Robust Algorithm for Multimodal Deception Detection

Sushma Venkatesh Raghavendra Ramachandra Patrick Bours
Norwegian University of Science and Technology (NTNU)
Gjøvik, Norway
Email:{sushma.venkatesh, raghavendra.ramachandra, patrick.bours}@ntnu.no

Abstract

Automatic deception detection from the video has gained a paramount of interest because of their applicability in various real-life applications. The recorded videos contain various information such as temporal variations of the face, linguistics and acoustics, which can be used together, to detect deception automatically. In this work, we proposed a new approach based on multimodal information like audio, linguistic (or text) and non-verbal features. The proposed multimodal deception detection framework is based on combining the decision from the audio, text and non-verbal features using majority voting. The proposed multimodal deception system is banked on the audio system based on Cepstral Coefficients (CC) and Spectral Regression Kernel Discriminant Analysis (SRKDA) of fixed length audio sequences. The text system is based on bag-of-n-gram features and the linear Support Vector Machine (SVM) classifier while the non-verbal features are classified using the AdaBoost classifier. Extensive experiments are carried out on a publicly available real-life deception video dataset to evaluate the efficacy of the proposed scheme. The obtained results on a 25-cross-fold validation have indicated a deception detection accuracy of 97% *out-performing both state-of-the-art techniques and human performance on the whole dataset.*

1. Introduction

Deception can be defined as the act which can intentionally mislead, conceal the truth, or promote false concepts or ideas for personal gain and advancement. Deception detection plays a vital role, especially with law enforcement agencies, addressing real-life applications such as airport transit passengers screening, trials in the court and interrogation of suspects. The conventional approach includes the decision made by the humans based on a series of questions and answers. However, such a decision usually is prone to error and thus judiciary systems in many countries will not accept this as a piece of evidence in court. These factors have motivated the development of techniques that can automatically detect deception from video. Further, the progress made in pattern recognition and applied machine learning has boosted the development of automatic deception detection techniques.

Figure 1: Illustration of video frame (a) facial expression while Truth (b) facial expression while Deceit

Early works in deception detection are based on verbal and non verbal behaviour that is observed by a psychology expert, based on questions and answers. The popular methods based on the cues mentioned above include Polygraph, Magnetic Resonance Imaging (fMRI) facial expression and wearable sensors [3]. Recent works are based on the analysis of the real-life video sequences that are captured based on questions and answers, since videos can capture multiple information, including visual and linguistic cues. These factors have motivated researchers to explore these multimodal cues using advanced machine learning techniques to detect the deception automatically.

Automatic deception detection using the real-life video sequences was introduced together with a publicly available dataset [5]. This dataset consists of 121 deceptive and truthful video clips, reflecting a real-life question and answer scenario. Figure 1 shows two example frames from the dataset introduced in [5]. This dataset provides different cues like video, audio and text (converted from au-

978-1-7281-1198-8/19/$31.00 ©2019 IEEE
DOI 10.1109/MIPR.2019.00108

dio) that can be used together to develop an automatic deception detection technique. Further, this dataset also provides the manual labels of the 39 different verbal and non-verbal cues. Analysis in [5], carried out on the multimodal information in this dataset, has shown a classification accuracy of 82%. In [7], the automatic analysis of facial micro-expression especially on the behaviour of eyebrows raising, together with audio and transcript modality on the same dataset [5] has indicated an improved performance. In [4], the 3D based visual features are extracted using Deep Convolutional Neural Network (DCNN). Further, the audio features are extracted using openSMILE and text features are extracted using DCNN. Finally, features from all three modalities are combined to make the final decision. In [6], the visual features based on the 2D appearance models are used to capture the micro movements from eye blink, eyebrow motion, wrinkle occurrence and mouth motion. Experiments are reported on a private dataset indicate a reasonable performance. In [2] a multi-view approach is proposed to capture the micro expression features from the video to detect deception.

