Multimodal Gender Detection

Mohamed Abouelenien
Computer and Information Science
University of Michigan, Dearborn
Dearborn, MI, USA 48128
zmohamed@umich.edu

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

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

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

ABSTRACT

Automatic gender classification is receiving increasing attention in the computer interaction community as the need for personalized, reliable, and ethical systems arises. To date, most gender classification systems have been evaluated on textual and audiovisual sources. This work explores the possibility of enhancing such systems with physiological cues obtained from thermography and physiological sensor readings. Using a multimodal dataset consisting of audiovisual, thermal, and physiological recordings of males and females, we extract features from five different modalities, namely acoustic, linguistic, visual, thermal, and physiological. We then conduct a set of experiments where we explore the gender prediction task using single and combined modalities. Experimental results suggest that physiological and thermal information can be used to recognize gender at reasonable accuracy levels, which are comparable to the accuracy of current gender prediction systems. Furthermore, we show that the use of non-contact physiological measurements, such as thermography readings, can enhance current systems that are based on audio or visual input. This can be particularly useful for scenarios where non-contact approaches are preferred, i.e., when data is captured under noisy audiovisual conditions or when video or speech data are not available due to ethical considerations.

CCS CONCEPTS

•Computing methodologies →Artificial intelligence;

KEYWORDS

gender detection; multimodal; thermal; physiological; linguistic; visual; vocal

ACM Reference format:
Mohamed Abouelenien, Verónica Pérez-Rosas, Rada Mihalcea, and Mihai Burzo. 2017. Multimodal Gender Detection. In *Proceedings of ICMI '17, Glasgow, UK, November 13–17, 2017,* 10 pages.
DOI: 10.1145/3136755.3136770

1 INTRODUCTION

The task of automatically identifying gender has gained a lot of attention recently due to ethical and security concerns. In particular, gender detection has a wide variety of applications, including human-computer interaction, surveillance purposes, computer forensics, statistical analysis for large scale text applications, collection of users' demographics, narrowing down database queries, and assessing consumer behavior.

Recent research on human-computer interaction and systems has shown that multimodal approaches can assist in addressing challenges such as noisy data and non-universality, which are frequently associated to the use of single modalities. This can be achieved by combining multiple data sources that might be available for a specific application and can potentially provide more robust or complimentary information.

An example of the challenges that deteriorate the performance when using single modalities can be seen in the visual data stream. For instance, disguise, wearing hats, sunglasses, fake beards and wigs, and accessories are among some of the countermeasures that can significantly affect the classification results during gender recognition from video data. Furthermore, the performance of visual recognition is highly dependent on the illumination conditions, time of day, shadows, and weather conditions. Another example is the unavailability of the verbal responses or their presence with noise and/or background conversations.

In contrast, modalities such as thermal imaging can bring important advantages over the visual data stream as it is robust against illumination conditions, disguised faces, and changes in pose. In addition, the thermal spectrum can capture anatomical information of the human face (unique to each individual), and can be used with or without the cooperation of human subjects.

In this paper, we explore and compare the potential benefits of using alternative modalities to detect gender in addition to the three most commonly used modalities, namely vision, acoustics, and language. In particular, we explore thermography by itself and in combination with audiovisual sources to detect gender. In

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
ICMI '17, Glasgow, UK
© 2017 ACM. 978-1-4503-5543-8/17/11…$15.00
DOI: 10.1145/3136755.3136770

addition, we also experiment with physiological data extracted from contact-based sensors to provide a reference value for the thermal readings.

The paper makes three main contributions. First, we collect a dataset of different responses from 51 males and 53 females using different devices. Second, we analyze the capability of five different modalities in detecting gender, namely the visual, linguistic, physiological, thermal, and acoustic modalities. Third, we analyze the performance of integrating features from different modalities assuming the availability of specific modalities for different applications.

The proposed approach can be particularly useful in a number of settings, including 1) applications where multiple modalities are present; 2) applications using real-world data that is noisy and poses challenges such as low quality audiovisual recordings, occluded faces, distances and angles of view, among others; 3) applications where privacy concerns restrict the access to a particular data stream, e.g., visual and audio information.

2 RELATED WORK

This section surveys previous work on gender detection, which has focused mainly on the visual, linguistic, acoustic and physiological modalities.

Different types of features and representations were extracted from the visual faces to identify gender. Global and local face representation approaches were compared using grey levels, principal component analysis, and local binary patterns for neutral, expressive and partially occluded faces to separate gender [4]. High accuracy rates were reported in [45] using the Webers local texture descriptor to detect gender from face images using the FERRET database [33]. [5] emphasized the effect of the dependencies among gender, age and pose facial attributes in the performance of gender classification systems. The significance of different facial regions in detecting gender was reported in [24], where fusion of multiple facial regions was utilized. [44] used laser scanning to obtain 3D human body shapes to show its effectiveness in detecting gender over using 2-D images or videos.

Most of the introduced methods focused on features derived from faces in constrained environments. More recently, the interest started to grow in detecting gender from faces in the wild and in unconstrained environments. Local binary patterns were used to describe faces for gender detection in unconstrained conditions using the Labeled Faces in the Wild dataset [17], where an improved performance was reported using Adaboost and support vector machine [38]. Deep-convolutional neural networks were recently used to improve gender detection from visual recordings in real-life conditions [18].

Gender detection has been extensively studied in textual sources to aid tasks such as authorship attribution, emotion recognition and deception detection [1, 2, 11, 19]. [28] analyzed several word categories related to linguistic, psychological, and cognitive processes in 14,00 text samples from 70 different studies and found important differences between male and female language. Other works analyzed lexical, discourse and syntactical features to automatically categorize written text by author gender [9, 13, 37, 47]. Topic differences in discourse due to gender were analyzed in [36]. Additionally, lexical differences in word usage between genders during telephone conversations were analyzed in [7].

