Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 14
TPCEE 2026

Analysis of the progress in lie detection technology based on multimodal information fusion

Xianghan Lv

Communication Engineering, School of Informatics, Xiamen University, 361005, China

37120242203967@stu.xmu.edu.cn

Abstract. Increasing the aim at accuracy, real-time transformation, and non-invasive deception detection contributes to the emergence of lie detection technology based on multimodal information fusion, which effectively breaks down barriers of uncertainty and limitations of a single model. This paper analyzes the novel system from three cores: audio-visual deep integration technology, facial micro-expressions and fine-grained feature extraction, and comprehensive machine learning combined with NLP fusion techniques. At the same time, the analysis deeply researches the use in the real world of this technology, and prospectively proposes a decolonization Artificial Intelligence (AI) framework and an audit game theory framework for the problems of data bias and Othello Error. Research has confirmed that lie detection technology based on multimodal information fusion greatly strengthened the impact of lie detection, successfully overcoming many problems of traditional single-modal approaches. To sum up, the analysis not only provided a more highly scalable non-invasive deception detection but also observably weakened subjective bias, thus laying the foundation for reliable, transparent, and practical use of the technology in the real world.

Keywords: lie detection technology; multimodal information fusion; novel system.

1. Introduction

Thanks to the development of medical technology, current systems can effectively detect lies by measuring physiological indicators such as heart rate and brain waves. Meanwhile, limited by the complexity of the environment and the pursuit of detection accuracy, real-time feedback and the interpretability of the results, there is unprecedented demand for non-invasive deception detection. Recent systematic reviews highlight that AI-enabled deception detection is rapidly expanding across fields like security and forensics, underscoring the critical need for more robust, non-invasive approaches [1]. In this context, lie detection technology based on multimodal information fusion has emerged. This novel system addresses the uncertainty and limitations of single-modal detection by combining diverse elements such as speech, facial expressions, and textual data to identify deceptive behavior [2]. In response to the lack of comprehensiveness in the analysis of related technologies in most related reviews, this study has been writing relevant innovative papers since 2025. At the beginning of this analysis, the author deeply analyzes the content and the practical use of three mainstream technologies: audio-visual deep integration, facial micro-expressions and fine-grained feature extraction, and comprehensive machine learning with NLP fusion. At the end of this analysis, the two major challenges currently faced by related technologies are proposed, along with corresponding solutions. Through this research, a relatively comprehensive analysis of lie detection technologies based on multimodal information fusion since 2025 has been achieved. At the same time, by analyzing the characteristics of different technologies, it can help lie detection practitioners choose different methods according to different scenarios.

2. Analysis of Multimodal Deception Detection Technologies

2.1. Audio-visual deep integration technology

As a pioneer breaking the limitations of traditional single-modal approaches, the technology of relation-aware visual deep integration systematically constructs a new analysis framework with the

Content from this work may be used under the terms of CC BY-NC 4.0 licence (https://creativecommons.org/licenses/by-nc/4.0/).
Published by Warwick Evans Publishing.

core of visual representation [3]. The main contribution of this technology is the efficient combination of deep learning and dynamic evaluation, which can not only capture macroscopic facial expressions but also keenly perceive imperceptible uncoordinated movements. In practical use, the model extracts visual modalities in parallel, precisely targeting micro-cues such as eye gaze, head pose, and limb movements. Then, the relation-aware module gets involved and dynamic weighting is applied to these cues, thus capturing the suspicious signals that are deeply connected to lies. In a complex test environment, this approach dramatically improves recognition reliability in a high-risk environment and successfully solves the data deviation caused by environmental differences.

The focus now shifts to the multi-modal deception analysis architecture based on the Transformer. The technology innovatively embeds a modal auto-completion mechanism to address environmental noise in the real world. When a modality is impacted by noise, the system can still get information from intact side modalities, reverse-deduce, and then fill the information gaps with its Transformer attention mechanism. Specifically, the first thing to do after multi-source data enters the system is modal alignment; when it comes to the feature encoding stage, the Transformer pinpoints and suppresses noise interference, generatively repairs incomplete data with cross-modal-attention networks, and refines a unified fusion feature in the end. The key to assessing this technology is its determination in environments with low signal-to-noise ratios and its performance decline threshold when a large area of modal data is missing (with a loss rate of up to 30% to 50%). After using a specially constructed real dataset, DECEPTlcON was trained; its performance in a real-world noise environment exceeded that of the traditional Convolutional Neural Network (CNN) model, greatly improving the reliability of use [4].