Based on the above-reported works, the video clips are provided with facial micro-expression along with the corresponding audio and text data. The early work [5] in this direction is to utilise the manual labelling of the human behaviour from the video that can be used for classification. Recently, automatic measurement of facial expressions [5, 4, 6, 7, 2] are explored using various techniques based on facial action units, extracting saliency features from the face, and end-to-end features using DCNN. However, these methods have indicated good results when the datasets are of good quality regarding facial pose and image resolution. Further, the cultural background of the data subject makes it very difficult to capture the facial micro-expression automatically. Further, it is also demonstrated in [5], that the use of only facial micro-expressions is not enough to achieve good performance. This fact is also justified in recent works [2] that have used only some part of the dataset to indicate the good performance. In this work, we propose a new framework for automatic deception detection based on micro expressions, audio and text data captured from the videos. The proposed scheme employs the Cepstral Coefficients (CC) features and the Spectral Regression Kernel Discriminant Analysis (SRKDA) classifier on the audio signal. The text features are extracted using a bag of n-grams together with a linear SVM classifier and the micro-expressions are classified using the AdaBoost algorithm. The final decision is made by combining the decision from three independent classifiers using majority voting. Extensive experiments are carried out on a publicly available dataset [5] with 121 deceptive and truthful video clips reflecting a real-life scenario. The results of the proposed methods are compared with the state-of-the-art methods and indicate an improved performance.

The rest of the paper is organised as follows: Section 2 presents the proposed method and Section 3 discuss the experimental results. Finally, in Section 4 conclusions are drawn.

2. Proposed Method

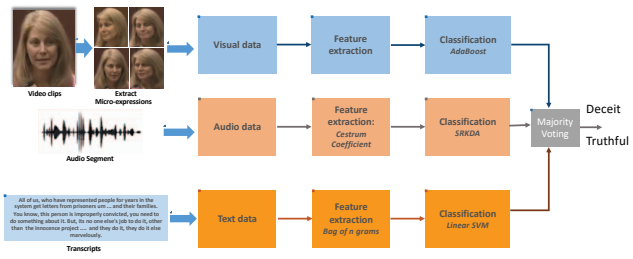

Figure 2: Block diagram of the proposed method

Figure 2 shows the block diagram of the proposed scheme that combines the multimodal information at decision level using the majority voting method. The proposed method can be structured into four main functional units, i.e. the Audio sub-system, the Text sub-system, the Micro-behaviour sub-system and the fusion sub-system.

Audio subsystem: In this work, we have employed the CC features together with the SRKDA classifier to reliably classify the audio as either truthful or deceitful. Since the goal is to extract the hidden cues from the audio, we have used fixed length chunks of the audio, corresponding to 3.2seconds each. Given the audio signal, we extract the CC features that are computed by windowing the audio signals on which the Fast Fourier Transform (FFT) is computed. The computed spectral signal is processed using the Mel filtering whose output is the weighted sum of the FFT magnitude spectral values which is then transformed using the non-linear logarithm function. Finally, 13 different CC values are calculated as follows:

$$CC_i = \sum_{j=1}^{23} f_j \times cos\left\{\frac{\pi \times i}{23}(j-0.5)\right\}, 0 \leq i \leq 12 \quad (1)$$

where, f_j represents the non-linear transformation on the mel filtering CC_i indicates the i^{th} cepstral coefficient. The final vector consists of 14 coefficients, being the log-energy coefficient and 13 cepstral coefficients. Thus, given the audio signal, we extract the CC features which has dimension 1×14.

In this work, we have employed the Spectral Regression Kernel Discriminant Analysis (SRKDA) [1] to classify the audio features as deceit or truthful. SRKDA basically works

on the data projected in space by performing discriminant analysis induced by non-linear mapping. For further computation, analysis of the subspace is performed using spectral graph analysis and regression. Employing both techniques for analysis leads to efficient computation by using regularization techniques. Basically SRKDA is employed to solve regularized least square problems that involve less computational load and time.

