Gender-based Multimodal Deception Detection

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

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

Bohan Zhao
Computer Science and Engineering
University of Michigan
Ann Arbor, MI 48109, USA
zhaoboha@umich.edu

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

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

ABSTRACT

This paper explores gender-based differences in multimodal deception detection. We introduce a new large, gender-balanced dataset, consisting of 104 subjects with 520 different responses covering multiple scenarios, and perform an extensive analysis of different feature sets extracted from the linguistic, physiological, and thermal data streams recorded from the subjects. We describe a multimodal deception detection system, and show how the two genders achieve different detection rates for different individual and combined feature sets, with accuracy figures reaching 80%. Our experiments and results allow us to make interesting observations concerning the differences in the multimodal detection of deception in males and females.

CCS Concepts

•**Computing methodologies** → **Artificial intelligence;**

Keywords

deception detection; multimodal; thermal; physiological; linguistic

1. INTRODUCTION

"This will be my last drink," "Sorry, I missed your call," "It was on sale," and "I am almost ready" are all common lies told by people to each other. A study[1] showed that the first two statements are commonly told by males, while the other two are usually told by females. These simple examples show that males and females often act and represent themselves differently, and the gender of a speaker can often be easily identified [13]. While some lies might seem acceptable, other lies could pose a risk and possibly lead to devastating effects. In addition, there could be some gender- or culture-specific behaviors that could be perceived as deceptive. Hence, as deception is increasingly being studied across different domains as well as different populations, the inclusion of demographic information and the association between deceiving behaviors and gender becomes increasingly important [23].

With the massive growth of online interactions, communication, and social media, relying on polygraph tests and subjects' behavioral analysis, as frequently done by law enforcement, cannot fully address the problem as these modalities might not be easy to acquire or integrate. Thus, a timely and relevant problem is the automatic identification of deception by using multiple modalities in addition to using demographic data such as gender, age, or location.

Previous work on the relation between gender and deception has explored whether the context of the lies made by each gender is easier to detect, or whether males or females are better lie detectors. It was found for instance that gender perception can have an impact on trustworthiness, as females are perceived as "more cooperative and less dominant than males" [6]; or that females are statistically better lie detectors compared to males [21]. However, previous work did not consider the problem of whether there are gender-based differences in the automatic detection of deception.

In this paper, we analyze the behaviors and trends associated with different genders as they act deceptively, as well as observe how the performance vary in detecting deceit with males and females. We make three main contributions. First we introduce a novel gender-balanced deception detection dataset consisting of 520 responses from 104 subjects covering three different scenarios. Second, we propose a multimodal gender-based deception detection system, including a novel tracking system of different thermal regions of interest in the face. Third, we conduct an extensive analysis of feature sets from different modalities, determine the capability of different features and regions of interest to indicate deceit, and analyze the relation between these patterns and gender. To our knowledge, this is the first deception detection approach that integrates linguistic, thermal, and physiological modalities, while also considering demographic information such as gender.

*Corresponding author
[1]http://datafication.com.au/

SAC 2017,April 03-07, 2017, Marrakech, Morocco

http://dx.doi.org/10.1145/3019612.3019644

2. RELATED WORK

2.1 Gender-based Deception

Gender differences in deception have been studied in different fields such as psychology, economics, language processing, and acoustics among others. Researchers in the economics field studied deceptive behavior to analyze the role that gender plays in economic modelling. Economic studies showed that male participants tend to be more deceptive to achieve monetary benefits [5]. Other economic studies reported no differences in deception by gender using the sender-receiver games, where the sender can deceive the receiver in order to secure higher incentive [3].

The psychological role of the gender was analyzed in relation to deception using different applications. Dating was particularly used to detect gender differences in committing and perceiving deception. Using different meeting conditions to detect dating deception, it was reported that males were generally more deceptive and altered their personality characteristics and physical appearance when they expected meeting a date [9].

Vision and speech analysis were used to detect gender-based clues of deception. Using visual facial features and speech aspects, a correlation between the lying cues and gender was reported. In particular, it was found that males reduce their leg and foot movements and use more qualifying statements when acting deceptively, unlike females [7].

Linguistic and speech processing was additionally used to differentiate between deception by males and females. Levitan et al. [12] showed that participants who are better at detecting lies are better deceivers, especially for females. However, using an interactive social media game platform, no significant difference was found between male and female deceivers [11].

2.2 Multimodal Deception

Different fields have analyzed clues of deception such as physiology, psychology, language, acoustics, and computer vision. Computational linguistic approaches have also covered the identification of deception on a variety of domains where computer mediated communication occurs, including chats, forums, online dating websites, social networks, as well as product reviews websites that were prone to have fake product reviews and spam content [23]. A data-driven method was proposed in [16] to build classifiers able to distinguish between deceptive and truthful essays covering three topics: opinions on abortion, opinions about death penalty, and feelings about a best friend.

