# Deep Learning Driven Multimodal Fusion For Automated Deception Detection

Mandar Gogate, Ahsan Adeel, Amir Hussain
CogBID Lab, Department of Computing Science and Mathematics
University of Stirling, Stirling FK9 4LA, UK
Email: {mgo,aad,ahu}@cs.stir.ac.uk

***Abstract*—Humans ability to detect lies is no more accurate than chance according to the American Psychological Association. The state-of-the-art deception detection methods, such as deception detection stem from early theories and polygraph have proven to be unreliable. Recent advancement in deception detection includes the application of advanced data analysis and machine learning algorithms. This paper presents a novel deep learning driven multimodal fusion for automated deception detection, incorporating audio cues for the first time along with the visual and textual cues. The critical analysis and comparison of the proposed deep convolutional neural network (CNN) based approach with the state-of-the-art multimodal fusion methods have revealed significant performance improvement up to 96% as compared to the 82% prediction accuracy reported in the recent literature.**

***Keywords*—Deception Detection, Multimodal Fusion, Deep Convolutional Neural Network**

## I. Introduction

Every deception detection study conducted since 1986 has demonstrated that humans ability to detect lies is no more accurate than chance [1][2][3][4][5][6]. There are very few people who claim to be really good at detecting deception. However, they are only correct somewhere around 60% of the time, even in that case, it is extremely risky to make them sit in a jury judging you [7]. Moreover, accurate deception detection is critical for police officers who are responsible to detain criminals most of the time and they must not detain innocent suspects.

The state-of-the-art practices in deception detection include deception detection stem based on early lying theories where it is assumed that liars exhibit stress-based cues when they are scared of being guilty [8]. In criminal settings, one of the standard deception detection methods include the polygraph test. The polygraph based deception detection requires the use of skin-contact devices and human expertise. However, the decisions are subject to error and bias [9] and it is not very difficult for subject to deceive these devices and human experts.

The advancement in deception detection could revolutionise military/public/private/law enforcement investigation performances. Given early findings, researchers have proposed new strategies to help investigation agencies and police to catch liars involved in the act. One of the advanced deception detection methods include the use of advanced machine learning algorithms using a number of modalities such as speech [10][11] and text [12]. However, one of the major challenges in automated deception detection is the generation or availability of corpuses. Most of the existing deception detection corpuses are based on acted or artificially collected data where subjects are asked to narrate stories in deceptive and truthful manner [4][5][13]. Consequently, such acted corpuses lack real-world evidence and true emotions.

Recently, the authors in [4] developed a new multimodal deception dataset based on real-life scenarios including both verbal and nonverbal features. In particular, the authors addressed the identification of deception in real-life trial by collecting videos from public court trials, where deceptive and truthful behaviours were fairly observable and verifiable. More details about the dataset are comprehensively presented in [4]. To analyse the dataset, the authors in [4] used two state-of-the-art classification algorithms: Decision Trees (DT) and Random Forest (RF) and reported the classification accuracy up to 75% based on visual and textual cues. In addition, the comparative results with human capability of detecting deception in trial hearings revealed outperformance of their proposed approach as compared to the human capability of identifying deceit. The authors in [5] extended the work proposed in [4] by using Support Vector Machine (SVM) and showed up to 82% deception detection accuracy. However, the authors in both [5] and [4] applied manual annotation and excluded the use of audio cues, which has proven to be very important feature for the optimisation of deception detection methodology.

This paper extends the work presented in [4] and [5] by fusing audio, visual, and textual cues and introducing a deep learning driven multimodal fusion for automated deception detection. In particular, the main contributions of this paper are:

1) First time use of audio cues (to the best of our knowledge) for deception detection, achieving the accuracy of 87.5%
2) Application of deep CNN to textual cues, achieving the accuracy of 83.78% as compared to the 60.33% accuracy of the state-of-the-art model [4]
3) Application of 3D deep CNN to visual cues, achieving the accuracy of 78.57% as compared to the 76.03% accuracy of the state-of-the-art model [4]
4) Fusion of audio (A), visual (V), and textual (T) features and comparison with the maximum achieved 82% accuracy of the state-of-the-art V+T multimodal fusion [5]. In particular, following fusion combinations are explored:
    - A+V+T early fusion, achieving the prediction accuracy of 96.42%

