Multimodal Deception Detection using Real-Life Trial Data

M. Umut Şen, Verónica Pérez-Rosas, Berrin Yanikoglu, Mohamed Abouelenien, Mihai Burzo, Rada Mihalcea

Abstract—Hearings of witnesses and defendants play a crucial role when reaching court trial decisions. Given the high-stakes nature of trial outcomes, developing computational models that assist the decision-making process is an important research venue. In this paper, we address the identification of deception in real-life trial data. We use a 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. In particular, three complementary modalities (visual, acoustic and linguistic) are evaluated for the classification of deception at the subject level. The final classifier is obtained by combining the three modalities via score-level classification, achieving 83.05% accuracy in subject-level deceit detection. To place our results in perspective, we present a human deception detection study where we evaluate the human capability of detecting deception using different modalities and compare the results to the developed system. The results show that our system outperforms the average non-expert human capability of identifying deceit.

real-life trial; deception detection; classification; multimodal; visual; acoustic; linguistic

1 Introduction

With thousands of trials and verdicts occurring daily in courtrooms around the world, there is a high chance of using deceptive statements and testimonies as evidence. 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.

The consequences of falsely accusing the innocents and freeing the guilty can be severe. For instance, in the U.S. alone there are tens of thousands 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 [1]. Hence, the need arises for a reliable and efficient system to aid the task of detecting deceptive behavior and discriminate between liars and truth-tellers.

Traditionally, law enforcement entities have made use of the polygraph test as a standard method to identify deceptive behavior. However, this approach becomes impractical in some cases, as it requires the use of skin-contact devices and human expertise to get accurate readings and interpretation. In addition, the final decisions are subject to error and bias not only from the device itself but also from human judgment [2], [3]. Furthermore, using proper countermeasures, offenders can deceive these devices as well as human experts.

Given the difficulties associated with the use of polygraph-like methods, machine learning-based approaches have been proposed to address the deception detection problem using several modalities, including text [4] and speech [5], [6]. 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 crowd-sourcing [7], [8], for instance by asking subjects to narrate stories deceptively and truthfully [7], by performing one-on-one interviews, or by participating in "mock crime" scenarios [8].

Despite their potential benefits, an important drawback in 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 or highly motivated to lie, thus making it difficult to generalize findings to real-life scenarios.

In this paper, we present a multimodal system that detects deception in real-life trial data using verbal, acoustic and visual modalities. The data consists of video clips

- *Verónica Pérez-Rosas (corresponding author) and Rada Mihalcea are with the Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI. (Emails: {vrncapr, mihalcea}@umich.edu)*
- *Mohamed Abouelenien is with the Department of Computer and Information Science, University of Michigan-Dearborn, Dearborn, MI. (Email: {zmohamed}@umich.edu)*
- *Mihai Burzo is with the Department of Mechanical Engineering, University of Michigan Flint, Flint, MI. (Email: {mburzo}@umich.edu)*
- *M. Umut Şen and Berrin Yanikoglu are with the Faculty of Engineering and Natural Sciences, Sabanci University, Istanbul, Turkey.*

This work is partially supported by a TÜBİTAK 2219 scholarship to B. Yanikoglu and 2211-A scholarship to M. Umut Şen.
(Emails: {umutsen,berrin}@sabanciuniv.edu)

1. www.uscourts.gov

obtained from real court trials, initially presented in [9].

Unlike previous work on this dataset, which focuses on detecting deception at the video-level, we aim to detect deception at the subject-level. We believe this is more in line with the ground-truth for this dataset since it was also obtained at the subject-level: defendants who are found guilty at the end of the trial are labeled as deceptive since they had not admitted to their guilt during the hearings. In the remainder of the paper, we will refer to this task as a subject-level deception classification.

Our main contributions are as follows:

- We introduce the subject-level deception detection as a novel take on the problem; we argue that it is also more appropriate for the problem given the way ground-truth is established.
- We explore the effectiveness of a diverse set of features extracted from the linguistic, visual, and acoustic channels, both separately and in combination (using early and late fusion methods).
- We repeat the experiment 3 times with different random seed for each test sample in the leave-one-out cross-validation, to obtain more reliable scores with the small dataset.
- We present a semi-automatic system that can identify deception with 83.05% accuracy using a combination of automatically extracted and manually annotated features, as well as a fully-automatic system that reaches almost 73% accuracy.
- We place our results in context by performing a study where humans evaluate the presence of deception in the real-life trial dataset.
- We present insights into the problem by analyzing the importance of features obtained manually and automatically, as well as the linguistic differences among deceptive and truthful subjects.

2 DATASET

During our experiments, we use a multimodal deception dataset obtained from real-life court trials. The dataset description is included here for completeness; further details can be found in [9].

2.1 Dataset Overview

The dataset consists of trial hearing recordings obtained from public sources. The videos were carefully selected to be of reasonably good audio-visual quality and portray a single subject with his/her face visible during most of the clip duration.

Videos are collected from trials with different outcomes: guilty verdict, non-guilty verdict, and exoneration. 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 whereas testimonies in favor of a guilty suspect are labeled as deceptive. Exoneration testimonies are collected as truthful statements.

TABLE 1
Distribution of gender in the two categories after aggregating individual videos.

	Female	Male	Total
Deceptive	11	13	24
Truthful	12	23	35
Total	23	36	59

The dataset includes several famous trials (including trials of Jodi Arias, Donna Scrivo, Jamie Hood, and others), police interrogations, and also statements from the "The Innocence Project" website.[2]

2.2 Subject-level Ground-truth

In the original dataset, the ground-truth was obtained at video level, by carefully identifying and labeling truthful and deceptive video clips from trial's recordings [9].

In this work, we focus on deception at the subject level for two reasons: 1) it is difficult to know the ground-truth of all video clips with certainty and 2) the ultimate goal is to determine whether an individual is being deceptive or not, rather than pinpoint exactly when s/he is lying. Note that subject-level decision is what human jurors are also asked to accomplish during real life trials consisting of several interrogation episodes.

To obtain subject-level ground truth, we only used the trial outcomes to indicate the subject as deceptive or not (deceptive in case of a guilty verdict vs not-deceptive in case of non-guilty verdict or exoneration). The resulting subject-level dataset has 59 instances, and the distributions of male vs female and deceptive vs truthful are given in Table 1.

Note that a subject-level deception detection system can be evaluated fairly, by comparing its predictions to the subject-level ground-truth, which is the trial outcome, with the assumption that the trial outcome is correct.

2.3 Transcriptions

The transcriptions are obtained using Amazon Mechanical Turk in the original dataset. In video clips where multiple speakers are portrayed (i.e., defendants or witnesses being questioned by attorneys), the AMT workers were asked to transcribe only the subject's speech, including word repetitions, fillers such as *um*, *ah*, and *uh*, and intentional silences encoded as ellipsis.

The final set of transcriptions consists of 8,055 words, with an average of 66 words per transcript. Table 2 shows transcriptions of sample deceptive and truthful statements.

2.4 Visual Behavior Annotations

Gesture annotations are also available in the dataset.[3] The annotation was conducted using the MUMIN [10] multi-

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

3. As done in the Human Computer Interaction Community, "gesture" is used as a broad term that refers to body movements, including facial expressions and hand gestures.

TABLE 2
Sample transcripts for deceptive and truthful clips in the dataset.

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.

Fig. 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*).

modal scheme, which includes several different facial expressions associated with overall facial expressions, eyebrows, eyes and mouth movements, gaze direction, as well as head and hand movements. Sample screenshots showing facial displays and gestures by deceptive and truthful subjects in the dataset are shown in Figure 1.