Using the visual modality, a noticeable decrease in the frequency of gestures was observed when subjects narrated stories deceptively compared to narrating the same stories truthfully [4]. Additionally, individuals acting truthfully produced more rhythmic pulsing gestures while those acting deceptively made more frequent speech prompting gestures [10].

Thermal imaging was also used in deception detection. It was shown that as the nervous system reacted with an act of deceit, a peripheral change in the blood flow distribution was detected towards the musculoskeletal tissue [17]. Tandem tracking and noise suppression methods were used to extract thermal features from the periorbital area without applying restrictions on the face movements of the subjects in order to improve deception detection rates [22]. Integrating features from multiple modalities such as the thermal, linguistic and physiological [1, 2] as well as linguistic and visual [19, 18] showed an overall improvement in detecting deceit.

3. ELICITING DECEPTIVE AND TRUTHFUL RESPONSES

3.1 Experimental Setup

Deceptive and truthful responses were collected using a setup consisting of a thermal camera, two visual cameras, and a microphone. In addition, we used four physiological bio-sensors including a blood volume pulse, skin conductance and skin temperature sensors, as well as an abdominal respiration band. The first three sensors were attached to the fingers of subject's non-dominant hand. The abdominal respiration sensor was placed to surround the thoracic region.

We used a FLIR SC6700 thermal camera with a resolution of 640x512 and 7.2 M electrons capacity, reaching a frame rate of approximately 100 frames/second. We also used two visual cameras. Subjects' verbal responses were recorded separately using a noise cancelling microphone.

Our experimental station consists of recording devices, the physiological sensors, two desktop computers, and a chair placed at a fixed distance from the cameras. Subjects were asked to sit comfortably and were told to respond truthfully and deceptively to three scenarios designed to elicit deceptive and truthful responses. The experimental setup and procedure were explained to the subjects and they were asked to avoid excessive movements to keep them in the field of view of the cameras.

3.2 Scenarios

Three scenarios were designed for the experiments, which were used successfully in previous research [16]. The subjects were informed of the topic matter before each individual recording. In two scenarios, namely "Abortion" and "Best Friend," subjects were allowed to speak freely first truthfully and then deceptively, while in the third scenario "Mock Crime" the subjects had to respond to questions asked by the interviewer.

Abortion. In this scenario participants were asked to provide first a truthful and then a deceptive opinion about their feelings regarding abortion and whether they think it is right or wrong and if it should be legalized. The experimental session consisted of two independent recordings for each case.

Best Friend. In this scenario subjects were instructed to provide an honest description of their best friend, followed by a deceptive description about a person they cannot stand. In the second part, they had to describe the individual they cannot stand as if he or she was their best friend. Therefore, in both cases, the person was described positively.

Mock Crime. In this scenario subjects were allowed to choose the truthfulness or deceitfulness of their responses in a mock crime scenario, where they presumably stole money. Specifically, a $20 bill was hidden in a box beside the participants. The subjects were told that the interviewer would leave the room and that it was their choice to steal the money or not. Additionally, they were told that the male interviewer would return back to the room to ask them questions regarding the missing bill in a one-on-one interview, and that they should make their own decisions whether they would lie to the interviewer or not. In addition, participants were told that, at the

end of the interview, the interviewer would attempt to guess if they were lying or telling the truth and they would receive additional monetary compensation if they manage to successfully deceive the interviewer.

The interview was conducted as follows:

1. Are the lights on in this room?
2. Regarding that missing bill, do you intend to answer each question truthfully about that?
3. Prior to 2016, did you ever lie to someone who trusted you?
4. Did you take that bill?
5. Did you ever lie to keep out of trouble?
6. Did you take the bill from the private area of the lab?
7. Prior to this year, did you ever lie for personal gain?
8. What was inside the white envelope?
9. Please describe step by step, in as much detail as you can, what you did while you were in the room and I was outside.
10. Do you know where that missing bill is right now?

Questions one to three were control questions that attempted to establish a baseline for all subjects. The remaining questions were designed to elicit information that might suggest deceptive mechanisms used by the participants.

Normalization. After recording their responses, the subjects were asked to relax and sit comfortably at the recording station for a one-minute recording with no activity on their side in order to establish a resting baseline. This recording was collected to account for interpersonal variations as described below.

