the second shows transcripts corresponding to the *Interviews* domain.

2.1 Data Collection

To collect real deception data, we start by identifying online multimedia sources where deceptive behavior can be observed and verified. We specifically target videos of people, on which we enforce some of the constraints imposed by current data processing technologies: the person in the video should be in front of the camera; her face should be clearly visible; 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 collect video clips from public real trials and interviews aired during television shows, where the truth or falsehood of the participant's statements ends up being known. Video clips from trials consist of statements from witnesses and defendants in the same trial. In order to have a clear distinction between deceptive and truthful trial videos portraying defendants, the process of labeling the trial relies on the verdict. Thus, clips with a guilty verdict are considered deceptive whereas clips with a non-guilty verdict or exoneration are labeled as truthful. Clips containing witness testimonies are labeled as truthful if their statements are verified by police investigations. Examples of trials included in our dataset are Jodi Arias, Andrea Sneiderman, and Amanda Hayes. Exoneree's statements were taken from "The Innocence Project" (`http://www.innocenceproject.org`).

Deceptive and truthful responses are also collected from TV shows and interviews. Examples of such shows are "Lie Witness," "Golden Balls," and the "American Film Institute" and "RevYOU" You-Tube channels. Deceptive videos portray scenarios where interviewees' responses were known to be a lie. For example, the interviewer asks a random individual on his opinion on a non-existing film where the interviewee fabricates a story. On the other hand, truthful videos are collected from individuals asked on their opinions on real movies.

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 includes 118 videos, including 59 that are labeled as deceptive and 59 labeled as truthful. Among them, 62 belong to the TV street interviews and shows category (*Interviews*) with 28 deceptive and 34 truthful video clips, and 56 belong to the trials category (*Trials*) with 31 deceptive and 25 truthful clips. The average length of the videos in the dataset is 27.28 seconds, with an average length of 33.02 seconds for the truthful clips and 21.54 seconds for the deceptive clips. Collected trial samples cover famous murder cases, while street interviews cover several topics such as movies, music, politics, and religion. The dataset contains 23 unique female and 39 unique male speakers, with their ages ranging approximately between 16 and 60 years.

2.2 Transcriptions and Nonverbal Behavior Annotations

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

First, all the video clips were manually transcribed. The transcription was performed by two transcribers using the Elan software (Wittenburg et al., 2006). We asked transcribers to include word repetitions and fillers such as um, ah, and uh, as well as long pauses that were marked using three consecutive dots. The final set of transcriptions contain 7835 words, with an average of 66 words per transcript. Table 1 shows transcriptions of sample deceptive and truthful statements from both trials and reality shows.

Second, we annotate the gestures[1] observed during the interactions in the video clips. We specifically focus on the annotation of facial displays and hand movements, as they have been previously found to correlate with deceptive behavior (Depaulo et al., 2003). The gesture annotation is performed using the MUMIN coding scheme (Allwood et al., 2007).

Gesture Category	Agreement	Kappa
Facial Expressions	72.88%	0.576
Eyebrows	80.51%	0.656
Eyes	68.64%	0.517
Gaze	61.40%	0.432
Mouth Openness	77.97%	0.361
Mouth Lips	82.20%	0.684
Head Movements	55.08%	0.420
Hand Movements	91.53%	0.858
Hand Trajectory	84.75%	0.753
Average	75.00%	0.584

Table 2: Gesture annotation agreement

In the MUMIN scheme, facial displays consist of several different facial expressions associated with eyebrows, eyes, gaze, and mouth. Smile, laughter, and scowl are also included, as well as general head and hand movements.

The multimodal annotation was performed by two annotators using the Elan software (Wittenburg et al., 2006). We decided to perform the gesture annotations at video level, rather than at utterance level, because the overall judgment of truthfulness and deceitfulness is based on the whole video content. 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 Table 3.

Table 3 shows the frequency counts associated with the nine gestures considered during the annotation. Note that the counts under each gesture add up to 118, reflecting the fact that for every gesture, the annotators had to choose one label for every video clip. When none of the labels applied, the "Other" category was selected. 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.

After all the video clips were annotated for gestures, the inter-annotator agreement was mea-

[1]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.

