Automated Deception Detection for Videos

Roddy MacSween
University of Cambridge
rlm72@cam.ac.uk

***Abstract*—We implement automated deception detection for videos, using a variety of classification algorithms and features from both the visual and textual modalities. In particular we use features from gaze data and a method of processing other visual features that captures the concept of expressions and microexpressions, neither of which have previous been applied to this task. Our system significantly outperforms humans at the task and achieves similar performance to the state-of-the-art computer approach, despite having lower reliance on human-annotated data.**

I. Introduction

The first known use of a deception detection system was recorded in China around 1000 BC [1]. A suspect of fraud would put dry rice in their mouth while being questioned; if the rice remained dry after being spat out they were found innocent. Using computer systems for deception detection is a considerably more recent phenomenon, but the broader topic has been studied heavily for a long time because of its widespread and important applications, for example use by the police and in courtrooms.

Computer deception detection is an interesting and challenging technical problem for several reasons. Unlike tasks such as face recognition, which are difficult for computers but easy for humans, deception detection is challenging for both (untrained humans barely outperform chance at it [2]). And unlike most classification problems, the data points in deception detection (or at least the dishonest ones) are generated by an adversarial process with the objective of making classification difficult.

Following the example of prior work in this area [3], [4], we take a multimodal approach to computer deception with classification done using features from both videos and associated transcripts. The system described here requires slightly less human input in producing features than existing approaches, but still performs similarly to the state-of-the-art. It also uses gaze as a novel source of features, which are found to predict deception well.

II. Related work

There is a substantial amount of research into the abilities of humans to detect deception. A consensus exists that untrained humans are generally unable to detect deception at a rate much greater than chance, and furthermore that training does not usually lead to improvement [2], [5]. However, there is evidence of a small number of "wizards" who can reliably detect deception with accuracy rates in the 70-100% range [6]. Studies have also been carried out on the use of physiological methods such polygraphs, with no clear consensus on their usefulness [7].

Especially relevant to our work is research by Paul Ekman [8], [9] into non-verbal leakage: ways a person can inadvertently communicate their emotional state through body language. In particular, he has studied microexpressions (facial expressions that last for a fraction of a second) which are thought to be particularly informative of deception. For this purpose he developed the Facial Action Coding System (FACS) which taxonomises facial expressions as combinations of Action Units (AUs) that represent actions of facial muscles. These are used as a source of features for our system.

Detecting deception in text based on linguistic markers has also been studied. For example, deceptive writing has been found to have a higher proportion of negative emotion words and a lower proportion of first-person singular pronouns than truthful writing [10].

Various approaches have been taken in the past to automated deception detection. We will only discuss those using a similar dataset to ours, i.e. unconstrained videos from real-life high-stakes situations. The particular dataset we use – the Real Life Deception Detection (RLDD) dataset – comes from a 2015 paper [3] and contains clips from recordings of trials and interviews connected to high-stakes criminal proceedings; an example frame from one clip is show in Figure 1.

Fig. 1. A frame from one of the (deceptive) videos in the RLDD dataset.

As well as producing the dataset, [3] also trained a classifier and tested human performance using it. A 2016 paper [11] using the dataset extracts raw features using the tool OpenFace as is done here. While these papers use classical machine learning methods (decision trees and SVMs), [12] applies neural networks to the task. Of special note is a 2017 paper [4] which is used as the state-of-the-art by our comparison of our results. Finally, [13] study a similar but not identical dataset of videos.

Although there are several previous papers that use this dataset, not all of them have a comparable methodology to ours for reasons discussed in section IV. Therefore we only compare our results those given in [4].

III. Method

As in previous work, we frame deception detection as a classification problem. The system consists of feature extraction for both the visual and textual modalities, the results of which are then fed into a classifier.

A. Dataset

The RLDD dataset used consists of 121 videos with average duration 28.0 seconds.The dataset is balanced between truthful and deceptive clips and contains 56 unique subjects. Some subjects appear in both truthful and deceptive clips. Transcriptions and manual annotations of gestures were also provided in the dataset; here we only use the latter.

