Deception Detection System with Joint Cross-Attention

[#]Peili Jiang[a], [#]Yunfan Wang[b], [#]Jiajun Li[c], [#]Ziyang Wang[d]

[a]New York University
College of Arts and Science
pj2097@nyu.edu

[b]The University of New South Wales
Faculty of Engineering
z5320828@ad.unsw.edu.au

[c]University of California-San Diego
Earl Warren College
jil186@ucsd.edu

[d]Carnegie Mellon University
Mellon College of Science
ziyangwa@andrew.cmu.edu

[#]These authors contributed equally to this work

Abstract

Recently, deception detection has been a new topic in biometrics. According to criminal psychology, there are many useful cues on human face and tone of voice, and we use these clues for deception detection. In this work, we propose a new deceptive system that combines 68 facial landmarks dis-placement, action unit (AU), and audio emotion unit (EU). It uses the clues proposed by criminal psychology to determine the facial expression changes in sequential frames and analyze the emotion changes with audio. Found that facial tensor may be a powerful feature for deception that uses 68 facial landmarks to calculate the point changes between sequential frames. Finally, we use the public deceptive dataset, Real-life, Bag-of-lies and private dataset MSPL-YTD to verify our system. Overall, the proposed method is a good candidate for intelligent deception detection.

1. Introduction

Despite the significant advancements in human cognition, our ability to discern deceit is merely as reliable as pure chance, akin to tossing a coin. This skill finds relevance among various groups, such as students, psychologists, judges, interviewers, and law enforcement professionals [1]. Especially in crime investigations, the precision in spotting lies is paramount for officers aiming to apprehend the guilty while safeguarding the innocent. Conventional lie detection theories suggest that dishonest individuals may inadvertently reveal certain cues due to the psychological burden of deceit. This belief has spurred researchers to hunt for consistent behavioral markers of dishonesty. Although some methods emphasize shifts in posture or minor limb movements, their efficacy remains questionable. Training law enforcement using extensive case studies is not only challenging but also fails to ensure unbiased judgment. Our research introduces a technique that leverages computer vision to identify deceit through facial indicators. This ensures consistent, bias-free evaluations.

While lie detection can employ both contact and non-contact methods, contact methods like the polygraph and fMRI focus on involuntary physiological responses such as heart rate and skin conductivity [2]. However, these methods face criticism for their reliability, especially when the individual being tested is aware of the deception assessment. The need for wearing equipment further limits their convenience.

Our research emphasizes non-contact deception detection methods. Reference [4] introduced eye-tracking technology that shifts its focus from emotional reactions, akin to the polygraph, to cognitive reactions. Techniques like voice risk or stress analysis [5-8] deploy computers to analyze vocal attributes to infer deceit. Pérez-Rosas et al. [9] integrated linguistic and gestural features, yet overlooked crucial visual cues like facial expressions. Jaiswal et al. [10] and others [28-30] incorporated visual, auditory, and textual data to discern micro-expressions. Nevertheless, many of these techniques, particularly voice risk analysis, falter in the presence of low volumes or loud background noises. Despite their innovations, a common limitation among them is their heavy reliance on verbal, non-verbal, and vocal indicators. The advent of visual devices allows for capturing invaluable facial and behavioral cues. Coupled with the right algorithms, this can significantly enhance detection capabilities.

Two essential attributes of any biometric identification system are its accuracy and user-friendliness [14, 15]. Our study delves into facial analysis for lie detection, addressing the challenges mentioned earlier. In our approach, we detect facial landmarks [11] and derive geometrical features from the face, examining facial action units and charting the temporal patterns of facial movements. These action units gauge movements based on specific emotion-related facial landmarks. Additionally, we harness geometrical attributes to represent subjects' physiological reactions. We use Joint cross-attention model [12] in late-fusion stage, and the SVM [13] is employed for recognition. Preliminary results validate the promising potential of our method in real-world deception detection scenarios.

2. Related Work

In the realm of deception detection, two primary avenues of exploration have emerged: Verbal and Non-verbal Deception Detection. The integrated model which combined verbal and nonverbal deception detection advents afterward.

2.1 Verbal Deception Detection

A wealth of research has delved into the identification of deceitful content in various domains, such as online dating websites, forums, social networks, and consumer report websites [39, 18, 42, 23, 21, 31, 24]. The efficacy of features derived from text analysis, including basic linguistic representations like n-grams and sentence count statistics, has been demonstrated, alongside more intricate linguistic features derived from syntactic CFG trees and part of speech tags [28, 13, 43]. Some studies have incorporated the analysis of psycholinguistic aspects using the Linguistic Inquiry and Word Count (LIWC) lexicon to build deception models through machine learning approaches, revealing the value of psycholinguistic information in automatic deceit identification. Furthermore, researchers have explored the link between text syntactic complexity and deception, hypothesizing that deceivers might employ simpler sentences to conceal their falsehoods and facilitate the recall of their lies [44].