Label	Count
Eyebrows	
Frown (Frowning)	17
Raise (Raising)	71
Other	30
Eyes	
X-open (Exaggerated opening)	17
Close-BE (Closing both)	7
Closing-E (Closing one)	1
Close-R (Closing repeated)	20
Other	73
Gaze	
Interlocutor	69
Up	7
Down	14
Side	24
Other	4

Label	Count
General Facial Expressions	
Smile	41
Scowl	13
Laugh (Laughter)	1
Other	63
Mouth Openness	
Close-M (Closed mouth)	26
Open-M (Open mouth)	92
Mouth Lips	
Up-C (Corners up)	61
Down-C (Corners down)	51
Protruded	1
Retracted	5
Hand Movements	
Both hands (Both-H)	31
Single hands (Single-H)	26
Other	61

Label	Count
Hand Trajectory	
Up (Upwards)	13
Down (Downwards)	5
Sideways	5
Complex	33
Other	62
Head Movements	
Down (Single nod)	3
Down-R (Repeated nods)	48
Forward (Move forward)	3
Back (Move backward)	3
Side-tilt (Single tilt)	8
Side-Tilt-R (Repeated tilts)	9
Side-Turn	9
Side-Turn-R (Shake repeated)	26
Waggle	3
Other	6

Table 3: Frequency counts for nine facial displays and hand gestures

sured. 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, 72.88% of the time the annotators agreed on the label assigned to the *General Face* category. On average, the observed agreement was measured at 75%, with a Kappa of 0.58 (macro-averaged over the nine categories), which reflects substantial agreement. Observed agreement for Head Movements and Gaze is noticeably lower than other categories, which can be attributed to a higher number of available gesture choices, as seen in Table 3.

3 Features of Verbal and Nonverbal Behaviors

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

3.1 Verbal Features

We implement three types of features, consisting of unigrams, psycholinguistic features, and syntactic complexity features.

Unigrams. We extract unigrams derived from the bag-of-words representation of the video transcripts. The unigram features are encoded as word frequencies and include all the words present in the transcripts.

Psycholinguistic Features. The Linguistic Word Count (LIWC) is a psycholinguistics lexicon that has been frequently used to incorporate semantic and psychological information into linguistic analysis (Pennebaker and Francis, 1999). It has been successfully used in previous work on deception detection (Newman et al., 2003; Mihalcea and Strapparava, 2009; Ott et al., 2011). We obtain features for each of the 80 psycholinguistic classes present in the lexicon by calculating the percentage of words in the transcription belonging to each class.

Syntactic Complexity. We also extract features to measure the syntactic complexity of the speech produced by the speakers in truthful and deceptive clips. This set of features is motivated by previous research that has suggested that deceivers' speech has lower complexity (Depaulo et al., 2003). We use the tool described in (Lu, 2010), which generates indexes of syntactic complexity, including general complexity metrics, length of production, and amount of coordination. The set of features consists of fourteen indexes including statistics related to T-units, which are linguistic units that include a main clause in addition to attached subordinate clauses. T-unit analysis is extensively used to analyze syntactic complexity in speech and written content. The set of features includes the mean length of sentence, mean length of T-

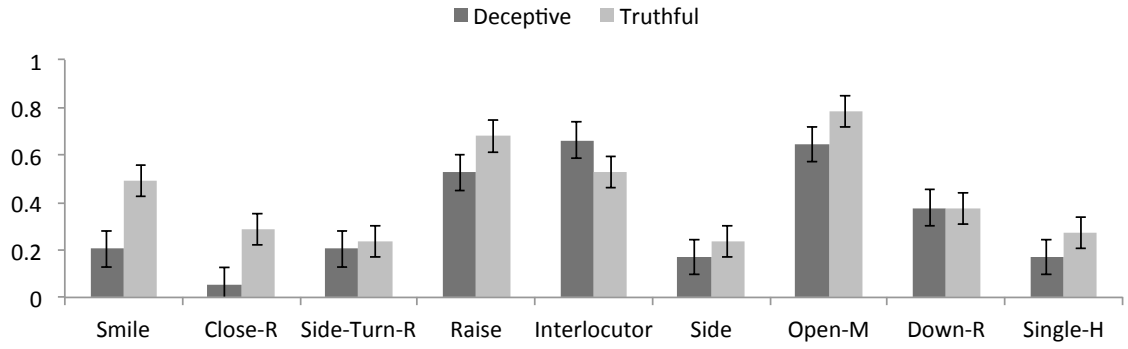

Figure 2: Distribution of nonverbal features for deceptive and truthful groups

unit, mean length of clause, clauses per sentence, verb phrases per T-unit, clauses per T-unit, dependent clauses per clause, dependent clauses per T-unit, T-units per sentence, complex T-unit ratio, coordinate phrases per T-unit, coordinate phrases per clause, complex nominals per T-unit, and complex nominals per clause.