This annotation was done at the video-level by identifying the facial displays and hand gestures that were most frequently observed during the entire clip duration. Two annotators independently labeled a sample of 56 videos. The inter-annotator agreement for this task is shown in Table 3. 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).

As a preliminary analysis, Figure 2 shows the percentages of all the non-verbal features for which we observe noticeable differences for the deceptive and truthful groups. The figure suggests eyebrow (rise) helps differentiate between the deceptive and truthful conditions. Twyman et al. reported that deceivers' right hand moves less during a mock crime experiment [11]. This coincides with our single and both hands movement analysis as depicted in Figure 2. ten Brinke and Porter [12] reported that deceptive people blink at a faster rate than genuinely distressed individuals; which also coincides with our findings that deceivers display more frequent occurrence of rapid eye closures, as seen in Fig. 2.). Interestingly, deceivers seem to shake their head (Side-Turn-R) and nod (Down-R) less frequently than truth-tellers while true-tellers seem to move their hands more frequently.

TABLE 3
Gesture annotation agreement

Gesture Category	Agreement	Kappa Score
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

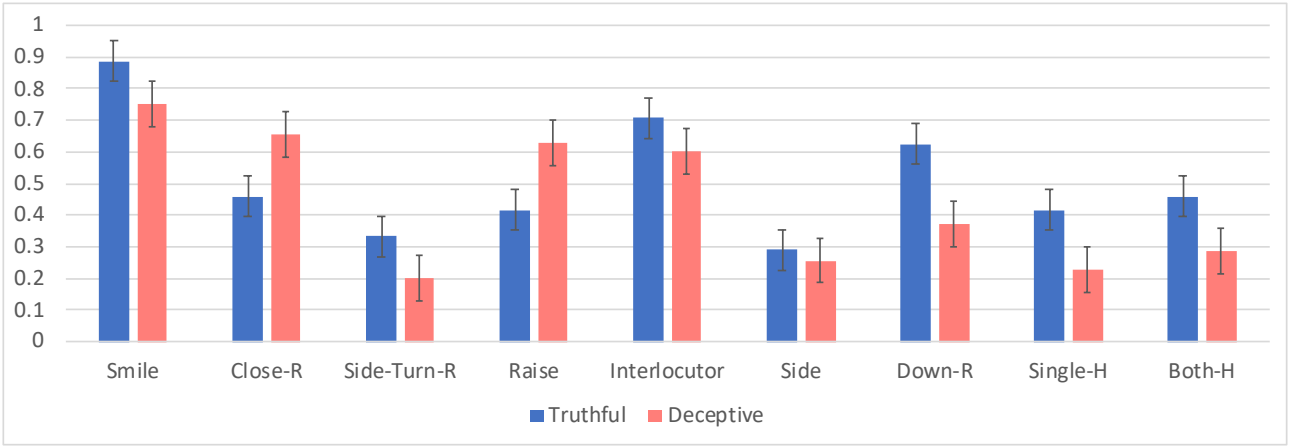

Fig. 2. Distribution of important visual features for deceptive and truthful groups: Smile, Close-R (closing eyes repeatedly), Side-Turn-R (head turning sides repeatedly), Raise (eyebrow raising), Interlocutor (gazing towards interlocutor), side (gazing to the sides), Down-R (moving the head downwards repeatedly), Single-H (single hand movement), Both-H (moving both hands)

3 Features for Deception Detection

Aiming to explore the subject-level deception detection with different levels of supervision, we conduct two main experiments using features obtained either manually or semi-automatically. We first present a semi-automatic system where the linguistic and visual feature extraction is done based on manual annotations, as described in Section 2. Second, we build a fully-automatic system that does not rely on human input. Finally, we compare the results with that of human performance on deception detection.

Given the multimodal nature of our dataset, we were interested to evaluate the usefulness of the linguistic, visual, and acoustic components of the recordings, both individually and in combination. Note that automatic temporal analysis of the videos would be significantly more complicated to accomplish and would require a larger dataset to prevent overfitting; hence it is outside of the scope of this paper. The feature extraction process is detailed below.

3.1 Linguistic Features

We experimented with linguistic features that have been previously found to correlate with deception cues [13], [14]. These features are derived from the text transcripts of the subjects' statements.

Unigrams We extract unigrams derived from the bag of words representation of each transcript. Each feature consists of frequency counts of unique words in the transcript. For this set, we keep only words with a frequency greater than or equal to 10. The threshold cut was experimentally obtained in a small development set.

LIWC We use features derived from the Linguistic Inquire Word Count (LIWC) lexicon [14]. These features consist of word counts for each of the 80 semantic classes in LIWC. For instance, the class "I" includes words associated with the self (e.g., I, me, myself); "Other" includes words associated with others (e.g., he, she, they); etc.

3.2 Annotated Visual Behaviour Features

One set of visual features are derived from the annotations performed using the MUMIN coding scheme described in Section 2.4. 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. The generated features represent nine different gesture categories listed in Table 3, covering 32 facial displays and 7 hand gestures.

Facial Displays. These are facial expressions or head movements displayed by the speaker during the deceptive or truthful interaction. They include overall facial expressions such as smiling and scowling; eyebrows, eyes and mouth movements (e.g. repeated eye closing or protruded lips); gaze direction (e.g. looking down or towards the interlocutor); and as well as head movements (e.g. repeated nodding or shaking) and hand movements.

Hand Gestures. The second broad category covers gestures made with the hands, including movements of one or both hands and their trajectories.

3.3 Automatically Extracted Visual Features

We automatically extract a second set of visual features consisting of assessments of several facial movements as described below:

Facial Action Units (FACS). These features denote the presence of facial muscle movements that are commonly used for describing and classifying expressions [15].

We use the OpenFace library [16] with the default multi-person detection model to obtain 18 binary indicators of Action Units (AUs) for each frame in our videos. These include: AU1 (inner brow raiser), AU2 (outer brow raiser), AU4 (brow lowerer), AU5 (upper lid raiser), AU6 (cheek raiser), AU7 (eyelid tightener), AU9 (nose wrinkler), AU10 (upper lip raiser), AU12 (lip corner puller), AU14 (dimpler), AU15 (lip corner depressor), AU17 (chin raiser), AU20 (lip stretcher), AU23 (lip tightener), AU25 (lips part), AU26 (jaw

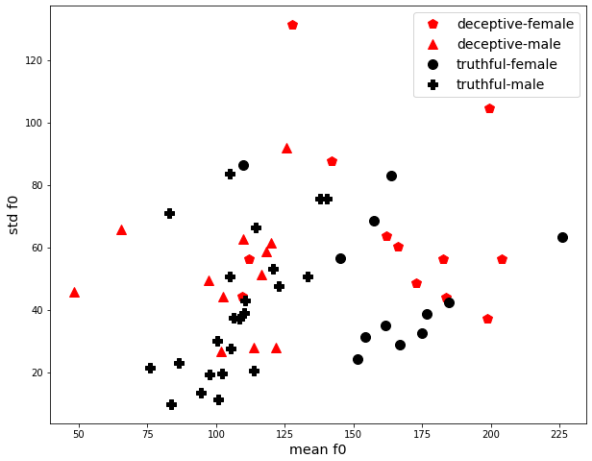

Fig. 3. Pitch standard deviation vs pitch mean by gender.