3.3 Development Data

A development data set was collected earlier from a different set of 30 subjects. We used the same scenarios described above, with a small difference in the "Mock Crime," where the subjects were pre-assigned to one of the deceptive or truthful conditions. Similar equipment and settings were used except for using an earlier model of the thermal camera (FLIR Thermovision A40). The participants consisted of 30 graduate and undergraduate students.

3.4 Gender-balanced Data

The experiments reported in this paper were conducted on a new gender-balanced dataset, introduced for the first time in this paper. The dataset consists of multimodal recordings collected from 104 undergraduate and graduate, students, with a gender distribution of 53 females and 51 males. The recordings were conducted using the settings and scenarios described above. All participants expressed themselves in English, were native and non-native English speakers, belonged to several ethnic backgrounds, and had an age range between 20 and 35 years.

4. METHODOLOGY

4.1 Verbal Cues of Deception

We extract several linguistic features that have been previously found to correlate with deception cues. These features are derived from the transcripts of the subjects' statements. For both the abortion and the best friend scenarios we use the full transcript, whereas for the mock crime scenario we remove the interviewer questions and concatenate the interviewee responses into a single chunk of text.

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.

Shallow and deep syntax: We extract a set of features derived from part-of-speech (POS) tags and production rules based on context-free grammar (CFG) trees as described in [8]. We use the Stanford parser to obtain both POS and CFG features. Our POS features are encoded as the frequency values of each POS tag occurring in the dataset. The CFG derived features consist of all lexicalized production rules combined with their grandparent node and are also encoded as frequency values.

LIWC derived features: We use features derived from the Linguistic Inquire Word Count (LIWC) lexicon. These features consist of word counts for each of the 80 semantic classes present in the LIWC lexicon.

Readability score features and syntactic complexity: This set of features consists of fourteen indexes representing the syntactic complexity of a sentence, covering frequency and lengths of sentences, t-units,[2] and clauses. To extract these features we use a tool provided by Lu et al. [14]. In addition, we also incorporate two standard text readability metrics, namely Flesch-Kincaid and Gunning Fog.

Length features: We also derive a set of features that indicate the length of responses over time. We use an utterance as a thought unit and estimate the number of utterances spoken during five equally distributed intervals. Finally, we count the number of words in the utterances spoken during each interval, which results in five features indicating the length of subject's responses over time. Note that we use the transcript to extract these features, thus we consider a sentence as the equivalent of a spoken utterance.

In order to identify differences in word usage between deceivers and true-tellers based on gender, we obtain the most dominant semantic word classes [15] using the LIWC lexicon. Table 1 shows the most dominant words classes used by deceivers and true-tellers from both genders. In this table we can observe that truthful behavior follows similar trends regardless of gender, for instance true-tellers use family, friends, optimism, and positive words. On the other hand, deceitful behaviors show important differences in word usage, more noticeably across gender where less overlap occurs among dominant classes. More interestingly, there are also noticeable differences in deceptive and truthful word usage within gender. For instance, male word usage suggests that lies told by men are frequently related to sports, job, eating, school, and money, whereas their truthful responses involve taking about friendship, past, optimism, and the self. On the other hand, female word usage suggests that female lies are related to others, religion, metaphors, future, and certainty, whereas females truths include words related to home, family, we, friends, and the self.

4.2 Non-verbal Cues: Physiological Features

The physiological features include raw physiological measurements of heart rate, respiration rate, skin conductance, and peripheral skin temperature using the sensors described in the experimental setup. In addition to raw measurements, we calculated their statistical descriptors including maximum and minimum values, means,

[2]Defined as the shortest grammatically allowable sentences into which writing can be split or a minimally terminable unit.

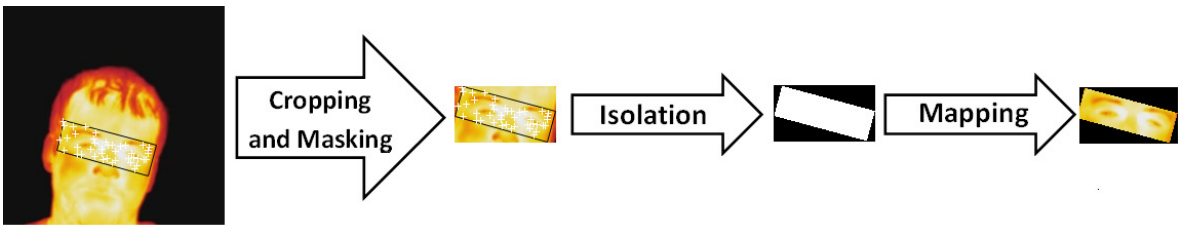

Figure 1: An overview of the region of interest extraction process where it is tracked, masked, isolated, and mapped to the original thermal video to extract thermal features.