While most prior research relied on controlled data collection settings, only a few works have ventured into the realm of real-life scenarios due

to the challenges associated with obtaining and verifying the nature of real-world data. An example of such research is presented by Fornaciari and Poesio, focusing on the identification of deception in statements made by witnesses and defendants in Italian courts. In line with this, our study delves into deception detection using real-life trial data and explores the utilization of multiple modalities for this purpose.

2.2 Non-verbal Deception Detection

Historically, non-verbal deception detection heavily relied on polygraph tests, which assessed physiological features like heart rate, respiration rate, and skin temperature. However, studies have shown that relying solely on physiological measurements can be biased and misleading. Recent approaches have explored non-verbal audio cues, audio-visual recordings, and thermal variations to detect deceit in scenarios ranging from casual games to criminal suspect interrogations. Hand gestures and facial expressions have also been scrutinized as indicators of deception. Researchers have tracked hand movements and used geometric features related to hand and head motion to detect deceit. Similarly, they've analyzed facial expressions, micro-expressions, face orientation, and facial expression intensity to identify signs of deceit.

Recent advancements have integrated features from multiple modalities to enhance deception detection performance. A multimodal deception dataset encompassing linguistic, thermal, and physiological features has been introduced, leading to the development of multimodal deception detection systems. However, only little work has yet addressed the challenge of deception detection in real-life data across multiple modalities, a gap that our study aims to bridge.

2.3 Integrated Deception Detection

Deception detection, a critical pursuit in various settings, has traditionally explored linguistic and non-verbal cues in isolation. While prior research has examined linguistic markers, such as sentence complexity and psycholinguistic aspects, alongside non-verbal indicators like facial expressions and physiological measurements, a unique and underexplored approach emerges when we integrate these disparate elements into a unified deception detection mechanism.

In diverse domains, ranging from formal courtroom proceedings to casual online interactions, the need for more precise deception detection mechanisms is evident. Language models have revealed the significance of linguistic cues in identifying deceit, yet their performance in isolation is limited. Simultaneously, non-verbal cues, such as micro-expressions and physiological changes, offer valuable insights into deception. However, both modalities face challenges when used independently, as witnessed in the inherent biases of polygraph tests and the potential pitfalls of linguistic analysis in real-life scenarios.

The novel approach we propose is the integration of linguistic and non-verbal deception detection models, drawing upon their respective strengths and compensating for their weaknesses. While there are few multimodal deception detection systems [16] that share the same concept as we do, by carefully combining and weighing the outcomes of both linguistic and non-verbal models, our integrated mechanism holds the promise of significantly improving deception detection accuracy. This innovative approach, applicable in a wide array of formal and informal contexts, is poised to be a transformative tool for distinguishing truth from deceit with unparalleled precision.

2.4 Data fusion of picture and audio using multimodality method

Endeavor for unveiling the importance of interplay between auditory visual cues in human perceptions can be dated back to as early as 1976 [28]. Starting from here, the significance of audio-visual integration in perception has been highlighted in the realm of cognitive science. The very first successful emotion recognition by combining cv recognition of mouth region and a deep belief for audio stream through deep learning approaches revealed the feasibility of combining sources of modalities that are distinct in nature to carry on cognitive understanding. [29] A self-supervised method of such combination is done in 2018. [30]

The dissection of video to audio and visual modalities grants computer better scene understanding ability. With the development of multimodal learning techniques, tasks of cognitive understanding are making break throughs in recent years. Though the metric of arousal and valence combination is selected to address the topic, the training model is of high flexibility and can be taken to different kinds of tasks. [31]

3. Datasets

Our goal is to build a multimodal deception detection system trained and tested on real-life data, which contains both video and audio data.

3.1 Real-life Trial Dataset

The Deceptive Behavior in Court Trials Dataset [9] is a comprehensive collection of occurrences of deception during court trial proceedings. This dataset is designed to facilitate the analysis of both verbal and non-verbal behaviors in relation to deception. The data collection process focused on identifying public multimedia sources with clear constraints to ensure the quality of the collected content. To create this dataset, the team targeted trial recordings where the defendant or witness was clearly identified and had their faces visible throughout most of the recording. Additionally, the visual and audio quality was a priority to identify facial expressions and hear verbal communication effectively.

The dataset covers three trial outcomes: guilty verdicts, non-guilty verdicts, and exoneration. Deceptive clips were collected from defendants during guilty verdict trials, while truthful videos came from witnesses in the same trials. Deceptive videos also include suspects denying a crime they committed, and truthful clips from the same suspects answering questions verified by the police as truthful. Exoneration testimonies were collected as truthful statements. The dataset comprises 121 videos, with 61 deceptive and 60 truthful trial clips, with an average video length of approximately 28 seconds.