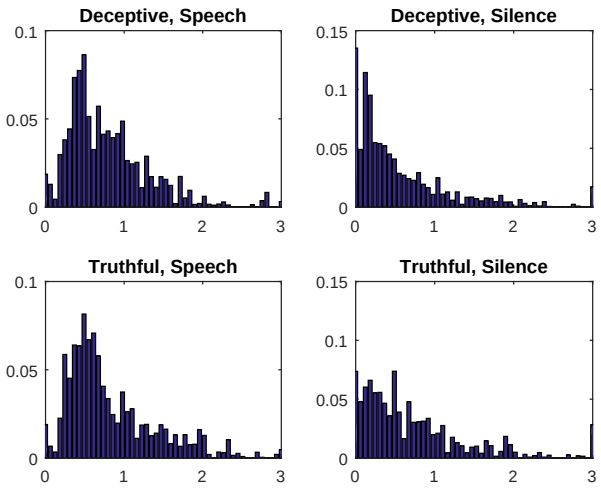

Fig. 4. Histograms of speech and silence length (measured in seconds) using 25 bins. In all cases, the last bin contains speech or silence segments with duration greater than 3 seconds.

drop), AU28 (lip suck), and AU45 (blink). We average these binary indicators through the frames and obtain a single AU feature for each video.

3.4 Acoustic Features

Previous work has suggested that pitch is an indicator of deceit, and showed that people tend to increase their pitch when they are being deceptive [17]. This motivated us to explore whether subjects will show particular pitch differences in their speech while telling the truth or deceiving.

In addition to pitch, we extracted acoustic features for voiced segments and pauses, based on previous findings showing that deceivers produce slightly shorter utterances and pause more frequently than true-tellers [18]. The extracted acoustic features are as follows.

Pitch. We derive features from pitch measurements in the audio portion of each video in the dataset. To estimate pitch, we obtained the fundamental frequency (f0) of the defendants' speech using the STRAIGHT toolbox [19]. Since f_0 is defined only over voiced parts of the speech, we remove unvoiced speech frames from our calculations. We then derive two features (mean and standard deviation) from the raw f_0 measurements: mean$-f_0$ and stdev$-f_0$.

Silence and Speech Histograms. To obtain these features, we run a voice activity detection (VAD) algorithm [20] to obtain the speech and silent segments in the subject's speech. Since the performance of VAD algorithms is affected by the segmentation threshold θ, i.e., high values of θ result on over-segmentation while low values produce under segmentation, we experiment with two values of θ to improve the VAD segmentation in our data: 0.01 and 0.2. After manual inspection, we observed that using a threshold of 0.2, the algorithm segment the audio into words rather than full sentences while a threshold of 0.01 produces full sentence segmentation. Using a VAD threshold of 0.2, with the intent of capturing short pauses, we extract the histograms (using 25 bins) of both voiced and silent segments as features.

Figure 3 shows the distribution of the mean and standard deviation of pitch frequencies for the deceptive and truthful groups by gender. As can be seen in this figure, pitch mean values depend on the gender, while standard deviation seems more correlated with deception. Figure 4 depicts the histograms of speech and silent lengths by deceptive and truthful subjects. Interestingly, the plot shows that deceptive individuals tend to make shorter pauses more frequently than truthful individuals.

3.5 Subject-level Feature Integration

Since our feature extraction is performed in each video clip separately for visual features and there are cases where there is more than one video for a single subject, we devised two strategies to aggregate the features across all videos from the same subject. First, taking the maximum values per feature across feature vectors corresponding to every subject's video. Second, averaging the feature values across feature vectors corresponding to each subject's video.

Taking the maximum of the feature values aims to represent single events (e.g., eyes blinking), even if it is observed in just one of the videos belonging to a subject. Averaging the feature values, on the other hand, aims to reduce potential noise introduced during the manual annotation.

During our initial experiments, we found that the averaging strategy outperforms the use of maximum values, hence the former is used during the rest of the experiments reported in the paper.

4 Classifiers

We chose the Random Forest (RF), Support Vector Machine (SVM) with Radial Basis Function kernel and Neural Network (NN) classifiers, due to their success in many other machine learning problems. For the RF and SVM, we use their implementations as available in Matlab. We use the PyTorch library for the implementation of the NN classifiers [21]. During our experiments, all classifiers are evaluated using accuracy and area under the curve (AUC) as our main performance metrics.

For the SVM classifiers, we performed parameter tuning over the training set using 4-fold cross-validation separately for each test instance. Specifically, we tune the penalty (C)

and the γ parameters of the RBF kernel using grid-search. We applied a 3×3 averaging filter to the resulting loss matrix of the grid search to smooth the parameter tuning results to reduce the noise that results from the low number of data points.

For the RF classifiers, we used the default value for the number of trees (100) and minimum leaf size of 3, without doing parameter optimization.

For the NN classifier, we used a two hidden layers network (100 and 500 nodes for the hidden layers) along with a softmax activation function and a cross-entropy loss function. L_2 regularization is applied with a weight of $1E-5$, to prevent over-fitting.

A strong advantage of using RF and the NN classifiers is that they are quite insensitive to the values of their meta-parameters. For instance, when evaluated with different number of hidden nodes in either layer $\{(10, 100), (100, 100), (500, 500), (500, 10), (100, 10), (10, 500)\}$, the NN showed a performance variation of only 1%.

5 Semi-Automatic Deception Detection

We develop a semi-automatic system using features derived from manually annotated modalities (visual and linguistic), along with automatically extracted features (speech). Thus, we run several comparative experiments using leave-one-out cross-validation where we test in a single test subject and train in the remaining ones. Furthermore, we run all experiments three times with different random seeds and report the mean and the standard deviation of the results.

5.1 Results for Individual Modalities

We initially conduct experiments using each feature set independently and then experiment with different feature combinations using the SVM, RF, and NN classifiers. Table 4 shows the results for individual and combined sets of features in each modality.

Among the different classifiers, the RF classifier is the best classifier for linguistic and acoustic features, while the NN performs best with the visual features. For the visual features, the best results are achieved with the facial displays, reaching an accuracy of 80.79% and an AUC score of 0.94. These results also constitute the best results across individual feature sets. For the acoustic features, the best performing feature is the $pitch\_stdv$, which represents the standard deviation of the subject's pitch, resulting in an accuracy of 71.19% and an AUC score of 0.79. The rest of the acoustic features obtain significantly lower performance than $pitch\_stdv$ alone. For the linguistic features, the classifier built with the unigram features outperformed both LIWC features alone and its combination with LIWC features. The highest accuracy with lexical features is 64.41% with the RF classifier.

5.2 Results for Combined Modalities

For the multi-modal approach, we conduct experiments using two different integration strategies of the three modalities in our dataset: early fusion and late fusion.

5.2.1 Early Fusion

First, we experiment with *early fusion* by concatenating the best performing feature sets from the three modalities and using the different classifiers. Results are shown in Table 5

During these experiments, the NN classifier consistently obtains the best results among different feature combinations as well as the lowest standard deviation through 3 repetitions of the experiments. Among the different combinations, the combination of features encoding the facial displays, pitch and silence and speech histograms achieve the highest accuracy (83.05%), improving the accuracy obtained with facial display features only by 2.26% points. However, in terms of the AUC, the combination of facial displays and the pitch standard deviation performs the best (0.95).

5.2.2 Late Fusion

Second we use *score-level fusion* with classifiers built for individual modalities. For these experiments, we use only the best classifiers and features, leaving out the SVM classifier and hand's gesture features. The aggregated score s_i is obtained as shown in Equation 1, where s_{ij} is the score of class c_i obtained with the classifier h_j and w_j is the weight assigned to the classifier h_j.

$$s_i = \sum_j w_j s_{ij} \tag{1}$$

We use different classifier weights for the facial displays using increments of 0.1 (the remaining weights are assigned equally to the other classifiers) and report results on the test set. Thus, the best scoring setting is obtained *a posteriori*.

Classification results obtained with this strategy are shown in Table 6. We observe that the best result (84.18%) is obtained using the NN classifier and the combination of visual features and acoustic features. This result is higher than the best result obtained with early fusion since it finds the best weights over the test set; but the improvement is very small. The best early fusion results are reported as the proposed system's result, throughout the paper.