3.2 Nonverbal Features

The nonverbal features are derived from the annotations performed using the MUMIN coding scheme as described in section 2.2. 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 Table 3 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 Table 3.

4 Experiments

We start our experiments with an analysis of the nonverbal 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 41 videos in the dataset that include the Smile feature (as shown in Table 3), out of which 12 are part of the deceptive set of 59 videos, and 29 are part of the truthful set (again, of 59 videos). Hence, the percentages for this feature are 20.33% in the deceptive class, and 49.13% in the truthful class. Figure 2 shows the percentages of all the nonverbal features for which we observe noticeable differences for the deceptive and truthful groups. As the figure suggests, facial displays seem to help differentiate between the deceptive and truthful conditions. For instance, we can observe that truth-tellers smile (Smile) and blink more (Close-R). Interestingly deceivers seem to make more eye contact (Interlocutor gaze) and nod (Side-Turn-R) more frequently than truth-tellers. This agrees with the findings in (Depaulo et al., 2003) that liars who are more motivated to get away with their lies (i.e., trials) are likely to increase their eye-contact behavior.

Feature Set	SVM	DT	RF
Unigrams	69.49%	76.27%	67.79%
Psycholinguistic	53.38%	50.00%	66.10%
Syntactic Complexity	52.54%	62.71%	53.38%
Facial Displays	78.81%	74.57%	67.79%
Hand Gestures	59.32%	57.62%	57.62%
Unigr.+Facial Disp.	71.18%	70.33%	68.64%
All Verbal	65.25%	63.55 %	57.62 %
All Nonverbal	75.42%	68.64%	72.03%
All Features	77.11%	69.49%	73.72%

Table 4: Deception classifiers using individual and combined sets of verbal and nonverbal features.

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

Feature Set	SVM
All	**77.11%**
– Hand gestures	74.57%
– Facial displays	64.40%
– Syntactic	76.27%
– Semantic	72.03%
– Unigrams	73.72%

Table 5: Feature ablation study.

We run our learning experiments on the real-deception dataset introduced earlier. Given the even distribution between deceptive and truthful clips, the baseline on this dataset is 50%. For each video clip, we create feature vectors formed by combinations of the verbal and nonverbal features described in the previous section. We build deception classifiers using three classification algorithms: Support Vector Machines (SVM), Decision Trees (DT), and Random Forest (RF).[2] We run several comparative experiments using leave-one-out cross-validation. Table 4 shows the accuracy figures obtained by the three classifiers on the major feature groups described in Section 3. As shown in this table, the facial displays classifier achieves the highest accuracy among the individual classifiers, followed by the unigrams classifier.

We also evaluate classifiers that rely on combined sets of features. The nonverbal features clearly outperform the verbal features, and the classifier that includes all the features improves over the classifiers that rely on all the verbal features or all the nonverbal features. Importantly, several of the classifiers improve significantly over the baseline.

4.1 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 SVM had the best performance in our initial set of experiments, we run all our analysis experiments only using this classifier. Table 5 shows the accuracies obtained when one feature group is removed and the deception classifier is built using the remaining features. From this table, we can again observe that Facial Displays contribute the most to the classifier performance, while Syntactic Features show the lowest contribution.