Text subsystem: In this work, we have employed the Bag-of-N-Grams (BoNG) as the feature representation and linear SVM as the classifier. The commonly used N-grams in text analysis includes uni-gram (N = 1), bi-grams (N = 2) and tri-grams (N = 3) represented as a vector. Since the text data is generated from the audio data, we have performed a series of pre-processing steps that included tokenising the text, erasing the punctuation, adding the part of speech and normalising the words. We then use BoNG with bi-grams and tri-grams which resulted in a feature set of dimension 11906. Figure 3 shows the word clouds from deceit and truthful text data used in this work. We then use linear SVM to perform the classification based on the BoNG features to obtain the final decision on each test text sample.

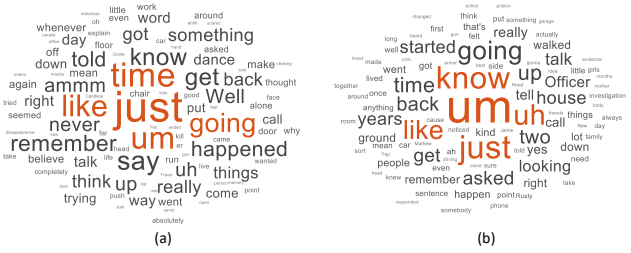

Figure 3: Word Cloud (a) Deceit text (b) Truthful text

Micro-behaviour subsystem: We have employed the binary labels of the micro-expression provided with the dataset [5] together with the Adaptive boosting (AdaBoost) classifier to detect deception. For each video, 39 different micro-expressions are provided, that quantifies both face and body micro movements. We have used the AdaBoost classifier by considering its high performance on the binary classification data. Given the test micro-expressions, the AdaBoost algorithm is used to obtain the final decision.

Decision level fusion: In this work, we employed majority voting to combine the binary decisions from the algorithms used on audio, text and micro-expression data. Given a test video, the multimodal decision on audio, text and micro-expression data are obtained as explained above. These decisions are concatenated to form a single vector, and the final decision is made based on majority voting.

3. Experiments and Results

In this section, we present the experimental results of the proposed scheme on the used dataset [5]. This dataset is based on real-life trial data, and consists of video clips of courtroom trials. The dataset consists of 121 videos, where 61 are deceptive and 60 are truthful video clips. The average video length corresponding to truthful clips is 28.3 seconds and for deceptive video clips it is 27.7 seconds. The dataset is enclosed with the corresponding audio, text and micro-expression for both the deceptive and the truthful set. We have followed a similar experimental protocol as provided in [7] for splitting the data in a training and a testing set, using a 25-fold cross-validation. The results are presented with the classification accuracy or correct classification rate (CCR %). Figure 4 illustrates the example video frames from publicly available multimodal deception dataset [5].

Figure 4: Example video frames from publicly available multimodal deception dataset [5] (a) Deceit (b) Truthful

In this work, we have carried out two different experiments to evaluate the performance of the proposed techniques, both on the independent modalities as well as on the multimodal modality. **Experiment 1:** presents the comparative performance evaluation on the individual modalities (i.e. audio, text, and micro-expressions) using both existing and the techniques employed in the proposed sub-systems. **Experiment 2:** This experiment will report the performance of the proposed systems on multimodal deception detection.

Table 1 shows the performance of the individual modalities on various algorithms. For the audio modality, we have evaluated three different features (Mel Frequency Cepstral Coefficients (MFCC), Log-Energy of MFCC, and CC) with three different classification schemes (linear SVM, SRKDA, and Long Short Term Memory). Based on the obtained results it can be observed that: (1) among three different features sets, CC has the best performance when SRKDA or LSTM classification is used. (2) Among the three different classifiers used, SRKDA has shown the best performance. (3) The best performance is noted with the CC-SRKDA where CCR = 76%.