Transcriptions of the video clips were obtained through crowdsourcing on Amazon Mechanical Turk, totaling 8,055 words with an average of 66 words per transcript. Non-verbal behavior annotations focused on gestures, including facial displays and hand movements, which were annotated using the MUMIN coding scheme, designed for interpersonal communication. This annotation process was conducted by two annotators using the Elan software, and inter-annotator agreement was ensured.

3.2 Bag-of-lies dataset

The Bag-of-Lies dataset [17] is a groundbreaking resource in the field of deception detection. Unlike most existing deception datasets that rely on subjective interviews with predetermined scenarios, Bag-of-Lies offers a unique approach. This dataset captures casual

deception in an objective, real-world context, integrating multiple modalities, including video, audio, EEG, and Eye Gaze data.

This innovative dataset comprises recordings from 35 participants who were free to choose whether to be truthful or deceitful while describing a series of images. The data includes 325 annotated recordings, evenly distributed between truth and lies. By allowing participants to naturally decide whether to deceive, Bag-of-Lies provides a novel perspective on deception research.

The dataset's materials include standard smartphone equipment for video and audio recording, an EEG headset, and an Eye Gaze tracker, emulating real-life scenarios where high-definition data may not be readily available.

3.3 MSPL-YTD dataset

The YTD-18M dataset [27], short for YouTube Video Dialogue Dataset with 18 million video segments, is a pioneering resource in the field of dialogue research. It was meticulously crafted through a robust collection process, aiming to provide a vast and diverse collection of video dialogues for research and analysis.

The dataset begins by extracting and filtering public YouTube videos, resulting in a pool of 20 million videos. These videos undergo further processing to create 18 million video segments. These segments are refined to ensure that they contain substantial dialogues and do not include harmful content, setting the stage for insightful analysis.

One of the dataset's notable features is its conversion of noisy video transcripts into well-structured dialogues. Instead of relying on speaker diarization systems with potential inaccuracies, a specialized converter model is trained to transform the transcripts into organized dialogues. This process is driven by the proven capabilities of GPT-3 models, leading to high-quality dialogue generation.

To align these dialogues with the corresponding video frames accurately, Dynamic Time Warping is employed, taking into account the original transcripts' timing information. This meticulous alignment minimizes errors and enhances the dataset's utility.

4. Methods

Fig. 1 illustrates the flow of the proposed method in the training phase. First, the face localization is employed for the face detection [11] and the facial landmark extraction [11] is adopted for the required 68 facial landmark points with notations {P1, P2, ..., P67, P68} as indicated in Fig. 2, which is able to assist in extracting the specific facial features expeditiously. A vector (F) is extracted for two kinds of characteristics that are considered as features in this work as follows. The feature vector F can be categorized into two different properties: (1) Action (A) and (2) Emotion (E) Units. The description of these features is further detailed as follows. Finally, the Joint cross-attention model is applied and the SVM is employed for deception classification.

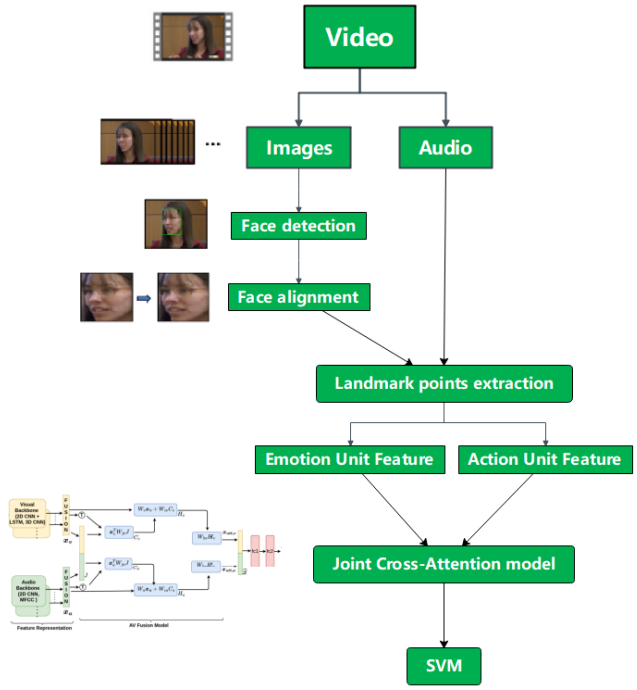


Fig. 1:The flow of the proposed method in training phase.

