Detecting Deceptive Behavior via Integration of Discriminative Features From Multiple Modalities

Mohamed Abouelenien, Verónica Pérez-Rosas, Rada Mihalcea, and Mihai Burzo

***Abstract*— Deception detection has received an increasing amount of attention in recent years, due to the significant growth of digital media, as well as increased ethical and security concerns. Earlier approaches to deception detection were mainly focused on law enforcement applications and relied on polygraph tests, which had proved to falsely accuse the innocent and free the guilty in multiple cases. In this paper, we explore a multimodal deception detection approach that relies on a novel data set of 149 multimodal recordings, and integrates multiple physiological, linguistic, and thermal features. We test the system on different domains, to measure its effectiveness and determine its limitations. We also perform feature analysis using a decision tree model, to gain insights into the features that are most effective in detecting deceit. Our experimental results indicate that our multimodal approach is a promising step toward creating a feasible, non-invasive, and fully automated deception detection system.**

***Index Terms*— Deception detection, multimodal processing, thermal features, linguistic features, physiological features.**

I. Introduction

DECEPTION is defined as an intentional attempt to mislead others [1]. Different instances of deceptive behavior occur on a daily basis such as intended lies, fabrications, omissions, misrepresentations, and others. Deceptive manners range from simple harmless lies to major threats. An increased national interest in the deception detection research exists, especially with the alarming security incidents that took place in several countries in the past two decades.

Whether in airports, courts, or police interrogations, the decisions concerning deceptive behaviors are subject to human errors and are usually biased. Additionally, traditional methods such as polygraph tests failed in multiple cases resulting in falsely accusing the innocent, or freeing up those guilty of committing crimes. Polygraph tests rely mainly on physiological measurements collected from the subjects, and in some cases, these measurements are influenced by the type of questions asked by the interviewer [2]–[4].

Because of this, alternative approaches were considered to improve deception detection [5]. In particular, biological, visual, linguistic, acoustic, and thermal features were extracted in order to develop more reliable screening systems. However, some of these modalities are invasive or unfeasible such as the biological and physiological modalities that require the use of devices such as MRIs or other invasive sensors. Moreover, it is inconvenient to employ such intrusive methodologies on millions of daily travelers and on every suspect in custody due to policy and economy considerations.

This article addresses the aforementioned drawbacks and introduces a multimodal approach to develop a reliable system for detecting deceit. Specifically, the paper makes three main contributions. First, we develop a dataset for detecting deception collected from 30 participants with multiple responses from each participant. The subjects were asked to discuss two different topics ("Abortion" and "Best Friend," described in detail below) in addition to participating in a "Mock Crime" scenario, while they were recorded using two web cameras, a thermal camera, and several physiological sensors. Second, we determine the region of the face that is most discriminative between deceptive and truthful behaviors using thermal imaging. This is specified by segmenting the participants' faces into five areas including the whole face, the forehead, the periorbital area, the cheeks and nose region, and the nose and then creating a heat map by tracking these segments over their entire responses. Third, we develop a novel multimodal system that integrates features from modalities such as thermal, linguistic, and physiological in order to automate and enhance the detection of deceptive manners, avoid the limitations associated with individual modalities and human judgment, and improve the efficiency of the decision making process. We further provide an extensive analysis of the most effective multimodal features in detecting deception by using a decision tree model and inspecting the nodes of the tree. To our knowledge, this is the first attempt to integrate these modalities and compare thermal face segments for improved deception detection.

The article is organized as follows. Section II provides an extensive survey of different modalities used to detect deceit. Section III explains the details of our hypothesis, the data collection process, and our experimental design. Section IV illustrates the feature extraction and fusion processes.

Manuscript received December 7, 2015; revised May 20, 2016 and November 14, 2016; accepted November 20, 2016. Date of publication December 13, 2016; date of current version February 22, 2017. This work was supported in part by the National Science Foundation under Award 1344257 and Award 1355633, in part by the John Templeton Foundation under Grant 48503, and in part by DARPA-BAA-12-47 DEFT under Grant 12475008. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation, the John Templeton Foundation, or the Defense Advanced Research Projects Agency. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Nitesh Saxena. *(Corresponding author: Mihai Burzo.)*

M. Abouelenien, V. Pérez-Rosas, and R. Mihalcea are with the Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI 48109 USA (e-mail: zmohamed@umich.edu; vrncapr@umich.edu; mihalcea@umich.edu).

M. Burzo is with Mechanical Engineering, University of Michigan–Flint, Flint, MI 48502 USA (e-mail: mburzo@umich.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIFS.2016.2639344

Section V discusses our experimental results. Finally, we conclude the paper and discuss future work in Section VI.

II. Related Work

Previous work on detecting deceptive behavior can be roughly divided into contact and non-contact approaches. Techniques relying on the extraction of physiological and biological measurements fall under the category of contact approaches. Non-contact deception detection approaches include verbal and acoustic, visual, and thermal techniques.

Older methodologies for detecting deceit, especially in law-enforcement, relied on polygraph tests and signals extracted from the nervous system. These tests looked for an increased activity in the nervous system that were determined using physiological measurements, such as heart rate, blood pressure, skin conductance, respiration rate, etc. Vrij [6] provided guidelines to detect deceit and concluded that polygraph tests were not sufficiently reliable in deception detection. Several other publications discussed the shortcomings of relying solely on polygraph tests [2], [3], [7], [8].

To examine alternative methods to improve deception detection rates, contact-based biological measurements, such as the functional magnetic resonance imaging (fMRI) technology were extracted from the human body to identify liars. Kozel *et al.* [9] captured fMRI images and recorded electrodermal activity signals while participants were responding to certain questions deceptively and truthfully. The study concluded that changes in the blood flow and certain brain waves were associated with deception.

As a consequence of the limitations of the contact invasive methods, deception detection research shifted towards non-contact, non-invasive methods. Visual body language was explored in order to detect deceit. In particular, spontaneous facial expressions and hand gestures were of special interest due to their usage to express people's emotions on daily basis. Ekman [10] analyzed micro- and squelched-expressions that can be associated with an act of deception.

Owayjan *et al.* [11] developed a lie detection system by applying geometric-based dynamic templates on recorded video frames of subjects acting deceptively to extract measurements from their facial micro-expressions. Pfister and Pietikäinen [12] introduced a temporal interpolation method to detect clues to lies from visual micro-expressions using kernel learning. They additionally published a publicly available database of micro-expressions.

Visual hand gestures were studied in relation to deceptive actions. Hillman *et al.* [13] examined subcategories of hand gestures and showed that individuals acting truthfully produced more rhythmic pulsing gestures while those acting deceptively made more frequent speech prompting gestures. Maricchiolo *et al.* [14] introduced a taxonomy of hand gestures related to multiple social contexts including deception.

Linguistic and acoustic analysis was the focus of many recent researches owing to its non-invasiveness and promising results to reveal clues of deceptive behavior. For instance, researchers studied verbal behaviors exhibited by people while deceiving [15], [16]. Speaking rate, energy, pitch, range as well as the identification of salient topics were found useful to distinguish between deceptive and non-deceptive speech [17]. Other work analyzed the number of words, sentences, self-references, affect, spatial, and temporal information associated with deceptive content [18]. Mihalcea and Strapparava [19] extracted salient linguistic features and found patterns of words that were correlated with deceptive text such as emphasizing certainty. Fornaciari and Poesio [20] studied the effect of having a more homogeneous sets of subjects on improving classification rates of deception. Feng *et al.* [21] explored syntactic stylometry for deception detection and showed that Context Free Grammar (CFG) parse trees achieved improved detection rates compared to shallow lexico-syntactic features.

Several efforts were additionally exerted in the direction of non-invasive approaches of detecting deception using thermal imaging. The relation between thermal measurements extracted from the subjects' faces and states of deception were investigated. Pavlidis [22] introduced a method to score polygraph tests based on features extracted from the facial area using thermal imaging. Garbey *et al.* [23] proposed a bioheat transfer model that described the geometry and anatomy of large blood vessels in the facial area using thermal images to indicate their relation to deceit. Warmelink *et al.* [24] extracted the maximum, minimum, and average temperatures from thermal images to use them as a lie detector in airports. Using a set of 51 subjects, their system was able to detect liars with accuracy above 60%. However, interviewers outperformed the system with above 70% accuracy.

In order to specify which thermal areas had higher correlation to deceptive behaviors, facial regions of interest were specified. Pavlidis and Levine [25], [26] applied thermodynamic modeling on thermal images to transform the raw thermal data from the periorbital area in the face to blood flow rates to detect deception. Zhou *et al.* [27] applied spatial and temporal smoothing components on thermal videos using a probabilistic template function for improved tracking of facial areas for deception and stress studies. Rajoub and Zwiggelaar [28] analyzed thermal faces by creating two regions of interest by manually identifying the corners of the eyes and tracking these regions over the recorded video frames. Their deception detection system performed well on within-person data but not on between-person scenarios. Jain *et al.* [29] employed a thermal camera with face detection, tracking, and landmark detection systems to track landmarks on the regions of interest in the face area. The method calculated the average temperature of the 10% hottest pixels of a window that included both tear ducts.