6 Fully-Automatic Deception Detection

We also conducted a set of experiments where we explore how well fully automatic feature extraction would work, for our task. Since our acoustic features are already obtained using automatic methods, we focus on the automatic extraction of linguistic and visual features.

We used the OpenFace library [16] with the default multi-person detection model, to obtain the facial action units (see Section 3.3) for the subject in the video. To address cases where the model identifies multiple persons in the frames, we select the person who is present in the majority of frames as the person of interest. We manually verified the result of this heuristic and confirmed that in most cases the selection corresponds to the main subject in the video. The software was unable to identify the subject's face in four videos in the dataset, due to the low video quality. These videos are nonetheless included in the evaluation, so as to measure the performance of the system under realistic conditions.

TABLE 4
Individual feature performance: accuracy (%) and AUC scores. Best results in each line are shown in bold.

Feature Set (dimension)	SVM		RF		NN	
	Accuracy	AUC	Accuracy	AUC	Accuracy	AUC
Visual						
Facial displays (32)	76.27± 0.00	0.8581	76.27 ± 1.69	0.9270	**80.79** ± 0.98	0.9416
Hand gestures (7)	50.28 ± 3.53	0.7232	**64.97** ± 3.91	0.6671	61.58 ± 0.98	0.6930
All visual (39)	58.19 ± 0.98	0.8641	77.40 ± 0.98	0.9187	**78.53** ± 1.96	0.9377
Acoustic						
Pitch (std-f_0) (1)	61.58 ± 0.98	0.6507	**71.19** ± 3.39	0.7939	51.41 ± 0.98	0.7427
Pitch (mean-f_0) (1)	54.24 ± 1.69	0.5223	53.11 ± 0.98	0.5465	**61.02** ± 0.00	0.5235
Sil.Sp.Hist (50)	57.63 ± 0.00	0.4159	**59.32** ± 2.94	0.7069	55.93 ± 1.69	0.6483
All Acoustic (52)	56.50 ± 2.59	0.5864	**63.28** ± 0.98	0.7059	61.02 ± 4.48	0.6589
Linguistic						
Unigrams (134)	53.11 ± 1.96	0.7275	**64.41** ± 4.48	0.6173	63.28 ± 0.98	0.7651
Unigrams - LIWC (100)	52.54 ± 4.48	0.5906	**63.84** ± 2.59	0.6764	55.93 ± 1.69	0.7729
All Linguistic (234)	53.11 ± 4.27	0.6765	**61.58** ± 2.59	0.6605	57.63 ± 1.69	0.7655

TABLE 5
Early fusion results using individual best performing features: accuracy and AUC scores. Best results are shown in bold.

Modalities	SVM		RF		NN	
	Accuracy	AUC	Accuracy	AUC	Accuracy	AUC
Facial Displays	76.27 ± 0.00	0.8581	76.84 ± 0.80	0.9270	**80.79** ± 0.98	0.9416
+ Pitch (std-f_0)	53.11 ± 5.45	0.7148	62.71 ± 1.69	0.6511	**82.49** ± 0.98	**0.9462**
+ Pitch (std-f_0) + Sil.Sp.Hist.	68.93 ± 0.98	0.8585	66.10 ± 11.86	0.8649	**83.05** ± 1.69	0.9166
+ All Acoustic	72.32 ± 0.98	0.8604	67.80 ± 2.94	0.8482	**82.49** ± 0.98	0.9153
+ LIWC	54.24 ± 1.69	0.8173	63.84 ± 1.96	0.7778	**75.14** ± 1.96	0.8961
+ Unigrams	55.93 ± 3.39	0.7928	63.28 ± 0.98	0.7310	**78.53** ± 0.98	0.8903
+ Pitch (std-f_0) + Sil.Sp.Hist. + LIWC	53.67 ± 3.53	0.8281	65.54 ± 1.96	0.7852	**75.14** ± 1.96	0.8780
+ All Acoustic + Unigrams	57.06 ± 0.98	0.8072	63.84 ± 3.53	0.7494	**75.71** ± 0.98	0.8778

TABLE 6
Late fusion results using best performing features and different classifier weight combinations. Face refers to facial displays and pitch refers to std-f_0. The results are obtained *a posteriori* and best results are shown in bold.

Score	RF			NN		
	+Acoustic	+Linguistic	+Acoustic +Linguistic	+Acoustic	+Linguistic	+Acoustic +Linguistic
$w_{face} = 1.0$	76.84 ± 0.80			80.79 ± 0.98		
$w_{face} = 0.9$	77.40 ± 0.98	77.97 ± 1.69	77.40 ± 0.98	81.92 ± 0.98	83.05 ± 0.00	81.92 ± 0.98
$w_{face} = 0.8$	77.40 ± 2.59	77.97 ± 1.69	77.40 ± 0.98	83.62 ± 1.96	79.66 ± 0.00	83.05 ± 0.00
$w_{face} = 0.7$	79.10 ± 2.59	76.84 ± 0.98	78.53 ± 1.96	**84.18** ± 0.98	80.79 ± 0.98	81.92 ± 1.96
$w_{face} = 0.6$	76.84 ± 0.98	76.27 ± 0.00	76.84 ± 1.96	**84.18** ± 1.96	79.66 ± 1.69	82.49 ± 0.98
$w_{face} = 0.5$	76.27 ± 1.69	68.93 ± 2.59	74.01 ± 5.18	81.92 ± 0.98	78.53 ± 0.98	83.62 ± 1.96

To extract the linguistic features, we applied Automatic Speech Recognition (ASR) to the videos using the Google Cloud Speech API [22] and obtained the corresponding transcriptions. Then, as in the manual system, we use these transcriptions to extract unigram features. One shortcoming of the automation here is that the transcriptions also contain the interviewer's speech. Furthermore, the ASR failed to recognize any speech for 10 videos, which correspond to three subjects in the dataset. The obtained transcriptions resulted in an average Word Error Rate (WER) of 0.603 and an insertion rate of 0.152.

The results of the automatic deception system are depicted in Table 7. We see that the performance obtained by classifiers build with automatic visual features falls behind the performance obtained when using manual annotations, while automatic extraction of the linguistic features results in a similar performance. As for combined modalities, we see that the best result, 72.88%, (obtained with the fully automatic system, score-level combination, and the NN classifier) is significantly lower than the best performance with the semi-automatic system, 83.05%. However, we would expect the performance gap would to be smaller when using videos that have better visual quality e.g., videos obtained with high-resolution cameras focused on the subject's face.

TABLE 7
Fully-automatic system: classification accuracies with individual and combined modalities (%)

Modality	SVM	RF	NN
Visual (Action Units)	53.67%	61.58%	57.63%
All Acoustic	56.50%	63.28%	61.02%
Linguistic (Unigrams)	57.06%	63.28%	71.75%
All (Early Fusion)	58.76%	68.36%	70.06%
All (Combiner)	56.50%	63.28%	**72.88**%

7 HUMAN PERFORMANCE

As part of our work analyzing the importance of multimodal features in deception detection, 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 instances for each modality in random order to each annotator, 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, which portray 59 different subjects. To calculate the agreement at the subject-level, we apply majority voting to the labels assigned by each annotator over all the clips belonging to the same subject. We resolve ties by randomly choosing between the deceptive and truthful labels. Table 8 shows the observed agreement and Kappa statistics among the three annotators for each modality.[4] We observe that the agreement for most modalities is rather low and the Kappa scores show mostly poor agreement. As noted before by Ott et al. [23], this low agreement can be interpreted as an indication that people are poor judges of deception.

