Deception Detection System with Joint Cross-Attention

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

[a] College of Arts and Science,New York University, USA. pj2097@nyu.edu
[b] Faculty of Engineering, The University of New South Wales, Australia. z5320828@ad.unsw.edu.au
[c] Earl Warren College, University of California-San Diego, USA. jil186@ucsd.edu
[d] Mellon College of Science, Carnegie Mellon University, USA. ziyangwa@andrew.cmu.edu

[*] Corresponding author email: pj2097@nyu.edu

[#] These authors contributed equally to this work

Abstract. In recent research, the field of biometrics has turned its attention towards detecting deception. Building on insights from criminal psychology, which highlights the significance of facial expressions and voice tones in uncovering deceit, this study introduces a novel system for detecting deception. This system integrates a trio of components: displacement of 68 facial landmarks, action units (AUs), and audio emotion units (EUs). It leverages criminal psychology's findings to track changes in facial expressions across consecutive frames and to assess shifts in emotions through audio analysis. A key discovery is the potential of the facial tensor, derived from the changes in the 68 facial landmarks across frames, as a robust indicator for deception detection. The system's effectiveness is evaluated using three datasets: the public datasets Real-life and Bag-of-lies, and a private dataset, MSPL-YTD. Overall, this new approach shows promise as an effective tool for intelligent deception detection.

Keywords: Deception Detection, Facial Landmarks Analysis, Emotion Units, Action Units, Joint Cross-Attention model

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 [3] 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 [4-7] deploy computers to analyze vocal attributes to infer deceit. Pérez-Rosas et al. [8] integrated linguistic and gestural features, yet overlooked crucial visual cues like facial expressions. Jaiswal et al. [9] and others [10-12] 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 [13, 14]. Our study delves into facial analysis for lie detection, addressing the challenges mentioned earlier. In our approach, we detect facial landmarks [15] 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 [16] in late-fusion stage, and the SVM [17] 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 Linguistic Deception Detection

Extensive research has focused on detecting deceptive content across diverse platforms, including online dating sites, social networks, forums, and consumer review websites. The effectiveness of text analysis features, ranging from fundamental linguistic elements such as n-grams and sentence counts, to more sophisticated linguistic attributes extracted from syntactic CFG trees and part-of-speech tags, has been proven. Certain research has integrated the examination of psycholinguistic facets by employing the Linguistic Inquiry and Word Count (LIWC) lexicon to construct deception models using machine learning methods. This has unveiled the significance of psycholinguistic data in the automated detection of deceit. Additionally, scholars have investigated the correlation between the syntactic complexity of text and deceit. They've posited that deceivers might utilize simpler sentence structures to obscure their falsehoods and aid in the easier recollection of their lies.

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. Fornaciari and Poesio provide an example of this kind of research, concentrating on spotting deception in statements given by witnesses and defendants in Italian court proceedings. Building upon this, our study delves into detecting deceit using authentic trial data and investigates the use of various modalities for this specific objective.

2.2 Nonlinguistic Deception Detection

In the past, non-verbal deceit detection primarily leaned on polygraph tests, evaluating physiological factors such as heart rate, respiration, and skin temperature. Nonetheless, research has highlighted that depending exclusively on physiological metrics can introduce biases and inaccuracies. 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 [18] 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 [10]. 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. [11] A self-supervised method of such combination is done in 2018. [12] 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. [19]

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 [8] 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 [20] 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 [21], 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 presents the workflow of the proposed method during its training phase. Initially, face detection is carried out using face localization [15], followed by the extraction of facial landmarks [15] to identify 68 specific facial points, labeled {P1, P2, ..., P67, P68}, as shown in Fig. 2. This process efficiently captures distinct facial features. The feature vector F, comprising two types of characteristics, forms the basis of this study. These characteristics are divided into two categories: (1) Action (A) Units and (2) Emotion (E) Units, with further details provided subsequently. In the final stage, a Joint cross-attention model is implemented, and Support Vector Machine (SVM) is used for classifying deception.

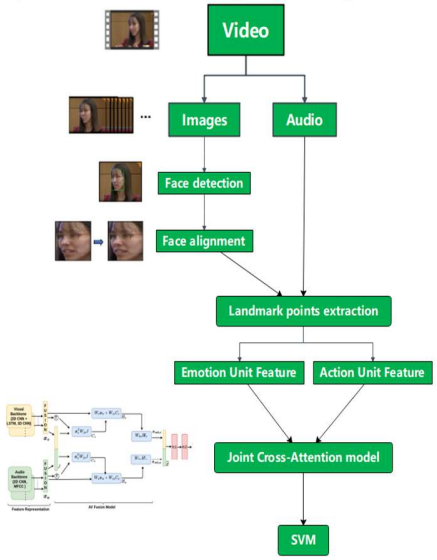

Fig. 1. The flow of the proposed method in training phase.