Recently, multimodal approaches have been suggested for improved deception detection by integrating features from different modalities from simulated data collected in lab settings [30] and from real-life data [31], [32]. Jensen *et al.* [33] integrated visual, acoustic, and verbal features such as head and hands position, pitch variety, and self-references using a multimodal approach for improved recognition of deceit. Nunamaker *et al.* [34] reviewed theoretical and technological methods for automated human credibility screening from acoustic, linguistic, physiological, and thermal imaging perspectives. Burgoon *et al.* [35] considered a combination of three verbal and nonverbal approaches

to detect deception including message feature mining, speech act profiling, and kinesics analysis.

III. Experimental Setup

A. Data Acquisition

We collect our deception dataset using: two Logitech web cameras with a resolution of 800x600 and a frame rate of 60 fps; thermal measurements using a FLIR Thermovision A40 thermal camera, with a resolution of 340x240 and a frame rate of 60 fps; physiological responses collected from four bio-sensors, namely blood volume pulse (BVP sensor), skin conductance (SC sensor), skin temperature (T sensor), and abdominal respiration (BR sensor). The voice of the participants is also recorded using a microphone embedded in one of the web cameras.

B. Participants

The participants were recruited from undergrad and graduate populations. All participants signed an informed consent form[1] and were informed about the goals of the study and their involvement. The participants consisted of 30 students, including five females and twenty-five males. All participants expressed themselves in English, had several ethnic backgrounds, and ages ranging between 22 and 38 years.

C. Truthful and Deceptive Responses

Before the beginning of the recording session, we described the experimental settings and procedures to the participants and instructed them to respond either truthfully or deceptively. In addition, participants were instructed to avoid excessive movements with their hands in order to prevent interference with the sensors' measurements. Following these simple restrictions helped us to obtain high quality data.

Aiming to elicit deceptive and truthful responses from the participants, we performed three experiments. The first applied a "Mock Crime" scenario where subjects stole an envelope containing money. This topic consisted of simple involvement from the interviewer by questioning the participants. The second and third experiments consisted of providing participants with two topics ("Abortion" and "Best Friend") for which they had to provide verbal responses freely while being recorded. These two topics did not include any involvement from the interviewer. While the deceptive and truthful conditions for the "Mock Crime" scenario were randomized across subjects, the "Abortion" and "Best Friend" scenarios were administered in a fixed order.

1) Mock Crime (MC): In this experiment, participants were assigned randomly to be deceptive or truthful. They were instructed to look for a hidden envelope on an office's desk. In the deceptive scenario the envelope contained a $20 dollar bill, while for the truthful case, the envelope was empty. Each participant was instructed to deny that he or she has seen or taken the money. Thus, participants who did not take the money were truthful while those who took the money were lying.

This was followed by a one-on-one interview conducted as follows:

1) Are the lights on in this room?
2) Regarding that missing bill, do you intend to answer truthfully each question about that?
3) Prior to 2012, 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 behind the white board. Please aim at a clear description of about 2-3 minutes.
10) Do you know where that missing bill is now?

2) Abortion (AB): This experiment consisted of asking participants to provide first a truthful response where they had to defend their point of view regarding abortion, and second a deceptive opinion about their feelings towards abortion. Participants were instructed to imagine they were participating in a debate, and asked to contribute a 2-3 minute speech of their truthful (or deceptive) opinion.

3) Best Friend (BF): This experiment consisted of asking participants 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 person they do not like as if he or she were their best friend. Therefore, in both cases, a person was described as the participants' best friend. Both descriptions were 2-3 minutes in length.

D. Hypothesis

This study was conducted based on our hypothesis that, as a person would act/speak deceptively under the given scenarios and without any involvement from the interviewer in the "Abortion" and "Best Friend" topics, there would be subtle changes in his/her physiological, thermal, and behavioral responses. Moreover, we prepared the "Mock Crime" scenario as to introduce partial human involvement via asking suspicious questions on the occurrence of a crime to investigate whether deceptive responses could be more accurately detected with interference. In general, we assumed that these subtle changes could be detected by extracting discriminant features from multiple modalities.

IV. Methodology

Features were extracted separately from each of the involved modalities as follows.

A. Linguistic Features

In order to incorporate information of the deceiver's language usage, we obtain speech transcriptions of the

[1]The study was approved by the IRB at the University of North Texas, where the data was collected.

recorded statements. This represents the linguistic component of our analysis. "Abortion" and "Best Friend" statements were entirely transcribed whereas "Mock Crime" interviews were partially transcribed. This was motivated by the fact that questions number 1 to 7 and number 10 were designed as yes/no questions, thus only questions 8-9 were transcribed.

In order to obtain linguistic features from the available transcriptions, we extract several sets of features that were successfully applied for deceit detection. First, we extract unigrams (UNI) derived from the bag of words representation of the words present in each topic transcriptions. Unigrams features are encoded as tf-idf (Term Frequency - Inverse Document Frequency) values, which consist of a frequency normalization that reflects the importance of each word in the deceptive and truthful statements. This feature set consist of 2424 unique words

Second, in order to obtain features that represent psychological processes occurring while people are deceiving, we extract features derived from the Linguistic Inquiry and Word Count (LIWC) lexicon [36]. This dictionary has been widely used for psycholinguistic analysis of deceptive texts and consists of 72 word classes relevant to physiological processes such as motion, cognition, affect, and linguistic processes, among others. A detailed description of these word-classes can be found in [37].

Third, we obtain linguistic features that represent shallow and deep syntax patterns associated to deception. Following [21], we extract a set of features derived from Part Of Speech (POS) tags and from production rules based on probabilistic Context Free Grammar (CFG) trees. We use the Berkeley parser to obtain both POS and CFG. Our POS features are encoded as the tf-idf values of each POS tag occurring in the dataset. The final set consists of 2,807 POS features and 1,339 CFG features.

In addition, we explore the use of linguistic features that measure transcription's syntactic complexity and reading difficulty in terms of readability scores. This is in line with previous research that has suggested that liars use less complex and less detailed sentences [1]. To extract these features, we use a tool provided by Lu [38], which generates indices of syntactic complexity based on the analysis of T-units, which are defined as the shortest grammatically allowable sentences into which writing can be split or a minimally terminable unit. T-units usually consist of a main clause plus all subordinate clauses and non-clausal structures that are attached to or embedded in it [39]. A total of 21 features are generated, including: mean length of sentence (MLS), mean length of T-unit (MLT), mean length of clause (MLC), clauses per sentence (C/S), among others. Two standard readability score indices, the Flesch-Kincaid and Gunning Fog, are calculated to represent transcriptions readability.

We conduct a set of experiments where we evaluate each linguistic feature set as well as their combinations using a machine learning approach. The classification results for truthful and deceptive statements obtained with a decision tree classifier implemented as described in Section V are shown in Figure 1.

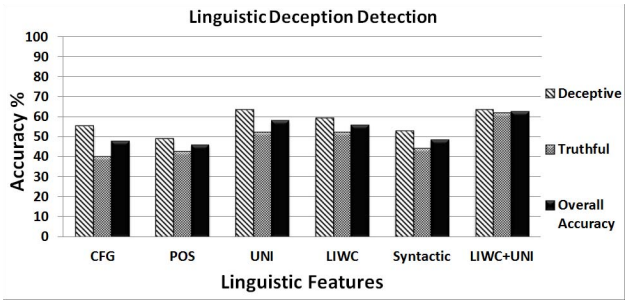


Fig. 1. Recall of the deceptive and truthful classes, and overall accuracy percentages for Context Free Grammar (CFG), Part of Speech Tags (POS), Unigrams (UNI), Linguistic Inquiry and Word Count (LIWC), and LIWC+UNI.

The graph shows the average recall of the deceptive and truthful classes, as well as the average overall accuracy percentages, using leave-one-instance-out cross validation. We can observe that, for the deceptive class, the classifiers built with the different feature sets achieve recall figures above the random baseline, thus indicating a good performance in the deceptive class prediction. However, we observe a lower performance for the truthful class using individual sets of features. To explore the benefit of integrating different linguistic cues in the deception detection task, we conduct an additional set of experiments using different combinations between the feature sets. From these experiments, the combination of unigrams and LIWC features provides the best trade-off between accuracy and recall, in the range of 61% to 63%, for both the deceptive and truthful classes. The remaining combinations did not show any improvements over the use of unigrams only. In the remainder of this work, a feature set of 2,496 features, consisting of unigrams and features derived from the LIWC lexicon, is used to represent the linguistic component of our system.

B. Thermal Features

The thermal feature extraction process was performed in three steps, face segmentation, tracking, and thermal map formation.