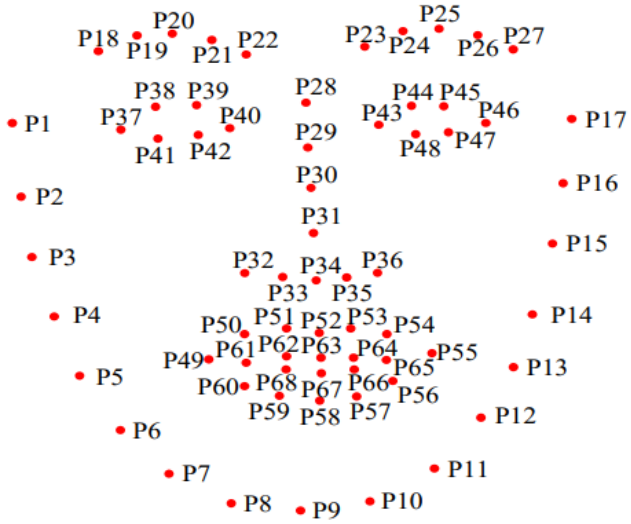


Fig. 2: 68 Facial landmarks.

4.1 Facial Action and Emotion Units

The Facial Action Coding System (FACS) [18] refers to a set of facial movements that correspond to a displayed emotion. Some examples of the Action Units are shown in Fig. 3. We can determine the displayed emotion of a participant using FACS. This is currently the only available technique for assessing emotions in real-time. Action Units have been employed to be potential observations for distinguishing the liars or truth-tellers in recent years. For instance, Porter et al. [19, 20] and Owayjan et al. [21] have shown that guilty suspects or liars make fake sadness or other emotions to cover their embarrassment when they are telling lies. Su et al. [22] have found that the potential indicators, e.g., eye blinking, eyebrow motion and mouth motion, can also distinguish liars or truth-tellers. Consequently, this study utilized FACS to measure the psychometric or deceptive tests as the true feeling in direct response of a participant.

Fig. 3: Examples of some action units extracted from Cohn and Kanades dataset [26].

The FACS can be categorized into two different properties: (1) Main Action Units (the feature A); (2) Emotions Units (the feature E). Each AU is associated with the facial movement and can affect in a motion of a part of the face or appearance changes in a facial region. In addition, multiple AUs can occur at the same time. To conclude, the proposed method aims to detect the Action Units and Emotions Units as summarized in Tables 1 and 2. Emotion Units are introduced when multiple Action Units show simultaneously. Subsequently, these potential deception indicators (AUs and EUs) are used to distinguish the deceptive and truthful suspects. These Action Units presence detection module is based on a recent state-of-the-art AU recognition framework [23, 24]. A more detailed description of the detection system can be found in Baltrusaitis et al. [24, 25]. The description of the method on feature extraction is detailed as follows.

First, the presence of each AU is extracted frame by frame, then subsequently calculate the

presence of each EU using AU simultaneously as shown in Table 2.

Table 1. Potential indicators of deception.

Action Unit	Description	Facial Region
AU1	Inner Brow Raiser	Eyebrows
AU2	Outer Brow Raiser	Eyebrows
AU4	Brow Lowerer	Eyebrows
AU5	Upper Lid Raiser	Eyes
AU6	Cheek Raiser	Eyes
AU7	Lid Tightener	Eyebrows + Eyes
AU9	Nose Wrinkler	Eyebrows + Nose
AU10	Upper Lip Raiser	Mouth
AU12	Lip Corner Puller	Mouth
AU14	Dimpler	Mouth
AU15	Lip Corner Depressor	Mouth
AU16	Lower Lip Depressor	Mouth
AU17	Chin Raiser	Mouth
AU20	Lip stretcher	Mouth
AU23	Lip Tightener	Mouth
AU26	Jaw Drop	Mouth
AU28	Lip Suck	Mouth
AU45	Blink	Eyes

Table 2. Potential indicators of deception (emotion units).

Emotion Unit	Description
AU6+12	Happiness / Joy
AU1+4+15	Sadness
AU1+2+5+26	Surprise
AU1+2+4+5+7+20+26	Fear
AU4+5+7+23	Anger
AU12+14	Contempt

Second, a binary sequence event is generated for each frame with each AU and EU, e.g., AU6, AU12 and AU6+AU12 (EU) as shown in Fig. 13. A binary sequence event where one represents the frame involving the presence of the Action Unit or Emotion Unit and zero is not involving the presence. Finally, the AU and EU are extracted as be our features. Notably, we extracted the features with a number of frames (α) in the Real-Life dataset, where α is later discussed. The features are separated into two parts which are the sum of the present event (the binary sequence shows one) and the sum of the change of the AU and EU event (the binary sequence shows one to zero or zero to one). Fig. 4 shows the sequence of the present event of the AU6, AU12, and AU6+12. The AU6+12 means that the Emotion Unit is happiness or joy which is involving the presence of AU6 and AU12 simultaneously as shown in Fig. 4. Each orange dot on the curve corresponds to a frame in Fig. 12. The sum of the presence of the AU6, AU12, and AU6+12 extracted features are 9, 21, and 6 with 60 frames, respectively. Moreover, the sum of the change of the AU6, AU12, and AU6+12 extracted features are 8, 8, and 4, respectively. In this paper, the facial action unit and emotion unit are used to generate visual vector features (A+E) for analysis of 46 dimensions.