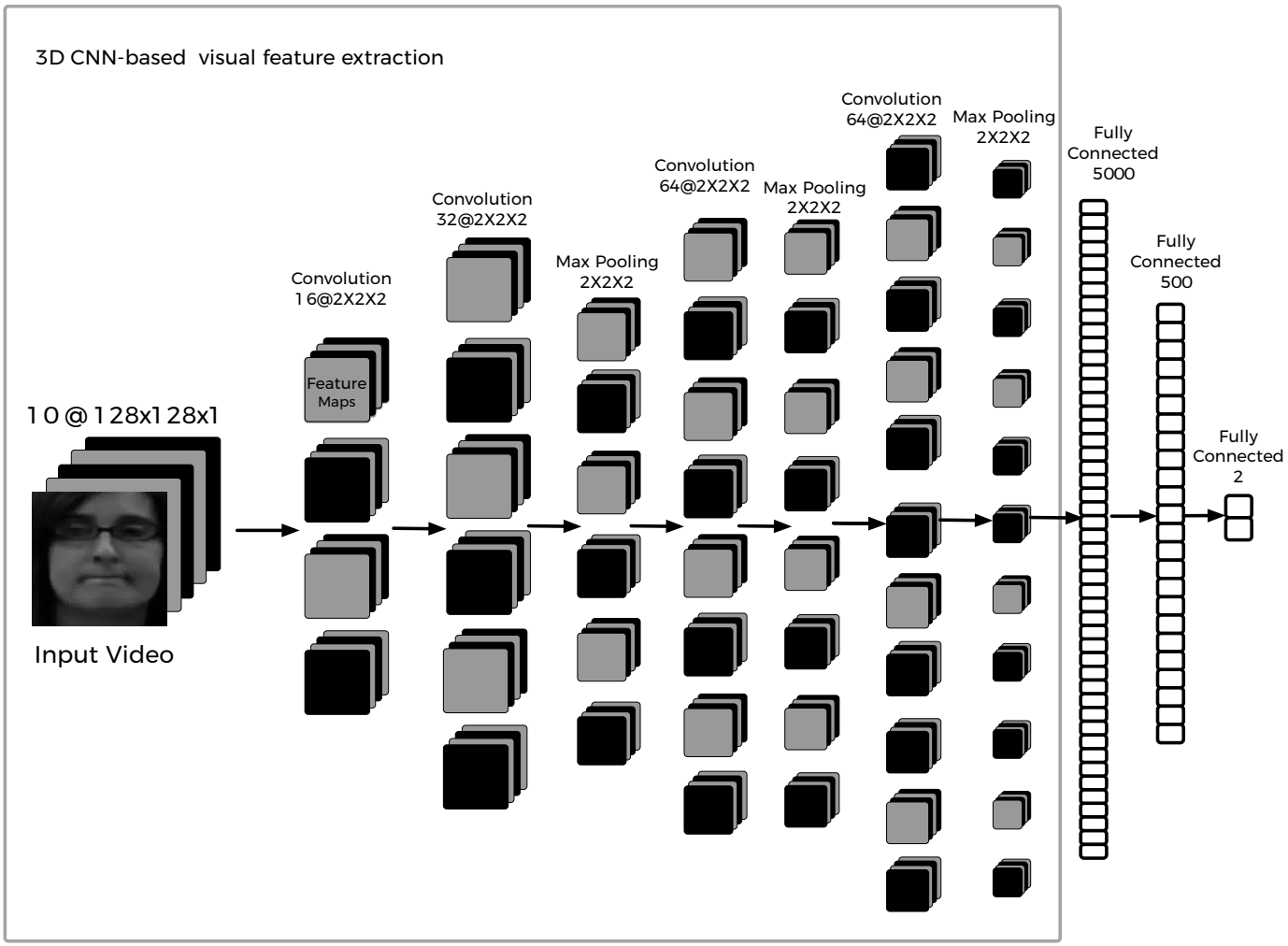


Fig. 1. 3D-CNN architecture: Visual Cues based Deception Detection

- A+V+T late fusion, achieving the prediction accuracy of 92.01%
- V+T early fusion, achieving the prediction accuracy of 91.89%
- A+T early fusion, achieving the prediction accuracy of 91.80%
- A+V early fusion, achieving the prediction accuracy of 89%
- A+T late fusion, achieving the prediction accuracy of 87%
- V+T late fusion, achieving the prediction accuracy of 86%
- A+V late fusion, achieving the prediction accuracy of 85%