The performance of specific acoustic features was analyzed to determine their capabilities of identifying gender. [42] analyzed the effect of the interaction of Glottal-pulse rate and vocal-tract length on identifying gender and reported an improvement using vocal-tract length in some cases. The phonetic differences between male and female speech were studied in [40]. The relation between pitch and speaker's sex was studied in [6]. Spectral and pitch features were used to identify gender using short speech from multilingual speakers [20].

Physical or behavioral features referred to as soft biometrics were surveyed in [34] to develop human and gender recognition systems. An overview of gender classification methods can be found in [26]. The study also analyzed the effect of face alignment on gender classification and reported an improved performance using support vector machines.

Fingerprints features as well as the temporal representations were added in order to improve the gender detection rates. Fingerprints were also used to specify gender especially for forensic purposes. Fingerprint ridge density was used to determine age and gender differences [43]. Bag-of-visual-words model was implemented to combine facial and fingerprint features in a bimodal gender recognition framework [22]. Bayesian hierarchical model was used to fuse fingerprint and face image representation to achieve better gender detection rates. Texture-based spatiotemporal representations were also used to describe and analyze faces for gender recognition by combining facial appearance and facial motion features [15].

Studies analyzed gender differences while responding to emotional stimuli using physiological signals such as event related potentials [48] and skin conductance responses[23]. EEG signals were also analyzed and used for age and gender classification [30]. However, to our knowledge there has not been an extensive analysis of gender differences using different physiological signals, and in particular the combination of the four measurements proposed in this paper.

Other types of visual recordings were explored in order to detect gender such as near-infrared images. Local binary pattern histograms were extracted from visual and near-infrared near frontal images to detect gender resulting in reasonable classification rates [10]. A recent attempt of using thermal imaging to improve gender detection can be found in [29]. The method used a combination of visual and thermal images of different body parts to detect gender in order to avoid the limitation of relying on visual images.

3 DATASET

For our experiments, we collected a multimodal dataset consisting of audiovisual, thermal and physiological recordings from 104 subjects. The dataset gender distribution is 51 males and 53 females. The dataset consists of 520 recordings, with five recordings per subject, one of which is a recording from a one-on-one interview with the subject while the remaining four are recordings of the subject providing his/her opinions towards a controversial topic.

The recordings have an average duration of 82s, with a standard deviation of 38s. The visual recordings consist of subject's frontal view where the upper body and hands are visible. The thermal recordings also contain the subject's frontal view but are focused on the face area. The physiological recordings, obtained from five different biosensors that were attached to the subject's hands and thoracic area, contain measurements from subject's blood volume pulse, skin conductance, skin temperature, and respiration rate.

3.1 Preprocessing

Before conducting our experiments, we preprocessed the five different modalities included in the dataset to enable our multimodal experiments.

First, starting with the audiovisual recordings, we extracted the audio stream for each subject recording. Since the thermal cameras can produce audio interference –due to mechanical noise– we applied speech enhancement methods to remove noise and improve the speech signal quality. We started by converting the audio signal from a stereo to a mono channel and to a uniform sample rate of 16k. We then applied the Mean Square Error estimation of spectral amplitude for audio denoising, as implemented in the Voicebox Speech Processing toolbox [8].

Then, we transcribed the subjects' statements via crowdsourcing with Amazon Mechanical Turk. The workers transcribed the subjects' statements using the audio recordings only. When recordings include interviews, we asked workers to transcribe the subject's speech only thus these transcripts do not contain the interviewer questions. The transcripts include word repetitions, word fillers, and long pauses. All transcriptions were manually verified to ensure quality. The final transcript set contains 64,016 words, with an average of 125 words per transcript.

To allow for the audio analysis of the recordings containing interviews, we processed the speech signal to isolate the subject's speech. We first applied automatic speech segmentation and clustering using the LIUM diarization toolkit [35]. Next, we used the resulting clustering to identify the speech segments belonging to the subject's speech. Finally, we proceeded to split the original recording using the automatic diarization output to obtain the speech segments corresponding to the participant only.

4 PROPOSED APPROACH

4.1 Verbal Cues

We start by extracting several linguistic features to explore language differences among genders. The features were derived from the transcripts of the subjects' statements. For the recordings containing subject's opinions on controversial topics we use the full transcript, whereas for the one-on-one interview we remove the interviewer questions and concatenate the interviewee responses into a single chunk of text. The features are as follows:

Unigrams: We extracted unigrams derived from the bag of words representation of each transcript. Each feature consists of frequency counts of unique words in the transcript.

LIWC derived features: We used features derived from the Linguistic Inquire Word Count (LIWC) lexicon. These features consisted of word counts for each of the 80 semantic classes in the LIWC lexicon [31]. For instance, the class "I" includes words associated with the self (e.g., I, me, myself); "Other" includes words associated with others (e.g., he, she, they); etc.

Syntactic complexity and readability features: This set of features consisted of fourteen indexes of sentence syntax complexity, including mean length of sentences, clauses, dependent clauses, and t-units,[1] as well as statistical descriptors of them. To extract these features, we used a tool provided by Lu et al. [25].

Shallow and deep syntax: We extracted a set of features derived from Part of Speech Tags (POS) and production rules based on context free grammars (CFG) trees using the Berkeley parser [32]. The CFG derived features consisted of all the lexicalized production rules (rules including child nodes) combined with their parent and grandparent node, e.g., *NNˆNP→friendship (in this example NN –a noun– is the grandparent node, NP –personal pronoun– the parent node, and "friendship" the child node). Features in this set were also encoded as frequency values.

Response Length Features: We designed 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 over the recording duration. Finally, we counted the number of words in the utterances spoken during each interval, which resulted 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.

For additional insight into language differences in gender, Table 1 presents the top ranked semantic LIWC classes associated with each gender, using the semantic word class scoring from [27]. In this table, we observe clear differences in word choices between genders. The top three word classes for males include sports, music and money whereas for females sleep, groom and inhibition are the most dominant word classes. These results are in line with previous studies on gender language differences [28].

Table 1: Results from LIWC word class analysis. Top ranked semantic classes associated to male and female subjects are shown.