In addition, we compare the performance of the three individual annotators and the developed systems, over the four different modalities in the dataset. As shown in Table 9, we observe a positive trend in human accuracy in the subject-level deceit detection when using multiple modalities. The trend could be explained by having more deception cues available to them. On average, the poorest accuracy is obtained on *text* only, followed by *Audio*, *Silent video*, and *Full video*, where the annotators have the highest performance. Interestingly, we notice a similar pattern for the developed systems, where we see that having a greater amount of multimodal cues does help to improve the system performance. The fully-automatic system outperforms the average human performance when using each modality individually and in combination (72.88% versus 71.79%). Furthermore, it achieves almost 30% reduction in error compared to the lowest performing human annotator's performance. The semi-automatic system further improves the results of the fully automatic system when using the three modalities (full video), thus suggesting that the feature fusion strategy is also an important aspect when building these models..

Overall, our study indicates that detecting deception is indeed a difficult task for humans and further verifies previous findings where the average human ability to spot liars was found to be slightly better than chance [24]. Moreover, the performance of the human annotators appears to be significantly below that of the developed systems.

4. Inter-rater agreement with multiple raters and variables. https://nlp-ml.io/jg/software/ira/

TABLE 8
Agreement among three human annotators on text, audio, silent video, and full video modalities.

Modality	Agreement	Kappa
Text	30.76%	0.014
Audio	53.84%	0.040
Silent video	53.84%	0.040
Full video	53.84%	0.050

TABLE 9
Classification accuracy of three annotators (A1, A2, A3) and the developed systems on the real-deception dataset over four modalities.

	Text	Audio	Silent video	Full video
A1	69.23%	69.23%	69.23%	61.53%
A2	53.84%	61.53%	61.53%	76.92%
A3	69.23%	76.92%	76.92%	76.92%
Average	64.10%	69.22%	69.22%	**71.79**%
Fully-autom. sys.	71.75%	63.28%	61.58%	**72.88**%
Semi-autom. sys.	64.41%	63.28%	80.79%	75.71%

8 Insights for Deception Detection

8.1 Visual features

We compute the feature importance scores using the *predictorImportance* function of Matlab [25] that bases its estimate on the performance change in the random forest classifier, with the use of each feature. Importance measures of visual AU features are depicted in Figure 5. We see that features describing actions of lips reveal substantial deception information (Upper Lip Raiser, Lip Stretcher, Lip Tightener, Lip Corner Depressor, Lip Corner Puller). In addition, (eye)Lid Tightener, Nose Wrinkler, Brow Lowerer and Inner Brow Raiser also have high importance scores.

8.2 Deception Language in Trials

To obtain insights into linguistic behaviors displayed by liars during court hearings, we explore patterns in word usage according to their ability to distinguish between the subjects' deceptive and truthful statements. We thus trained a binary Naive Bayes (NB) classifier that discriminates between liars and true-tellers using the unigram features obtained from the subject's statements. We then use the NB model to infer the expected probabilities of each word given its class label. We then sort the words by importance using the following scoring formula:

$$s_i = E[f_i|class = deceptive]/E[f_i|class = truthful], \quad (2)$$

In this equation, the expectation E of the word f_i is compared across the deceptive and truthful classes. Note that expectation values are obtained from the resulting NB model rather than empirically from the dataset. The words that are more strongly associated with the deceptive and truthful groups are shown below :

Deceptive Words: not, he, do, 'm, would, his, no, an, mean, with, uh, just, n't, at, but, want, did, if, a, her, any, very, never , . . .

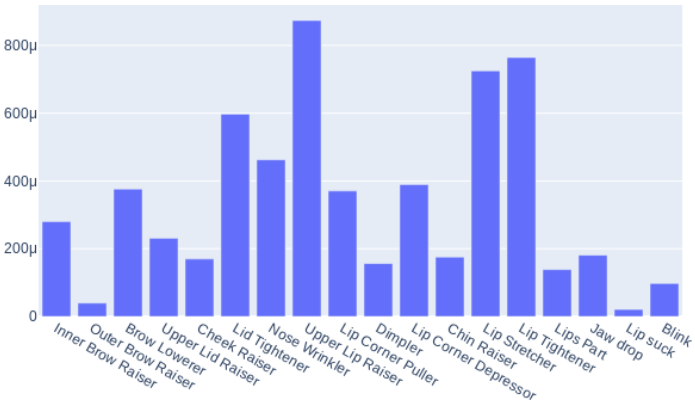

Fig. 5. Visual feature importance for automatically extracted AU features.

Truthful Words: . . ., by, so, then, other, was, had, all, through, started, up, on, the, years, two, my, when, of, to, from, um.

In each set, words are shown in decreasing score order i.e., from most deceptive (”not”) to most truthful (”um”). We see that negative words such as “not”, “no” and “n’t” have higher scores, suggesting that deceptive subjects often focus on denying the accusations, whereas truthful subjects are more focused on explaining past events. This coincides with the meta-analysis work of Hauch et al. which shows that deceptive statements have slightly more negative utterances than truthful statements.

Also extreme quantifiers (i.e. ”any”, ”never”, ”very”) occur more frequently in deceptive statements. Houch et al. investigated the effect of certainty on deception and, altough certainty indicating words did not have significant effects on deception, they revelead that ”deceptive accounts contained slightly fewer tentative words (such as ’may’, ’seem’, ’perhaps’) than truthful accounts” [26]. They commented on the possibility of liers’ motivation to appear credible. Our findings do not coincide exactly, but they are in the same direction.

Newman et al. have found that deceivers have a tendency to use fewer self-referencing expressions, such as ”I”, ”my”, ”mine” [6]. This coincides with our findings, because self-referencing words do not appear among the most deceptive words; while the word ”my” is one of the most truth-indicating words.

Interestingly, the word “uh” indicates deception whereas the word “um” indicates truthfulness despite both words having the function of pausing.

9 Related Work

9.1 Verbal Deception Detection

Initial work on deception detection focused on statistical methods to identify verbal cues associated with deceptive behavior. Bachenko et al. selected 12 linguistic indicators of deception, including lack of commitment to a statement or declaration, negative expressions, and inconsistencies with respect to verb and noun forms [27]. They extracted and analyzed the effect of these indicators on deception for a textual database of criminal statements, police interrogations, depositions and legal testimony. Hauch et al. conducted a meta-study covering 44 studies with a total of 79 linguistic deception cues and obtained a robust analysis of verbal deceptive indicators [26].

To date, works on verbal-based deception detection have explored the identification of deceptive content in a variety of domains, including online dating websites [28], [29], forums [30], [31], social networks [32], and consumer report websites [23], [33]. 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 [7], and also more complex linguistic features derived from syntactic CFG trees and part of speech tags [4], [34]. Some studies have also incorporated the analysis of psycholinguistics aspects related to the deception process. Some research work has relied on the Linguistic Inquiry and Word Count (LIWC) lexicon [14] to build deception models using machine learning approaches [7], [35] 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 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 [36].

There is a also significant amount of social science literature that statistically analyzes verbal indicators for deception. Burns et al. extracted LIWC indicators from transcriptions of a set of 911 calls [37]. They fed these indicators as features to machine learning classifiers and obtained an accuracy of 84%. Burgoon et al. examined linguistic and acoustic features extracted from a company’s quarterly conference call recordings using the Structured Programming for Linguistic Cue Extraction (SPLICE) toolkit [38]. They analyzed the strategic and nonstrategic behaviors of deceivers by annotating utterances as prepared (presentation) and unprepared (Q&A) responses and reported significant differences between these two, in terms of deceptive feature statistics. Larcker and Zakolyukina also applied linguistic analysis on conference call recordings from CEOs and CFOs and obtained significantly better deception prediction than a random guess [39], [40]. Fuller et al. analyzed verbal cues developed by Zhou et al. [41], [42] and their revised framework using written statements prepared by suspects and victims of crimes on military bases [43]. Braun et al. used LIWC indicators to investigate deceptive statements made by politicians labeled by editors of the politifact.com website and reported deceptive linguistic indicators in interactive and scripted settings separately [44].