Table 1: Results from LIWC word class analysis. Top ranked semantic classes associated to deceptive and truthful statements provided by male and female participants are shown.

Male			
Truthful		Deceptive	
Class	Score	Class	Score
Religion	1.45	Death	2.92
Friends	1.38	Groom	2.32
I	1.29	Anxiety	2.17
Hear	1.29	Inhibition	2.10
Past	1.23	Job	1.95
Self	1.22	Eating	1.88
Up	1.18	Metaphor	1.76
Body	1.18	You	1.65
Time	1.16	Sports	1.58
Sexual	1.16	School	1.52
Common verbs	1.15	Money	1.35
Physical	1.14	Humans	1.35
Optimism	1.14	Anger	1.28
Positive feeling	1.10	Occupation	1.25
Female			
Truthful		Deceptive	
Class	Score	Class	Score
Optimism	1.66	Death	3.08
Home	1.60	Metaphor	2.69
Up	1.50	You	2.13
Family	1.38	Religion	2.04
Sad	1.36	Groom	1.90
We	1.29	Music	1.70
Anxiety	1.28	Eating	1.36
Friends	1.27	Anger	1.34
Similes	1.26	Money	1.29
Leisure	1.26	Humans	1.28
Down	1.26	Assent	1.28
Communication	1.23	Job	1.27
Self	1.21	Certain	1.18
Achieve	1.19	Sports	1.16

power means, standard deviations, and mean amplitudes (epochs). The features were extracted at a rate of 2,048 samples per second. The final set consists of a total of 59 physiological features, which include 40 features extracted from raw measurements with the blood volume pulse sensor (BVP), five skin conductance features, five skin temperature features, and seven respiration rate features. Moreover, two features are extracted from the BVP and the respiration rate sensors combined, namely, the mean and heart rate max-min difference, which is a measure of breath to heart rate variability. The final vector for each response is averaged over all the samples. Moreover, we divided the raw signal of each sensor into five stages and calculated the average of each stage in order to capture the dynamics of the signals as the response progressed, resulting in five additional temporal features.

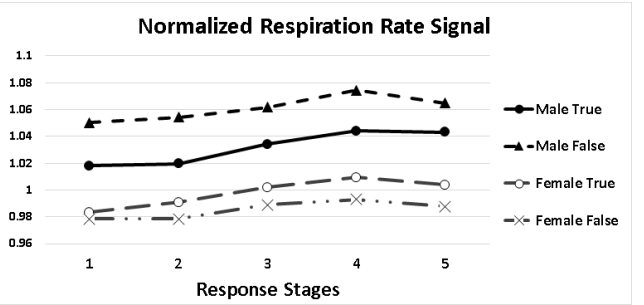

Figure 2: Average respiration rate through the responses for truthful and deceptive males and females.

Figure 2 shows how the average respiration rate signal varies for males and females as they respond truthfully and deceptively. Interestingly, the respiration rate increases as males respond deceptively while it increases as females respond truthfully. The same opposite patterns were observed for the skin conductance signal. However, similar trends for males and females were observed for the BVP and skin temperatures signals.

4.3 Non-verbal Cues: Thermal Features

Thermal features are extracted in order to determine whether certain thermal patterns occur as a subject acts deceptively and whether these patterns vary between males and females. Specifically, we extract thermal features in three steps: region of interest segmentation, tracking, and thermal map formation.

Segmenting and tracking regions of interest. First we located five regions of interest (ROI) manually from the first frame of each recorded thermal videos by determining their bounding boxes and exporting the raw thermal videos to a supported video format for tracking. The regions are the whole face, forehead, periorbital area, cheeks (including the nose), and the nose. Interesting points are then detected in each region using Shi-Tomasi corner detection algorithm. These points are located in areas with sharper changes in colors, i.e., temperatures. Following this, the detected points are tracked through the entire response using the fast Kanade-Lucas-Tomasi (KLT) tracking algorithm [20]. Our thermal videos were recorded at a rate of 100 frames per second and all recorded frames were used to extract the thermal features.

Following the tracking process and displacement estimation, we apply a geometric transformation, which globally estimates the interesting points transformation based on similarity and maps the

interesting points between the frames. Once the points are mapped, the new boundary box is geometrically determined. We allow a maximum distance of five pixels between the tracked point and its projection on the next frame. Furthermore, if the number of points matched between two successive frames is less than 95%, we consider that there is a chance of occurrence of occlusion. In this case, we discard the tracking of the current frame and proceed to the next one. An overview of the process is shown in Figure 1.

Thermal features extraction. The locations of the bounding boxes of the polygon containing the ROI of each frame are then mapped back into the raw thermal data to extract features from the actual temperatures. We set a temperature of 80.5 F as a threshold below which any temperature is set to zero as to eliminate the background behind the subjects. The ROI is then cropped from the raw thermal video.