Male		Female	
Class	Score	Class	Score
Sports	2.18	Sleep	1.83
Music	1.58	Groom	1.83
Money	1.51	Inhibition	1.81
TV	1.42	Anxiety	1.63
Job	1.29	Sad	1.55
Family	1.27	Similes	1.43
Sexual	1.25	You	1.38
Body	1.21	Home	1.34
School	1.21	Eating	1.31
Anger	1.20	Positive Feeling	1.29
Article	1.18	Certain	1.20
Physical	1.17	Inclusive	1.20

4.2 Vocal Cues

To incorporate vocal behavior into the analysis, we extracted a set of prosodic features to capture speech patterns from both genders.

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

We extracted these features using OpenEar [12]. We used a predefined feature set, EmoBase, which consists of a set of 988 prosodic features frequently used for emotion recognition tasks. The features are derived from 25 low-level speech descriptors including intensity, loudness, 12 Mel-frequency cepstral coefficients (MFCC), pitch (F0), probability of voicing, F0 envelope, zero-crossing rate, and 8 line spectral frequencies. The feature extraction was conducted at audio-frame level every 10ms with a 25ms Hamming window. For the recordings containing interviews, we first extracted the acoustic features on the subject's speech segments only (obtained as described in section 3.1) and then we averaged each feature over the different segments.

4.3 Visual Cues

To specify visual differences between males and females, we extracted two type of visual features, facial and hand gestures, and global facial representation.

4.3.1 Facial and Hand Gestures. In order to incorporate gesture information into the analysis, we annotated the subjects' facial displays as well as head and hand movements using the MUMIN coding scheme [3]. To conduct the annotation, we first split each visual recording into video segments of 20 seconds length and then we annotated each segment using Amazon Mechanical Turk. We requested four different hits for each video clip to annotate gestures for the head and general face; mouth; eyes; and hands. The annotation was done on silent video so that the annotators could focus on identifying the predominant gesture over each segment.

Each segment was annotated by three independent annotators. To ensure the quality of the annotations and to avoid spam, we checked that the annotators correctly responded to instructions in a control video showed randomly during the annotation. We rejected annotations from workers who failed the quality control more than twice. After the annotations were completed, we assigned gesture labels at segment and video level using majority voting over the labels provided by the three annotators. For those cases where a majority could not be identified, the final label was randomly selected among the three available annotations.

To identify differences in gesture behavior among subjects, we analyzed the percentage of occurrences of each gesture clustered by gender. Figure 1 presents the box plots corresponding to the gestures for which we observed larger variations between genders, particularly in facial expressions such as smile, scowl, and laugh, as well as eyebrow, eyes, and lips gestures. This analysis suggests that the proposed gestures are potentially useful predictors of the subject's gender.

Next, from the gesture annotations at video level, we derived 40 binary features that indicated whether the subjects elicited the given gesture during the full recording. We opted for a binary representation due to differences in recording lengths.

In addition, we extracted temporal variations in visual behavior, by computing the number of times a gesture change occurred in the series of video segments throughout the subject's response. We computed this feature for each of the nine gesture categories in the MUMIN scheme.

The final set of visual features includes 40 binary features of the facial and hand gestures for each response, as well as a set of nine features representing the dynamics of the gestures in each response.

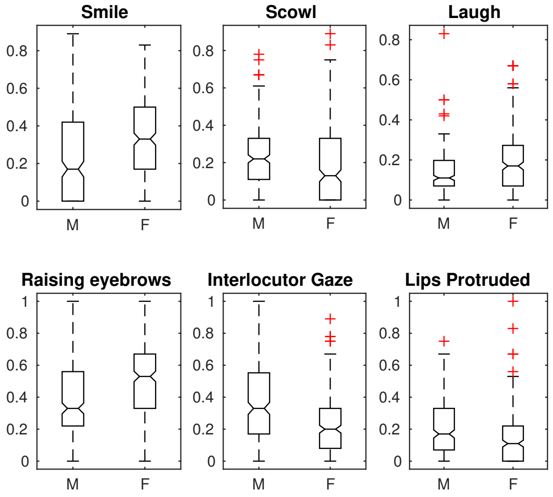

Figure 1: Differences in facial displays by males (M) and females (F)

4.3.2 Global Facial Representation. We use the Eigenfaces approach [39] to derive global face features that can be used to model facial differences between genders. The features were extracted using the Principal Component Analysis (PCA) technique. To extract the Eigenfaces, each individual face image is projected into a lower dimensional space and was expressed as a weighted summation of the Eigenface vectors. To recognize a new unlabeled face image, the image is projected into the new PCA space.

Hence, we randomly sampled images from the male and female video recordings. The faces in the images were automatically detected using the Viola-Jones algorithm [46]. The training set of gender-based faces was used to create the new PCA space and the faces to be tested were projected into the new space. It should be noted that we are aware of other recent methods that utilized deep learning to predict gender from faces as listed earlier; however, these approaches provide pretrained models, which cannot be used in our case for a fair comparison with other modalities. Moreover, given the size of our collected dataset, deep learning might not perform effectively.

4.4 Physiological Sensors

The physiological recordings were processed with the Biograph Infiniti Physiology[2] suite to obtain four raw physiological measurements from the subject's blood volume pulse, skin conductance, skin temperature, and respiration rate, as well as statistical measurements derived from them.[3] The statistical descriptors include the maximum and minimum values, means, power means, standard deviations, and mean amplitudes (epochs). In addition, we obtained

[2] http://thoughttechnology.com/index.php/software/physiology-suite-sa7970.html
[3] All measurements were obtained at a sampling rate of 2048 samples per second.

features derived from inter-beat intervals (IBI) measurements such as the minimum and maximum amplitudes and their intervals. The final feature set consists of 59 physiological features including 40 BVP features, five SC features, seven RR features, five ST features, and two features extracted from the BVP and the RR sensors combined, namely, the mean and heart rate max-min difference, which is a measure of breath to heart rate variability.