1) Segmenting and Tracking Regions of Interest: First, we manually segmented the face area of each thermal video response into five areas including the whole face, the forehead, the periorbital area, the cheeks and nose region, and the nose by itself. These regions will be referred to as regions of interest (ROI) in the rest of this paper. Each response was preceded by at least a minute of no activity which was used for the normalization and thermal correction process as explained below. The locations of the pixels of the boundary boxes surrounding each ROI were manually specified from one frame in the beginning of the video. The next step was tracking the ROIs over the participants' entire deceptive/truthful response. We localized interesting points in each ROI using Shi-Tomasi corner detection algorithm [40]. To detect these points, the method calculates the weighted sum of square difference (SSD) between two images. In this case, as the method compares an image patch $I_1(x_i)$ and a shifted version of this image, $I_1(x_i + \Delta u)$, an auto-correlation function S is

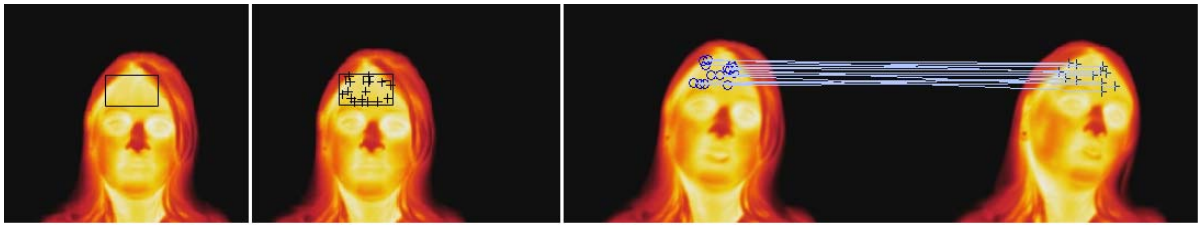


Fig. 2. An overview of the tracking process including manually determining the bounding box of the ROI, detecting interesting points, and matching/tracking the points throughout the thermal video.

employed.

$$S(\Delta u) = \sum_i w(x_i)(I_0(x_i + \Delta u) - I_0(x_i))^2 \quad (1)$$

where u is the displacement vector and $w(x_i)$ is a window function. The function is approximated using Taylor Series expansion into

$$S(\Delta u) \approx \sum_i w(x_i)(\nabla I_0(x_i).\Delta u)^2 \quad (2)$$

where,

$$\nabla I_0(x_i) = (\frac{\partial I_0}{\partial x}, \frac{\partial I_0}{\partial y})(x_i)$$

We used a Gaussian filter of fixed size to smooth the calculated gradient. S can be rewritten as:

$$S(\Delta u) = \Delta u^T V \Delta u \quad (3)$$

V denotes the auto-correlation matrix. The interesting corner points to be tracked were located using the variation in S by computing the minimum eigenvalues from V. The detected points were then tracked using a fast Kanade-Lucas-Tomasi (KLT) tracking algorithm [41]. The algorithm assumed a small displacement between the pixels in a frame at times t and $t + \tau$, which suited our tracking requirements. It then obtained the displacement which minimized the computer error between the current and following frames. The implementation additionally calculated the Forward-Backward Error [42] by tracking the points back and force through the frames in order to eliminate outliers and avoid the uncertainty associated with some points using a total of 1000 frames uniformly selected from each video to improve efficiency.

Following the tracking process and displacement estimation, geometric transformation [43], which globally estimated the interesting points transformation based on similarity, was applied to map the interesting points between the frames. Once the points were mapped, the new boundary box was geometrically determined. We discarded the tracking of the current frame and proceeded to the next one if the number of matched points was less than 95%. An overview of tracking process can be seen in Figure 2.

2) Thermal Features Extraction: The tracked ROIs were cropped in order to extract meaningful thermal features to discriminate between deceptive and truthful states. The tracked regions could in some cases take the shape of a polygon. The rectangular region masking the boundaries of the polygon-shaped ROI was geometrically determined and cropped.

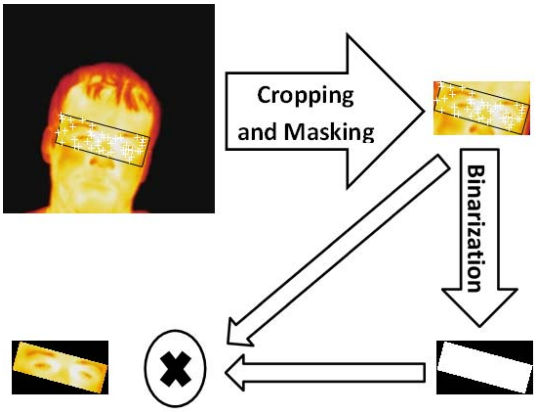


Fig. 3. An example of extracting the periorbital ROI from a tilted face during tracking. The ROI is masked, cropped, binarized, and finally multiplied by the cropped image to blacken and eliminate external regions.

The backgrounds of the cropped ROIs were additionally discarded by thresholding the values of the pixels into (0) for black or (1) for white to eliminate their effect on the quality of the features using an image binarization technique.

This transformation formed a holistic shape of the ROIs which was then multiplied by the original image to eliminate the background. The cropping and binarization processes are shown in Figure 3. Once the ROIs were cropped into their final form, a complete thermal map that defined the heat distribution of each ROI was created.

The thermal map was created using the gray scale level, and Hue Saturation Value (HSV) channels by extracting the maximum pixel value corresponding to the highest temperature in the ROI, the average of the pixels values of the ROI, the minimum pixel value corresponding to the lowest temperature in the ROI, the maximum/minimum pixels range which measured the difference between the maximum and minimum temperatures, the mean of the 10% highest pixel values corresponding to the mean of 10% highest temperatures in the ROI, and a histogram of 255 bins (zero-valued pixels were excluded) over the values of the pixels in the ROI to form a total of 260 thermal features for the gray scale ROI and 780 thermal features for all HSV channels. The histograms were normalized to form a probability distribution over all bins.

As different individuals could have varying skin temperatures in normal conditions, a thermal correction process was followed in order to treat data from different participants equally and improve the quality of the extracted features.

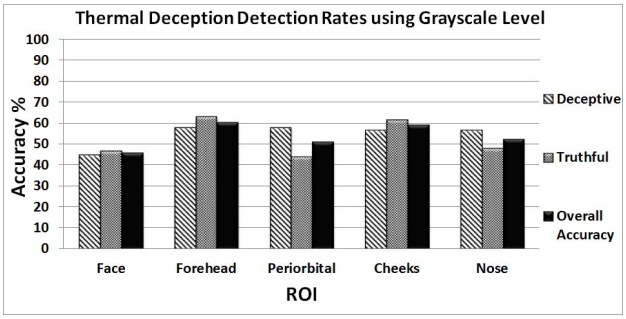


Fig. 4. Recall of the deception and truthfulness classes, and overall accuracy % for different ROIs using thermal features from gray scale level.

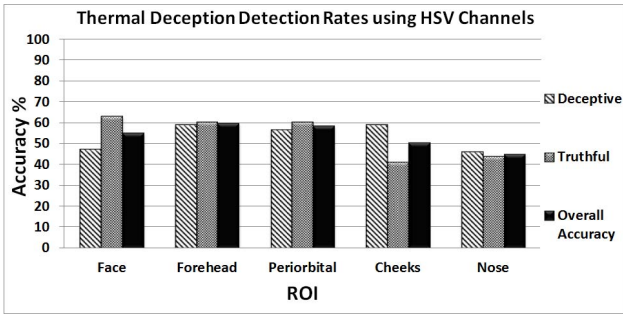


Fig. 5. Recall of the deception and truthfulness classes, and overall accuracy % for different ROIs using thermal features from the HSV channels.

Similarly, the same set of thermal features was extracted from a one minute recording of no activity preceding each response forming a thermal baseline. Hence, thermal correction was achieved by dividing the features extracted from the actual responses by their thermal baseline.

3) Deceptive Face Segments: Figure 4 compares the performance of different ROIs using features extracted from the gray scale pixels. The figure shows that the forehead region outperforms all other ROIs in the overall and individual class performances. The periorbital region provides an improved deception detection rate but deteriorated truthfulness detection rate. The cheeks region also achieves enhanced performance that is close to 60% overall accuracy.

Figure 5 illustrates the deceptive and truthful classes recall as well as the overall accuracy using the HSV thermal features extracted from different ROIs. The HSV features extracted from the forehead and periorbital regions achieve the highest accuracy of approximately 60% overall accuracy, with slight improvement for the forehead region. The face features exceed 60% recall for the truthful class; however, the deceptive class receives less than 50% recall. The cheeks and nose regions achieve the poorest overall performance.

Based on the previous observations, the forehead and periorbital regions were found to exhibit a relatively consistent and improved performance compared to all other regions. While most of the previous research focused on the periorbital area, the forehead offers competitive and sometimes improved performance. Hence, we decided to use features from the two most promising indicators of deception for integration with features from the other modalities.