While most of the data used in related research was collected under controlled settings, only a 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 presented by Vrij and Mann (2001) was the first study, to the best of our knowledge, on a real-life high-stake scenario including police interviews of murder suspects [45]. Ten Brinke et al. worked on a collection of televised footage from individuals pleading to the

public community for the return of a missing relative [12]. The work closest to ours is presented by Fornaciari and Poesio [46], 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.

9.2 Non-verbal Deception Detection

Earlier approaches to 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 [2], [3], [47] indicated that relying solely on such physiological measurements can be biased and misleading. Chittaranjan et al. [48] created audio-visual recordings of the "Are you a Werewolf?" game to detect deceptive behavior 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 [49] 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 [50] proposed a methodology using psychologically informed mind-reading to evaluate statements from suspects, witnesses, and innocents.

Facial expressions also play a critical role in the identification of deception. Ekman defined micro-expressions as relatively short involuntary expressions, which can be indicative of deceptive behavior [51]. Moreover, these expressions were analyzed using smoothness and asymmetry measurements to further relate them to an act of deceit [52]. Ekman and Rosenberg [53] developed the Facial Action Coding System (FACS) to taxonomize facial expressions and gestures for emotion- and deceit-related applications. Bartlett et al. [54] introduced a real-time system to identify deceptive behavior from facial expressions using FACS. Tian et al. [55] considered features such as face orientation and facial expression intensity. Owayjan et al. [56] extracted geometric-based features from facial expressions, and Pfister and Pietikainen [57] developed a micro-expression dataset to identify expressions that are clues for deception. Blob analysis was used to detect deceit by tracking the hand movements of subjects and extracting color features using hierarchical Hidden Markov Model [58], [59]. Meservy et al. [60] used individual frames as well as videos to extract geometric features related to the hand and head motion to identify deceptive behavior. Caso et al. [61] identified particular hand gestures that can be related to an act of deception using data collected from simulated interviews including truthful and deceptive responses. Cohen et al. [62] determined that fewer iconic hand gestures were a sign of a deceptive narration using data collected from participants with truthful and deceptive responses. To further analyze the characteristics of hand gestures, a taxonomy of such gestures was developed for multiple applications such as deception and social behaviour [63]. Hillman et al. [64] determined that increased speech prompting gestures were associated with deception while increased rhythmic pulsing gestures were associated with truthful behavior. Vrij and Mann analyzed visual and acoustic features on a dataset of police interviews of murder suspects and reported that convicted subjects "showed more gaze aversion, had longer pauses, spoke more slowly and made more non-ah speech disturbances" when lying than telling the truth [45]. Ten Brinke et al. manually extracted codings depicting speech, body language and emotional facial expressions for a collection of televised footage in which individuals pleading to the public community for the return of a missing relative [12]. They report informative codings that reflect deception, e.g. liars use fewer words but more tentative words.

Recently, features from different modalities were integrated to find a combination of multimodal features with superior performance [65], [66]. An extensive review of approaches for evaluating human credibility using physiological, visual, acoustic, and linguistic features is available in [67]. Burgoon et al. [65] combined verbal and non-verbal features such as speech act profiling, feature mining, and kinetic analysis for improved deception detection rates. Jensen et al. [66] extracted features from acoustic, verbal, and visual modalities following a multimodal approach. Mihalcea and Burzo [68] developed a multimodal deception dataset composed of linguistic, thermal, and physiological features. Nunamaker et al. [67] provided a review of approaches for evaluating human credibility using physiological, visual, acoustic, and linguistic features. A multimodal deception dataset consisting of linguistic, thermal, and physiological features was introduced in [69], which was then used to develop a multimodal deception detection system that integrated linguistic, thermal, and physiological features from human subjects to create a reliable deception detection system [70], [71].

10 Comparison to State-of-Art

Our work extends the work of Pérez-Rosas et al., where the real-life trial dataset was first presented, together with a video-clip level deception detection system [9].

iv) Different from the earlier work, our evaluations are conducted using 3 repetitions for each test sample in the leave-one-subject-out cross-validation, to obtain more robust results. Furthermore, we obtained both accuracy and AUC metrics.

v) Finally, our work obtained improved results on the deception detection task (83.05% accuracy and 0.95 AUC with feature-level fusion and 84.18% accuracy and 0.94 AUC with score-level fusion), which are also more reliable due to the cross-validation settings.

Among the studies that report results on this database, Jaiswal et. al. used the OpenFace toolkit [72] to extract visual features and the OpenSmile toolkit [73] to extract acoustic features which are then fed to an SVM classifier [74]. They report a 78.95% accuracy with feature-level fusion, after excluding videos (21 of 121) that are either too short or portray many people such that OpenFace is unable to recognize the subject.

Wu et. al. labeled short segments of video clips to train a micro-expression classifier whose outputs are fed to the deception classification system [75]. They report that even though the micro-expression classifier has low performance,

its output probabilities are useful to improve the performance of the overall system. They also use GloVe (Global Vectors for Word Representation) embeddings [76] for the linguistic representation and MFCC features for the acoustic modality. They report an AUC score of 0.92 obtained with a Logistic Regression classifier on a subset of the dataset (104 videos), pruning videos with either significant scene change or human editing.

The resulting semi-automatic system achieves an AUC score of 0.98 and an accuracy of 96.14%, thus obtaining the best results reported so far on this dataset. However as the authors also acknowledge, there is a possibility that the results may not generalize as well on larger datasets, due to overfitting or learning the idiosyncrasies of the small dataset.

Karimi et al. developed a multimodal deception detection system with automated features [77]. They employ CNNs followed by a Long-Short Term Memory (LSTM) model to extract the temporal information in the visual and vocal input, along with an attention mechanism focusing on the frames that include visual cues of deception. Their system achieves an accuracy of 84.16% for video-level classification.

In summary, existing research on this dataset has approached the problem at the video-level only, obtaining classification performances ranging from 78% to 97%. However, the experimental evaluations are not fully compatible, it is difficult to compare their results directly. For instance, in some work, the videos where the subject is not clearly seen are removed from the dataset; and a subject-based cross validation is note performed in others.

Our results are also not directly comparable with the state-of-art since we detect deception at the subject-level rather than at the video-level. Nonetheless, our best figures obtained with the semi-automatic system (AUC of 0.9462 obtained with a feature-level combination and an AUC of 0.9323 obtained with a score-level combination of all modalities) are on par with the results of the semi-automatic system of [78].

11 Conclusions

In this paper, we presented a study of multimodal deception detection using real-life high-stake occurrences of deceit. We use a dataset from public real trials to perform both qualitative and quantitative experiments. We built classifiers relying on the individual or combined sets of verbal and non-verbal features and showed that a system using score-level combination can detect deceptive subjects with an accuracy of 84.18%. Our analysis of non-verbal behaviors occurring in deceptive and truthful videos brought insight into the gestures that play a role in deception. Additional analyses showed the role played by the various feature sets used in the experiments.

We also performed a study of the human ability to detect deception with single or multimodal data streams of real-life trial data. The study revealed high disagreement and low deception detection accuracies among human annotators. Our automatic system using all the modalities outperformed the average non-expert human performance by more than 6% points, and the lowest human annotator's performance by more than 11% points.

In the future, we will work on improving automatic gesture identification and automatic speech transcription, with the goal of taking steps towards a real-time deception detection system.

