A Survey on Multimodal Deception Detection

Berat Biçer
Department of Computer Engineering
Bilkent University
Ankara, Turkey
berat.bicer@bilkent.edu.tr

***Abstract*—In this paper, we provide a survey on the topic of multimodal deception detection which is an active field of research in recent years. 14 recent papers that have significant impact on the field have been thoroughly studied with respect to the modality they employ, fusion approaches, and the datasets they use; their contributions are provided in comparison to one another.**

***Index Terms*—affective computing, deception, multimodality**

I. Overview

Deception can be defined as the act of completely or partially hiding the truth. It is one of the most extensively studied topics of research in psychology and computer science alike. Recent advances on multimodal approaches are driven by the findings of [1] and [2], which state that deception has unintentional behavioral and expressional leaks such as changes in facial expressions, head pose/movement, eye gaze, hand movement/gesture, facial action units and vocal features of the speech. These findings suggest that it is possible to automatically detect deceitful behavior through the fusion of multiple modalities that are effected by deceitful actions. Research in this direction has two major goals: First, researchers are interested in which modality is the most informative in detection of deceit. This is significant since it may be expensive to utilize many modalities and extracting relevant features, which limits model performance. Second, since there are multiple modalities, it is expected that each one has a varying salience in detecting deceit. Therefore, selecting the best way to fuse the modalities either through data or decision level fusion is important since it has great impact on model performance. Overall, the current studies focuses these aspects of the problem and either devise new modality specific approaches for improving performance, or propose new fusion techniques. The rest of the paper studies the work in these categories.

II. Modality and Improvements

Research in modalities are focused on three major aspects of deceit: facial, verbal, and behavioral cues. Facial features are studied in [3], [4], and [5]. As one of the leading studies, authors in [5] introduces a high-stakes lie database as well as the baseline method of deceit detection composed of facial action unit detection and localization, followed by eye blink, eyebrow and mouth movement detection. In [3], authors tackle the problem of data scarcity and unnatural nature of available data collected in controlled environments. The authors argue that capturing and studying high stakes lies (such as those in court trials) in their natural setting is difficult. Also, they acknowledge the issue of modality fusion and propose a face-focused cross-stream network across spatial and temporal streams for joint learning of deceit from multiple modalities with meta learning for tackling overfitting. Face expression branch of the network is a facial recognition model based on recurrent CNNs (R-CNN) generalized for facial expression recognition. The temporal stream, which runs on body motion data, employs improved dense trajectories [6] for feature extraction and the two parts are fused based on correlation analysis. Next, the meta learner maps the two-class classification problem of deceit detection into a multi-class classification problem by combining a deceptive sample with four truthful samples and learning to rank their truth values based on their correlation. Lastly, they train a generative model to confuse the classifier using the training data. Final classifier distinguishes true and fake samples as well as truthful and deceitful ones. Most importantly, the authors show that their model can be generalized to other affective computing problems such as emotional expression recognition. In [4], authors aim to construct the 3D facial model of the subject using 2D face images extracted from videos and use this model for feature extraction and deceit prediction. Authors are the first to apply 2D-to-3D reconstruction to deceit detection.

In speech modality, authors in [7] propose a semi-supervised autoencoder for combining classification and categorization of unlabelled data, derived by the fact that feature distribution of deceptive speech is different from other phonetic features, and deception detection only needs to determine whether the sample is deceitful. They also propose a Chinese speech dataset to help the research in phonetic modality.

Research in combining different modalities include proposition of new methods and datasets benchmarks, as authors in [8] did. They introduce a new task, Box-of-lies, based on a game played on television where participants try to convince the opposite person that they are describing the hidden object in front of them correctly. The authors extract linguistic features from conversation transcripts, non-verbal behavioral cues (such as head shake, single head nod, etc.), and statistical features (such as number of truthful statements). They then train a random forest classifier on these features and compare their results on human performance of deceit detection (which is around 57% as described in [9]). Real-life Trial Dataset(RLT), as introduced in [10], is another

study that introduces a new high-stakes dataset. It contains 121 videos, 60 truthful and 61 deceitful, that are recorded in high-stakes criminal trials. RTL has been employed to evaluate newly proposed methods since its publication and is among the few high-stakes datasets available. [11] is the next study that introduces a low-stakes deception database, Miami University deception detection database (MUD), in a controlled environment. These studies are important since these datasets are among the fundamental benchmarks in deception detection.

Many researches focus on, along with creating new benchmark datasets, new approaches on deception detection. One such study is [12], and the follow-up paper [13]. The authors propose the DEV framework for end-to-end deceptive video detection. They study the effects of different modalities, temporal information, and the need for human-readable solutions. The network first applies preprocessing to videos for feature extraction. Visual features are a sequence of images enhanced with an attention mechanism. Vocal features are extracted from speech windows using some pretrained CNN-based audio model. These features are then fused into a single feature vector for each sample. For the classifier, rather than training for binary classification it is trained to separate three samples, two belonging to the same class where similarity between samples with the same labels are closer compared to the third one. The proposed network automatically handles feature extraction and delivers interpretable results due to attention layers.