C. Physiological Features

Biograph Infiniti Physiology suite was used to obtain physiological assessments for heart rate, blood volume pulse, respiration rate, and skin conductance in a rate of 2048 samples per second. The physiological feature set consists of raw measurements and their statistical descriptors, including maximum and minimum values, means, power means, standard deviations, and mean amplitudes (epochs). In addition, we obtain features derived from inter-beat intervals (IBI) measurements such as the minimum and maximum amplitudes and their intervals. The final set consists of a total of 60 physiological features.

D. Feature Fusion and Deception Classification

We used two levels of fusion to create our multimodal dataset, feature-level (early) and decision-level (late) fusion. The early fusion was performed by averaging and integrating the features of each response from different modalities. Late fusion was performed by training each of the three modalities separately and combining their decisions on test instances using majority voting. For the classification process, a decision tree classifier was used as recommended in [18] and [44] for deception detection.

We followed a leave-one-instance-out cross validation scheme to report the average overall accuracy and the average recall of the deceptive and truthful classes. The performance of individual modalities was compared to multimodal fusion to clarify whether features integration in fact improved the performance. In this case, to further identify the statistical significance of the improvement, we used the Poisson Binomial Test (PBT) [45]. The method measures the probability a certain algorithm/dataset provides higher prediction capability than others using zero-one loss for each instance in the dataset. The measured probability additionally shows if the size of the data is enough to draw conclusions. We considered a probability that is equal to or greater than 0.65 to be significant. Moreover, the nodes of the trained decision tree were visualized and specified in order to determine the most discriminative features for detecting deceit.

V. Experimental Results

As described earlier, our multimodal dataset consists of 149 instances collected from 30 participants using three different topics with a distribution of 76 deceptive instances and 73 truthful instances. The best feature sets from the linguistic and thermal modalities are fused with features extracted from the physiological modality. The performance of individual and fused modalities is also evaluated and compared on each topic separately and all topics combined. To explore whether a model trained on deceptive and truthful data from one domain (topic) can successfully identify deceit in a different domain, a cross-topic learning scheme is also evaluated using individual and integrated modalities.

A. Feature Fusion

1) Learning From Individual Topics: The classification performance of the three individual topics is compared in order

TABLE I

THE RECALL AND OVERALL ACCURACY PERCENTAGES FOR INDIVIDUAL AND INTEGRATED MODALITIES WITH THE USAGE OF GRAY SCALE THERMAL FEATURES FOR THE THREE INDIVIDUAL TOPICS. BEST RESULTS ARE HIGHLIGHTED IN BOLD

Modalities	Phys	Ling	Thermal Forehead	Thermal Periorbital	Ling+Phys	Phys+Thrm Forehead	Phys+Thrm Periorbital	Ling+Thrm Forehead	Ling+Thrm Periorbital	Ling+Phys+ Forehead	Ling+Phys+ Periorbital
"Abortion"											
Deceptive	60.0	**80.0**	56.67	30.0	76.67	53.33	33.33	60.0	**80.0**	60.0	**80.0**
Truthful	33.33	**80.0**	56.67	46.67	**80.0**	53.33	40.0	63.33	70.0	60.0	70.0
All Accuracy	46.67	**80.0**	56.67	38.33	78.33	53.33	36.67	61.67	75.0	60.0	75.0
"Best Friend"											
Deceptive	**63.33**	43.33	46.67	50.0	50.0	43.33	53.33	56.67	33.33	53.33	36.67
Truthful	46.67	33.33	46.67	60.0	43.33	40.0	**63.33**	33.33	40.0	30.0	40.0
All Accuracy	55.0	38.33	46.67	55.0	46.67	41.67	**58.33**	45.0	36.67	41.67	38.33
"Mock Crime"											
Deceptive	50.0	31.25	62.50	62.50	56.25	**68.75**	50.0	50.0	37.50	62.50	43.75
Truthful	**69.23**	30.77	46.15	23.08	**69.23**	61.54	**69.23**	46.15	53.85	61.54	**69.23**
All Accuracy	58.62	31.03	55.17	44.83	62.07	**65.52**	58.62	48.28	44.83	62.07	55.17

to indicate if certain scenarios induce more arousal to liars than true tellers. Table I lists the recall of the deceptive and truthful classes as well as the overall accuracy for each topic using the gray scale thermal features with 11 combination of individual and feature-fused modalities. The linguistic features clearly outperform all other modalities for the "Abortion" topic reaching an accuracy and recall of 80%. The fusion of linguistic features with physiological and periorbital thermal features achieves above 70% accuracy. The overall accuracy of the individual physiological and periorbital thermal modalities is below that of random guessing.

For the "Best Friend" topic, the best overall performance is achieved by the fusion of physiological and periorbital thermal features while the best deceptive class recall is achieved by the physiological features. To compare between the performance achieved by the best multimodal fusion and the best single modality, a PBT test is used. The probability p the multimodal set (Phys+Thrm Periorbital) is a better indicator of deceit than the single modality (Phys) is 0.61, and than the single modality (Thermal Periorbital) is 0.58, which is not significantly better.

The best performance for the "Mock Crime" scenario is obtained by integrating the physiological and the forehead thermal features. Modalities involving physiological features are able to more accurately detect truthful instances in this scenario. Using the PBT test, the multimodal set (Phys+Thrm Forehead) provides better prediction of deceit than the single modality (Phys) with a probability of 0.64.

In general, the "Abortion" topic exhibits the best performance of all three topics. The improved performance can be related to the emotional effect this topic can have on some subjects (especially on females), which is most noticeable in the linguistic features. On the other hand, subjects might not have been emotionally involved in the "Best Friend" topic, given that they talked positively either way. With the exception of the linguistic features performance in the "Abortion" topic, the multimodal fusion of different modalities outperforms the employment of single modalities. The performance exceeds the 50% random guessing baseline in 8 out of 11 cases for the "Abortion" topic and 7 out of 11 for "Mock Crime." On the contrary, the overall accuracy is below that of random guessing in 8 out of 11 cases for the "Best Friend."

Table II shows the same metrics as Table I for the three individual topics except for the replacement of the gray scale thermal features with the HSV features. For the "Abortion" topic, the linguistic features performance still stands out compared to the other modalities. The performance of "Best Friend" is also deteriorated and the best overall accuracy is claimed by the physiological modality and is slightly above random guessing. On the other hand, the top "Mock Crime" performance is slightly improved reaching almost 70% accuracy.

In general, the same trends are observed with the integration of HSV features. "Abortion" and "Mock Crime" performance is improved compared to "Best Friend." This improvement is related to the linguistic features performance for the "Abortion" topic and the fusion of physiological and forehead thermal features for "Mock Crime." The deteriorated performance of the linguistic features for "Best Friend" can be attributed to the fact that some feelings towards best friends can translate into negative words in the participants' responses. Using the PBT test for "Mock Crime," the fusion of (Phys+Thrm Forehead) provides better capability of indicating deceit compared to the single modality (Phys) (p = 0.68).

2) Learning From Combined Topics: To get a conclusive performance on all the collected data, the decision tree classifier was trained with the data collected from all three topics combined. Table III lists the recall of the deceptive and truthful classes in addition to the overall accuracy for all instances

TABLE II

THE RECALL AND OVERALL ACCURACY PERCENTAGES FOR INDIVIDUAL AND INTEGRATED MODALITIES WITH THE USAGE OF HSV THERMAL FEATURES FOR THE THREE INDIVIDUAL TOPICS. BEST RESULTS ARE HIGHLIGHTED IN BOLD

Modalities	Phys	Ling	Thermal Forehead	Thermal Periorbital	Ling+Phys	Phys+Thrm Forehead	Phys+Thrm Periorbital	Ling+Thrm Forehead	Ling+Thrm Periorbital	Ling+Phys+ Forehead	Ling+Phys+ Periorbital
"Abortion"											
Deceptive	60.0	**80.0**	36.67	16.67	76.67	36.67	30.0	36.67	76.67	36.67	76.67
Truthful	33.33	**80.0**	50.0	46.67	**80.0**	43.33	40.0	53.33	76.67	50.0	76.67
All Accuracy	46.67	**80.0**	43.33	31.67	78.33	40.0	35.0	45.0	76.67	43.33	76.67
"Best Friend"											
Deceptive	**63.33**	43.33	50.0	40.0	50.0	46.67	43.33	50.0	50.0	50.0	50.0
Truthful	46.67	33.33	53.33	50.0	43.33	**56.67**	33.33	36.67	43.33	36.67	46.67
All Accuracy	**55.0**	38.33	51.67	45.0	46.67	51.67	38.33	43.33	46.67	43.33	48.33
"Mock Crime"											
Deceptive	50.0	31.25	43.75	50.0	56.25	**68.75**	37.50	18.75	50.0	62.50	31.25
Truthful	**69.23**	30.77	23.08	30.77	**69.23**	**69.23**	30.77	7.69	53.85	61.54	38.46
All Accuracy	58.62	31.03	34.48	41.38	62.07	**68.97**	34.48	13.79	51.72	62.07	34.48