References

[1] S. Gross and R. Warden, "Exonerations in the united states, 1989 - 2012," National Registry of Exonerations, Tech. Rep., 2012.

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

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

[4] 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*, ser. ACL '12. Stroudsburg, PA, USA: Association for Computational Linguistics, 2012, pp. 171–175. [Online]. Available: http://dl.acm.org/citation.cfm?id=2390665.2390708

[5] 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*, 2005, pp. 1833–1836.

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

[7] 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.

[8] I. Pavlidis, N. Eberhardt, and J. Levine, "Human behaviour: Seeing through the face of deception," *Nature*, vol. 415, no. 6867, 2002.

[9] V. Pérez-Rosas, M. Abouelenien, R. Mihalcea, and M. Burzo, "Deception detection using real-life trial data," in *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*. ACM, 2015, pp. 59–66.

[10] 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*, vol. 41, no. 3-4, pp. 273–287, 2007. [Online]. Available: http://dx.doi.org/10.1007/s10579-007-9061-5

[11] N. W. Twyman, A. Elkins, and J. K. Burgoon, "A rigidity detection system for the guilty knowledge test," in *HICSS-44 Symposium on Credibility Assessment and Information Quality in Government and Business*. Citeseer, 2011.

[12] L. ten Brinke and S. Porter, "Cry me a river: Identifying the behavioral consequences of extremely high-stakes interpersonal deception." *Law and Human Behavior*, vol. 36, no. 6, p. 469, 2012.

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

[14] J. Pennebaker and M. Francis, "Linguistic inquiry and word count: LIWC," 1999, erlbaum Publishers.

[15] P. Ekman, W. V. Friesen, and J. C. Hager, "Facial action coding system: The manual on cd rom," *A Human Face, Salt Lake City*, pp. 77–254, 2002.

[16] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency, "Openface 2.0: Facial behavior analysis toolkit," in *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE, 2018, pp. 59–66.

[17] L. A. Streeter, R. M. Krauss, V. Geller, C. Olson, and W. Apple, "Pitch changes during attempted deception." *Journal of personality and social psychology*, vol. 35, no. 5, p. 345, 1977.

[18] L. ten Brinke, D. Stimson, and D. R. Carney, "Some evidence for unconscious lie detection," *Psychological Science*, p. 0956797614524421, 2014.

[19] H. Kawahara, T. Takahashi, M. Morise, and H. Banno, "Development of exploratory research tools based on tandem-straight," in *Proceedings: APSIPA ASC 2009: Asia-Pacific Signal and Information Processing Association, 2009 Annual Summit and Conference*. Asia-Pacific Signal and Information Processing Association, 2009 Annual Summit and Conference, International Organizing Committee, 2009, pp. 111–120.

[20] Z.-H. Tan and B. Lindberg, "Low-complexity variable frame rate analysis for speech recognition and voice activity detection," *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, no. 5, pp. 798–807, 2010.

[21] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in PyTorch," in *NIPS Autodiff Workshop*, 2017.

[22] "Cloud speech-to-text recognition," https://cloud.google.com/speech-to-text/, accessed: 2019-10-14.

[23] 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*, ser. HLT '11. Stroudsburg, PA, USA: Association for Computational Linguistics, 2011, pp. 309–319.

[24] M. Aamodt and H. Custer, "Who can best catch a liar? a meta-analysis of individual differences in detecting deception," *Forensic Examiner*, vol. 15, no. 1, pp. 6–11, 2006.

[25] MATLAB, *version 7.10.0 (R2010a)*. Natick, Massachusetts: The MathWorks Inc., 2010.

[26] V. Hauch, I. Blandón-Gitlin, J. Masip, and S. L. Sporer, "Are computers effective lie detectors? a meta-analysis of linguistic cues to deception," *Personality and social psychology Review*, vol. 19, no. 4, pp. 307–342, 2015.

[27] J. Bachenko, E. Fitzpatrick, and M. Schonwetter, "Verification and implementation of language-based deception indicators in civil and criminal narratives," in *Proceedings of the 22nd International Conference on Computational Linguistics (COLING 2008)*, 2008, pp. 41–48.

[28] 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*, ser. CSCW '10. New York, NY, USA: ACM, 2010, pp. 5–8. [Online]. Available: http://doi.acm.org/10.1145/1718918.1718921

[29] R. Guadagno, B. Okdie, and S. Kruse, "Dating deception: Gender, online dating, and exaggerated self-presentation," *Comput. Hum. Behav.*, vol. 28, no. 2, pp. 642–647, Mar. 2012. [Online]. Available: http://dx.doi.org/10.1016/j.chb.2011.11.010

[30] D. Warkentin, M. Woodworth, J. Hancock, and N. Cormier, "Warrants and deception in computer mediated communication," in *Proceedings of the 2010 ACM Cnference on Computer Supported Cooperative Work*. ACM, 2010, pp. 9–12.

[31] A. N. Joinson and B. Dietz-Uhler, "Explanations for the perpetration of and reactions to deception in a virtual community," *Social Science Computer Review*, vol. 20, no. 3, pp. 275–289, 2002.

[32] 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*, vol. 50, no. 1, pp. 1–4, 2013.

[33] 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.

[34] Q. Xu and H. Zhao, "Using deep linguistic features for finding deceptive opinion spam," in *Proceedings of COLING 2012: Posters*. Mumbai, India: The COLING 2012 Organizing Committee, December 2012, pp. 1341–1350. [Online]. Available: http://www.aclweb.org/anthology/C12-2131

[35] 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*. Avignon, France: Association for Computational Linguistics, April 2012, pp. 15–22. [Online]. Available: http://www.aclweb.org/anthology/W12-0403

[36] 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)*. Sofia, Bulgaria: Association for Computational Linguistics, August 2013, pp. 944–953. [Online]. Available: http://www.aclweb.org/anthology/P13-1093

[37] M. B. Burns and K. C. Moffitt, "Automated deception detection of 911 call transcripts," *Security Informatics*, vol. 3, no. 1, p. 8, 2014.

[38] J. Burgoon, W. J. Mayew, J. S. Giboney, A. C. Elkins, K. Moffitt, B. Dorn, M. Byrd, and L. Spitzley, "Which spoken language markers identify deception in high-stakes settings? evidence from earnings conference calls," *Journal of Language and Social Psychology*, vol. 35, no. 2, pp. 123–157, 2016.

[39] D. F. Larcker and A. A. Zakolyukina, "Detecting deceptive discussions in conference calls," *Journal of Accounting Research*, vol. 50, no. 2, pp. 495–540, 2012.

[40] R. Bloomfield, "Discussion of detecting deceptive discussions in conference calls," *Journal of Accounting Research*, vol. 50, no. 2, pp. 541–552, 2012.

[41] L. Zhou, J. K. Burgoon, J. F. Nunamaker, and D. Twitchell, "Automating linguistics-based cues for detecting deception in text-based asynchronous computer-mediated communications," *Group decision and negotiation*, vol. 13, no. 1, pp. 81–106, 2004.

[42] L. Zhou, J. K. Burgoon, D. P. Twitchell, T. Qin, and J. F. Nunamaker Jr, "A comparison of classification methods for predicting deception in computer-mediated communication," *Journal of Management Information Systems*, vol. 20, no. 4, pp. 139–166, 2004.

[43] C. M. Fuller, D. P. Biros, J. Burgoon, and J. Nunamaker, "An examination and validation of linguistic constructs for studying high-stakes deception," *Group Decision and Negotiation*, vol. 22, no. 1, pp. 117–134, 2013.