Figure 2 shows the box plots of each sensor raw output clustered by gender. The figure suggests noticeable differences of physiological responses between males and females, particularly, in electrodermal response (skin conductance) and respiration rate.

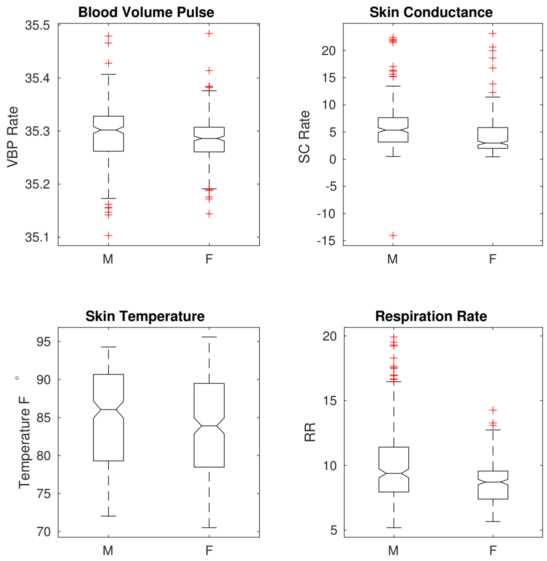

Figure 2: Average raw values of physiological measurements when clustered by gender (males M, females F).

4.5 Thermal Imaging

Thermal features were extracted in order to determine whether certain thermal patterns vary between males and females. We derived feature vectors that represent the thermal signature of the subject's face over different regions of interest (ROI) as described below.

Segmenting and Tracking the ROI. We started by manually locating five ROIs from the first frame of each recorded thermal video by specifying their bounding boxes. The ROIs included the whole face; forehead; periorbital (eyes); cheeks including the nose; and nose. Interesting points were then detected in this ROI using Shi-Tomasi corner detection algorithm. These points were located where sharper changes in temperatures existed. Following this, the detected points were tracked through the entire response using a fast Kanade-Lucas-Tomasi (KLT) tracking algorithm [41]. The points tracking was performed by estimating the displacement between successive frames.

Following the tracking process and displacement estimation, geometric transformation [16] was applied, which globally estimated the interesting points transformation based on similarity in order to map the interesting points between the frames. Once the points were mapped, the new boundary box was geometrically specified. The maximum distance we allowed between the tracked point and its projection on the next frame was five pixels. Moreover, if the number of points matched between two successive frames is less than 95%, a chance of occurrence of occlusion was considered. Hence, we discarded the tracking of the current frame and proceeded to the next one.

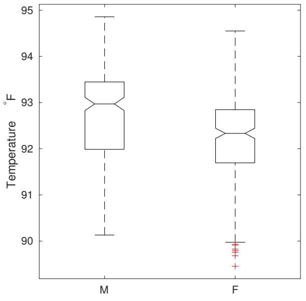

Figure 3: Difference in the average thermal temperatures extracted from the faces of the males (M) and females (F).

Thermal Feature Extraction. The locations of the bounding boxes containing the ROI of each frame were cropped from the raw thermal video, and a thermal map was created to represent the heat distribution in each ROI. In particular, we extracted statistical measurements on the full response-level 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 of each ROI. Moreover, we extracted a histogram of 20 bins of the non-zero temperatures in each ROI. Furthermore, we extracted temporal features by dividing each response into five equal stages and computing the statistical features from each stage. The features extracted from the males' and females' responses in addition to the five temporal thermal features presented the final feature set for each ROI.

Figure 3 shows the box plots of the mean thermal temperatures extracted from the whole faces of the male and female subjects. The figure clearly indicates that males' facial temperatures are higher compared to females, on average. As the notch marks clearly do not overlap, the gender-based thermal differences are significant at the 95% confidence interval.

5 RESULTS AND DISCUSSION

5.1 Experimental Setup

The experiments are conducted using feature sets from individual modalities as well as integrated modalities. A decision tree classifier is used to predict binary gender using a leave-one-subject-out cross validation scheme in order to avoid any bias. In this scheme, the five instances belonging to each subject are reserved for testing and all other instances are used for training during each fold. We report the overall accuracy as well as the recall of the male and female classes. Moreover, we conduct feature analysis by visualizing the decision tree model developed for the best performing set of features and determining the specific features that are most capable of detecting gender.

5.2 Individual Modalities

Table 2: Overall accuracy and class recall using the five thermal regions of interest including the face, forehead, periorbital (eyes), cheeks, and nose. The best results are highlighted in bold.

Metric	Accuracy	Male Recall	Female Recall
Baseline	51.0	49.0	51.0
Face	**70.2**	**71.0**	**69.4**
Forehead	45.2	44.7	45.7
Eyes	51.9	53.3	50.6
Cheeks	69.6	70.6	68.7
Nose	60.4	61.2	59.6

Table 2 lists the overall accuracy and class recall using the five thermal ROIs including the face, forehead, periorbital (eyes), cheeks, and nose. The table indicates that the thermal measurements from the whole facial area are the most capable of indicating gender. Interestingly, the cheeks region achieves very close performance and reaches approximately 70% accuracy, indicating the capability of this region to differentiate between males and females. On the other hand, the forehead and eyes regions are not capable of identifying gender and provide performance similar to that of random guessing. It can be noted that there exists a slight improvement in the males' recall compared to the females' recall; however the difference is not significant. Table 3 shows the overall accuracy and class recall using the physiological signals including the blood volume pulse, skin conductance, skin temperature, respiration rate, and "All" signals combined. The table indicates that the blood volume pulse and skin conductance signals achieve the best performance among the individual signals. The accuracy of all the individual signals stands below 60%. However, when all the physiological features are integrated, the accuracy is boosted to close to 70% accuracy, indicating that the specific combination of different signals can be indicative of gender.

Table 3: Overall accuracy and class recall using the physiological signals including the blood volume pulse, skin conductance, skin temperature, respiration rate, and "All" signals combined. The best results are highlighted in bold.