A thermal map is created for each response to define the heat distribution in each ROI normalized by the resting baseline to account for the inter-personal variations. This was performed by extracting response-level statistical measurements such as the minimum, maximum, mean, and standard deviation of the frame-level mean, maximum, minimum, standard deviation, and the average of the 10% hottest temperatures. This results in a thermal vector of size 20 for each of the 520 thermal recordings and each region of interest. We also extract temporal features by dividing each response into five equal stages, and calculating statistical features from each stage. The statistical features extracted from the entire responses in addition to the five temporal thermal features will be used to train our model.

4.4 Classification

After the features are extracted from each modality, we test their performance individually and combined. The fusion is performed by integrating the features collected from all the multimodal streams to form a single feature vector, which is then used to make a decision about the classification of the stream. For the classification process, a decision tree classifier is used, and we follow a leave-one-subject-out cross validation scheme, meaning that all the five instances belonging to the same subject are reserved for testing while all the instances belonging to other subjects are used for training. We report the overall average accuracy and recall of the deceptive and truthful classes for individual and combined genders as well as for individual and combined topics.

5. RESULTS AND DISCUSSION

5.1 Experiments on Development Data

The development dataset consists of a different set of 149 instances collected from 30 participants using the three topics with a distribution of 76 deceptive instances and 73 truthful instances. Using this data, we conducted an analysis on different sets of features from different modalities to analyze their capability of generally indicating deceit. We report the performance of different feature sets below. The results obtained on this dataset determine which sets are to be integrated together on the new gender-balanced dataset.

The classification results for truthful and deceptive statements obtained using linguistic features are shown in Figure 3. The figure shows the average recall of the deceptive and truthful classes, as well as the average overall accuracy percentages. We also attempt combinations among the feature sets; for these experiments only

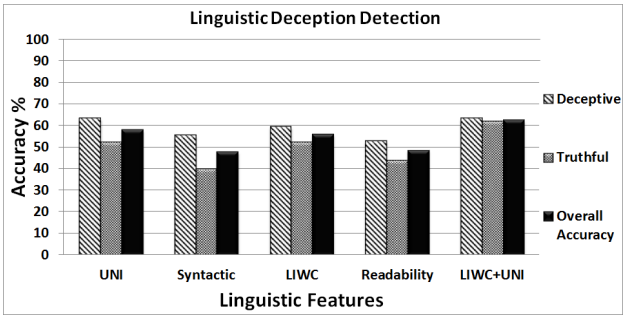

Figure 3: Results on development data: Recall of the deceptive and truthful classes, and overall accuracy percentages for five different linguistic feature sets.

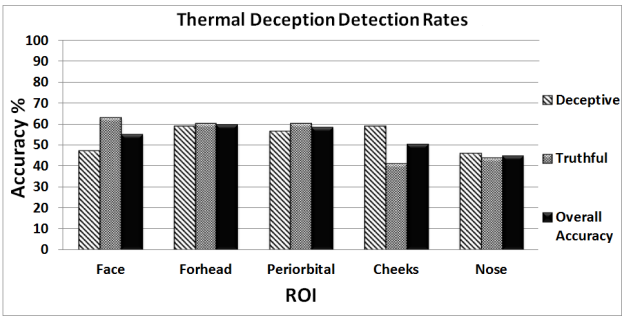

Figure 4: Results on development data: Recall of the deception and truthfulness classes, and overall accuracy % for different ROIs using thermal features.

the combination of unigrams and LIWC derived features is reported in Figure 3 as it attained the best overall accuracy and best recall in the range of 61% to 63% for the deceptive and truthful classes. Hence this particular combination will be used for the feature fusion on the new dataset.

Figure 4 illustrates the deceptive and truthful classes recall as well as the overall accuracy using the thermal features extracted from different ROIs. The features from the forehead region achieve the best performances followed by the features extracted from the periorbital regions, with approximately 60% overall accuracy. Hence, the features extracted from the forehead region will be used for fusion in the new dataset.

The performance of a per-sensor analysis using the physiological data in the development set is close to random guessing in most cases. The best accuracy is achieved by the respiration rate sensor at 53.7%, followed by the fusion of all physiological sensors at 53.0%. Hence, the features from the respiration rate sensor will be used for fusion in the new dataset.