Subsequently, all of the extracted Action Unit and Emotion Unit features are fed to Joint Cross-Attention model and then the SVM for deception classification.

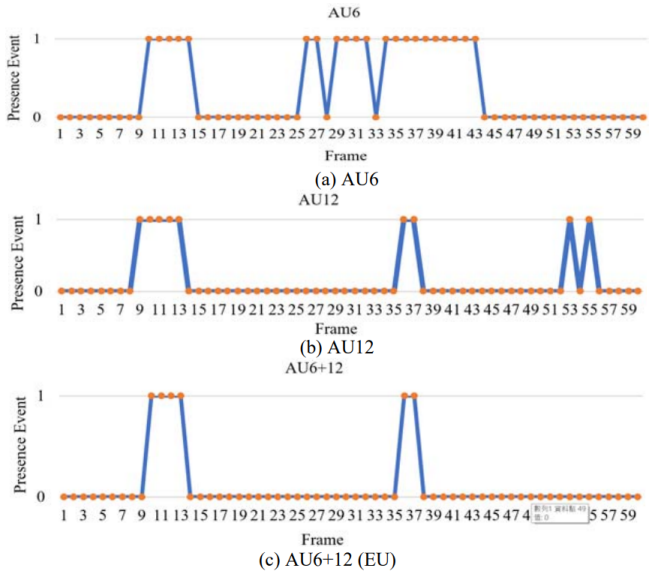


Fig. 4: A binary sequence event where one represents the frame involving the presence of the AU or EU and zero does not involving the presence.

4.2 Joint Cross-Attention model proposed for A-V fusion

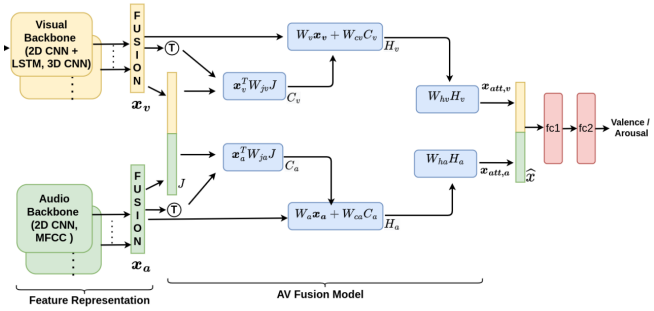


Fig. 5: Joint cross-attention model proposed for A-V fusion

While achieving A-V fusion through unified multimodal training is possible, it has been observed that the performance of multimodal systems often falls short of that of individual modalities. Several factors contribute to this, including differences in learning dynamics between A and V modalities, variations in noise levels across modalities, specialized input representations, among others. Consequently, we have independently trained deep learning models for each A and V modality to extract their respective features. These extracted features are then fed into the JCA fusion model for A-V fusion, which produces results for Emotion Unit and Action Unit and then the results are fed to the SVM for deception classification.

In the context of a given video sequence, the relevance of the V modality varies across different video clips, with the A modality being more relevant in some instances. Since multiple modalities convey diverse and complementary information for valence and arousal prediction compared to a single modality, their complementarity can be harnessed effectively through A and V fusion. To achieve reliable fusion, we rely on a cross-attention based fusion mechanism to efficiently encode intermodal information while preserving intra-modal characteristics. While cross-attention is typically applied across the features of individual modalities, our approach explores cross-attention in a joint framework. Specifically, we obtain a joint A-V feature representation by concatenating A and V features and attending to both individual A and V features. This joint representation allows each modality's features to attend to both itself and the other modality, facilitating the capture of semantic intermodal relationships across A and V. Additionally, the heterogeneity between A and V modalities can be significantly reduced by using the combined feature representation in the cross-attentional module, resulting in further improvements in system performance.

A joint cross-attention model for A-V fusion has been proposed. Conventionally, cross-attention has been applied across the features of individual modalities. In the proposed model, we use cross-attention in a joint learning framework and obtain the joint A-V feature representation by concatenating the A and V features to attend to the individual A and V features.