TABLE III

THE RECALL AND OVERALL ACCURACY PERCENTAGES FOR INDIVIDUAL AND INTEGRATED MODALITIES WITH THE USAGE OF GRAY SCALE AND HSV THERMAL FEATURES FOR ALL TOPICS COMBINED. BEST RESULTS ARE HIGHLIGHTED IN BOLD

Modalities	Phys	Ling	Thermal Forehead	Thermal Periorbital	Ling+Phys	Phys+Thrm Forehead	Phys+Thrm Periorbital	Ling+Thrm Forehead	Ling+Thrm Periorbital	Ling+Phys+ Forehead	Ling+Phys+ Periorbital
All Topics Grayscale											
Deceptive	60.53	63.16	57.90	57.90	50.0	**68.42**	60.53	63.16	56.58	64.47	53.95
Truthful	45.21	61.64	**63.01**	43.84	58.90	56.16	47.95	47.95	56.16	56.16	61.64
All Accuracy	53.02	**62.42**	60.40	51.01	54.36	**62.42**	54.36	55.71	56.38	60.40	57.72
All Topics HSV											
Deceptive	60.53	63.16	59.21	56.58	50.0	53.95	55.26	**68.42**	35.53	67.11	31.58
Truthful	45.21	61.64	60.27	60.27	58.90	58.90	57.53	68.49	50.69	**69.86**	45.21
All Accuracy	53.02	62.42	59.73	58.39	54.36	56.38	56.38	**68.46**	42.95	**68.46**	38.26

using gray scale and HSV thermal features. Using gray scale thermal features, the best overall accuracy is achieved by the fusion of physiological and forehead thermal features. The individual linguistic modality additionally achieved similar performance.

Using HSV thermal features, the fusion of the forehead thermal features with only linguistic and with both linguistic and physiological features clearly outperforms all other modalities and reaches a correct detection rate right below 70%. Their fusion is shown to provide significantly higher ability of discriminating between deceptive and truthful responses compared to the single linguistic modality using the PBT test (p = 0.69).

Based on our experimental results, it is clear that the best overall performance emerges from the fusion of the forehead features with linguistic features. The fusion of physiological features with linguistic features also leads to a similar performance. We also note that the performance of the individual physiological modality is distinctly lower than the individual linguistic and thermal modalities. Our analysis indicates that the increase in blood flow can be better detected in the forehead compared to the periorbital region, which can be due to the presence of hair areas such as the eyebrows and eyelashes, the dilation of the eyes, and blinking, which affect the thermal energy emitted from the periorbital area.

Moreover, over all three topics, it can be noted that the best performance figures are most of the time due to the use of multiple modalities as opposed to individual modalities (Table III). This is an important result, as the long term goal is to develop systems that are robust and which can be applied to

TABLE IV

PERCENTAGE IMPROVEMENT OF THE OVERALL ACCURACY ON LEARNING FROM INSTANCES FROM ALL TOPICS OVER LEARNING SOLELY FROM THE COMBINATION OF AB ("ABORTION") + BF ("BEST FRIEND") INSTANCES. "IMP" DENOTES THE PERCENTAGE IMPROVEMENT

Modalities	AB + BF	All Topics	Imp %
Phys	56.67	53.02	-6.44
Ling	52.50	62.42	18.90
Ling+Phys	53.33	54.36	1.93

	Gray Scale			HSV		
Modalities	AB + BF	All Topics	Imp %	AB + BF	All Topics	Imp %
Thermal Forehead	49.17	60.40	22.84	56.67	59.73	5.4
Thermal Periorbital	45.83	51.007	11.29	45.0	58.39	29.76
Phys+Thrm Forehead	56.67	62.42	10.15	52.50	56.38	7.38
Phys+Thrm Periorbital	49.17	54.36	10.56	46.67	56.38	20.81
Ling+Thrm Forehead	47.50	55.71	17.28	53.33	68.46	28.37
Ling+Thrm Periorbital	48.33	56.3	16.49	31.67	42.95	35.62
Ling+Phys+Forehead	51.67	60.40	16.90	51.67	68.46	32.49
Ling+Phys+Periorbital	44.17	57.72	30.68	32.50	38.26	17.72

any data regardless of its domain; Table III shows the behavior of our system in the presence of such multiple-domain data.

To observe whether the involvement of the interviewer in "Mock Crime" affects the performance of our deception detection system, we compare the overall accuracy when learning from all the topics versus learning from the combination of "Abortion" + "Best Friend" topics. Table IV presents the percentage improvement achieved by learning from all topics over learning solely from the combination of AB ("Abortion") + BF ("Best Friend") instances. The table indicates a consistent improvement when the "Mock Crime" instances are included, except for one case involving the individual physiological modality. The improvement is significant in multiple cases and exceeds 20% in 7 out of 19 cases. Although the improvement is evident, several other factors might have contributed to it. For instance, the additional increase in the training data size can generally improve the performance. The increased performance of the "Abortion" topic compared to "Best Friend," especially for the linguistic features, can deteriorate their performance combined. This deterioration could be compensated by adding the "Mock Crime" instances. Further analysis can be conducted with the collection of more data.

3) Cross-Topic Training: In order to develop a system that can reliably detect deceptive behaviors, it needs to be trained on data from multiple domains and scenarios. However, it is unfeasible to collect data involving all possible scenarios. Hence, testing such a system on domains that are not used for training is critical. Table V lists the deceptive and truthful classes recall, and the overall accuracy using gray scale thermal features to test instances of one topic while the classifier is trained on instances of the other two topics. For example, (Test "Abortion") in the table demonstrates the classification recall and accuracy of "Abortion" after training the classifier with instances from "Best Friend" and "Mock crime".

The table indicates a general deterioration in performance for the "Abortion" topic by cross-referencing with the individual "Abortion" topic performance in Table I. For the "Best Friend" topic, the performance alternates providing 7 out of 11 higher accuracies using cross-topic learning scheme. For most modalities, the "Mock crime" topic exhibits an improved performance using the cross-learning approach, especially with the individual forehead and periorbital thermal features. This can be attributed to the enlarged data size used for training when testing the "Mock Crime" topic. In most cases, the linguistic features provide a significant improved performance for one class on the expense of the other class.

Similarly, Table VI lists the performance with the employment of HSV thermal features using cross-topic learning scheme. Cross-referencing with individual topics performance in Table II confirms the trend of which the thermal features exhibit improved performance for each of the tested topics, while including linguistic features whether individually or fused improves the performance of one class on the expense of the other.

There was an improved performance in 8 out of 11 cases for the "Best Friend" topic when trained on instances from the two other topics. The fusion of physiological and thermal periorbital features is determined to be a statistically better indicator of deceit than the sole thermal periorbital modality using the PBT test for the "Mock Crime" dataset ($p = 0.72$).

In general, the results indicate that the thermal features benefit from the increased size of training data used in cross-topic learning regardless of the topic. On the other hand, the linguistic features appear to be dependent on the topic they are extracted from. For example, the unigrams extracted using words such as "friend," "like," and "steal" from the "Best Friend" and "Mock Crime" topics are not able to improve or keep the outstanding linguistic performance on the "Abortion" topic. If a model trained on such features is tested on a completely different matter, the results deteriorate significantly resulting in an imbalanced performance between the deceptive and truthful classes. Evidently, to be able to robustly use the linguistic modality on a new domain, a large amount of data from multiple domains and scenarios needs to be collected.

B. Leave-One-Subject-Out Cross Validation

In the majority of the experiments reported in this paper, we use a leave-one-instance-out strategy; under this strategy, for the "Abortion" and "Best Friend" topics, statements drawn from the same subject are shared between training and test. However, we also wanted to test the effect of a leave-one-subject-out strategy. Using this strategy, all the instances belonging to one subject are reserved for testing while all the other instances of all the other subjects are used for training; therefore, statements drawn from the same subject are never shared between training and test.