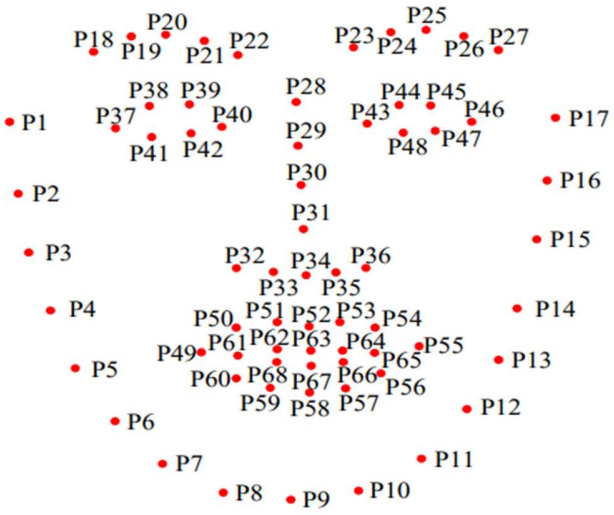

Fig. 2. 68 Facial landmarks.

4.1 Facial Action and Emotion Units

The Facial Action Coding System (FACS) [22] is a system that maps specific facial movements to the emotions they represent. Examples of these Action Units (AUs) are depicted in Fig. 3. FACS is a unique method for real-time emotion assessment, particularly useful in identifying whether someone is lying or telling the truth. Recent research, including studies by Porter et al. [33, 34] and Owayjan et al. [25], indicates that individuals who are lying often feign emotions like sadness to hide their guilt. Su et al. [26] have identified behaviors like blinking, eyebrow movement, and mouth motion as indicators of deceit. This study employs FACS in psychometric or deception testing to capture a participant's genuine emotional response.

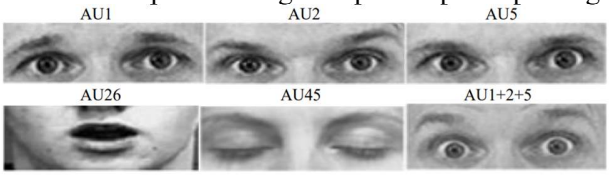

Fig. 3. Action units **extracted** from Cohn and Kanades dataset [30].

FACS is divided into two main components: (1) Main Action Units (Feature A) and (2) Emotion Units (Feature E). Each AU is linked to a specific facial movement, affecting either a part or an entire facial region, and it's possible for multiple AUs to be active at once. The aim of this research is to identify these Action Units and Emotion Units, as outlined in Tables 1 and 2. Emotion Units emerge from the simultaneous occurrence of multiple Action Units. These AUs and EUs serve as indicators for identifying deceitful and honest individuals. This detection process is based on an advanced AU recognition framework [27, 28], further explained in Baltrusaitis et al. [28, 29].

Fig. 3. Action units **extracted** from Cohn and Kanades dataset [30].

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	Up-per Lid Raiser	Eyes
AU6	Cheek Raiser	Eyes
AU7	Lid Tightener	Eyebrows + Eyes
AU9	Nose Wrinkler	Eyebrows + Nose
AU10	Upper Lip Raiser	Mouth
AU12	Lip Comer 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

The method involves isolating each Action Unit (AU) frame by frame, followed by calculating each Emotion Unit (EU) using the corresponding AU, as detailed in Table 2. For every frame, we generate a binary sequence for each AU and EU, such as AU6, AU12, and their combination AU6+AU12 (EU), as illustrated in Fig. 4. In this sequence, a '1' signifies the presence and a '0' the absence of an AU or EU. The crucial final step is the extraction of AU and EU features, with a focus on the number of frames utilized in the Real-Life dataset. These features are categorized into two groups: the total count of occurrences where AUs/EUs are present (represented by '1' in the binary sequence), and the total count of transitions in these occurrences (changing from '1' to '0' or vice versa). Fig. 4 shows the occurrence of AU6, AU12, and the combination AU6+12, where the latter suggests emotions like happiness or joy. Each orange dot on the

graph represents a frame, as shown in Fig. 4. The occurrences and transitions of AU6, AU12, and AU6+12 are quantified across 60 frames. These details are then utilized to construct visual vector features (A+E) for a 46-dimensional analysis.

In the final stage, all the extracted Action Unit and Emotion Unit features are fed into a Joint Cross-Attention model and analyzed using a Support Vector Machine (SVM) to classify deception.