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

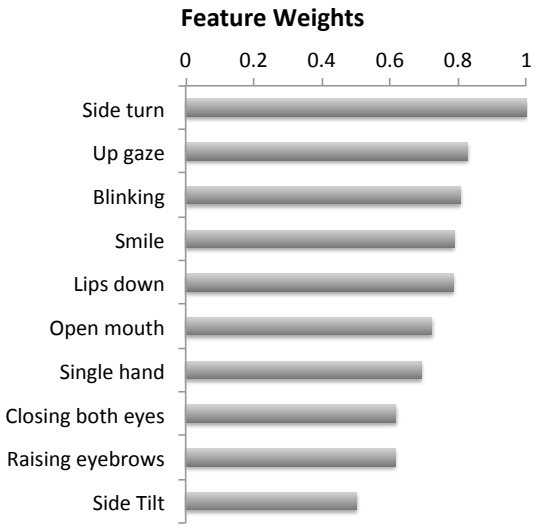

Figure 3: Weights of top nonverbal features

For a closer look at the contribution of individual features included in the group of *Facial Displays*, we analyzed the absolute values of the weights assigned by the learning algorithm to the features in this group. Figure 3 shows the features normalized with respect to the largest feature weight. The five most predictive features are the presence of side turns, up gazes, blinking, and smiling, which we previously identified as possible indicators of deception. This further confirms our initial hypothesis that gestures associated with human interaction are an important component of human deception.

We also analyze the contribution of the linguistic features. Using the linguistic ethnography method (Mihalcea and Pulman, 2009), we obtain the most dominant LIWC word classes associated with deceptive and truthful transcripts extracted from trials and interviews clips. Results are shown in Table 6. Interestingly, the most dominant classes in truthful clips, regardless of being from interviews or trials, correspond to words related to Family, Home, and Humans. This suggests that truth-tellers show similar word usage when interviewed on a real scenario. On the other hand, dominant classes associated to deceivers are less consistent as they discuss aspects related to the topic being discussed. For instance, while being interviewed about a non-existing movie, deceivers talk about their Past, Assent, and use Motion words in order to support their lies. In contrast, while being on trial stating their (false) innocence, they use Anxiety, Anger, and negative emo-

Truthful			
Interviews		Trials	
Class	Score	Class	Score
Metaphor	2.98	You	3.99
Money	2.74	Family	3.07
Inhibition	2.74	Home	2.45
Home	2.13	Humans	1.87
Humans	2.02	Posemo	1.81
Family	1.96	Insight	1.64
Deceptive			
Interviews		Trials	
Class	Score	Class	Score
Assent	4.81	Anger	2.61
Past	2.59	Anxiety	2.61
Sexual	2.00	Certain	2.28
Other	1.87	Death	1.96
Motion	1.68	Physical	1.77
Negemo	1.44	Negemo	1.52

Table 6: LIWC word classes most strongly associated with deception and truth.

tion words (class Negemo). In line with earlier observations (Mihalcea and Strapparava, 2009), deceptive texts include more words that reflect certainty (class Certain, with words such as *completely, truly, always*) and more references to others (class Other, with words such as *she, day, him*).

4.2 Domain Experiments

We perform three sets of experiments to determine the role played by the domain. The first set of experiments uses only the *Interviews* video clips (62 in total), and the results are shown in the left column of Table 7. The second set uses only the *Trials* instances (56 in total), with results shown in the right column of Table 7. Finally, we also perform cross-domain experiments, with the training data drawn from one domain and the test data from the other. The results of these experiments are shown in Table 8. Given the uneven distribution of the truthful and deceptive video clips in two domains, the baselines are 54.83% for the *Interviews* domain (34 truthful, 28 deceptive), and 55.35% for the *Trials* domain (25 truthful, 31 deceptive).

What we learn from these experiments is that the domain does matter. Despite the smaller dataset, the experiments run on one domain at a time lead to results that are higher than the ones obtained with more data but with a mix of domains. The cross-domain experiments also support this argument, as the performance drops significantly when there is no overlap in domain between the training and the test instances. Overall, in all our machine learning experiments, the combined classifier that makes use of all the verbal and nonverbal features achieves the best trade-off between performance and robustness, as it always leads to the best or second best performance across all the experiments using individual or combined feature sets. While a classifier based on an individual feature set can sometime lead to a better performance (e.g., the Facial Displays classifier has better performance when all the video clips are used), that same classifier may not perform well in another setting (e.g., the Facial Displays classifier is significantly below the All Features classifier in the domain experiments).

Feature Set	Interviews	Trials
Baseline	54.83%	55.35%
Unigrams	75.80 %	82.14%
Psycholinguistics	59.67%	50.00%
Syntactic Complexity	54.83%	60.71%
Facial Displays	70.96%	80.35%
Hand Gestures	56.45%	48.21%
Unigr.+Facial Disp.	70.96%	76.78%
All Verbal	70.96%	64.28%
All Nonverbal	67.14%	83.92%
All features	79.03%	82.14%

Table 7: Deception classifiers for the *Interviews* and *Trials* domains, using a SVM classifier trained on individual and combined sets of verbal and nonverbal features.