In the training mode, we let X_a and X_b be two sets of deep feature vectors extracted for A and V modalities in response to a given input video sub-sequence S of fixed size, where $X_a = \{x_a^1, x_a^2, ..., x_a^L\} \in \mathbb{R}^{d_a \times L}$ and $X_v = \{x_v^1, x_v^2, ..., x_v^L\} \in \mathbb{R}^{d_v \times L}$.

Here, L represents the number of non-overlapping fixed-size clip samples extracted uniformly from S, d_a and d_v represent the feature dimension of A and V representations. Now we obtain the joint representation of A-V features by concatenating the A and V feature vectors and denote it by J, i.e.,

$$J = [X_a; X_v] \in \mathbb{R}^{d \times L}.$$

Here, $d = d_a + d_v$ is the feature dimension. Then, we can compute the joint correlation matrix C_a across the A features Xa and combined A-V features J as:

$$C_a = tanh(\frac{X_a^T W_{ja} J}{\sqrt{d}})$$

where $W_{ja} \in \mathbb{R}^{L \times L}$ represents learnable weight matrix across A and joint A-V features. Similarly, the joint correlation matrix C_v is given by:

$$C_v = tanh(\frac{X_v^T W_{jv} J}{\sqrt{d}})$$

For the A modality, the joint correlation matrix C_a and the corresponding A features X_a are combined using the learnable weight matrices W_{ca} and W_a respectively to compute the

attention weights of the A modality, which is given by:

$$H_a = ReLu(W_a X_a + W_{ca} C_a^T)$$

Similarly for the V modality, we have:

$$H_v = ReLu(W_v X_v + W_{cv} C_v^T)$$

Next, the attention maps are used to compute the attended features of A and V modalities as:

$$X_{att,a} = W_{ha} H_a + X_a$$
$$X_{att,v} = W_{hv} H_v + X_v$$

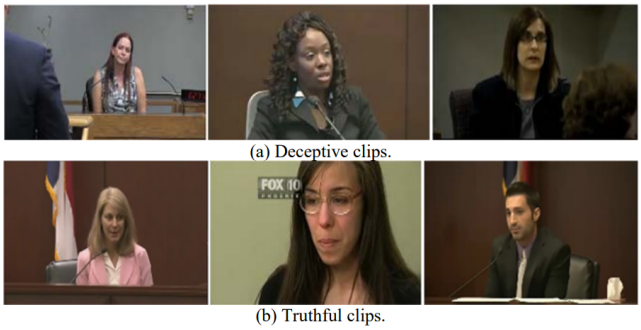
The attended A and V features, $X_{att,a}$ and $X_{att,v}$ are further concatenated to obtain the A-V features representation as:

$$X_{att} = [X_{att,v}; X_{att,a}]$$

Finally, the A-V features we obtained are fed to the fully connected layers (fc1 and fc2) for the results of Emotion Unit and Action Unit.

5. Experiments

In the following experiments, two datasets, the Real-Life Trail Data [9] and the MSPYTD, are adopted to evaluate the performance of deception detection using the proposed algorithm. The Real-Life dataset comprises of 28 deceptive and 38 truthful videos are from identified public multimedia sources, where some sample screenshots are shown in Fig. 6.



(a) Deceptive clips.

(b) Truthful clips.

Fig. 6: Sample screenshots showing facial displays from Real-Life Trail clips.

The average deceptive and truthful video lengths are 24.96 seconds and 27 seconds, respectively. The dataset composes of 21 female and 35 male speakers and are aged between 16 to 60 years. As mentioned in [9], three different trial results were utilized to correct label video clip as deceptive or truthful: guilty verdict, non-guilty verdict, and exoneration. For guilty verdicts, deceptive and truthful videos were collected from a defendant and witnesses in a trial, respectively. In some cases, deceptive videos are collected from a suspect denying a crime he committed while truthful clips are taken from the same suspect when answering questions concerning some facts that were verified by the police as truthful. Exoneration testimonies are assembled as truthful statements. On the other hand, the MSP-YTD dataset consists of 145 videos including 62 deceptive and 83 truthful videos sourced from various YouTube channels. The average video lengths of the deceptive and truthful are 9.9 sec. and 5.1 sec., respectively. The database consists of 15 female and 20 male participants. The video includes a clip of celebrity called a press conference but was verified to be a fraud later on by the police, a clip of polygraph testing the participants were lying or not and a clip of children to lie cause by some incidents. Some sample screenshots of the MSP-YTD dataset are shown in Fig. 7.

(a) Deceptive videos.

(b) Truthful videos.

Fig. 7: Sample screenshots showing facial displays from MSP-YTD dataset.