It is to be noted that all the proposed multimodal fusion approaches have outperformed the state-of-the-art approaches presented in [4][5]. The rest of the papers is organised as follows: Section 2 provides a brief overview of unimodal feature extraction. Section 3 describes early and late multimodal fusion approaches. Section 4 presents the used dataset, experimental results, and detailed analysis. Finally Section 5 concludes this paper.

## II. Unimodal Feature Extraction

### A. Visual Feature Extraction: 3D-CNN

Visual Features are extracted from the videos using a 3D-CNN [14] shown in Fig. 1. 3D-CNN was used due to its inherent ability to extract both spatial (intra frame) as well as temporal (inter frame/contextual) features from video.

In the literature, 3D-CNN has been widely used for classification of volumetric data achieving state-of-the-art performance on various tasks such as human action recognition [14] and video classification [15]. This ability motivated us to adopt 3D-CNN in our framework.

Let $video \in \mathbb{R}^{f \times h \times w \times c}$, where f is the number of frames, h is the height of the frame, w is the width of the frame and c is the number of channels in an image (in our case c=1 since we consider only black & white frames). Let convolution $kernel \in \mathbb{R}^{k_m \times k_d \times k_h \times k_w \times c}$ where $k_h$ and $k_w$ are height and width of the kernel respectively, $k_d$ is depth of the kernel and $k_m$ is the number of feature maps. The max pooling $window \in \mathbb{R}^{p_d \times p_h \times p_w}$ where $p_d$, $p_h$ and $p_w$ are temporal depth, height, and width of the pooling window respectively.