Metric	Accuracy	Male Recall	Female Recall
Baseline	51.0	49.0	51.0
Blood Volume	57.5	58.4	56.6
Skin Conductance	58.3	58.4	58.1
Skin Temperature	50.8	45.5	55.8
Respiration Rate	52.9	53.3	52.5
All	**68.7**	**67.1**	**70.2**

Table 4: Overall accuracy and class recall using linguistic features including the unigrams (Unigrams), language inquiry and word count (LIWC), readability scores (Readability), context free grammars (CFG), part of speech tags (POS), and "All" linguistic features combined. The best results are highlighted in bold.

Metric	Accuracy	Male Recall	Female Recall
Baseline	51.0	49.0	51.0
Unigrams	57.3	55.3	59.2
LIWC	54.4	47.1	**61.5**
Readability	55.8	54.1	57.4
CFG	57.9	57.3	58.5
POS	55.0	56.1	54.0
All	**58.7**	**58.8**	58.5

Table 4 lists the overall accuracy and class recall using linguistic features including the unigrams (UNI), language inquiry and word count (LIWC), readability scores (Read.), context free grammars (CFG), part of speech tags (POS), and "All" linguistic features combined.

Context free grammars and unigrams have better capability of determining gender. Moreover, different linguistic features provide performance that is above the baseline of random guessing. We experimented with several combinations and noticed that the integration of unigrams and LIWC achieve higher performance. However, for fair comparison and due to space limit, we provide the results for combining all linguistic features together, which exceed the accuracy of all individual linguistic sets. The males and females recall are very close in the majority of the cases.

Table 5: Overall accuracy and class recall using the visual features including the face gestures, hand gestures, both gestures, Eigenfaces, and all features combined. The best results are highlighted in bold.

Metric	Accuracy	Male Recall	Female Recall
Baseline	51.0	49.0	51.0
Face Gestures	57.3	56.5	58.1
Hand Gestures	48.1	50.6	45.7
Both Gestures	53.3	52.2	54.3
Eigenfaces	**71.7**	**69.4**	**74.0**
All	**71.7**	**69.4**	**74.0**

Table 5 lists the overall accuracy and class recall using the visual features including the face gestures, hand gestures, both gestures, Eigenfaces, and all features combined. The table shows that the Eigenfaces method provides a significantly improved performance compared to other gestures. This indicates that global facial features have higher capability of indicating gender compared to facial gestures and other body movements. Facial gestures provide performance that is above the baseline, however, the hand gestures are not capable of detecting gender. Hence, the combination of both facial and hand gestures does not achieve improved performance. The integration of all visual features achieves identical performance to the Eigenfaces approach, which indicates that the decision tree model only utilized the Eigenfaces features. The females recall is slightly better than the males recall using facial-related features.

Table 6: Overall accuracy and class recall using the vocal features including the MFCC and EmoBase features. The best results are highlighted in bold.

Metric	Accuracy	Male Recall	Female Recall
Baseline	51.0	49.0	51.0
MFCC	**80.4**	**80.8**	80
EmoBase	79.2	77.3	**81.1**

For the acoustic modality we experimented with the Emobase set as well as a subset containing the MFCC features only. Table 6 illustrates the overall accuracy and recall figures when using these features. The table shows that the acoustic features have superior performance compared to other features from other modalities, exceeding 80% accuracy. In particular, the MFCC features exhibit a slight improvement compared to using the full Emobase set.

5.3 Integrated Modalities

In order to analyze whether the integration of several modalities provides improved performance in detecting gender, we experiment with different combinations of features. To avoid any bias and due to the absence of development data, we integrate all the feature sets of a chosen modality before combining them with other modalities. In other words, we use the "All" feature set for the linguistic, physiological, and visual modalities, the whole face for the thermal modality, and the Emobase for the vocal modality.

As it is infeasible to experiment with all possible combinations, we chose the combinations that might exist in real-life situations, taking also into consideration that some physiological signals can be extracted from thermal videos as a non-contact approach. For instance, recent work has shown that thermography can be used to extract heart and respiration rate [14, 21]. Thus, the selected combinations are the thermal and physiological features, thermal and acoustic features, and the thermal and linguistic features where visual cameras are prohibited, vocal and linguistic features in applications where speech is only available, thermal and visual features for surveillance applications, thermal, physiological, and visual as a non-contact approach with the unavailability of speech, and all five modalities combined.

Table 6 lists the overall accuracy and class recall using different combination of modalities such as thermal (Face) and physiological (All) {Thrm+Phys}, vocal (EmoBase) and linguistic (All) {Voc+Ling}, thermal (Face) and visual (All) {Thrm+Vis}, thermal (Face) and vocal (EmoBase) {Thrm+Voc}, thermal (Face) and linguistic (All) {Thrm+Ling}, thermal (Face), physiological (All), and visual (All) {Thrm+Phys+Vis}, and "All Modalities" combined.

The table shows that following a multimodal approach does not provide significant differences as compared to the use of single modalities. In particular, the combination of thermal and physiological, vocal and linguistic, thermal and visual suffered a slight decrease in performance compared to the best performing individual modality in the combination. On the other hand, the integration of the thermal and vocal, thermal and linguistic, thermal, physiological, and visual, and all modalities combined exhibit a slight improved performance compared to the best performing individual modality in the integration. The best overall accuracy among all individual and combined modalities is achieved by integrating all five modalities with a relative improvement of 1.8% compared to the best performing modality (Vocal Emobase) in the combination. This potentially indicates that there might be some benefit in following a multimodal approach. However, based on the application it might be infeasible to process all five modalities.

Table 7: Overall accuracy and class recall using different combination of modalities such as thermal (Face) and physiological (All) {Thrm+Phys}, vocal (EmoBase) and linguistic (All) {Voc+Ling}, thermal (Face) and visual (All) {Thrm+Vis}, thermal (Face) and vocal (EmoBase) {Thrm+Voc}, thermal (Face) and linguistic (All) {Thrm+Ling}, thermal (Face), physiological (All), and visual (All) {Thrm+Phys+Vis}, and "All Modalities" combined . The best results are highlighted in bold.