Training	Test	SVM
Trials	Interviews	58.06%
Interviews	Trials	58.92%

Table 8: Cross-domain classification results using a SVM classifier trained on all the features

5 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.

We conduct a study where we evaluate the human ability to identify deceit when exposed to four different modalities: *Text*, consisting of the language transcript; *Audio*, consisting of the audio track of the clip; *Silent video*, consisting of only the video with muted audio; and *Full video*, where

Modality	Agreement	Kappa
Text	58.80%	0.047
Audio	66.70%	0.288
Silent video	52.00%	0.065
Full Video	61.60%	0.191

Table 9: Agreement among three human annotators on text, audio, silent video, and full video modalities.

	Text	Audio	Silent video	Full video
A1	54.24%	58.47%	50.85%	63.00%
A2	55.93%	67.80%	45.76%	68.00%
A3	65.25%	70.34%	55.93%	71.00%
Sys.	65.75%	NA	75.42%	77.11%

Table 10: Performance of three annotators and the developed automatic system (Sys) on the real-deception dataset over four modalities.

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.

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. Furthermore, 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.

Three annotators labeled all the 118 video clips in the dataset. Since four modalities were extracted from each video, each annotator annotated a total of 412 instances. Annotators were not offered a monetary reward and we considered their judgments to be honest as they participated voluntarily in this experiment. Table 9 shows the observed agreement and Kappa statistics among the three annotators for each modality.[3] The agreement for most modalities is rather low and the Kappa scores range between slight to fair agreement. As noted before (Ott et al., 2011), this low agreement can be interpreted as an indication that people are poor judges of deception.

We also determine each annotator's performance for each modality. The results, shown in Table 10, additionally support the argument that human judges have difficulty performing the deception detection task. An interesting, yet perhaps unsurprising observation is that the human performance increases with the availability of modalities. The 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 (Aamodt and Custer, 2006). Moreover, the performance of the human annotators appears to be significantly below that of our system.

6 Related Work

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 (Toma and Hancock, 2010; Guadagno et al., 2012), forums (Warkentin et al., 2010; Joinson and Dietz-Uhler, 2002), social networks (Ho and Hollister, 2013), and consumer report websites (Ott et al., 2011; Li et al., 2014). 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 (Mihalcea and Strapparava, 2009), and also more complex linguistic features derived from syntactic CFG trees and part of speech tags (Feng et al., 2012; Xu and Zhao, 2012). Research work has also relied on the LIWC lexicon to build deception models using machine learning approaches (Mihalcea and Strapparava, 2009; Ángela Almela et al., 2012) and showed that the use of psycholinguistic information is 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 (Yancheva and Rudzicz, 2013).

Nonverbal Deception Detection. Earlier approaches to nonverbal deception detection relied

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

on polygraph tests to detect deceptive behavior. These tests are mainly based on such physiological features such as heart rate, respiration rate, skin temperature. Several studies (Vrij, 2001; Gannon et al., 2009; Derksen, 2012) indicated that relying solely on physiological measurements can be biased and misleading. Chittaranjan et al. (Chittaranjan and Hung, 2010) created an audio visual recording 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. For hand gestures, blob analysis was used to detect deceit by tracking the hand movements of the subjects (Lu et al., 2005; Tsechpenakis et al., 2005), or using geometric features related to the hand and head motion (Meservy et al., 2005). Caso et al. (Caso et al., 2006) identified particular hand gestures that can be related to the act of deception using data from simulated interviews. Cohen et al. (2010) found that fewer iconic hand gestures were a sign of a deceptive narration, and Hillman et al. (2012) 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.,) for deception and social behavior. Facial expressions also played a critical role in the identification of deception. (Ekman, 2001) 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 (Ekman, 2003). Tian et al. (Tian et al., 2005) considered features such as face orientation and facial expression intensity. Owayjan et al. (Owayjan et al., 2012) extracted geometric-based features from facial expressions, and Pfister and Pietikainen (Pfister and Pietikäinen, 2012) 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 (Burgoon et al., 2009; Jensen et al., 2010). A multimodal deception dataset consisting of linguistic, thermal, and physiological features was introduced in (Pérez-Rosas et al., 2014), which was then used to develop a multimodal deception detection system (Abouelenien et al., 2014). An extensive review of approaches for evaluating human credibility using physiological, visual, acoustic, and linguistic features is available in (Nunamaker et al., 2012).