In experiments, best results were obtained using a 10-

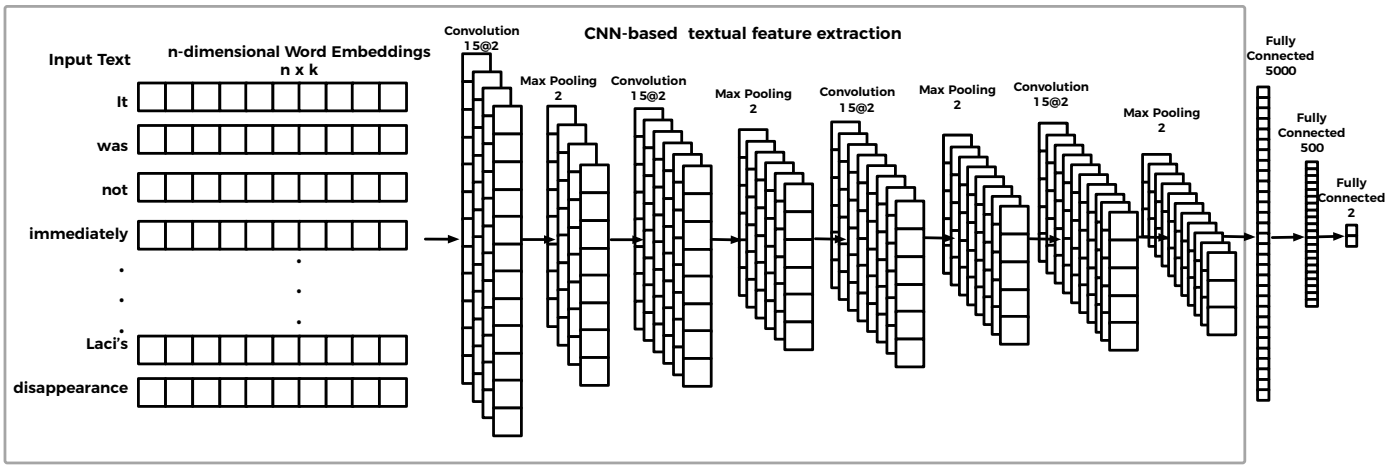


Fig. 2. CNN architecture: Textual Cues based Deception Detection

layered 3D-CNN architecture as shown in Fig. 1. In the first convolution layer, 16 feature maps ($k_m$) with kernel size of 2 x 2 x 2 ($k_d$ x $k_h$ x $k_w$) are used. In the second convolution layer, 32 feature maps ($k_m$) with kernel size of 2 x 2 x 2 ($k_d$ x $k_h$ x $k_w$) are used. First two layers are followed by a max pooling layer with window size 1 x 2 x 2 ($p_d$ x $p_h$ x $p_w$). In the fourth convolution layer, 64 feature maps ($k_m$) with kernel size of 2 x 2 x 2 ($k_d$ x $k_h$ x $k_w$) are used. In the subsequent layer, max pooling with window size 2 x 2 x 2 ($p_d$ x $p_h$ x $p_w$) is used. The max pooling is followed by convolution with 64 feature maps ($k_m$) with kernel size of 2 x 2 x 2 ($k_d$ x $k_h$ x $k_w$). The final convolution is followed by a max pooling layer with window size 1 x 2 x 2 ($p_d$ x $p_h$ x $p_w$). This feature extraction framework is followed by fully connected layers of size 5000, 500 and 2. The activation values of final max pooling layer were used as features for fusion experiments.

### B. Textual Feature Extraction: text-CNN

For extracting features from textual modality, a CNN model has been used as shown in Fig. 2. Each utterance is represented as a concatenation vector of constituent words. Each utterance is either trimmed with a window of 100 words or zero padded at the end depending on the number of words in it. Words are converted to vectors using 300-dimensional GloVe word representation [16], which is trained on 840 billion words obtained from web crawling. The final CNN model has the input dimension of $300 \times 100$ .

Convolution filters are applied to these concatenated utterances. The network has total 11 layers: 4 convolution layers, 4 max pooling and 3 fully connected layers. Each convolution layer has a filter size equal to 2 and 15 feature maps. Each convolution layer is followed by a max polling layer with window size 2. The last max pooling layer is followed by fully connected layers of size 5000, 500 and 2. Rectified Linear Unit [17] was used as a activation function for fully connected layers of size 5000 and 500. For final layer, softmax activation was employed. The network learns the abstract representation of the utterances with implicit semantic information, which spans over number of words with each successive layer and finally the entire utterance.

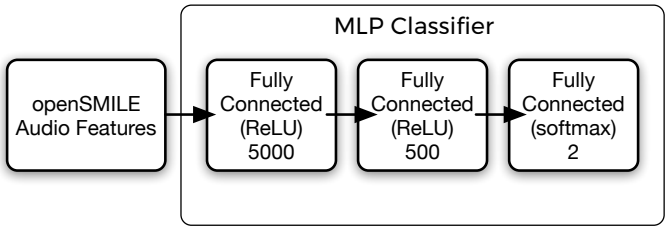


Fig. 3. Acoustic Cues based Deception Detection

### C. Acoustic Feature Extraction: openSMILE

The audio features, such as pitch and voice intensity, are extracted using widely used open-source software openSMILE [18]. The features are extracted at frame rate of 30Hz and 100ms sliding window. Voice normalisation is performed using Z-standardisation. The extracted features consist of several low level descriptors (e.g. Mel-frequency cepstral coefficients (MFCCs), intensity, pitch, loudness etc.) and their statistics (e.g. mean, variance, skewness, root quadratic mean, etc.)

In total 6373 features are extracted using the state-of-the-art feature set for paralinguistic recognition, specifically Interspeech 2013 Computational Paralinguistics ChallengE (ComParE) feature set. This feature extraction framework is followed by fully connected layers of size 5000, 500 and 2 as shown in Fig. 3.