[44] M. T. Braun, L. M. Van Swol, and L. Vang, "His lips are moving: Pinocchio effect and other lexical indicators of political deceptions," *Discourse Processes*, vol. 52, no. 1, pp. 1–20, 2015.

[45] A. Vrij and S. Mann, "Telling and detecting lies in a high-stake situation: the case of a convicted murderer," *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition*, vol. 15, no. 2, pp. 187–203, 2001.

[46] T. Fornaciari and M. Poesio, "Automatic deception detection in Italian court cases," *Artificial Intelligence and Law*, vol. 21, no. 3, pp. 303–340, 2013.

[47] M. Derksen, "Control and resistance in the psychology of lying," *Theory and Psychology*, vol. 22, no. 2, pp. 196–212, 2012.

[48] 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)*, March 2010, pp. 5334–5337.

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

[50] P. A. Granhag and M. Hartwig, "A new theoretical perspective on deception detection: On the psychology of instrumental mind-reading," *Psychology, Crime & Law*, vol. 14, no. 3, pp. 189–200, 2008.

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

[52] E. Paul, "Darwin, deception, and facial expression," *Annals of the New York Academy of Sciences*, vol. 1000, no. EMOTIONS INSIDE OUT: 130 Years after Darwin's The Expression of the Emotions in Man and Animals, pp. 205–221, 2003.

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

[54] M. Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. Fasel, and J. Movellan, "Automatic recognition of facial actions in spontaneous expressions," *Journal of Multimedia*, vol. 1, no. 6, pp. 22–35, 2006.

[55] Y. Tian, T. Kanade, and J. Cohn, "Facial expression analysis," in *Handbook of Face Recognition*. Springer New York, 2005, pp. 247–275.

[56] 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)*, Dec 2012, pp. 33–38.

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

[58] 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)*, ser. HICSS '05. Washington, DC, USA: IEEE Computer Society, 2005, pp. 20–29.

[59] 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*, July 2005, pp. 824–827.

[60] 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*, vol. 20, no. 5, pp. 36–43, September 2005.
[61] 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*, vol. 30, no. 1, pp. 1–19, 2006.
[62] D. Cohen, G. Beattie, and H. Shovelton, "Nonverbal indicators of deception: How iconic gestures reveal thoughts that cannot be suppressed," *Semiotica*, vol. 2010, no. 182, pp. 133–174, 2010.
[63] F. Maricchiolo, A. Gnisci, and M. Bonaiuto, "Coding hand gestures: A reliable taxonomy and a multi-media support," in *Cognitive Behavioural Systems*, ser. Lecture Notes in Computer Science, A. Esposito, A. Esposito, A. Vinciarelli, R. Hoffmann, and V. Muller, Eds. Springer Berlin Heidelberg, 2012, vol. 7403, pp. 405–416.
[64] J. Hillman, A. Vrij, and S. Mann, "Um ... they were wearing ...: The effect of deception on specific hand gestures," *Legal and Criminological Psychology*, vol. 17, no. 2, pp. 336–345, 2012.
[65] 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*, vol. 10, no. 1, pp. 103–112, March 2009.
[66] M. Jensen, T. Meservy, J. Burgoon, and J. Nunamaker, "Automatic, multimodal evaluation of human interaction," *Group Decision and Negotiation*, vol. 19, no. 4, pp. 367–389, 2010.
[67] 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)*, June 2012, pp. 202–211.
[68] R. Mihalcea and M. Burzo, "Towards multimodal deception detection – step 1: Building a collection of deceptive videos," in *Proceedings of the 14th ACM International Conference on Multimodal Interaction*, ser. ICMI '12. New York, NY, USA: ACM, 2012, pp. 189–192.
[69] V. Pérez-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.
[70] 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*, ser. ICMI '14. Istanbul, Turkey: ACM, 2014, pp. 58–65.
[71] M. Abouelenien, V. Pérez-Rosas, R. Mihalcea, and M. Burzo, "Detecting deceptive behavior via integration of discriminative features from multiple modalities," *IEEE Transactions on Information Forensics and Security*, vol. 12, no. 5, pp. 1042–1055, 2016.
[72] T. Baltrušaitis, P. Robinson, and L.-P. Morency, "Openface: an open source facial behavior analysis toolkit," in *Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on*. IEEE, 2016, pp. 1–10.
[73] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in opensmile, the munich open-source multimedia feature extractor," in *Proceedings of the 21st ACM international conference on Multimedia*. ACM, 2013, pp. 835–838.
[74] M. Jaiswal, S. Tabibu, and R. Bajpai, "The truth and nothing but the truth: Multimodal analysis for deception detection," in *Proceedings of the 2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW)*, 2016.
[75] Z. Wu, B. Singh, L. S. Davis, and V. Subrahmanian, "Deception detection in videos," *arXiv preprint arXiv:1712.04415*, 2017.
[76] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1532–1543.
[77] H. Karimi, J. Tang, and Y. Li, "Toward end-to-end deception detection in videos," in *2018 IEEE International Conference on Big Data (Big Data)*. IEEE, 2018, pp. 1278–1283.
[78] G. Krishnamurthy, N. Majumder, S. Poria, and E. Cambria, "A deep learning approach for multimodal deception detection," *arXiv preprint arXiv:1803.00344*, 2018.

M. Umut Sen Mehmet Umut Sen is a PhD student in Electronics Engineering Department of Sabanci University, Istanbul. He holds MSc and BSc degrees also from the same department, receiving them in 2009 and 2011 respectively. His research interests include Text Categorization, Machine Learning and Speech Processing..

Verónica Pérez-Rosas is a Assistant Research Scientist at University of Michigan. She received her Ph.D. in Computer Science and Engineering from the University of North Texas in 2014. Her research interests include machine learning, natural language processing, computational linguistics, affect recognition, and multimodal analysis of human behavior. Her research focuses on developing computational methods to analyze, recognize, and predict human affective responses during social interactions. She has authored papers in leading conferences and journals in Natural Language Processing and Computational linguistics, and served as a program committee member for multiple international journals and conferences in the same fields.

Berrin Yanikoglu is Professor of Computer Science and Director of the Center of Excellence in Data Analytics (VERIM), at Sabanci University, Istanbul, Turkey. She received a double major in Computer Science and Mathematics from Bogazici University, Turkey in 1988 and her Ph.D. degree in Computer Science from Dartmouth College, USA in 1993. Prof. Yanikoglu worked at Rockefeller University, Xerox Imaging Systems and IBM Almaden Research Center, before joining Sabanci University in 2000. Her research interests lie in machine learning with applications to image/video understanding; in particular multimodal deception detection, sentiment analysis, biometric verification and privacy, and handwriting recognition. Along with her students, she has developed award-winning signature verification systems.

Mohamed Abouelenien 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 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 interests include multimodal deception detection, multimodal sensing of thermal discomfort and drivers' alertness levels, emotion and stress analysis, machine learning, ensemble learning, image processing, face and action recognition, and natural language processing. Abouelenien has published in several top venues including IEEE, ACM, and Springer. 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.

Mihai Burzo 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, 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", the best paper award at the Semitherm conference in both 2013 and 2006, the Young Engineer of the Year from the North Texas Section of ASME (2006), a Leadership Award from SMU (2002), and a Valedictorian Award (1995).

Rada Mihalcea is a Professor in the Computer Science and Engineering department at the University of Michigan. Her research interests are in computational linguistics, multimodal behavior analysis, and computational social sciences. She has published more than 200 papers in these and related areas, and she co-authored two books published by the Cambridge University Press and SAGE respectively. She is the recipient of a National Science Foundation CAREER award (2008) and a Presidential Early Career Award for Scientists and Engineers (2009). In 2013, she was made an honorary citizen of her hometown of Cluj-Napoca, Romania.