Table VII lists the average accuracy as well as the recall of the deceptive and truthful classes for the "Abortion,"

TABLE V

THE RECALL AND OVERALL ACCURACY PERCENTAGES FOR INDIVIDUAL AND INTEGRATED MODALITIES WITH THE USAGE OF GRAY SCALE THERMAL FEATURES FOR CROSS-TOPIC LEARNING SCHEME. TEST "ABORTION" INDICATES THAT THE "ABORTION" INSTANCES ARE TESTED WHILE THE CLASSIFIER IS TRAINED USING INSTANCES FROM "BEST FRIEND" AND "MOCK CRIME" AND SO ON. BEST RESULTS ARE HIGHLIGHTED IN BOLD

Modalities	Phys	Ling	Thermal Forehead	Thermal Periorbital	Ling+Phys	Phys+Thrm Forehead	Phys+Thrm Periorbital	Ling+Thrm Forehead	Ling+Thrm Periorbital	Ling+Phys+ Forehead	Ling+Phys+ Periorbital
Test "Abortion", train "Best Friend" + "Mock Crime"											
Deceptive	46.67	26.67	56.67	66.67	**76.67**	50.0	43.33	33.33	73.33	40.0	73.33
Truthful	50.0	**70.0**	56.67	50.0	40.0	30.0	53.33	**70.0**	43.33	56.67	33.33
All Accuracy	48.33	48.33	56.67	**58.33**	**58.33**	40.0	48.33	51.67	**58.33**	48.33	53.33
Test "Best Friend", train "Abortion" + "Mock Crime"											
Deceptive	50.0	46.67	50.0	43.33	20.0	40.0	**53.33**	**53.33**	46.67	13.33	20.0
Truthful	36.67	43.33	70.0	46.67	80.0	70.0	40.0	70.0	73.33	**83.33**	76.67
All Accuracy	43.33	45.0	60.0	45.0	50.0	55.0	46.67	**61.67**	60.0	48.33	48.33
Test "Mock Crime", train "Abortion" + "Best Friend"											
Deceptive	43.75	**87.50**	56.25	68.75	75.0	37.50	75.0	75.0	**87.50**	75.0	**87.50**
Truthful	61.54	7.69	61.54	**69.23**	53.85	61.54	46.15	46.15	23.08	38.46	23.08
All Accuracy	51.72	51.72	58.62	**68.97**	65.52	48.28	62.07	62.07	58.62	58.62	58.62

TABLE VI

THE RECALL AND OVERALL ACCURACY PERCENTAGES FOR INDIVIDUAL AND INTEGRATED MODALITIES WITH THE USAGE OF HSV THERMAL FEATURES FOR CROSS-TOPIC LEARNING SCHEME. TEST "ABORTION" INDICATES THAT THE "ABORTION" INSTANCES ARE TESTED WHILE THE CLASSIFIER IS TRAINED USING INSTANCES FROM "BEST FRIEND" AND "MOCK CRIME" AND SO ON. BEST RESULTS ARE HIGHLIGHTED IN BOLD

Modalities	Phys	Ling	Thermal Forehead	Thermal Periorbital	Ling+Phys	Phys+Thrm Forehead	Phys+Thrm Periorbital	Ling+Thrm Forehead	Ling+Thrm Periorbital	Ling+Phys+ Forehead	Ling+Phys+ Periorbital
Test "Abortion", train "Best Friend" + "Mock Crime"											
Deceptive	46.67	26.67	70.0	56.67	76.67	53.33	46.67	76.67	63.33	**83.33**	43.33
Truthful	50.0	**70.0**	36.67	46.67	40.0	36.67	56.67	30.0	40.0	36.67	50.0
All Accuracy	48.33	48.33	53.33	51.67	58.33	45.0	51.67	53.33	51.67	**60.0**	46.67
Test "Best Friend", train "Abortion" + "Mock Crime"											
Deceptive	50.0	46.67	60.0	50.0	20.0	60.0	50.0	**63.33**	33.33	**63.33**	33.33
Truthful	36.67	43.33	63.33	53.33	**80.0**	63.33	53.33	63.33	56.67	63.33	56.67
All Accuracy	43.33	45.0	61.67	51.67	50.0	61.67	51.67	**63.33**	45.0	**63.33**	45.0
Test "Mock Crime", train "Abortion" + "Best Friend"											
Deceptive	43.75	**87.50**	25.0	68.75	75.0	25.0	81.25	43.75	68.75	43.75	68.75
Truthful	61.54	7.69	76.92	69.23	53.85	76.92	76.92	**92.31**	38.46	84.62	38.46
All Accuracy	51.72	51.72	48.28	68.97	65.52	48.28	**79.31**	65.52	55.17	62.07	55.17

"Best Friend," and "Mock Crime" topics, as well as for "All Topics" combined, using individual and different combinations of the physiological, linguistic, and HSV Forehead thermal modalities, in a leave-one-subject-out cross validation scheme. Note that for the "Mock Crime" topic, the leave-one-instance-out and leave-one-subject-out validation schemes give identical results, given that each subject has only one "Mock Crime" instance.

By comparing the results in this table to the "Abortion" and "Best Friend" results in Table II and "All Topics" HSV results in Table III, it can be seen that that the results follow the same trend. The best overall accuracy and recall are attained

TABLE VII

THE RECALL AND OVERALL ACCURACY PERCENTAGES USING LEAVE-ONE-SUBJECT-OUT CROSS VALIDATION FOR INDIVIDUAL AND INTEGRATED MODALITIES WITH THE USAGE OF HSV THERMAL FORHEAD FEATURES FOR THE THREE INDIVIDUAL TOPICS. BEST RESULTS ARE HIGHLIGHTED IN BOLD

Modalities	Phys	Ling	Thrm	Ling+ Phys	Phys+ Thrm	Ling+ Thrm	Ling+Phys +Thrm
"Abortion"							
Deceptive	60.0	**83.33**	40.0	80.0	40.0	56.67	50.0
Truthful	56.67	**83.33**	53.33	**83.33**	50.0	76.67	73.33
All Accuracy	58.33	**83.33**	46.67	81.76	45.0	66.67	61.67
"Best Friend"							
Deceptive	**66.67**	53.33	53.33	50.0	53.33	56.67	60.0
Truthful	**63.33**	50.0	60.0	56.67	60.0	40.0	43.3
All Accuracy	**65.0**	51.67	56.67	53.33	56.67	48.33	51.67
"Mock Crime"							
Deceptive	50.0	31.25	43.75	56.25	**68.75**	18.75	62.50
Truthful	**69.23**	30.77	23.08	**69.23**	**69.23**	7.69	61.54
All Accuracy	58.62	31.03	34.48	62.07	**68.97**	13.79	62.07
All Topics HSV							
Deceptive	57.90	68.42	51.32	60.53	53.95	**69.74**	68.42
Truthful	47.95	54.80	58.90	54.80	54.80	**67.12**	**67.12**
All Accuracy	53.02	61.75	55.03	57.72	54.36	**68.46**	67.79

by the same modalities or fusion in all the cases except for the recall of the truthful class in "Best Friend." Moreover, the best overall accuracy for "All Topics" is the same, although there are minor differences in the recall figures. Also for "All Topics," the individual and combined modalities exhibit very close figures using both validation schemes with a maximum absolute overall accuracy difference of 4% using the individual thermal forehead modality. However, a difference can be seen in the individual "Abortion" and "Best Friend" results. It can be noticed by comparison to Table II that all the overall accuracy results using leave-one-subject-out have improved. This is expected as mentioned earlier due to the presence of an instance with the opposite label in each fold using leave-one-instance-out. Furthermore, 11 out of 14 accuracy figures are above the random guessing baseline using the leave-one-subject-out cross validation scheme compared to 7 out of 14 earlier.

C. Decision Fusion

In addition to feature fusion, explored in the previous section, we also experimented with decision fusion as a way to combine the various modalities. A classification model is created for each of the three modalities separately, and a final decision is made by combining the individual decisions of the three models using majority voting.

TABLE VIII

PERCENTAGE IMPROVEMENT OF THE DECEPTIVE AND TRUTHFUL CLASSES RECALL, AND THE OVERALL ACCURACY USING DECISION FUSION (DF) OVER USING FEATURE FUSION (FF) FOR ALL THREE MODALITIES. "IMP" DENOTES THE PERCENTAGE IMPROVEMENT

Modalities	Ling+Phys+Forehead			Ling+Phys+Periorbital		
	FF	DF	Imp	FF	DF	Imp
Grayscale						
Deceptive	64.47	63.16	-2.03	53.95	64.47	19.50
Truthful	56.16	58.90	4.88	61.64	52.05	-15.56
All Accuracy	60.4	61.07	1.11	57.72	58.39	1.16
HSV						
Deceptive	67.11	67.11	0.0	31.58	65.79	108.33
Truthful	69.86	52.05	-25.49	45.21	60.27	33.31
All Accuracy	68.46	59.73	-12.75	38.26	63.09	64.90

Table VIII illustrates the percentage improvement of the deceptive and truthfulness recall, and the overall accuracy using decision fusion over those achieved with feature fusion. The table shows that fusing the modalities using the forehead gray scale thermal features achieves a slight improvement in the truthful class recall and the overall accuracy and deterioration in the recall of the deception class. Using the forehead HSV features, the performance drops drastically with decision fusion. However, there is a large improvement in the performance using decision fusion over feature fusion with the periorbital HSV thermal features.

Overall, we cannot draw a final conclusion whether decision fusion is preferred over feature fusion for improved deception detection rates. While in some cases there is a significant improvement, in other cases there is a drastic deterioration in performance.