## III. Multimodal Fusion

Recently the multimodal fusion has gained attention of many researchers mainly due to its ability to outperform unimodal systems. There are two most widely used strategies in multimodal fusion: (1) Feature Level or Early Fusion (2) Decision Level or Late Level. This section has discussed various approaches for fusing the information including early fusion and late fusion.

### A. Early fusion

In early or feature-level fusion, the features are first extracted from input data using either state-of-the-art feature extraction algorithms or deep neural network based automated

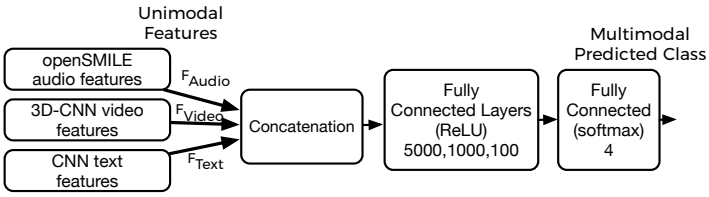


Fig. 4. Early (Feature-level) Fusion

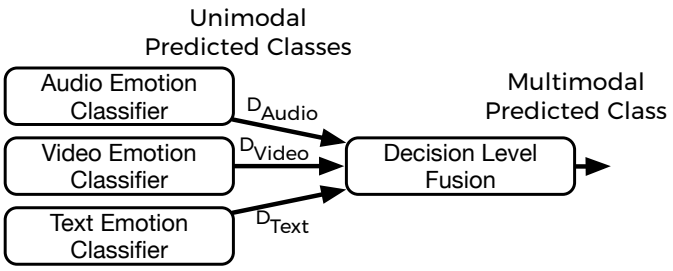


Fig. 5. Late (Decision Level) Fusion

feature extraction. Then, features are concatenated and feeded into a classifier. Let $F_m$ be the feature extracted from modality $m$. For example, as depicted in Fig. 4, the audio ($F_{Audio}$), visual ($F_{Video}$) and textual ($F_{Text}$) features are extracted using a openSMIlE feature extractor [18], 3D-CNN and CNN respectively. The features are then concatenated and feeded into an MLP classifier to classify the multimodal input into 2 categories (truthful or deceptive).

The feature level fusion is advantageous because it utilises the correlation between multiple features at an early stage which often leads to better task accomplishment. However, it is challenging to combine multimodal features as each modal is acquired at different frame rate. In addition, uneven features dimension lead to non-uniform neural network weights distribution which further leads to poor collective learning.

### B. Late (Decision-Level) fusion

In the late fusion, the unimodal classifiers are used to identify local class prediction for each modality. The local predictions are fused into a single vector which is further classified to obtain the final decision.

The late fusion strategy has numerous advantages over early fusion. For instance in early fusion, it is challenging to fuse modalities which are of different sizes/dimensions but in late fusion local decision have same representation which makes the fusion much easier. However, the major disadvantage of late fusion lies in its poor utilisation of the feature-level correlation between modalities.

In decision level fusion, local decisions ($D_{modality}$) are fused together in one feature vector. For example, as depicted in Fig. 5(a), separate audio, video and text classifiers can be trained to obtain $D_{Audio}$, $D_{Video}$ and $D_{Text}$. In the experiments, we concatenated all the unimodal predicted labels and an MLP classifier is trained to obtain the final predicted label.

## IV. Experimental Results

### A. Real-life Trial Dataset

In this paper, a Real-life Trial corpus [4] has been used. The dataset contains 121 real-life trial videos (61 deceptive and 60 truthful). The medium of conversation in all videos is English. Each video is labeled into two categorical labels: deceptive and truthful. A subset of the dataset is used to train (70% training dataset) the classifier and rest of the dataset is used to test the performance of the trained classifier in face of new context (30% testing dataset).

### B. Unimodal Deception Detection

*1) Video:* The utterance videos are first converted into grey scale from RGB. For each of the utterance, frames are extracted, normalised and combined into a single five dimensional vector (number of utterances $\times$ number of frames $\times$ frame height $\times$ frame width $\times$ number of channels). The combined vector is then fed to a 3D-CNN architecture depicted in Fig. 1. The network is trained using RMSProp optimiser [17]. The trained 3D-CNN network weights are saved and used for extracting features for fusion experiments.