7 Conclusions

In this paper we presented a study of multimodal deception detection using real-life occurrences of deceit. We introduced a novel dataset covering recordings from public real trials and street interviews, and used this dataset to perform both qualitative and quantitative experiments. Our analysis of nonverbal 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 nonverbal features, and showed that we can achieve accuracies in the range of 77-82%.

Additional analyses showed the role played by the various feature sets used in the experiments, and the importance of the domain. To place our results in perspective and better understand the difficulty of the task, we performed a study of human ability to detect deception, which revealed high disagreement among the annotators. Our automatic system outperforms the human detection of deceit by 6-15%.

To our knowledge this is the first work to automatically detect instances of deceit using both verbal and nonverbal features extracted from real deception data. In order to develop a fully automated deception deception system, our future work will address the use of automatic gesture and facial expression identification and automated speech transcription. Our goal is to move forward towards a real-time deception detection system.

The dataset introduced in this paper is publicly available from http://lit.eecs.umich.edu.

Acknowledgments

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.

References

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

Mohamed Abouelenien, Veronica Pérez-Rosas, Rada Mihalcea, and Mihai Burzo. 2014. Deception detection using a multimodal approach. In *Proceedings of the 16th International Conference on Multimodal Interaction*, pages 58–65, New York, NY, USA. ACM.

Jens Allwood, Loredana Cerrato, Kristiina Jokinen, Costanza Navarretta, and Patrizia 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.

Ángela Almela, Rafael Valencia-García, and Pascual 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*, pages 15–22, Avignon, France, April.

Judee K. Burgoon, Douglas P. Twitchell, Matthew L. Jensen, Thomas O. Meservy, Mark Adkins, John Kruse, Amit V. Deokar, Gabriel Tsechpenakis, Shan Lu, Dimitris N. Metaxas, et al. 2009. Detecting concealment of intent in transportation screening: A proof of concept. *IEEE Transactions on Intelligent Transportation Systems*, 10(1):103–112, March.

Letizia Caso, Fridanna Maricchiolo, Marino Bonaiuto, Aldert Vrij, and Samantha Mann. 2006. The impact of deception and suspicion on different hand movements. *Journal of Nonverbal Behavior*, 30(1):1–19.

Gokul Chittaranjan and Hayley 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)*, pages 5334–5337, March.

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

Bella Depaulo, Brian Malone, James Lindsay, Laura Muhlenbruck, Kelly Charlton, and Harris Cooper. 2003. Cues to deception. *Psychological Bulletin*, pages 74–118.

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

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

Paul 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.

Song Feng, Ritwik Banerjee, and Yejin Choi. 2012. Syntactic stylometry for deception detection. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 171–175, Jeju Island, Korea, July. Association for Computational Linguistics.

Theresa Gannon, Anthony Beech, and Tony Ward, 2009. *Risk Assessment and the Polygraph*, pages 129–154. John Wiley and Sons Ltd.

Rosanna E. Guadagno, Bradley M. Okdie, and Sara A. Kruse. 2012. Dating deception: Gender, online dating, and exaggerated self-presentation. *Comput. Hum. Behav.*, 28(2):642–647, March.

Jackie Hillman, Aldert Vrij, and Samantha Mann. 2012. Um they were wearing : The effect of deception on specific hand gestures. *Legal and Criminological Psychology*, 17(2):336–345.

Julia Hirschberg, Stefan Benus, Jason M. Brenier, Frank Enos, Sarah Friedman, Sarah Gilman, Cynthia Girand, Martin Graciarena, Andreas Kathol, Laura Michaelis, et al. 2005. Distinguishing deceptive from non-deceptive speech. In *In Proceedings of Interspeech 2005 - Eurospeech*, pages 1833–1836.

Shuyuan Mary Ho and Jonathan M Hollister. 2013. 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.

Matthew Jensen, Thomas Meservy, Judee Burgoon, and Jay Nunamaker. 2010. Automatic, multimodal evaluation of human interaction. *Group Decision and Negotiation*, 19(4):367–389.