In our simulation, the 3-fold cross-validation is applied to the datasets for performance comparison. The detection accuracies discussed in

this section highly depend on the detected landmark points. It suggests that a failure in detecting the landmark points can lead to inability to detection. To this end, Table 3 shows the successful detection probabilities of the landmark points under these datasets. As can be seen, a reliable result is achieved by using facial landmarks, and it ensures a consistent cognition between the values shown in the following experiments as well as the actual performance of the proposed method. The length of the videos is ranged from just a few seconds to about thirty seconds. A paragraph is divided into *K*-segments and divided once every 100 frames. The proposed method mentioned in Section 4 extracts the feature F every fragment; consequently, the feature at *k*th fragment of *n*th video F_n^k is extracted in all videos. Notably, the size of *K* is different in each video. Finally, the majority decision of the SVM classifier is applied to classify each feature F_n^1, …, F_n^k. The feature F_n^1, …, F_n^k are then utilized to classify the *n* video by the majority decision. The accuracy (ACC) is adopted as the metric for the evaluation as follows.

Table 3. Probability of landmark points detection.

Datasets	Probability of detection
Real-life trail data [9]	0.9755
MSP-YTD dataset	0.9843

$ACC = N_{corr}/N_{all}$, where Ncorr denotes the number of the correct classification video, Nall is the number of videos.

In this work, the 3 cross-validation methods are utilized to randomly divide the film into 3 groups of data for training and testing of SVM. Because the lengths of the videos are not unified, the features of honesty or lying in each film are different. The best accuracy can be derived from these random samples that were generated by the random combinations of two datasets, Real-Life Trail Data [9] and MSP-YTD. The overall accuracy is computed through 3 folds of databases. The more balanced distribution of the positive and negative samples in each fold of data, the more reliable and robust training model for testing we can obtain from SVM classification. The original frame sizes are 640×480 and the format of a color image frame is 24-bit in an RGB system. All gray level frames are used, by transferring the RGB system to the YCbCr system. It is used for the proposed system for the deception detection of features in real-time. The experimental environment is established using a CPU i7-10750H, 32 GB RAM, Microsoft Windows 10 and Open CV, OpenFace, Visual studio 2022. There are chosen as the software development platform. The frame rate for the proposed system is 50 FPS (Frame per Second).

6. Conclusion

This study presents a deception detection system with a Joint Cross-Attention model. The experimental results show the joining of the classifications SVM with the feature sets “A + E” and Joint Cross-Attention can achieve optimal performance on two datasets, Real-Life Trail Data and MSP-YTD. According to the experiments, both features ‘A’ and ‘E’ positively contribute to the system accuracy, and the joining of Joint Cross-Attention model can yield an additional improvement. As documented in the experimental results, the proposed method can be a very promising candidate for the practical application of deception detection. Future possible improvements can be made to explore more robust features for further enhancing the performance on the excessive head movement or considering speech as well.

7. References

[1] F. Charles, Jr. Bond, and M. B. DePaulo, "Accuracy of deception judgments," *Personality and Social Psychology Review*, Vol. 10, 2006, pp. 214-234.

[2] R. Adelson, "Detecting deception," *Monitor on Psychology*, Vol. 37, 2004, p. 70.

[3] Office of Technology Assessment, United States Congress, *Scientific Validity of Polygraph Testing: A Research Review and Evaluation*, University Press of the Pacific, Washington, 1983.

[4] "Education psychologists use eye-tracking method for detecting lies," psychologialscience.org, 2012.

[5] F. Horvath, J. McCloughan, D. Weatherman, and S. Slowik, "The accuracy of auditors' and layered voice analysis (LVA) operators' judgments of truth and deception during police questioning," *Journal of Forensic Sciences*, Vol. 58, 2013, pp. 385-392.

[6] K. R. Damphousse, "Voice stress analysis: Only 15 percent of lies about drug use detected in field test," *NIJ Journal*, Vol. 259, 2008, pp. 8-12.

[7] J. D. Harnsberger, H. Hollien, C. A. Martin, and K. A. Hollien, "Stress and deception in speech: evaluating layered voice analysis," *Journal of Forensic Sciences*, Vol. 54, 2009, pp. 642-650.

[8] H. Hollien, J. D. Harnsberger, C. A. Martin, and K. A. Hollien, "Evaluation of the NITV CVSA," *Journal of Forensic Sciences*, Vol. 53, 2008, pp. 183-193.

[9] V. Perez-Rosas, M. Abouelenien, R. Mihalcea, and M. Burzo, "Deception detection using real-life trial data," in *Proceedings of ACM International Conference on Multimodal Interaction*, 2015, pp. 59-66.

[10] M. Jaiswal, S. Tabibu, and R. Bajpai, "The truth and nothing but the truth: multimodal analysis for deception detection," in *Proceedings of IEEE International Conference on Data Mining Workshops*, 2017, pp. 938-943.

[11] T. Baltrusaitis, P. Robinson, and L.-P. Morency, "OpenFace: An open source facial behavior analysis toolkit," in *Proceedings of IEEE Winter Conference on Applications of Computer Vision*, 2016, pp. 1-10.