2.2. Micro-Expressions and Fine-Grained Feature Extraction

The first framework to be considered is the Global-Local Feature Fusion Network (GLFNet) technology [5]. This solution innovatively establishes a channel for both global and local features: on one hand, the global branch is responsible for overseeing the macroscopic limb posture of the subject and the background environment; on the other hand, the local branch uses the attention mechanism to tightly lock onto the emergence areas of micro-expressions such as the corners of the eyes and mouth. This design effectively addresses a persistent limitation of traditional models, which often miss subtle deceit cues in large-scale scenes. Video frames are preprocessed and simultaneously distributed to the global encoder and local interest area extractor; the cross-attention module takes on the responsibility of aligning the dual features and performing deep interaction later; the fusion layer outputs the final judgment of the deceit probability. The accuracy of local feature anchoring, the dilution effect of global context on noise, and the classification robustness under variable focal lengths and different shooting angles collectively construct the core evaluation indicators of this technology. Thanks to its keen sense of granularity, GLFNet excels in detecting surveillance videos in long-range or uncontrolled environments, significantly improving the capture rate of concealed emotions.

2.3. LLM-Driven Deception Detection

The author then delves into Large Language Model (LM) DeceptionNet, which deeply combines powerful semantic parsing power and behavioral patterns of large language models [6]. Its core innovation is establishing a bridge connecting textual meaning and physiological stress signals; thus, it can accurately detect a lack of narrative logic or avoidance of vocabulary usage caused by making up lies. During operation, the system first transforms speech into structured text, then assigns a pre-trained language model to investigate the underlying logical currents behind the text; at the same time, the stress readings of the visual modality are also undergoing rigorous comparison and analysis. Once the "affirmation" exposed at the semantic level collides with the "anxiety" leaked at the visual level, the model will label it as deceitful behavior. The detection rate of semantic conflicts, the concealment level of first-person pronouns in false narratives, and the depth of understanding of long and complex insurance claim texts by the model are the three main criteria for evaluating this network. In scenarios

such as credit approval and insurance claims, LM DeceptionNet can effectively identify false statements, with its detection accuracy for complex logical deceit far exceeding traditional machine learning models

The technology of comprehensive machine learning and LLM techniques has transformed into a powerful auxiliary tool for judicial experts evaluating key testimonies [7]. Using this framework, the system can analyse multidimensional language and motion data streams at the same time, aiming to replace traditional detection methods with inherent defects in their underlying logic. By systematically applying sorting algorithms to transcripts of court proceedings and visual cues, the model can subtly recognise minor indicators such as fragmented sentence structures, emotionalized expressions, and abnormal cognitive load. This design completely replaces black-box decision-making mechanisms in the past, greatly raises the reference value of objective evidence during the judicial determination stage by systematically reflecting specific behavioral characteristics and deceptive intentions.

The technology of fusing natural language processing and facial micro-features deeply combines language semantic analysis with physiological characterization, aiming to address the complex challenge of interpersonal interactions in anti-fraud scenarios. Based on this framework, the system extracts deep clues in unscripted polygene data, using lightweight Convolutional Neural Networks (CNNs) to synchronously process and cross-validate linguistic cues and visual cues. The system accurately analyses the match degree between language content and real-time visual feedback. As soon as these two types of clues show a clear deviation and inconsistency, the warning will be triggered. On dual test sets, this multimodal fusion model achieved high accuracy, building a solid defense line based on objective data for security prevention [8].

3. Analysis of Applications

3.1. Analysis of Audio-Visual Deep Integration Applications

3.1.1. Relation-Aware Visual Integration in Interrogations.

This powerful technology mainly serves law enforcement officers, security analysts, or forensic experts, specializing in interrogating suspects who try to conceal their true selves in high-pressure environments. In the process, investigators have no need to urge suspects to wear any intrusive contact devices; using high-definition cameras and sensitive microphones is sufficient to achieve the extraction of facial video frames and audio spectra simultaneously. The system abandons traditional surface-level lie detection characteristics and turns to use the technology of global multimodal feature fusion, which provides a technical standard for rapid autofiltering of massive video information, representing a comprehensive leap of existing lie detector techniques [9].

Compared to the traditional method of single-point breakthrough, this system can strip out facial action units, acoustic features, and semantic transcripts from complex audio-visual streams at the same time. Furthermore, it can achieve high-precision generalization across large-scale heterogeneous datasets without requiring manually predefined rigid baselines. An advanced attention mechanism deployed in the model enables it to adaptively align subtle and imperceptible deceptive cues across different modalities, capturing micro-expressions and vocal tremors that human evaluators might miss.