One of the first attempts in combining visual and speech-based features is published in [14]. The authors employed RLT dataset described before as well as their own benchmark. They employ uni-grams, psycho-linguistic features (lexicon used to incorporate semantic and psychological information into linguistic analysis [15]), and syntactic complexity (based on the observation that syntactic complexity of truthful speech is more complex that deceitful one [16]) as verbal features. For nonverbal features, they extract facial displays (facial expressions and head movements) and hand gestures. They trained a SVM classifier over all features, obtaining and accuracy of 82.14% final accuracy. This study is important because it explores the effects of different features on deceit detection. Next study [17] uses 3D-CNN to extract visual features, Word2Vec [18] for extracting vector representations of the transcript of the input video, and openSMILE [19] for auditory features. Lastly, they concatenate these features with 39-dimensional micro-expression labels provided by [10]. After feature-level fusion, a MLP is trained for binary classification. The contribution of this paper comes from their accuracy which is 96.14% (considering they employ off-the-shelf components for feature extraction, this accuracy is arguably impressive). Next study [20] uses improved dense trajectories [6] for visual features, MFCC features extracted from speech segments for auditory features, and [21] for textual features. They then use Gaussian mixture models(GMM) for predicting facial micro-expressions, and use these predictions as features for classification. In summary, this paper treats low-level features as intermediary features which is used for extracting high-level features, obtaining 92% classification accuracy. The intermediary feature translation is the most significant contribution of this paper. Lastly, authors in [22] discover a multi-view learning approach to improve feature-level fusion, which assumes different modalities contain complimentary information and by minimizing the disagreement between these modalities final error is also reduced. This provides insight to the contribution of each modality to the deception detection task, eliminating the need for feature selection. They obtain 99% classification accuracy, which implies that the task of multimodal deception detection problem is solved (this is open to discussion).

III. Conclusion

Research in multimodal deception detection focuses both dataset creation and proposing new methods for feature extraction and fusion. Our research showed that the most fundamental issue with the current state of the field is the lack of a large-scale deceit database that has high quality(resolution and framerate for videos, etc.) samples and good labels (almost all datasets published defines the problem as a binary classification, however, formalizing the problem as a multi-class classification task is also possible where labels represent the level of truth a sample contains). Even though the task at hand seems to have been solved (due to reaching accuracy levels around 99%), the ongoing research in the field suggests more work is required to say for certain.

References

[1] M. Hartwig and C. F. Bond Jr, "Lie detection from multiple cues: A meta-analysis," *Applied Cognitive Psychology*, vol. 28, no. 5, pp. 661–676, 2014.

[2] M. R. Morales, S. Scherer, and R. Levitan, "Openmm: An open-source multimodal feature extraction tool." in *INTERSPEECH*, 2017, pp. 3354–3358.

[3] M. Ding, A. Zhao, Z. Lu, T. Xiang, and J.-R. Wen, "Face-focused cross-stream network for deception detection in videos," 2018.

[4] M. Ngô, B. Mandira, S. F. Yilmaz, W. Heij, S. Karaoglu, H. Bouma, H. Dibeklioglu, and T. Gevers, "Deception detection by 2d-to-3d face reconstruction from videos," *CoRR*, vol. abs/1812.10558, 2018. [Online]. Available: http://arxiv.org/abs/1812.10558

[5] 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*, vol. 147, pp. 52–68, 2016.

[6] H. Wang and C. Schmid, "Action recognition with improved trajectories," in *Proceedings of the IEEE international conference on computer vision*, 2013, pp. 3551–3558.

[7] H. Fu, P. Lei, H. Tao, L. Zhao, and J. Yang, "Improved semi-supervised autoencoder for deception detection," *PloS one*, vol. 14, no. 10, 2019.

[8] F. Soldner, V. Pérez-Rosas, and R. Mihalcea, "Box of lies: Multimodal deception detection in dialogues," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 1768–1777. [Online]. Available: https://www.aclweb.org/anthology/N19-1175

[9] C. F. Bond Jr and B. M. DePaulo, "Accuracy of deception judgments," *Personality and social psychology Review*, vol. 10, no. 3, pp. 214–234, 2006.

[10] V. Pérez-Rosas, M. Abouelenien, R. Mihalcea, and M. Burzo, "Deception detection using real-life trial data," in *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, ser. ICMI '15. New York, NY, USA: ACM, 2015, pp. 59–66. [Online]. Available: http://doi.acm.org/10.1145/2818346.2820758

[11] E. P. Lloyd, J. C. Deska, K. Hugenberg, A. R. McConnell, B. T. Humphrey, and J. W. Kunstman, "Miami university deception detection database," *Behavior Research Methods*, vol. 51, no. 1, pp. 429–439, Feb 2019. [Online]. Available: https://doi.org/10.3758/s13428-018-1061-4

[12] H. Karimi, "Interpretable multimodal deception detection in videos," 10 2018, pp. 511–515.

[13] H. Karimi, J. Tang, and Y. Li, "Toward end-to-end deception detection in videos," *2018 IEEE International Conference on Big Data (Big Data)*, pp. 1278–1283, 2018.

[14] 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 the 2015 Conference on Empirical Methods in Natural Language Processing*. Lisbon, Portugal: Association for Computational Linguistics, Sep. 2015, pp. 2336–2346. [Online]. Available: https://www.aclweb.org/anthology/D15-1281

[15] J. W. Pennebaker and L. A. King, "Linguistic styles: Language use as an individual difference." *Journal of personality and social psychology*, vol. 77, no. 6, p. 1296, 1999.

[16] B. M. DePaulo, J. J. Lindsay, B. E. Malone, L. Muhlenbruck, K. Charlton, and H. Cooper, "Cues to deception." *Psychological bulletin*, vol. 129, no. 1, p. 74, 2003.

[17] G. Krishnamurthy, N. Majumder, S. Poria, and E. Cambria, "A deep learning approach for multimodal deception detection," 2018.

[18] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, 2013, pp. 3111–3119.

[19] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in opensmile, the munich open-source multimedia feature extractor," in *Proceedings of the 21st ACM international conference on Multimedia*. ACM, 2013, pp. 835–838.

[20] Z. Wu, B. Singh, L. S. Davis, and V. S. Subrahmanian, "Deception detection in videos," 2017.

[21] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.

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