[12] P. Pudil, J. Novovicova, and J. Kittler, "Floating search methods in feature selection," *Pattern Recognition Letters*, Vol. 15, 1994, pp. 1119-1125.

[13] R. E. Fan, K. W. Chang, C. J. Hsieh, X. R. Wang, and C. J. Lin, "LIBLINEAR: a library for large linear classification," *Journal of Machine Learning Research*, 2008, pp. 1871-1874.

[14] C.-H. Hsia, J.-M. Guo, and C.-S. Wu, "Finger-vein recognition based on parametricoriented corrections," *Multimedia Tools and Applications*, Vol. 76, 2017, pp. 25179-25196.

[15] E. Turki, R. Alabboodi, and M. Mahmood. "A proposed hybrid biometric technique for patterns distinguishing," *Journal of Information Science and Engineering*, Vol. 36, 2020, pp. 337-345.

[16] Li-Wei Hsiao, Jing-Ming Guo, Gi-Luen Huang, Yi-Fang Hsieh, Chih-Hsien Hsia and Herleeyandi Markoni. "Face Expression and Tone of Voice for Deception System," *The Annual International Conference on System Science and Engineering*, 2020, #1165.

[17] V. Gupta, M. Agarwal, M. Arora, T. Chakraborty, R. Singh and M. Vatsa, "Bag-of-Lies: A Multimodal Dataset for Deception Detection," *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Long Beach, CA, USA, 2019, pp. 83-90, doi: 10.1109/CVPRW.2019.00016.

[18] P. Ekman and W. V. Friesen, "Measuring facial movement," *Environmental Psychology and Nonverbal Behavior*, Vol. 1, 1976, pp. 56-75.

[19] S. Porter, L. Brinke, and B. Wallace, “Secrets and lies: involuntary leakage in deceptive facial expressions as a function of emotional intensity,” *Journal of Nonverbal Behavior*, Vol. 36, 2012, pp. 23-27.
[20] L. Brinke, S. Porter, and A. Baker, “Darwin the detective: observable facial muscle contractions reveal emotional high-stakes lies,” *Evolution and Human Behavior*, Vol. 33, 2012, pp. 411-416.
[21] M. Owayjan, A. Kashour, N. A. Haddad, M. Fadel, and G. A. Souki, “The design and development of a lie detection system using facial micro-expressions,” in *Proceedings of IEEE International Conference on Advances in Computational Tools for Engineering Applications*, 2012, pp. 33-38.
[22] L. Su and D. L. Martin, “High-stakes deception detection based on facial expressions,” in *Proceedings of IEEE International Conference on Pattern Recognition*, 2014, pp. 2519-2524.
[23] M. F. Valstar, T. Almaev, J. M. Girard, G. McKeown, M. Mehu, L. Yin, M. Pantic, and J. F. Cohn, “Fera 2015-second facial expression recognition and analysis challenge,” in *Proceedings of IEEE International Conference and Workshops on Automatic Face and Gesture Recognition*, Vol. 6, 2015, pp. 1-8.
[24] T. Baltrušaitis, M. Mahmoud, and P. Robinson, “Cross-dataset learning and personspecific normalisation for automatic action unit detection,” in *Proceedings of IEEE International Conference on Automatic Face and Gesture Recognition*, Vol. 6, 2015, pp. 1-6.
[25] Y. I. Tian, T. Kanade, and J. F. Cohn, “Recognizing action units for facial expression analysis,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 23, 2001, pp. 97-115.
[26] T. Kanade, J. F. Cohn, and Y. Tian, “Comprehensive database for facial expression analysis,” in *Proceedings of IEEE International Conference on Automatic Face and Gesture Recognition*, 2000, pp. 46-53.
[27] Han, S., Hessel, J., Dziri, N., Choi, Y., & Yu, Y. (2023). CHAMPAGNE: Learning Real-world Conversation from Large-Scale Web Videos. arXiv preprint arXiv:2303.09713.
[28] MCGURK, H., MACDONALD, J. Hearing lips and seeing voices. *Nature* **264**, 746–748 (1976). https://doi.org/10.1038/264746a0
[29] Kahou, Samira Ebrahimi, et al. "Emonets: Multimodal deep learning approaches for emotion recognition in video." *Journal on Multimodal User Interfaces* 10 (2016): 99-111.
[30] Owens, Andrew & Efros, Alexei. (2018). Audio-Visual Scene Analysis with Self-Supervised Multisensory Features.
[31]Rajasekar, Gnana Praveen, et al. "A joint cross-attention model for audio-visual fusion in dimensional emotion recognition." *arXiv preprint arXiv:2203.14779* (2022).