B. Visual features

The first step in producing visual features is using OpenFace [14] to extract FACS AUs and gaze information from videos. This produces an 18-dimensional vector for each frame in each video, containing the intensity of each AU that OpenFace implements detection for at that frame.

Then framewise AU data is combined to produce one feature vector per video in several ways. Firstly, the simple mean of each AU intensity is taken.

Secondly, we build up *expression vectors* based on occurrences of consecutive frames where the same AU is present (has intensity above some threshold hyperparameter). For each AU and video, this first requires finding all pairs of frame indices (i, j) such that the AU is present in frame k for all k where $i \leq k < j$, which is done using a Finite State Machine. Then two 18-dimensional vectors $\mathbf{v}$ and $\mathbf{w}$ are produced, where $\mathbf{v}_{AU_x}$ is the number of frame indices (i, j) for AU_x with $i - j \leq L$ for some hyperparameter L, and $\mathbf{w}_{AU_x}$ is the corresponding number for when $i - j > L$. This split is intended to capture the difference between microexpressions and normal length expressions, since research suggests [8] the former are more informative of deception. Finally, $\mathbf{v}$ and $\mathbf{w}$ are concatenated and normalised by video length to give the 36-dimensional expression vector.

Thirdly, *overlap vectors* are produced. These are similar to expression vectors, except that the vectors are indexed over pairs of distinct AUs, with frame indices (i, j) representing sequences where both AUs are present. Interval trees [15] are used to efficiently find these intervals. These features are intended to capture deception cues that involve multiple AUs, for example a genuine smile involves AUs 6 and 12 whereas in a fake smile only the latter is present [16].

These approaches of using AUs are broadly similar to those in existing work, although the specific technique of expression and overlap vectors is novel. However, we also use features constructed from gaze information; this approach has not been tried in previous automated deception detection systems. OpenFace provides data about gaze in the form of a vector representing the direction the subject is looking for each frame, in a coordinate system where the z-axis is the direction the subject is facing. The 4-dimensional *gaze vectors* for each video consist of the mean and standard deviation of the x and y components of this vector across the frames.

We also experimented with using features based on pupil dilation (average pupil radius in each frame), as psychological research suggests this could be informative about deception [17]. However, these did not differ significantly between deceptive and truthful videos and therefore were not used.

C. Textual features

Additionally, we extracted features from the transcripts for the videos. We used simple tf–idf features [18].

D. Feature selection

After all features had been extracted, features selection was performed to disregard redundant and uninformative features. The first step here was to remove all features with zero variance. Then the k most informative features (measured by ANOVA F-value) were selected, with the value of k being optimised by hyperparameter search.

E. Classification

We experimented with a variety of classification algorithms: SVMs (with linear and radial basis function kernels), k-nearest neighbours, random forests and neural networks. The NN architecture is specified in Figure 2.

In a multimodal setting such as this, classification can be done with either early fusion (concatenating feature vectors from all modalities and using a single classifier) or late fusion (performing classification for each modality separately, and then combining the results using another classifier). Both approaches have been used on this dataset; here we focused primarily on early fusion but did also explore late fusion (using logistic regression for the final layer).

F. Evaluation

Performance of the system was measured using AUC and accuracy with 10-fold cross validation, to allow comparison with other work using the dataset. Since some subjects appeared in multiple videos in the dataset, a custom cross-validation algorithm was required that ensured each subject only appeared in one fold, as without this the system could have learnt to classify videos by recognising the subjects rather than by detecting deception. Hyperparameters were optimised using grid search.

IV. Results

Table I shows the AUC achieved using early fusion for each combination of feature source and classifier. Results where simple AUC mean intensity features were used are not given, as performance was significantly worse when those were included, possibly due to high redundancy between those and expression features.

Fig. 2. Architecture of the neural network used. The input size is 80 in the diagram but varied in practice depending on the number of feature sources. Each dropout layer had a rate of 0.3. The activation function used was ReLU, except for at the last layer where a sigmoid is used. The `None` in the dimension specifications indicates that it accepts batches of any size..