*2) Audio:* The audios are extracted from videos and fed into openSMILE [18] feature extraction software. The features are extracted at frame rate of 30Hz and 100ms sliding window. Voice normalisation is performed using Z-standardisation. In total, 6373 features are extracted using state-of-the-art feature set for paralinguistic recognition, specifically Interspeech 2013 Computational Paralinguistics ChallengE (ComParE) feature set. These features are combined and fed into an MLP classifier as depicted in Fig. 3.

*3) Text:* The transcription of spoken words provided with the dataset are represented as a concatenation vector of constituent words. Each utterance is either trimmed with a window of 100 words or zero padded at the end depending on the number of words. Words are converted to vectors using 300-dimensional GloVe word representation [16] trained on 840 billion words from common web crawling. The concatenated word representations are then fed to CNN architecture depicted in Fig. 2. The network is trained using RMSProp optimiser [17]. The trained CNN network weights are saved and used for extracting features for fusion experiments.

The transcription of spoken words provided with the dataset are represented as a concatenation vector of constituent words. Each utterance is either trimmed with a window of 100 words or zero padded at the end depending on the number of words. Words are converted to vectors using 300-dimensional GloVe word representation [16], trained on 840 billion words obtained from web crawling. The final CNN model has the input dimension of $300\times1000$. The concatenated word representations are then fed into a CNN architecture depicted in Fig. 2. The network is trained using RMSProp optimiser [17]. The trained CNN network weights are saved and used for extracting features for fusion experiments.

The architectures are trained and validated using TensorFlow library and NVIDIA Titan Xp GPU with 12GB of GDDR5X memory.

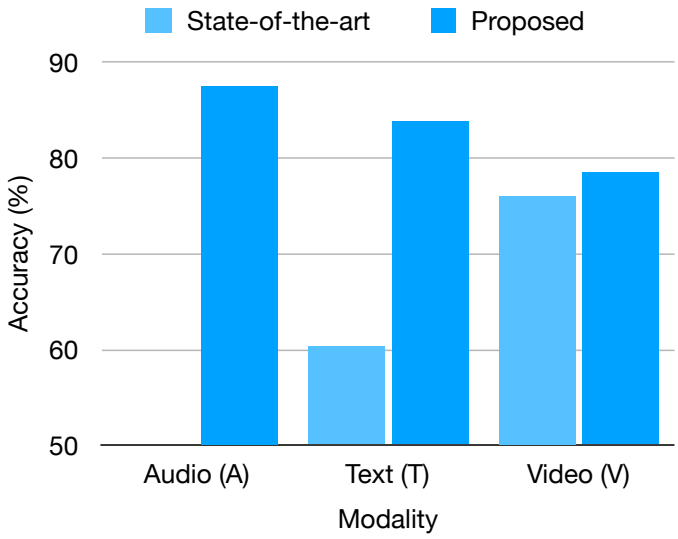


Fig. 6. Prediction Accuracy: Comparison of proposed unimodal deception detection with state-of-the-art approach

TABLE I. PREDICTION ACCURACY: COMPARISON OF PROPOSED UNIMODAL DECEPTION DETECTION WITH THE STATE-OF-THE-ART

| **Modality** | Prez-Rosas et al. [4] | **Proposed (Unimodal)** |
|---|---|---|
| Video (V) | 76.33% | **78.57%** |
| Audio (A) | - | **87.5%** |
| Text (T) | 60.33% | **83.78%** |

TABLE II. PREDICTION ACCURACY: COMPARISON OF PROPOSED LATE AND EARLY MULTIMODAL FUSION WITH THE STATE-OF-THE-ART

| **Modality** | **Fusion** | | **State-of-the-art** |
|---|---|---|---|
| | **Late** | **Early** | [5] |
| A + V | 85% | 89.1% | - |
| V + T | 87% | 91.8% | 82% |
| A + T | 86% | 91.9% | - |
| A + V + T | 92% | 96.4% | - |

### C. Performance of Unimodal Architectures

The unimodal deception detection classifier for video, text, and audio are depicted in Figs. 1, 2 and 3, respectively. The accuracy performance of the unimodal classifiers are compared with the state-of-the-art deception detection classifier proposed in [5]. The results are summarised in Table I. It is to be noted that for text and video modalities, the proposed unimodal architectures have outperformed the state-of-the-art approaches by 23% and 2% respectively. In addition, audio modality achieved the accuracy of 87.5% which is more than the accuracy of both visual and textual modalities.