The application effect reflects that the framework greatly reduces the deviation caused by human subjective assessment while smartly filtering redundant background noise. By deeply excavating potential multimodal contradictions, the system can stably output high-precision ratings in highly deceptive complex samples, ensuring the high availability and operational robustness of the system in a dynamic social environment.

3.1.2. Applications of Transformer-Based Architectures.

This system is established for on-site law enforcement personnel, border security examiners, and frontline investigators to deal with uncooperative interviewees in noisy, uncontrolled, and high-risk "in-the-wild" environments. When facing a complex environment filled with crowd noise, traffic noise, and unpredictable systemic interference, law enforcement officers use portable integrated devices to collect behavioral and physiological signals in real time. The technology of fusing behavioral and physiological multimodal data is technically designed for these high-risk investigation and enforcement environments, significantly enhancing the analysis of objective evidence [10].

By integrating neurophysiological indicators such as electrocardiograms (ECG) and eye-gaze tracking with dynamic facial behavior analysis, this technique completely breaks the reliance on signals that are single and easy to control in the past. When processing unscripted natural interactions, the system utilizes a dynamic score-level fusion strategy to deeply cross-compare audio, visual, and internal physiological features, which effectively prevents suspects from deliberately masking unimodal features after receiving counterintelligence training.

Large-scale subject validation demonstrates that this multimodal comprehensive assessment mechanism substantially overcomes the errors and blind spots of single models. Even in environments with great external interference or when subjects exhibit strong emotional resistance, the system can still infer the true psychological state from cross-modal contextual correlation. This ensures that the system maintains extremely high robustness, operational reliability, and objective deployment value in complex and ever-changing real-world social scenarios around the clock.

3.2. Micro-expressions and Fine-grained Application Analysis

This technology primarily empowers customs border inspection personnel, national security examiners, or clinical psychologists, specifically for high-risk subjects with extreme emotional control ability. During brief and high-pressure examination intervals, subjects typically maintain a neutral expression, only revealing extremely slight twitches lasting less than 0.5 seconds.

Inspectors utilize this dual-branch network system: first, extracting optical flow features; then, the "local" branch, like a microscope, focuses on specific areas of the eyes and corners of the mouth for the tiniest muscle tremors; at the same time, the "global" branch scans the entire facial contour to extract macroscopic baselines. Ultimately, the system dynamically adapts and fuses the local micro-tremors with the global macroscopic state [11].

This dual-track collaborative fusion mechanism perfectly overcomes the critical flaw of micro-expressions, which are easily missed due to their extremely small amplitude and uneven distribution, enabling security personnel to accurately detect the real cracks exposed by the subjects' calm camouflage out of fear.

3.3. Application Analysis of Technique Driven by Large Language Models

3.3.1. Analysis of Application Technologies Related to LM DeceptionNet.

This technology is widely used by cross-border investigation agencies, large-scale e-commerce risk control platforms, or exit and entry management departments, directly targeting suspects with different mother tongues. In the face of large-scale cross-language written records or digital statements, investigators do not need to build massive foreign language databases from scratch.

The system directly loads the pre-trained universal multilingual large model and performs a minimal "contextual fine-tuning" with a small number of samples [12]. This technology directly transfers the known lie-detection knowledge to the target language task, automatically overcoming language barriers, and locking in the subconscious linguistic features shared by liars (such as a sharp decrease in first-person pronouns, a concealed increase in specific negative emotional words, etc.).

3.3.2. Application of Comprehensive Machine Learning.

This technology transforms into a powerful assistant for judges, juries, or judicial experts, exclusively tasked with evaluating witnesses providing key testimony and defendants. In court, requiring absolute transparency and procedures, review staff lock inputs, video frames, audio acoustic coefficients, and transcribed text into an integrated machine learning architecture.

The model not only outputs an opaque probability score, but combine Explainable Artificial Intelligence (XAI) frameworks to clarify the importance of specific visual and action clues in high-risk environments. The system progressively maps specific behavioral features to deceptive intentions, explicitly identifying subtle indicators such as abnormal cognitive load and suppressed facial movements [13].

This technique completely solves the legal problems of the opaque black box in traditional AI lie detectors. It can not only provide true judgment, but can establish logical and highly interpretable quantitative standards at the same time, functioning much like an explainable written judgment. This transparent deduction endows the AI conclusions with the authority of judicial review, greatly enhancing their objective evidentiary reference value during the conviction phase.

3.3.3. Application of Fusing NLP and Facial Micro-Features.

This technology is widely adopted by cybersecurity defense experts and financial quantitative risk control systems to combat scammers who set up strategy traps in complex interpersonal interactions. In anti-fraud scenarios or virtual business negotiations where human manipulation is prevalent, defense systems must extract deep cues from unscripted multi-source data.