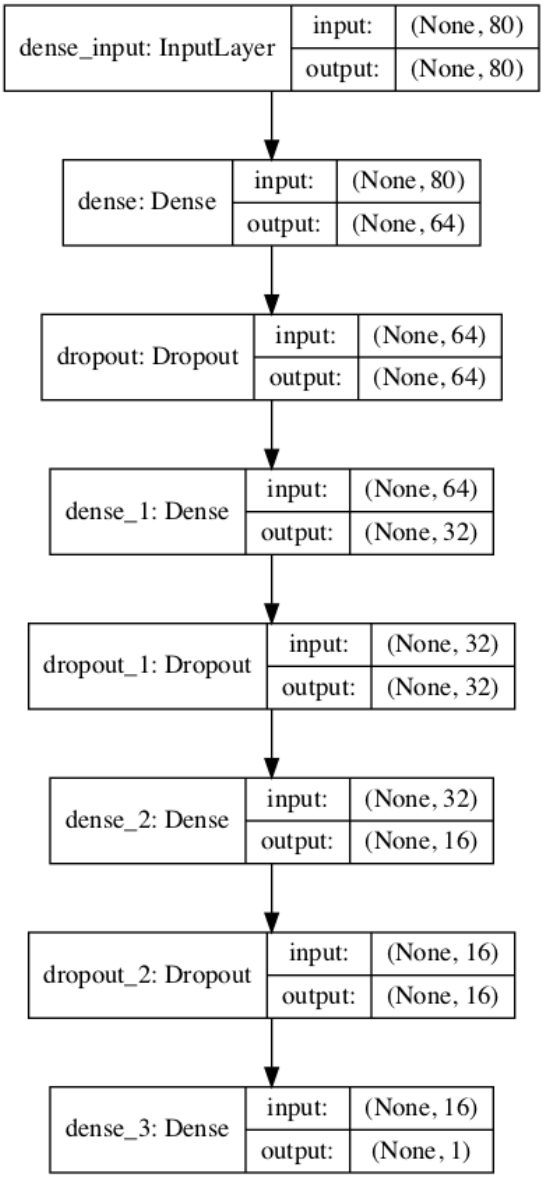

Figure 3 examines how AUC changes when each feature source is included or excluded. As expected, most feature sources positively contribute to AUC. However, overlap features are an exception, and this remains true (to a lesser extent) even when expression features are omitted as in Figure 4 so redundancy with those cannot be a full explanation. But some individual overlap features are very predictive of deception, suggesting that they could be useful if a more sophisticated method of features selection were used.

One particularly notable result is that gaze is a useful feature, both in isolation and in combination with other features, confirming the hypothesis in [11].

Although there are several existing papers that also use this dataset, many of them are not a good point of comparison for this system; some of them don't ensure subjects are grouped into cross-validation folds which could cause inflated results, and others use human-annotated microexpressions as features which may give them an advantage over our fully-automated approach. For this reason, the relevant state-of-the-art we compare with is the DARE (Deception Analysis and Reasoning Engine) presented in [4].

Figure 5 compares the AUCs achieved by each system in the visual and textual modalities separately, and with features from both combined. Performance for DARE in the visual modality is broken up into the case where only low-level IDT features are used and the case where higher-level microexpression features are added. This is because although DARE's microexpression features are produced automatically, this is done using a classifier that was trained on human-annotated microexpressions from the same dataset. In contrast, our system has a fixed general notion of an expression and does not incorporate hardcoded human concepts of kinds of

E	O	G	T	K-SVM	L-SVM	k-NN	RF	NN
			✓	0.67	0.69	0.69	0.62	0.70
		✓		0.69	0.69	0.66	0.53	**0.76**
		✓	✓	0.72	**0.71**	0.69	0.64	0.69
	✓			0.30	0.41	0.48	0.37	0.41
	✓		✓	0.64	0.68	0.71	0.65	0.71
	✓	✓		0.69	0.70	0.65	0.56	0.57
	✓	✓	✓	0.69	0.69	**0.75**	0.60	0.70
✓				0.67	0.63	0.63	0.71	0.47
✓			✓	0.69	0.68	0.73	0.67	0.70
✓		✓		**0.74**	0.63	0.68	0.69	0.72
✓		✓	✓	0.72	0.70	0.68	0.73	0.69
✓	✓			0.64	0.59	0.63	0.69	0.48
✓	✓		✓	0.67	0.66	0.74	0.57	0.65
✓	✓	✓		0.71	0.70	0.70	0.62	0.55
✓	✓	✓	✓	0.71	0.70	0.68	**0.79**	0.71

Table I
AUCs achieved using early fusion. Features are E(xpressions), O(verlaps), G(aze) and T(ranscripts). Classifiers are (rbf) kernel SVM, linear SVM, k-nearest neighbours, random forest and neural network. The best result for each classifier is given in bold.