### D. Performance Comparison of Proposed Multimodal Early and Late Fusion Approaches

In this subsection, the proposed multimodal early and late fusion approaches are compared with the state-of-the-art fusion approaches for deception detection. The simulation results are presented in Fig. 7 and Table III. It is to be noted that the state-of-the-art approach presented in [4] and [5] considered only visual and textual multimodal cues.

In contrast, this paper has considered all the possible fusion combinations including A+V, A+T, T+V, A+V+T. In all the fusion combinations, early fusion outperformed late fusion and state-of-the-art approaches, whereas the late fusion outperformed the state-of-the-art approach as well. In particular, the early fusion performed 4%, 5%, 5%, and 4% better than late fusion in the case of A+V, A+T, T+V, A+V+T respectively. In comparison with the state-of-the-art T+V fusion, the proposed late and early fusion outperformed it by 5% and 9.2% respectively.

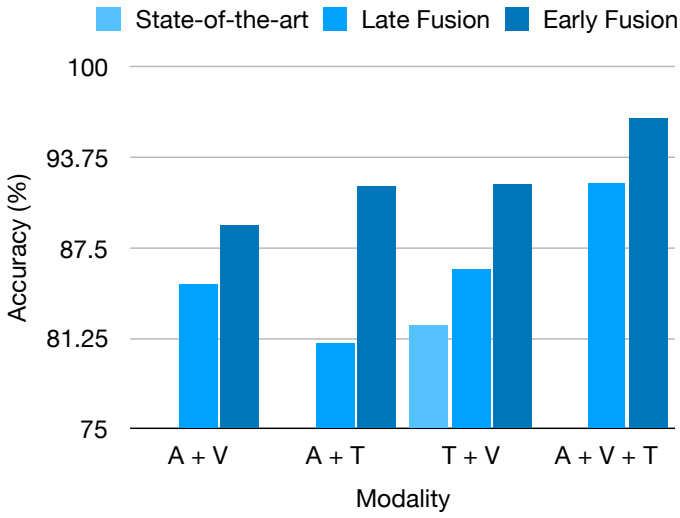


Fig. 7. Prediction accuracy: Comparison of proposed early and late multimodal fusion approaches with the state-of-the-art[5] Note: A + V, A + T and A + V + T were not reported in [5]

Similar trend is evident in Table III, where precision, recall, and F1 scores for both late and early fusions are presented. Consequently, in all the experiments bimodal and trimodal deception classifiers have outperformed unimodal classifiers.

## V. CONCLUSION AND FUTURE WORK

In the recent literature, researchers have proposed several deception detection approaches based on manual annotation, incorporating only textual and visual cues. To the best of our knowledge, this work is the first attempt to develop a fully automated multimodal deception detection approach, fusing audio, visual and textual features. The simulation results and critical performance analysis of the proposed unimodal deception detection models showed that: (1) the audio based deception detection model achieved the prediction accuracy of 87.5% (2) the automated extracted textual cues based deep CNN approach achieved the prediction accuracy of 83.78% as compared to the prediction accuracy of 60.33% presented in [4] (3) the visual based 3D deep CNN achieved the accuracy of 78.57% as compared to the 76% accuracy in manual annotation based approach [4]. The simulation results of our proposed multimodal early and late fusion approaches, incorporating audio, visual, and textual features achieved the highest accuracy of 92% and 96% as compared to the prediction accuracy of 82% achieved by the state-of-the-art approach, where only visual and textual cues were considered. In future, we intend to extend our work by applying advanced attention-based deep learning approaches.