D. Decision Tree Model

To find the most discriminant features capable of indicating deception using fusion of the forehead HSV thermal modality, linguistic modality, and physiological modality, a decision tree model is created using all 149 instances. Figure 6 displays the constructed tree model. The modality type of the feature utilized for node splitting is shown beside each node in the tree. The tree provides a visualization of which features are used to discriminate between deceptive and truthful instances.

The model selected the root node and its left child from the thermal histogram of the V channel. The right child of the root is selected from the linguistic unigram features. With a total of 14 features used in constructing the tree, eight features are selected from the thermal S and H channels, and six features are selected from the linguistic unigrams and LIWC. The LIWC features build the lowest levels of the tree. To evaluate the performance of these specific features, we ran the classification process again selecting only these 14 features

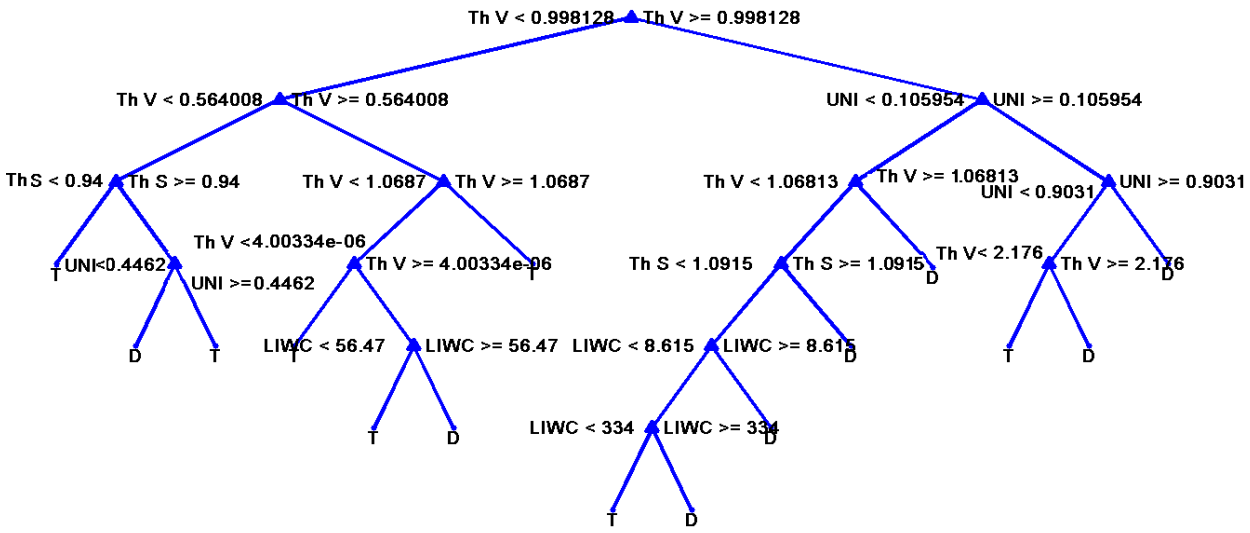


Fig. 6. Decision tree model created using all 149 instances. "D" denotes the deceptive class and "T" denotes the truthful class. The modality type of each feature selected for node splitting is shown beside the node. "Th V" and "Th S" denote the thermal V channel and S channel, respectively. "UNI" and "LIWC" denote the linguistic unigrams and LIWC features, respectively.

per instance using leave-one-instance-out cross validation. The recall was boosted significantly to reach **90.79%** and **87.67%** for the deceptive class and truthful class, respectively. The overall accuracy increased to **89.26%**.

In order to determine the reason for this prominent improvement, we analyze the characteristics of these features. For the thermal S channel, there are two features selected from the histogram bins of medium saturation. All six features from the histogram of the thermal V channel are selected from the last 60 bins which correspond to the re-occurrence of higher valued-pixels representing higher temperatures. The majority of the deceptive leaf nodes in the tree occur when the values of these thermal features increase. Evidently, an increase in the thermal heat emission from the forehead due to increased blood flow exists when a person acts deceptively. The three unigram words used in constructing the tree are "I", "Have", and "Great" starting from the top to the bottom of the tree. Interestingly, it can be noted that an increased self-referencing is an indication of deceit. Additionally, usage of exaggeration words can discriminate between deceptive and truthful behaviors.

VI. Conclusions and Future Work

In this paper, we proposed a novel multimodal deception detection approach that integrated features extracted from physiological, linguistic, and thermal modalities. Our results were promising, especially by fusing the linguistic and thermal modalities, which can pave the way to a non-invasive yet accurate deception detection system. Additionally, we were able to determine which thermal region in the face was most capable of detecting deceit by dividing the face into five segments and analyzing them. While previous work mostly focused and analyzed the periorbital area for this purpose, we demonstrated that extracting features from the forehead could be a better indicator of deception. This could be partially due to the effect of the hair areas found in the eyebrows and eyelashes among other factors.

The linguistic features and in particular the Unigrams and LIWC played a critical role in discriminating between deceptive and truthful responses. The physiological features were effective in some cases, however, in other cases they did not contribute in realizing an improved performance. It can be concluded that following a multimodal approach by integrating features from different modalities outperformed relying solely on single modalities reaching an overall accuracy of approximately 70%. A problem existed with the usage of linguistic modalities when tested on a domain not used for training as elucidated in our cross-topic learning scenarios. A variety of domains are essential for extracting effective linguistic features. Other modalities, particularly the thermal, benefited from the cross-topic learning scheme with the increase of the size of the training set.

Visualization of the tree model constructed from our data showed that specific thermal and linguistic features were prominent indicators of deceit. In particular, the distribution of higher thermal temperatures in the forehead along with an increased usage of self-referencing and exaggeration words were able to accurately discriminate between deception and truthfulness. We are currently in the process of collecting more data, which we expect will lead to further improvements in deception detection rates.

Acknowledgments

The authors would like to thank the reviewers for their thoughtful suggestions.

References

[1] B. M. Depaulo *et al.*, "Cues to deception," *Psychol. Bull.*, vol. 129, no. 1, pp. 74–118, 2003.
[2] M. Derksen, "Control and resistance in the psychology of lying," *Theory Psychol.*, vol. 22, no. 2, pp. 196–212, 2012.

[3] T. A. Gannon, A. R. Beech, and T. Ward, *Risk Assessment and the Polygraph*. Hoboken, NJ, USA: Wiley, 2009, pp. 129–154.

[4] C. F. Bond and B. M. DePaulo, "Accuracy of deception judgments," *Personality Social Psychol. Rev.*, vol. 10, no. 3, pp. 214–234, 2006.

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

[6] A. Vrij, *Detecting Lies and Deceit: The Psychology of Lying and the Implications for Professional Practice* (Wiley Series in the Psychology of Crime, Policing and Law). Hoboken, NJ, USA: Wiley, 2001.

[7] B. Verschuere, V. Prati, and J. De Houwer, "Cheating the lie-detector: Faking in the autobiographical implicit association test," *Psychol. Sci.*, vol. 20, no. 4, pp. 410–413, 2009.

[8] G. Maschke and G. Scalabrini. (2005). *The Lie Behind the Lie Detector*. [Online]. Available: http://antipolygraph.org

[9] F. A. Kozel *et al.*, "A pilot study of functional magnetic resonance imaging brain correlates of deception in healthy young men," *J. Neuropsychiatry Clin. Neurosci.*, vol. 16, no. 3, pp. 295–305, Aug. 2004.

[10] P. Ekman, *Telling Lies: Clues to Deceit in the Marketplace, Politics and Marriage*. New York, NY, USA: Norton, 2001.

[11] M. Owayjan, A. Kashour, N. Al Haddad, M. Fadel, and G. Al Souki, "The design and development of a lie detection system using facial micro-expressions," in *Proc. 2nd Int. Adv. Comput. Tools Eng. Appl. (ACTEA)*, Dec. 2012, pp. 33–38.

[12] T. Pfister and M. Pietikäinen, *Electronic Imaging Signal Processing Automatic Identification of Facial Clues to Lies*. Bellingham, WA, USA: SPIE, Jan. 2012.

[13] J. Hillman, A. Vrij, and S. Mann, "Um . . . they were wearing . . .: The effect of deception on specific hand gestures," *Legal Criminol. Psychol.*, vol. 17, no. 2, pp. 336–345, 2012.

[14] F. Maricchiolo, A. Gnisci, and M. Bonaiuto, "Coding hand gestures: A reliable taxonomy and a multi-media support," in *Cognitive Behavioural Systems Series Lecture Notes in Computer Science*, vol. 7403, A. Esposito, A. Esposito, A. Vinciarelli, R. Hoffmann, and V. Müller, Eds. Berlin, Germany: Springer, vol. 7403, 2012, pp. 405–416.