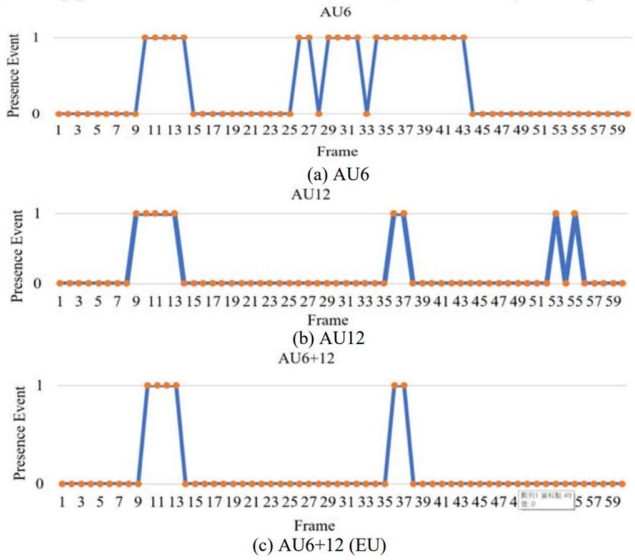

Fig. 4. A binary sequence event where 1 represents the frame involving the presence of the AU or EU and 0 does not involve the presence.

4.2 Joint Cross-Attention model

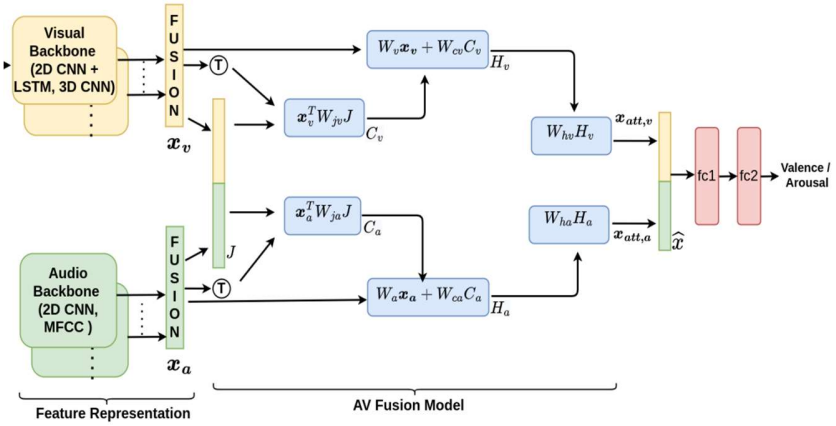

Fig. 5. Joint cross-attention model proposed for $A-V$ fusion

In the Fig. 5, there shows a learning model that can combine multiple sources of information together and achieves a 2x1 vector as an outcome.

It is feasible to achieve audio-visual $(A-V)$ fusion through integrated multimodal training, yet it's been noted that such multimodal systems often underperform compared to systems using single modalities. This underperformance can be attributed to several factors, such as differing learning dynamics between audio

(A) and visual (V) modalities, variable noise levels in each modality, and the need for modality-specific input representations. Therefore, to capture the distinct characteristics of each modality, we trained deep learning models separately for the audio and visual components. This approach allows us to extract unique features from each. These features are subsequently integrated using the Joint Concatenation Approach (JCA) fusion model for A-V fusion. The fusion model outputs data for both Emotion Units and Action Units, which are then processed using a Support Vector Machine (SVM) for the purpose of deception detection.

In varying segments of a video sequence, the visual (V) modality's significance fluctuates, with the audio (A) modality sometimes playing a more pivotal role. Recognizing that combining multiple modalities offers richer and more complementary insights for predicting valence and arousal than a single modality, we leverage the synergistic potential of $A-V$ fusion. Our strategy for effective fusion is grounded in a cross-attention fusion method, designed to adeptly encode cross-modal information while retaining the unique properties of each modality. Unlike conventional approaches where cross-attention is applied separately to each modality's features, our method adopts a unified approach. We generate a combined A-V feature set by merging audio and visual features, applying cross-attention to both. This unified feature set enables each modality to interact with not only its own features but also those of its counterpart, thus enhancing the detection of meaningful connections between the A and V modalities. Moreover, this approach significantly diminishes the disparities between the audio and visual modalities within the cross-attention framework, contributing to notable enhancements in overall system efficacy.

In our approach to audio-visual ($A-V$) fusion, we employ a joint cross-attention model. Within this framework, we combine audio (A) and visual (V) features to create a unified $A-V$ feature representation. This combined representation is then processed using cross-attention, allowing it to focus on both the individual audio and visual features within this joint context.