5.2 Experiments on Gender-balanced Data

We compare the performance of different feature sets for each modality in addition to comparing the performance of the fused sets. In all cases the temporal features are included with their corresponding linguistic, face segment, and sensor feature sets. The gender-balanced dataset consists of 520 instances collected from 104 participants using three different topics with a distribution of 255 deceptive instances and 265 truthful instances as well as a distribution

Table 2: Average accuracy percentage using different sets of linguistic features for both genders, males only, and females only, as well as for "Abortion," "Best Friend," "Mock Crime," and "All Topics" combined. The best results for all topics are highlighted in bold, and the results for the feature sets found to work best on the development data are underlined.

Linguistic	Abortion			Best Friend			Mock Crime			All Topics		
	Both	Male	Female	Both	Male	Female	Both	Male	Female	Both	Male	Female
Unigrams	57.7	51.0	63.2	59.1	51.0	58.5	47.1	**74.5**	**81.1**	58.5	**63.5**	60.8
Syntax	58.7	43.1	**64.2**	51.0	52.9	**61.3**	51.0	39.2	41.5	51.3	45.9	57.0
LIWC	55.3	**56.9**	54.7	**63.0**	52.0	57.5	50.0	66.7	49.1	50.4	52.5	57.0
Readability	56.3	50.0	51.9	53.4	52.0	50.9	**53.8**	58.8	62.3	59.2	51.0	54.3
LIWC+Uni	**59.1**	**56.9**	63.2	54.8	**59.8**	46.2	45.2	64.7	**81.1**	**59.4**	54.9	**73.6**

Table 3: Average accuracy percentage using different sets of physiological features for both genders, males only, and females only, as well as for "Abortion", "Best Friend", "Mock Crime", and "All Topics" combined. The best result for all topics are highlighted in bold, and the results for the feature sets found to work best on the development data are underlined.

Physiological	Abortion			Best Friend			Mock Crime			All Topics		
	Both	Male	Female	Both	Male	Female	Both	Male	Female	Both	Male	Female
Blood Volume Pulse	51.0	50.0	**58.5**	35.6	43.1	34.0	46.2	52.9	45.3	49.6	49.4	46.8
Respiration Rate	35.6	29.4	50.9	43.3	**52.0**	42.5	**59.6**	**64.7**	60.4	46.5	46.3	54.7
Skin Conductance	**63.0**	**54.9**	49.1	**52.9**	**52.0**	**52.8**	**59.6**	49.0	**75.5**	49.6	**58.4**	58.1
Skin Temperature	59.1	51.0	**58.5**	51.0	41.2	43.4	46.2	56.9	67.9	46.9	50.2	**60.8**

of 255 responses from males and 265 from females. The data distribution results in a majority baseline performance for deception detection of slightly less than 51%.

5.3 Individual Modalities

5.3.1 Linguistic

Table 2 lists the average accuracy percentage using different linguistic sets for individual and combined topics as well as individual and combined genders. The feature set found to work best on the development data (unigrams combined with LIWC features) achieves the highest accuracy as well in 6 out of 12 cases (50%). Overall, an improved accuracy is achieved by the females' instances compared to the males' instances in 14 out of 19 cases, excluding the one case where the performance of both genders is below that of the baseline. The female instances performance exceeds 80% using the unigrams and temporal features. Unlike males, the syntax features provide some deception clues for females, especially for "Abortion" and "Best Friend". However, it is interesting to note that using instances of both genders combined, which results in a larger dataset of double the size, the performance does not improve. In fact, it deteriorates in many cases. The same trend can be observed for the performance of all topics combined, which can be related to the integration of multi-domain features. Moreover, lies are better detected for "Abortion" and "Mock Crime" for females compared to "Best Friend". This is the case for males with only "Mock Crime".

5.3.2 Physiological

Table 3 lists the average accuracy percentage using different physiological sets for individual and combined topics as well as individual and combined genders. The respiration rate features found to work best on the development data achieve the second best performance overall following the skin conductance features. Interestingly, both sensors signals also showed opposite trends for males and females as shown earlier in Figure 2. However, it can be noted that in general the performance deteriorates compared to the linguistic features performance. While the best performing sensor is domain-specific for females, it seems that the skin conductance sensor is more consistent in providing deception clues for males. The female group once again achieves improved performance for the individual and combined topics, which can be seen in 8 out 13 cases, excluding three cases where both genders are below the baseline. As with the linguistic features the performance does not improve by combining genders.

5.3.3 Thermal

Table 4 lists the average accuracy percentage using different thermal sets for individual and combined topics as well as individual and combined genders. The feature set found to work best on the development data (forehead features) achieves the highest accuracy among individual and combined topics and among genders in four cases, which is the highest among all five ROIs. It can be noted that the whole face have higher capability of indicating deceit in females. On the other hand, the periorbital area features are more indicative of deceit for males. The females group achieved a clear improved performance compared to males in the "Mock Crime" scenario, especially using the face and cheeks regions.