[15] D. Howard and C. Kirchhübel, "Acoustic correlates of deceptive speech—An exploratory study," in *Engineering Psychology and Cognitive Ergonomics* (Series Lecture Notes in Computer Science). Berlin, Germany: Springer, 2011, pp. 28–37.

[16] A. Vrij, P. Granhag, and S. Porter, "Pitfalls and opportunities in nonverbal and verbal lie detection," *Psychol. Sci. Public Interest*, vol. 11, no. 3, pp. 89–121, Dec. 2010.

[17] J. Hirschberg *et al.*, "Distinguishing deceptive from non-deceptive speech," in *Proc. Interspeech Eurospeech*, 2005, pp. 1833–1836.

[18] T. Qin, J. K. Burgoon, J. P. Blair, and J. F. Nunamaker, "Modality effects in deception detection and applications in automatic-deception-detection," in *Proc. 38th Hawaii Int. Conf. Syst. Sci.*, 2005, p. 23b.

[19] R. Mihalcea and C. Strapparava, "The lie detector: Explorations in the automatic recognition of deceptive language," in *Proc. ACL-IJCNLP Conf. Assoc. Comput. Linguistics*, Aug. 2009, pp. 309–312.

[20] T. Fornaciari and M. Poesio, "On the use of homogenous sets of subjects in deceptive language analysis," in *Proc. Workshop Comput. Approaches Deception Detection*, Stroudsburg, PA, USA, 2012, pp. 39–47.

[21] S. Feng, R. Banerjee, and Y. Choi, "Syntactic stylometry for deception detection," in *Proc. 50th Annu. Meeting Assoc. Comput. Linguistics*, 2012, pp. 171–175.

[22] I. T. Pavlidis, "Lie detection using thermal imaging," *Proc. SPIE*, vol. 5405, Apr. 2004, pp. 270–279.

[23] M. Garbey, A. Merla, and I. Pavlidis, "Estimation of blood flow speed and vessel location from thermal video," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 1. Jun. 2004, pp. I-356–I-363.

[24] L. Warmelink, A. Vrij, S. Mann, S. Leal, D. Forrester, and R. P. Fisher, "Thermal imaging as a lie detection tool at airports," *Law and Human Behavior*, vol. 35, no. 1, pp. 40–48, 2011.

[25] I. Pavlidis and J. Levine, "Monitoring of periorbital blood flow rate through thermal image analysis and its application to polygraph testing," in *Proc. 23rd Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, vol. 3. Oct. 2001, pp. 2826–2829.

[26] I. Pavlidis and J. Levine, "Thermal image analysis for polygraph testing," *IEEE Eng. Med. Biol. Mag.*, vol. 21, no. 6, pp. 56–64, Nov. 2002.

[27] Y. Zhou, P. Tsiamyrtzis, P. Lindner, I. Timofeyev, and I. Pavlidis, "Spatiotemporal smoothing as a basis for facial tissue tracking in thermal imaging," *IEEE Trans. Biomed. Eng.*, vol. 60, no. 5, pp. 1280–1289, May 2013.

[28] B. Rajoub and R. Zwiggelaar, "Thermal facial analysis for deception detection," *IEEE Trans. Inf. Forensics Security*, vol. 9, no. 6, pp. 1015–1023, Apr. 2014.

[29] U. Jain, B. Tan, and Q. Li, "Concealed knowledge identification using facial thermal imaging," in *Proc. IEEE Int. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2012, pp. 1677–1680.

[30] M. Abouelenien, V. Pérez-Rosas, R. Mihalcea, and M. Burzo, "Deception detection using a multimodal approach," in *Proc. 16th ACM Int. Conf. Multimodal Interact. (ICMI)*, Nov. 2014, pp. 58–65.

[31] V. Pérez-Rosas, M. Abouelenien, R. Mihalcea, Y. Xiao, C. Linton, and M. Burzo, "Verbal and nonverbal clues for real-life deception detection," in *Empirical Methods Natural Language Processing*. Lisbon, Portugal: Association for Computational Linguistics, Sep. 2015, pp. 2336–2346.

[32] V. Pérez-Rosas, M. Abouelenien, R. Mihalcea, and M. Burzo, "Deception detection using real-life trial data," in *Proc. 17th ACM Int. Conf. Multimodal Interact. (ICMI)*, Nov. 2015, pp. 59–66.

[33] M. L. Jensen, T. O. Meservy, J. K. Burgoon, and J. Nunamaker, "Automatic, multimodal evaluation of human interaction," *Group Decision Negotiation*, vol. 19, no. 4, pp. 367–389, 2010.

[34] J. F. Nunamaker, Jr., J. K. Burgoon, N. W. Twyman, J. G. Proudfoot, R. Schuetzler, and J. S. Giboney, "Establishing a foundation for automated human credibility screening," in *Proc. IEEE Int. Conf. Intell. Secur. Inform. (ISI)*, Jun. 2012, pp. 202–211.

[35] J. K. Burgoon *et al.*, "Detecting concealment of intent in transportation screening: A proof of concept," *IEEE Trans. Intell. Transp. Syst.*, vol. 10, no. 1, pp. 103–112, Mar. 2009.

[36] J. Pennebaker and M. Francis, *Linguistic Inquiry and Word Count: LIWC*. Mahwah, NJ, USA: Erlbaum Publishers, 1999.

[37] Y. R. Tausczik and J. W. Pennebaker, "The psychological meaning of words: LIWC and computerized text analysis methods," *J. Language Social Psychol.*, vol. 29, no. 1, pp. 24–54, 2010.

[38] X. Lu, "Automatic analysis of syntactic complexity in second language writing," *Int. J. Corpus Linguistics*, vol. 15, no. 4, pp. 474–496, 2010.

[39] K. W. Hunt, "Early blooming and late blooming syntactic structures," in *Evaluating Writing: Describing, Measuring, Judging*, C. R. Cooper and L. Odell, Eds. 1977, pp. 91–106.

[40] J. Shi and C. Tomasi, "Good features to track," in *Proc. IEEE CVPR*, Jun. 1994, pp. 593–600.

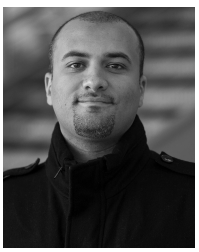
[41] S. N. Sinha, J.-M. Frahm, M. Pollefeys, and Y. Genc, "GPU-based video feature tracking and matching," Univ. North Carolina Chapel Hill, Chapel Hill, NC, USA, Tech. Rep. TR 06-012, 2006.

[42] Z. Kalal, K. Mikolajczyk, and J. Matas, "Forward-backward error: Automatic detection of tracking failures," in *Proc. 20th Int. Conf. Pattern Recognit. (ICPR)*, Aug. 2010, pp. 2756–2759.

[43] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision* (Series Cambridge Books Online). Cambridge, MA, USA: Cambridge Univ. Press, 2003.

[44] T. Qin, J. Burgoon, and J. F. Nunamaker, "An exploratory study on promising cues in deception detection and application of decision tree," in *Proc. 37th Annu. Hawaii Int. Conf. Syst. Sci.*, Jan. 2004, pp. 23–32.

[45] A. Lacoste, F. Laviolette, and M. Marchand, "Bayesian comparison of machine learning algorithms on single and multiple datasets," in *Proc. 15th Int. Conf. Artif. Intell. Statist. (AISTATS)*, vol. 22. 2012, pp. 665–675.



Mohamed Abouelenien received the Ph.D. degree in computer science and engineering from the University of North Texas. He is currently a Post-Doctoral Research Fellow with the Electrical Engineering and Computer Science Department, University of Michigan, Ann Arbor. He has published in several top venues including IEEE, ACM, Springer, and SPIE. His areas of interest include multimodal deception detection, the 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. He served as a Program Committee Member for multiple international conferences. He also served as a Reviewer of IEEE TRANSACTIONS and Elsevier journals.

Verónica Pérez-Rosas received the Ph.D. degree in computer science and engineering from the University of North Texas in 2014. She is currently a Post-Doctoral Research Fellow with the University of Michigan. She has authored papers in leading conferences and journals in natural language processing and computational linguistics. Her research interests include machine learning, natural language processing, computational linguistics, affect recognition, the multimodal analysis of human behavior, developing computational methods to analyze, recognize, and predict human affective responses during social interactions. She served as a Program Committee Member for multiple international conferences in the same fields. She also has served as a reviewer for IEEE INTELLIGENT SYSTEMS, Springer, and *PlosOne* journals.

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

Mihai Burzo is an Assistant Professor of Mechanical Engineering with the University of Michigan–Flint. He has published over 50 articles in peer reviewed journals and conference proceedings. His research interests include heat transfer in microelectronics and nanostructures, the thermal properties of thin films of new and existing materials, the multimodal sensing of human behavior, and the computational modeling of forced and natural heat convection. He is a 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 2013 and 2006, the Young Engineer of the Year from the North Texas Section of ASME in 2006, the Leadership Award from SMU in 2002, and a Valedictorian Award in 1995.