In the training mode, from a given video sub-sequence input S with fixed size , we let X_a and X_v be two sets of deep feature vectors extracted for A and V modalities of S, defined as

$$X_a = \{x_a^1, x_a^2, \dots, x_a^L\} \in \mathbb{R}^{d_a \times L} \text{ and } X_v = \{x_v^1, x_v^2, \dots, x_{av}^L\} \in \mathbb{R}^{d_v \times L}$$

In this context, L signifies the quantity of non-overlapping,fixed-size clip samples uniformly derived from the sub-sequence S. Additionally, D_a and D_v correspond to the feature dimensions of the audio (A) and visual (V) representations, respectively. We then create a combined representation of these $A-V$ features by concatenating the respective feature vectors of A and V, which is represented as J . Therefore, J is defined as:

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

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

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

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:

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

In the case of the audio (A) modality, we integrate the joint correlation matrix C_a and the associated audio features X_a using learnable weight matrices W_{ca} and W_a respectively. This integration is employed to calculate the attention weights for the A modality using the ReLu function.

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

Similarly for the case of the video modality (V), we have:

$$H_v = ReLu(W_v X_a + W_{cv} C_{av}^T)$$

We further utilize the attention maps obtained to calculate the attended features of the A and V modalities in the following manner:

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

Subsequently, the attended A and V features are combined, resulting in the generation of the $A - V$ feature representation as follows:

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

The audio-visual features we gathered are subsequently inputted into the fully connected layers (fc1 and fc2), which yield the outcomes for the Emotion Unit and Action Unit.

5. Experiments

In the following experiments, two datasets, the Real-Life Trail Data [8] 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.

Fig. 6. Sample **s**creenshots showing facial displays from Real-Life Trail clips.

The average lengths of deceptive and truthful videos in the dataset are 24.96 seconds and 27 seconds, respectively. This dataset includes 21 female and 35 male speakers, ranging in age from 16 to 60 years. As detailed in source [8], three types of trial outcomes were used to accurately classify video clips as deceptive or truthful: guilty verdicts, non-guilty verdicts, and exonerations. For guilty verdicts, deceptive videos were obtained from defendants, while truthful videos were collected from witnesses in a trial. Sometimes, deceptive videos were gathered from suspects denying a crime they committed, whereas truthful clips were taken from the same suspects when they responded to questions about facts verified as truthful by the police. Exoneration testimonies were categorized as truthful statements. In contrast, the MSP-YTD dataset comprises 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, we use 3-fold cross-validation on the datasets to compare performance. The detection accuracies mentioned in this part are greatly influenced by the accuracy of landmark point detection. This implies that failing to detect these landmark points might result in detection failure. Consequently, Table 3 presents the probabilities of successfully detecting landmark points in these datasets. The data indicates that using facial landmarks yields dependable results, ensuring a consistent understanding between the values observed in subsequent experiments and the real-world effectiveness of the method we propose. 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 kth fragment of nth 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 N_{corr} denotes the number of the correct classification video, N_{all} 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 [8] 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 introduces a deception detection system that incorporates a Joint Cross-Attention model. The findings from the experiments indicate that combining the classifications of SVM with the feature sets "$A + E$" and Joint Cross-Attention leads to optimal results on the Real-Life Trail Data and MSP-YTD datasets. The results demonstrate that both 'A' (Action Units) and 'E' (Emotion Units) features significantly enhance the system's accuracy, and the integration of the Joint Cross-Attention model further boosts its performance. The experimental evidence suggests that this proposed method holds great promise for practical use in deception detection. Looking ahead, there are opportunities for improvement, such as developing more sophisticated features to better handle scenarios with excessive head movement or incorporating speech analysis into the system.

Acknowledgement

Peili Jiang, Yunfan Wang, Jiajun Li, and Ziyang Wang contributed equally to this work, they should be considered as co-first author.

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

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] MCGURK, H., MACDONALD, J. Hearing lips and seeing voices. *Nature* 264, 746–748 (1976). https://doi.org/10.1038/264746a0

[11] Kahou, Samira Ebrahimi, et al. "Emonets: Multimodal deep learning approaches for emotion recognition in video." *Journal on Multimodal User Interfaces* 10 (2016): 99-111.

[12] Owens, Andrew & Efros, Alexei. (2018). Audio-Visual Scene Analysis with Self-Supervised Multisensory Features.

[13] 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.

[14] 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.

[15] 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.
[16] P. Pudil, J. Novovicova, and J. Kittler, “Floating search methods in feature selection,” *Pattern Recognition Letters*, Vol. 15, 1994, pp. 1119-1125.
[17] 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.
[18] 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.
[19]Rajasekar, Gnana Praveen, et al. "A joint cross-attention model for audio-visual fusion in dimensional emotion recognition." *arXiv preprint arXiv:2203.14779* (2022).
[20] 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.
[21] 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.
[22] P. Ekman and W. V. Friesen, “Measuring facial movement,” *Environmental Psychology and Nonverbal Behavior*, Vol. 1, 1976, pp. 56-75.
[23] 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.
[24] 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.
[25] 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.
[26] 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.
[27] 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.
[28] 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.
[29] 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.
[30] 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.