Similar to the linguistic and physiologic modalities, creating a larger dataset composed of both genders' instances as well as combining instances from all topics deteriorated the performance in multiple cases. This consistent observation indicates that learning from each gender separately is beneficial to the deception detection process as they might use different linguistic, thermal, and physiological signatures.

5.4 Integrated Modalities

Table 5 lists the average accuracy percentage using the fusion of different modalities for individual and combined topics as well as individual and combined genders. Different combinations of the linguistic (represented by a set of combined unigrams and LIWC features), thermal (forehead features), and physiological (respiration rate features) modalities are tested. The table provides some

Table 4: Average accuracy percentage using different sets of thermal features for both genders, males only, and females only, as well as for "Abortion", "Best Friend", "Mock Crime", and "All Topics" combined. The best result for all topics are highlighted in bold, and the results for the feature sets found to work best on the development data are underlined.

Thermal	Abortion			Best Friend			Mock Crime			All Topics		
	Both	Male	Female	Both	Male	Female	Both	Male	Female	Both	Male	Female
Face	51.0	42.2	54.7	**54.8**	52.9	50.9	60.6	45.1	**66.0**	53.3	52.5	**53.2**
Forehead	**53.8**	54.9	**56.6**	50.5	**69.6**	**55.7**	59.6	51.0	52.8	52.3	43.5	51.7
Periorbital	51.9	**58.8**	47.2	53.8	42.2	52.8	53.8	**52.9**	50.9	51.9	51.8	48.3
Cheeks+Nose	50.0	57.8	51.9	47.6	58.8	53.8	53.8	41.2	**66.0**	**53.8**	**52.9**	52.1
Nose	53.4	47.1	50.9	47.1	47.1	48.1	**63.5**	45.1	43.4	49.8	47.5	**53.2**

Table 5: Average accuracy for the feature fusion using combinations of the best feature sets for the linguistic (LIWC combined with unigrams), thermal (forehead), and physiological (respiration rate) modalities. Best results for "All Topics" are highlighted in bold.

Fusion	Abortion			Best Friend			Mock Crime			All Topics		
	Both	Male	Female	Both	Male	Female	Both	Male	Female	Both	Male	Female
Linguistic+Thermal	**63.5**	54.9	57.5	**61.1**	55.9	47.2	58.7	56.9	69.8	**56.7**	**55.7**	62.6
Linguistic+Physiological	58.7	56.9	63.2	56.3	**60.8**	52.8	**59.6**	54.9	56.6	56.0	52.9	**66.4**
Thermal+Physiological	55.8	65.7	**67.0**	50.0	48.0	**62.3**	54.8	**64.7**	**71.7**	50.2	52.5	50.9
Linguistic+Thermal+Physiological	58.2	**70.6**	55.7	53.8	52.9	50.0	52.9	58.8	54.7	56.2	50.2	60.0

Table 6: Distribution of the choice of the 104 subjects on whether to lie or be truthful in the "Mock Crime" scenario as well as the interviewer correct prediction rate.

	Male	Female
Truthful	30 (58.8%)	27 (51%)
Deceptive	21 (41.2%)	26 (49%)
Overall No. of Responses	51	53
Interviewer prediction	45.1%	73.6%

interesting observations. The integration of features from multiple modalities provides the classifier with richer information, which is reflected in an overall improved performance compared to individual modalities, especially and consistently for "Abortion". The best performing fusion is achieved in four cases with the combination of thermal and linguistic as well as thermal and physiological features. Moreover, the results are consistently above random guessing in the vast majority of cases.

Following the same trend, the average accuracy achieved by learning from the female instances outperforms that of the male instances in individual and combined topics. In particular, this is the case in 10 out of 16 cases. Once again combining topics as well as integrating features from both genders seem to lose patterns that are domain- and gender-specific. Additionally, it seems that deception can more easily be identified in the "Abortion" and "Mock Crime", followed by "Best Friend".

5.5 Human Detection of Deception

In the one-on-one interview for the "Mock Crime" scenario, the subjects made their own decisions on whether to lie to the interviewer. Hence, the subjects had to choose one of four options; steal the money and deny it, steal the money and admit taking it, leave the money and falsely claim they took it, or leave the money and say the truth. At the end of the scenario, the same interviewer would predict whether the subjects took the money and the subjects would receive a higher incentive if the interviewer was wrong.

Table 6 shows that a relatively larger fraction of males choose to be truthful in "Mock Crime". However, the distribution of females choosing to lie is approximately equal to the number of truthful females. Interestingly, the interviewer is clearly better at identifying deceptive and truthful responses from females compared to males with a correct prediction relative improvement of 63.2%. This agrees with the classifier performance for "Mock Crime" in most cases using individual and combined modalities.