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

Jiwei Li, Myle Ott, Claire Cardie, and Eduard Hovy. 2014. 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.

Shan Lu, Gabriel Tsechpenakis, Dimitris Metaxas, Matthew Jensen, and John Kruse. 2005. 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.

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

Fridanna Maricchiolo, Augusto Gnisci, and Marino Bonaiuto. In Anna Esposito, AntoniettaM. Esposito, Alessandro Vinciarelli, Rdiger Hoffmann, and VincentC. Mller, editors, *Cognitive Behavioural Systems*, pages 405–416. Springer Berlin Heidelberg.

Thomas Meservy, Matthew Jensen, John Kruse, Douglas Twitchell, Gabriel Tsechpenakis, Judee Burgoon, Dimitris Metaxas, and Jay Nunamaker. 2005. Deception detection through automatic, unobtrusive analysis of nonverbal behavior. *IEEE Intelligent Systems*, 20(5):36–43, September.

Pamela Meyer. 2010. *Liespotting: Proven Techniques to Detect Deception*. New York: St. Martin's.

Rada Mihalcea and Stephen Pulman. 2009. Linguistic ethnography: Identifying dominant word classes in text. In *Computational Linguistics and Intelligent Text Processing*, pages 594–602. Springer.

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

Matthew L. Newman, James W. Pennebaker, Diane S. Berry, and Jane M. Richards. 2003. Lying words: Predicting deception from linguistic styles. *Personality and Social Psychology Bulletin*, 29.

Jay F. Nunamaker, Judee K. Burgoon, Nathan W. Twyman, Jeffrey Gainer Proudfoot, Ryan M. Schuetzler, and Justin Scott Giboney. 2012. Establishing a foundation for automated human credibility screening. In *2012 IEEE International Conference on Intelligence and Security Informatics (ISI)*, pages 202–211, June.

Myle Ott, Yejin Choi, Claire Cardie, and Jeffrey 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, pages 309–319, Stroudsburg, PA, USA. Association for Computational Linguistics.

Michel Owayjan, Ahmad Kashour, Nancy Al Haddad, Maurice Fadel, and Ghinwa Al Souki. 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)*, pages 33–38, Dec.

James W. Pennebaker and Martha E. Francis. 1999. Linguistic inquiry and word count: LIWC. Erlbaum Publishers.

Verónica Pérez-Rosas, Rada Mihalcea, Alexis Narvaez, and Mihai Burzo. 2014. A multimodal dataset for deception detection. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014), Reykjavik, Iceland, May 26-31, 2014.*, pages 3118–3122.

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

Madeline Smith, Jeffrey Hancock, Lindsay Reynolds, and Jeremy Birnholtz. 2014. Everyday deception or a few prolific liars? the prevalence of lies in text messaging. *Computers in Human Behavior*, 41(0):220–227.

Ying-Li Tian, Takeo Kanade, and Jeffrey F. Cohn. 2005. Facial expression analysis. In *Handbook of Face Recognition*, pages 247–275. Springer New York.

Catalina L. Toma and Jeffrey T. 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, pages 5–8.

Gabriel Tsechpenakis, Dimitris Metaxas, Mark Adkins, John Kruse, Judee K. Burgoon, Matthew L. Jensen, Thomas Meservy, Douglas P. Twitchell, Amit Deokar, and Jay F. Nunamaker. 2005. Hmm-based deception recognition from visual cues. In *IEEE International Conference on Multimedia and Expo, 2005. ICME 2005*, pages 824–827, July.

Aldert 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.

Darcy Warkentin, Michael Woodworth, Jeffrey T Hancock, and Nicole Cormier. 2010. Warrants and deception in computer mediated communication. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work*, pages 9–12. ACM.

Peter Wittenburg, Hennie Brugman, Albert Russel, Alex Klassmann, and Han Sloetjes. 2006. Elan: a professional framework for multimodality research. In *Language Resources and Evaluation*, volume 2006.

Qiongkai Xu and Hai Zhao. 2012. Using deep linguistic features for finding deceptive opinion spam. In *Proceedings of COLING 2012: Posters*, Mumbai, India, December.

Maria Yancheva and Frank Rudzicz. 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)*, pages 944–953, Sofia, Bulgaria, August. Association for Computational Linguistics.