Metric	Accuracy	Male Recall	Female Recall
Baseline	51.0	49.0	51.0
Thrm+Phys	69.2	69.8	68.7
Voc+Ling	78.8	76.9	80.8
Thrm+Vis	70.6	69.0	72.1
Thrm+Voc	79.6	76.1	**83.0**
Thrm+Ling	70.4	71.4	69.4
Thrm+Phys+Vis	72.5	66.7	78.1
All Modalities	**80.6**	**78.0**	**83.0**

5.4 Feature Analysis

In order to analyze the specific features that play an important role in predicting gender, we visualize the decision tree model constructed from the best performing set of features, which is achieved by integrating all five modalities together as can be seen in Figure 4.

The figure shows that tree nodes are constructed from four of the five modalities excluding the linguistic features. As expected the vocal features played a crucial role in the construction of the tree, including the root node and nine other nodes out of 17 nodes. Three nodes are composed using the thermal features starting from the second level of the tree. Two nodes are built using each of the visual and physiological features.

The vocal features include the pitch (F0) at the root node and the first level of the tree indicating that the pitch features provide the best separation between males and females. Moreover, four MFCC, two LSP, one PCM, and one voice probability features are used to build the nodes of the tree at different levels. Two thermal features are extracted from the histogram of 20 bins and one feature presents the maximum temperature extracted from the faces of the subjects, which indicates that males and females have different thermal distribution among their faces and reach different levels of maximum temperatures.

The two visual features were extracted from the Eigen vectors, which emphasizes the importance of the global facial features to indicate gender. Finally, the two physiological features present in the tree are the respiration rate and skin temperature signals.

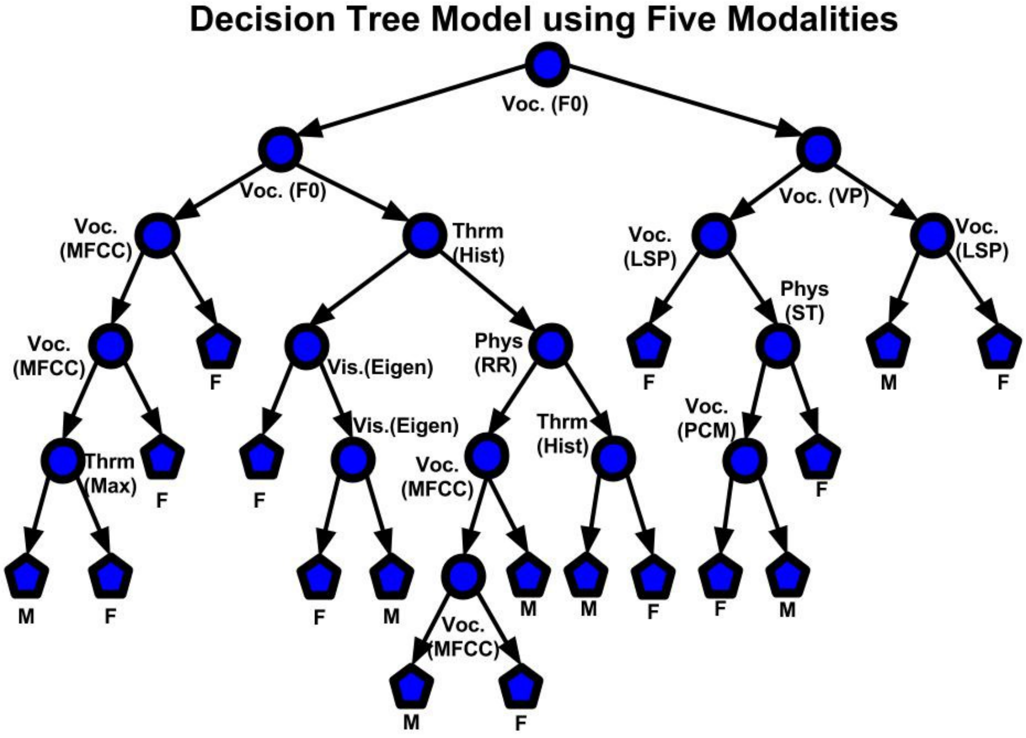

Figure 4: A visualization of the decision tree model constructed from the five modalities.

This might be surprising given that these two signals exhibited the lowest performance when used individually. However, their integration provides richer information that can improve gender prediction.

6 CONCLUSION

In this paper we analyzed different modalities for detecting gender taking into consideration the wide variety of applications that require gender detection such as surveillance purposes, computer forensics, electronic marketing, statistical analysis, demographic information collection, among others.

In particular we developed a dataset of male and female responses, explored the potential of utilizing for the first time five different modalities to detect gender including the visual, linguistic, physiological, thermal, and vocal modalities, and evaluated the performance of integrating different modalities together for real-life applications.

The experiments showed that the vocal features outperform other modalities in identifying gender especially using the pitch features. The visual and thermal modalities came second. The global facial features in particular were capable of separating between genders. Moreover, the thermal features in the whole facial area and in the cheeks region represent good clues for detecting gender, which presents an analysis that was not explored before. The physiological features provide very close performance to that of the visual and thermal features especially using the individual blood volume pulse and skin conductance signals.

While different combinations of modalities did not exhibit a significant improvement over using individual modalities, the best performance achieved using the five modalities shed some light on the specific integrated features that can potentially detect gender with higher reliability. These features include the pitch, MFCC, LSP, PCM, and voice probability vocal features, the maximum temperature as well as the temperature distribution in the face for the thermal features, the facial global representations, and the combination of the respiration rate and skin temperature. On the other hand, the linguistic features and the gestures were not as useful compared to the rest of the features. These sets of features can be further explored in future work to identify gender based on the availability of specific modalities as well as the requirements of different real-life applications.

ACKNOWLEDGMENTS

This material is based in part upon work supported by National Science Foundation awards #1344257 and #1355633, by DARPA-BAA-12-47 DEFT grant #12475008, and by the Michigan Institute for Data Science. 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 Defense Advanced Research Projects Agency, or the Michigan Institute for Data Science.