6. CONCLUSION

In this paper, we introduced a novel gender-balanced deception dataset that includes a large number of subjects and three different scenarios. We developed the first gender-based deception detection system using a multimodal approach. In particular, we provided an extensive analysis of different linguistic and physiological feature sets. Additionally we tracked and extracted thermal features from different facial areas using a novel tracking approach.

Our experimental analysis led to interesting observations concerning the differences in lying behaviors among genders. A predominant trend is that deception appears to be more easily detectable in females. In particular, it was easier to detect deception in females using individual and combined modalities with accuracy figures of above 80% . This is supported as well with the human performance in predicting deception for females.

Our analysis showed that different linguistic, physiological, and thermal patterns were observed with each gender. While there were specific feature sets that performed well with both genders such as the thermal features from the forehead area and the combination of LIWC features and Unigrams, there were other sets that performed well only with females such as the thermal features from the whole face and syntax-based features.

Additionally, linguistic features derived from the LIWC categories as well as the respiration rate and skin conductance signal showed different trends between males and females as they acted truthfully and deceptively. Moreover, the performance is negatively affected by the combination of instances from different domains as well

as different genders. Clearly the gender-specific thermal, linguistic, and physiological signatures were blended in this combination, which was reflected in the learning process. In conclusion, it would be beneficial to consider the gender differences in deception in order to improve the detection performance. Furthermore, following a multimodal approach by integrating features from different modalities enriched the learning process.

In the future we are planning to add the visual modality to enhance our gender-based deception detection system and explore the role of facial expressions and gestures in this task as well as to observe additional gender- and cultural-based deceptive behavior trends.

7. ACKNOWLEDGMENTS

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

8. REFERENCES

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

[2] M. Abouelenien, V. Pérez-Rosas, B. Zhao, R. Mihalcea, and M. Burzo. Detecting deceptive behavior via integration of discriminative features from multiple modalities. *IEEE Transactions on Information Forensics and Security*, 2016.

[3] J. Childs. Gender differences in lying. *Economics Letters*, 114(2):147 – 149, 2012.

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

[5] J. Conrads, B. Irlenbusch, R. M. Rilke, and G. Walkowitz. Lying and team incentives. *Journal of Economic Psychology*, 34:1 – 7, 2013.

[6] A. Dreber and M. Johannesson. Gender differences in deception. *Economics Letters*, 99(1):197 – 199, 2008.

[7] J. P. T. Fatt. Detecting deception through non verbal cues: gender differences. *Equal Opportunities International*, 17(2):1–9, 1998.

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

[9] R. E. Guadagno, B. M. Okdie, and S. A. Kruse. Dating deception: Gender, online dating, and exaggerated self-presentation. *Computers in Human Behavior*, 28(2):642 – 647, 2012.

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

[11] S. M. Ho and J. M. Hollister. Guess who?: An empirical study of gender deception and detection in computer-mediated communication. In *Proceedings of the 76th ASIS&T Annual Meeting: Beyond the Cloud: Rethinking Information Boundaries*, ASIST '13, pages 117:1–117:4, Silver Springs, MD, USA, 2013. American Society for Information Science.

[12] S. I. Levitan, M. Levine, J. Hirschberg, N. Cestero, G. An, and A. Rosenberg. Individual differences in deception and deception detection. In *The Seventh International Conference on Advanced Cognitive Technologies and Applications*, pages 52–56, France, March 2015.

[13] H. Liu and R. Mihalcea. Of men, women, and computers: Data-driven gender modeling for improved user interfaces. In *International Conference on Weblogs and Media*, 2007.

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

[15] R. Mihalcea and S. Pulman. Linguistic ethnography: Identifying dominant word classes in text. In *Computational Linguistics and Intelligent Text Processing*, pages 594–602. Springer, 2009.

[16] 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.

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

[18] 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*, ICMI '15, pages 59–66, New York, NY, USA, 2015. ACM.

[19] V. Pérez-Rosas, M. Abouelenien, R. Mihalcea, Y. Xiao, C. J. Linton, and M. Burzo. Verbal and nonverbal clues for real-life deception detection. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 2336–2346, 2015.

[20] S. N. Sinha, J.-m. Frahm, M. Pollefeys, and Y. Genc. Gpu-based video feature tracking and matching. Technical report, The University of North Carolina, 2006.

[21] P. Tilley, J. F. George, and K. Marett. Gender differences in deception and its detection under varying electronic media conditions. *2014 47th Hawaii International Conference on System Sciences*, 1:24b, 2005.

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

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