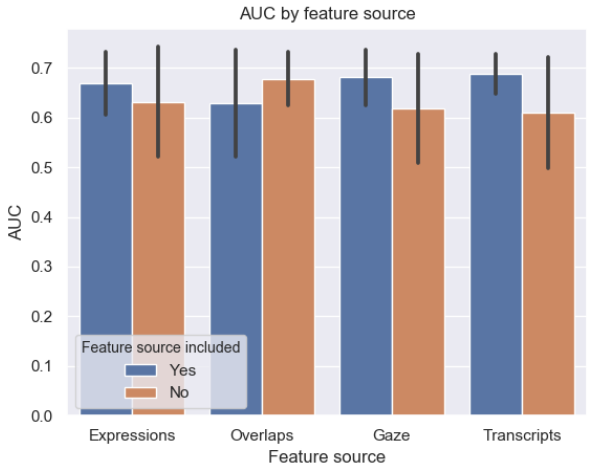

Fig. 3. Variation in AUC based on which feature sources are included.

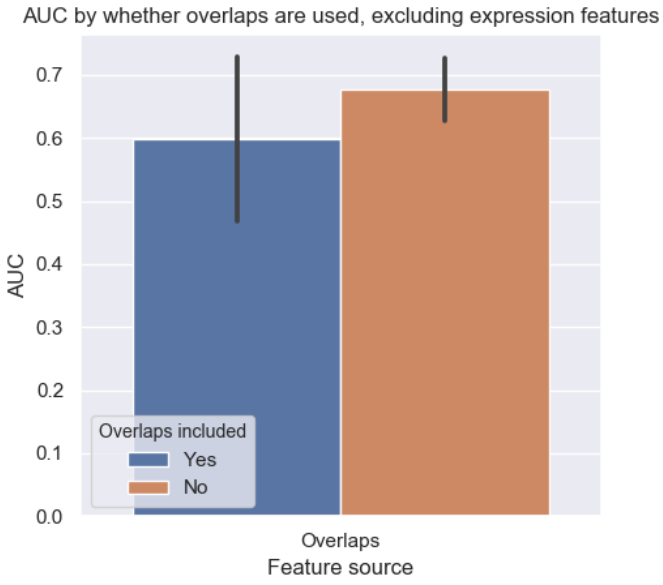

Fig. 4. Variation in AUC based on whether overlaps are included or not, with expression features excluded in both cases.

expressions. Therefore performance of this system with visual features is compared to both.

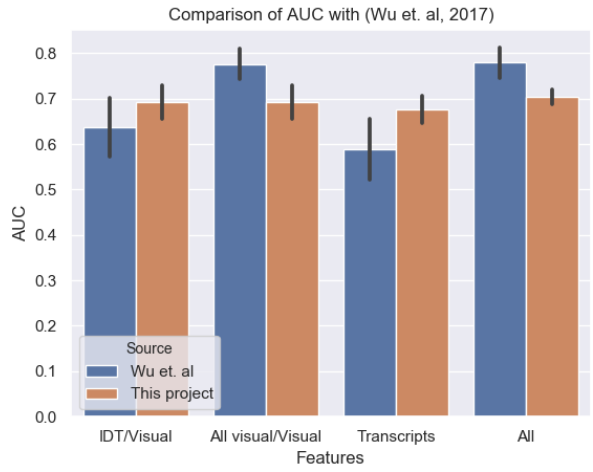

Fig. 5. Comparison between AUCs achieved in this project and those produced by DARE. The values for this project are the early fusion results using expression and gaze features.