REFERENCES

[1] Mohamed Abouelenien, Veronica Pérez-Rosas, Rada Mihalcea, and Mihai Burzo. 2014. Deception Detection Using a Multimodal Approach. In *Proceedings of the 16th International Conference on Multimodal Interaction (ICMI '14)*. ACM, Istanbul, Turkey, 58–65.

[2] M. Abouelenien, V. Pérez-Rosas, R. Mihalcea, and M. Burzo. 2017. Detecting Deceptive Behavior via Integration of Discriminative Features From Multiple Modalities. *IEEE Transactions on Information Forensics and Security* 12, 5 (May 2017), 1042–1055. DOI: https://doi.org/10.1109/TIFS.2016.2639344

[3] Jens Allwood, Loredana Cerrato, Kristiina Jokinen, Costanza Navarretta, and Patrizia Paggio. 2007. The MUMIN coding scheme for the annotation of feedback, turn management and sequencing phenomena. *Language Resources and Evaluation* 41, 3-4 (2007), 273–287. DOI: https://doi.org/10.1007/s10579-007-9061-5

[4] Yasmina Andreu, Pedro Garcia-Sevilla, and R.A. Mollineda. 2014. Face gender classification: A statistical study when neutral and distorted faces are combined for training and testing purposes. *Image and Vision Computing* 32, 1 (2014), 27–36. DOI: https://doi.org/10.1016/j.imavis.2013.11.001

[5] Juan Bekios-Calfa, J. Buenaposada, and Luis Baumela. 2014. Robust gender recognition by exploiting facial attributes dependencies. *Pattern Recognition Letters* 36 (2014), 228 – 234. DOI: https://doi.org/10.1016/j.patrec.2013.04.028

[6] J. Bishop and P. Keating. 2009. Perception of pitch location within a speakerfis range: Fundamental Frequency, voice quality and speaker sex. *The Journal of the Acoustical Society of America* 132, 2 (2009), 1100–1112. http://dx.doi.org/10.1111/j.1749-818X.2009.00125.x

[7] Constantinos Boulis and Mari Ostendorf. 2005. A Quantitative Analysis of Lexical Differences Between Genders in Telephone Conversations. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics (ACL '05)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 435–442. DOI: https://doi.org/10.3115/1219840.1219894

[8] Michael Brookes. 2003. VOICEBOX: Speech Processing Toolbox for MATLAB. (2003).

[9] D. John Burger, John Henderson, George Kim, and Guido Zarrella. 2011. Discriminating Gender on Twitter. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 1301–1309. http://aclweb.org/anthology/D11-1120

[10] C. Chen and A. Ross. 2011. Evaluation of gender classification methods on thermal and near-infrared face images. In *2011 International Joint Conference on Biometrics (IJCB)*. 1–8. DOI: https://doi.org/10.1109/IJCB.2011.6117544

[11] Na Cheng, R. Chandramouli, and K. P. Subbalakshmi. 2011. Author Gender Identification from Text. *Digit. Investig.* 8, 1 (July 2011), 78–88. DOI: https://doi.org/10.1016/j.diin.2011.04.002

[12] Florian Eyben, Martin Wöllmer, and Björn Schuller. 2009. OpenEAR Introducing the Munich open-source emotion and affect recognition toolkit. In *2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*. IEEE, 1–6.

[13] Aparna Garimella and Rada Mihalcea. 2016. *Proceedings of the Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media (PEOPLES)*. The COLING 2016 Organizing Committee, Chapter Zooming in on Gender Differences in Social Media, 1–10. http://aclweb.org/anthology/W16-4301

[14] Travis Gault and Aly Farag. 2013. A fully automatic method to extract the heart rate from thermal video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 336–341.

[15] Abdenour Hadid and Matti Pietikäinen. 2009. Combining Appearance and Motion for Face and Gender Recognition from Videos. *Pattern Recogn.* 42, 11 (Nov. 2009), 2818–2827. DOI: https://doi.org/10.1016/j.patcog.2009.02.011

[16] R. Hartley and A. Zisserman. 2003. *Multiple View Geometry in Computer Vision*. Cambridge University Press.

[17] Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. 2007. *Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments*. Technical Report 07-49. University of Massachusetts, Amherst.

[18] G. Levi and T. Hassncer. 2015. Age and gender classification using convolutional neural networks. In *2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. 34–42. DOI: https://doi.org/10.1109/CVPRW.2015.7301352

[19] Ita Sarah Levitan, Yocheved Levitan, Guozhen An, Michelle Levine, Rivka Levitan, Andrew Rosenberg, and Julia Hirschberg. 2016. *Proceedings of the Second Workshop on Computational Approaches to Deception Detection*. Association for Computational Linguistics, Chapter Identifying Individual Differences in Gender, Ethnicity, and Personality from Dialogue for Deception Detection, 40–44. DOI: https://doi.org/10.18653/v1/W16-0806

[20] Sarah Ita Levitan, Taniya Mishra, and Srinivas Bangalore. 2016. Automatic Identification of Gender from Speech. In *Speech Prosody*.

[21] Gregory F Lewis, Rodolfo G Gatto, and Stephen W Porges. 2011. A novel method for extracting respiration rate and relative tidal volume from infrared thermography. *Psychophysiology* 48, 7 (2011), 877–887.