TABLE III. EARLY VS. LATE FUSION: COMPARISON OF PREDICTION ACCURACY, PRECISION, RECALL AND F1 SCORE

| Modality | Late Fusion | | | | Early Fusion | | | |
|---|---|---|---|---|---|---|---|---|
| | Accuracy(%) | Precision | Recall | F1-score | Accuracy(%) | Precision | Recall | F1-score |
| A + V | 85.0% | 0.88 | 0.85 | 0.84 | 89.1% | 0.9 | 0.87 | 0.88 |
| V + T | 87.0% | 0.87 | 0.86 | 0.86 | 91.8% | 0.91 | 0.90 | 0.90 |
| A + T | 86.0% | 0.83 | 0.81 | 0.81 | 91.9% | 0.92 | 0.89 | 0.90 |
| A + V + T | 92% | 0.92 | 0.92 | 0.92 | 96.42% | 0.96 | 0.95 | 0.95 |

## ACKNOWLEDGMENT

This research is funded by the University of Stirling IMPACT Collaborative Research Scholarship. We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp GPU used for this research. The authors would also like to gratefully acknowledge the support of UK Engineering and Physical Sciences Research Council (EPSRC) Grant No. EP/M026981/1 (CogAVHearinghttp://cogavhearing.cs.stir.ac.uk).

## REFERENCES

[1] P. Granhag and L. Strömwall, *The Detection of Deception in Forensic Contexts*. Cambridge University Press, 2004. [Online]. Available: https://books.google.co.uk/books?id=5GjTLtibmXYC

[2] P. Ekman and M. O'sullivan, "Who can catch a liar?" *American psychologist*, vol. 46, no. 9, p. 913, 1991.

[3] S. Mann, A. Vrij, and R. Bull, "Detecting true lies: police officers' ability to detect suspects' lies." *Journal of applied psychology*, vol. 89, no. 1, p. 137, 2004.

[4] 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, Seattle, WA, USA, November 09 - 13, 2015*, 2015, pp. 59–66. [Online]. Available: http://doi.acm.org/10.1145/2818346.2820758

[5] V. Pérez-Rosas, M. Abouelenien, R. Mihalcea, Y. Xiao, C. J. 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, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, 2015, pp. 2336–2346. [Online]. Available: http://aclweb.org/anthology/D/D15/D15-1281.pdf

[6] M. Abouelenien, V. Pérez-Rosas, R. Mihalcea, and M. Burzo, "Detecting deceptive behavior via integration of discriminative features from multiple modalities," *IEEE Transactions on Information Forensics and Security*, vol. 12, no. 5, pp. 1042–1055, 2017.

[7] L. Zimmerman, "Deception detection," in *American Psychological Association, Vol 47, No. 3*, 2016, p. 46. [Online]. Available: http://www.apa.org/monitor/2016/03/deception.aspx

[8] R. Adelson, "Detecting deception," in *American Psychological Association, Vol 35, No. 7*, 2004, p. 70. [Online]. Available: http://www.apa.org/monitor/2016/03/deception.aspx

[9] A. Vrij, *Detecting lies and deceit: Pitfalls and opportunities*. John Wiley & Sons, 2008.

[10] J. Hirschberg, S. Benus, J. M. Brenier, F. Enos, S. Friedman, S. Gilman, C. Girand, M. Graciarena, A. Kathol, L. Michaelis *et al.*, "Distinguishing deceptive from non-deceptive speech," in *Ninth European Conference on Speech Communication and Technology*, 2005.

[11] 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.

[12] S. Feng, R. Banerjee, and Y. Choi, "Syntactic stylometry for deception detection," in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2*. Association for Computational Linguistics, 2012, pp. 171–175.

[13] R. Mihalcea and C. Strapparava, "The lie detector: Explorations in the automatic recognition of deceptive language," in *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*. Association for Computational Linguistics, 2009, pp. 309–312.

[14] S. Ji, W. Xu, M. Yang, and K. Yu, "3d convolutional neural networks for human action recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 1, pp. 221–231, 2013.

[15] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2014, pp. 1725–1732.

[16] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1532–1543. [Online]. Available: http://www.aclweb.org/anthology/D14-1162

[17] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural computation*, vol. 18, no. 7, pp. 1527–1554, 2006.

[18] 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.