For the textual modality alone, this system outperforms DARE ($p = 0.022$). This is surprising, since DARE uses a more sophisticated type of feature (GloVe vectors). Possibly this can be explained by use of better hyperparameters on our part. DARE achieves significantly better results in both cases where it uses microexpression features ($p = 0.0068$, $p = 0.00090$ for All visual/Visual and All/All respectively) but our system is non-significantly better when compared with IDTs. Overall our results appear comparable with the state-of-the-art as represented by DARE.

V. Conclusion

We have presented a system for automated deception detection in videos with multi-modal data, which uses novel features (expression/overlap and gaze vectors) in the visual modality. It performs similarly to the state-of-the-art even though it requires slightly less human input, and therefore outperforms human performance on the dataset (which was studied in [3] and found to be similar to chance in keeping with results from previous studies of human deception detection).

Acknowledgements

I would like to thank my supervisor Dr Marwa Mahmoud for her help and guidance.

References

[1] M. Vicianova, "Historical techniques of lie detection," *Europe's Journal of Psychology*, vol. 11, no. 3, 2015.

[2] J. Charles F. Bond and B. M. DePaulo, "Accuracy of deception judgments," *Personality and Social Psychology Review*, vol. 10, no. 3, pp. 214–234, 2006. PMID: 16859438.

[3] V. Pérez-Rosas, M. Abouelenien, R. Mihalcea, and M. Burzo, "Deception detection using real-life trial data," in *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, ICMI '15, (New York, NY, USA), pp. 59–66, ACM, 2015.

[4] Z. Wu, B. Singh, L. S. Davis, and V. S. Subrahmanian, "Deception detection in videos," *CoRR*, vol. abs/1712.04415, 2017.

[5] R. Bull, *Training to detect deception from behavioural cues: attempts and problems*, p. 251–268. Cambridge University Press, 2004.

[6] M. O'Sullivan and P. Ekman, *The wizards of deception detection*, p. 269–286. Cambridge University Press, 2004.

[7] N. R. Council, *The Polygraph and Lie Detection*. Washington, DC: The National Academies Press, 2003.

[8] P. Ekman and M. O'Sullivan, "Who can catch a liar?," *The American psychologist*, vol. 46, pp. 913–20, 10 1991.

[9] P. Ekman and W. V. Friesen, "Nonverbal leakage and clues to deception," *Psychiatry*, vol. 32, no. 1, pp. 88–106, 1969. PMID: 27785970.

[10] M. L. Newman, J. W. Pennebaker, D. S. Berry, and J. M. Richards, "Lying words: Predicting deception from linguistic styles," *Personality and Social Psychology Bulletin*, vol. 29, no. 5, pp. 665–675, 2003. PMID: 15272998.

[11] R. B. Mimansa Jaiswal, Sairam Tabibu, "The truth and nothing but the truth: Multimodal analysis for deception detection," *3D Digital Imaging and Modeling, International Conference on*, pp. 938–943, 2016.

[12] H. Karimi, "Interpretable multimodal deception detection in videos," in *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, ICMI '18, (New York, NY, USA), pp. 511–515, ACM, 2018.

[13] L. Su and M. D. Levine, "High-stakes deception detection based on facial expressions," in *2014 22nd International Conference on Pattern Recognition*, pp. 2519–2524, Aug 2014.

[14] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L. Morency, "Openface 2.0: Facial behavior analysis toolkit," in *2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018)*, pp. 59–66, May 2018.

[15] M. d. Berg, O. Cheong, M. v. Kreveld, and M. Overmars, *More Geonetric Data Structures*, ch. 10, pp. 220–226. Santa Clara, CA, USA: Springer-Verlag TELOS, 3rd ed. ed., 2008.

[16] 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. Spontaneous Facial Behaviour Analysis.

[17] B. M. DePaulo and W. L. Morris, *Discerning lies from truths: behavioural cues to deception and the indirect pathway of intuition*, p. 15–40. Cambridge University Press, 2004.

[18] scikit-learn developers, "Tf-idf term weighting." https://scikit-learn.org/stable/modules/feature_extraction.html#tfidf-term-weighting, 2018. [Online; accessed 3-May-2019].