[22] X. Li, X. Zhao, Y. Fu, and Y. Liu. 2010. Bimodal gender recognition from face and fingerprint. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. 2590–2597. DOI: https://doi.org/10.1109/CVPR.2010.5539969

[23] C. A.and Papadelis C.and Vivas Ana B.and Klados M. A.and Kourtidou-Papadeli C.and Pappas C.and Ioannides A. A.and Bamidis P. D. Lithari, C.and Frantzidis. 2010. Are Females More Responsive to Emotional Stimuli? A Neurophysiological Study Across Arousal and Valence Dimensions. *Brain Topography* 23, 1 (01 Mar 2010), 27–40. DOI: https://doi.org/10.1007/s10548-009-0130-5

[24] L. Lu, Z. Xu, and P. Shi. 2009. Gender Classification of Facial Images Based on Multiple Facial Regions. In *2009 WRI World Congress on Computer Science and Information Engineering*, Vol. 6. 48–52. DOI: https://doi.org/10.1109/CSIE.2009.871

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

[26] E. Makinen and R. Raisamo. 2008. Evaluation of Gender Classification Methods with Automatically Detected and Aligned Faces. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 30, 3 (March 2008), 541–547. DOI: https://doi.org/10.1109/TPAMI.2007.70800

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

[28] Matthew L Newman, Carla J Groom, Lori D Handelman, and James W Pennebaker. 2008. Gender differences in language use: An analysis of 14,000 text samples. *Discourse Processes* 45, 3 (2008), 211–236.

[29] Dat Tien Nguyen and Kang Ryoung Park. 2016. Body-Based Gender Recognition Using Images from Visible and Thermal Cameras. *Sensors* 16, 2 (2016). DOI: https://doi.org/10.3390/s16020156

[30] P. Nguyen, D. Tran, X. Huang, and W. Ma. 2013. Age and gender classification using EEG paralinguistic features. In *2013 6th International IEEE/EMBS Conference on Neural Engineering (NER)*. 1295–1298. DOI: https://doi.org/10.1109/NER.2013.6696178

[31] J. Pennebaker and M. Francis. 1999. Linguistic Inquiry and Word Count: LIWC. (1999). Erlbaum Publishers.

[32] Slav Petrov, Leon Barrett, Romain Thibaux, and Dan Klein. 2006. Learning accurate, compact, and interpretable tree annotation. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 433–440.

[33] P.Jonathon Phillips, Harry Wechsler, Jeffery Huang, and Patrick J. Rauss. 1998. The FERET database and evaluation procedure for face-recognition algorithms. *Image and Vision Computing* 16, 5 (1998), 295 – 306. DOI: https://doi.org/10.1016/S0262-8856(97)00070-X

[34] Daniel Reid, Sina Samangooei, Cunjian Chen, Mark Nixon, and Arun Ross. 2013. Soft biometrics for surveillance: an overview. *Machine learning: theory and applications. Elsevier* (2013), 327–352.

[35] Mickael Rouvier, Grégor Dupuy, Paul Gay, Elie Khoury, Teva Merlin, and Sylvain Meignier. 2013. *An open-source state-of-the-art toolbox for broadcast news diarization*. Technical Report. Idiap.

[36] Ruchita Sarawgi, Kailash Gajulapalli, and Yejin Choi. 2011. Gender attribution: tracing stylometric evidence beyond topic and genre. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning*. Association for Computational Linguistics, 78–86.

[37] Alexandra Schofield and Leo Mehr. 2016. *Proceedings of the Fifth Workshop on Computational Linguistics for Literature*. Association for Computational Linguistics, Chapter Gender-Distinguishing Features in Film Dialogue, 32–39. DOI: https://doi.org/10.18653/v1/W16-0204

[38] Caifeng Shan. 2012. Learning Local Binary Patterns for Gender Classification on Real-world Face Images. *Pattern Recognition Letters* 33, 4 (March 2012), 431–437. DOI: https://doi.org/10.1016/j.patrec.2011.05.016

[39] Maha Sharkas and Mohamed Abouelenien. 2008. Eigenfaces vs. fisherfaces vs. ICA for face recognition; a comparative study. In *2008 9th International Conference on Signal Processing*. 914–919. DOI: https://doi.org/10.1109/ICOSP.2008.4697276

[40] Adrian P. Simpson. 2009. Phonetic differences between male and female speech. *Language and Linguistics Compass* 3, 2 (2009), 621–640. DOI: https://doi.org/10.1111/j.1749-818X.2009.00125.x

[41] Sudipta N. Sinha, Jan-michael Frahm, Marc Pollefeys, and Yakup Genc. 2006. *GPU-based Video Feature Tracking and Matching*. Technical Report. The University of North Carolina at Chapel Hill.

[42] David Smith and Roy D. Patterson. 2005. The Interaction of Glottal-Pulse Rate and Vocal-Tract Length in Judgements of Speaker Size, Sex, and Age. *Journal of the Acoustical Society of America* 118, 5 (2005), 3177fi?!3186.

[43] Pattanawit Soanboon, Somsong Nanakorn, and Wibhu Kutanan. 2016. Determination of sex difference from fingerprint ridge density in northeastern Thai teenagers. *Egyptian Journal of Forensic Sciences* 6, 2 (2016), 185 – 193. DOI: https://doi.org/10.1016/j.ejfs.2015.08.001 Advances in Forensic Anthropology.

[44] J. Tang, X. Liu, H. Cheng, and K. M. Robinette. 2011. Gender Recognition Using 3-D Human Body Shapes. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 41, 6 (Nov 2011), 898–908. DOI: https://doi.org/10.1109/TSMCC.2011.2104950

[45] I. Ullah, M. Hussain, G. Muhammad, H. Aboalsamh, G. Bebis, and A. M. Mirza. 2012. Gender recognition from face images with local WLD descriptor. In *2012 19th International Conference on Systems, Signals and Image Processing (IWSSIP)*. 417–420.

[46] Paul Viola and MichaelJ. Jones. 2004. Robust Real-Time Face Detection. *International Journal of Computer Vision* 57, 2 (2004), 137–154.

[47] Adam Vogel and Dan Jurafsky. 2012. *Proceedings of the ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries*. Association for Computational Linguistics, Chapter He Said, She Said: Gender in the ACL Anthology, 33–41. http://aclweb.org/anthology/W12-3204

[48] Marzia Del Zotto and Alan J. Pegna. 2015. Processing of masked and unmasked emotional faces under different attentional conditions: an electrophysiological investigation. *Frontiers in Psychology* 6 (2015), 1691. DOI: https://doi.org/10.3389/fpsyg.2015.01691