For the text modality, we have evaluated five different algorithms including the proposed method. Based on the

Table 1: Performance of algorithms on individual modality

Modalities	Algorithm	CCR(%)
Audio	MFCC-SVM [7] [4]	64
	CC-SVM	66
	LE-SVM	72
	MFCC-LSTM	46
	CC-LSTM	58
	LE-LSTM	50
	MFCC-SRKDA	64
	CC-SRKDA	**76**
	LE-SRKDA	68
Text	LSTM	44
	SRKDA	68
	SVM [7]	24
	Bag-of-words - SVM	66
	Bag-of-N-Grams - SVM	**84**
Micro-Expressions	**AdaBoost**	**88**
	Random Forest	82
	SVM [7]	75

obtained results, it can be noted that the use of Bag-of-N-Grams (BoNG) features together with the linear SVM classifier has indicated the best performance with CCR = 84%. This further justifies that the use of bi-gram and tri-gram features together can improve the performance.

The performance of the Micro-expressions is tested with three different classifiers, as presented in Table 1. Since the micro-expressions are coded in the binary, we have evaluated three different binary classifiers. Based on the obtained results, the AdaBoost classifier gave the best performance with CCR = 88%. Thus, based on the obtained results, the proposed individual sub-systems on all three modalities have indicated the best performance. The obtained results have further motivated us to combine the decisions from these three subsystems using majority voting to detect the multimodal deception.

Table 2: Performance of proposed method and existing methods on multimodal data

Algorithms	Correct Classification Rate (CCR) (%)
Proposed Method	**97**
[7]	87.73
[5]	82
[4]	96.14

Table 2 shows the performance of the proposed multimodal deception detection scheme together with the existing methods. It is important to note that, we have performed 25-fold cross validation while existing methods (in Table 2) have reported the results for 10-fold cross-validation. Based on the obtained results, the proposed methods show the best performance with CCR = 97%. The obtained results justify the efficacy of the proposed scheme on real-life deception detection.

4. Conclusion

Automatic deception detection from videos is emerging as the potential solution as there exist certain patterns that can be captured using advanced machine learning approaches. In this paper, we present a novel approach for multimodal deception detection, based on audio, text, and micro-expressions. The audio subsystem employed in the proposed approach is based on the CC features with the SRKDA classifier, the text sub-system is based on the Bag-of-N-Grams (BoNG) features and linear SVM, and the micro-expression features are classified with the AdaBoost classifier. Finally, the decision from all three sub-systems is fused using majority voting to make the final decision. Extensive experiments are carried out on a publicly available multimodal deception dataset with 121 video clips. Obtained results have demonstrated the improved performance of the proposed scheme when compared with existing systems.

References

[1] D. Cai, X. He, and J. Han. Speed up kernel discriminant analysis. *The VLDB JournalThe International Journal on Very Large Data Bases*, 20(1):21–33, 2011.

[2] N. Carissimi, C. Beyan, and V. Murino. A multi-view learning approach to deception detection. In *13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018)*, pages 599–606, May 2018.

[3] P. A. Granhag, A. Vrij, and B. Verschuere. *Detecting deception: Current challenges and cognitive approaches*. John Wiley & Sons, 2015.

[4] G. Krishnamurthy, N. Majumder, S. Poria, and E. Cambria. A deep learning approach for multimodal deception detection. *arXiv preprint arXiv:1803.00344*, 2018.

[5] V. Pérez-Rosas, M. Abouelenien, R. Mihalcea, Y. Xiao, C. Linton, and M. Burzo. Verbal and nonverbal clues for real-life deception detection. In *Proceedings of Conference on Empirical Methods in Natural Language Processing*, 2015.

[6] L. Su and M. Levine. Does lie to me lie to you? an evaluation of facial clues to high-stakes deception. *Computer Vision and Image Understanding*, 147:52 – 68, 2016.

[7] Z. Wu, B. Singh, L. S. Davis, and V. Subrahmanian. Deception detection in videos. *arXiv preprint arXiv:1712.04415*, 2017.