After receiving radio and video input, the system uses multimodal fusion technology to synchronously process and cross-validate language clues and spatiotemporal visual cues. During the dynamic interview, the framework accurately analyses the degree of match between language content, audio transcript features and real-time visual feedback [14].

This technique achieves an upgrade in detection dimension by conducting cross-modal comparison forwardly instead of relying on a single modality. When these distinct modalities exhibit obvious deviation and incoordination, the defense system can instantly trigger an alert, automatically block high-risk transactions, and provide an extremely stable, forward-looking anti-fraud moat based on objective data for high-end financial negotiations and security screening.

4. Challenges and responses

4.1. Challenges

The practical implementation of multimodal lie detection technology still faces two major challenges. The first is the lack of cross-cultural universality: currently, the training data sources for lie detection models are too singular, heavily reliant on the "WEIRD" group (Western, Educated, Industrialized, Rich, and Democratic populations). This serious data bias usually leads to erroneous judgments when facing subjects from different cultures, greatly restricting the cross-cultural applicability and reliability of lie detection.

The other problem is the “Othello error”: systems find it difficult to distinguish the tension of honest people under pressure from the panic of liars trying to conceal deception. In a real high-pressure interrogation session, honest individuals may become extremely tense due to immediate stress or concern about potential consequences. This causes them to exhibit the same physiological manifestations as deceivers, such as heart rate surges, micro-expressions, and a loss of emotional control. Since genuine liars may also exhibit similar stress responses, the high overlap in physiological indicators is likely to lead traditional models to erroneously convict the innocent, resulting in miscarriages of justice.

4.2. Responses

In response to data bias, the author put forward the “Decolonization AI Framework” to address it. Its core is eliminating the dominance of the “WEIRD” data at its source by introducing tons of data that is outside of “WEIRD”. In the training of the algorithm, the framework reshapes the weights of different emotional expressions across cultures, creatively embedding "cultural adjustment factors" directly into the training functions. In the use of this framework, the system is not limited to the analysis of facial features, but also incorporates the collection of geographic coordinates and local social advocacy of participants as core elements. This approach highly aligns with the research advocacy of the psychology community to eliminate AI bias, effectively expanding the global applicability of the model.

To meet the challenge of the "Othello Error," the author introduced the "Audit Game Theory Framework." This project built up dynamic defense mechanism to filter out interfering signals: to start with, the system sets a "pressure baseline under honest conditions," and if the participants' disarray under high-pressure questions does not significantly breach this baseline, the "Othello Error" can be preliminarily ruled out; then, the AI uses game theory to calculate the "tension cost" of the participants—if abnormal behavior is proportionally related to the potential severe punishment they face (which conforms to the fear logic of normal people), the algorithm will actively dilute the weight of this abnormal feature in the final judgment; the system's focus of investigation will firmly shift from single expressions or intonations that are easily distorted by fleeting emotions to deeper, more difficult to disguise cross-sentence logical consistency at last. This deep analysis mechanism, based on the perspective of game theory, significantly raises the accuracy of judgment in chaotic situations.

5. Conclusion

To sum up, this analysis thoroughly analyzed the content and use of three typical Lie Detection Technologies based on Multimodal Information Fusion. Using the combination of audio-visual deep integration technology, facial micro-expressions and fine-grained feature extraction, and comprehensive machine learning with NLP fusion, the practical performance of the lie detection system has risen unprecedentedly, successfully breaking down the barrier of using a single modality.

Relying on the innovative mechanisms of the "Decolonization AI Framework" and the "Audit Game Theory Framework," the analysis further optimizes the handling of challenges such as "WEIRD" data bias and "Othello errors," thereby achieving a double win in cross-cultural inclusiveness and logical rigidity. These highly extensible automatic detection and investigation solutions have significantly diluted the subjective color of human judgment, thereby ensuring that this technology can achieve more reliable, transparent, and highly practical deployment in the complex and diverse real-world scenarios.

References

[1] S.L. King, T. Neal, Applications of AI-enabled deception detection using video, audio, and physiological data: a systematic review. IEEE Access 12, 135207–135240 (2024).

[2] A. D’Ulizia, A. D’Andrea, P. Grifoni, F. Ferri, Analysis, evaluation, and future directions on multimodal deception detection. Technologies 12, 71 (2024).

[3] X. Li, et al., Detecting deceptive behavior via learning relation-aware visual representations. IEEE Trans. Inf. Forensics Secur. 20, 1–14 (2025).

[4] B. Biçer, H. Dibeklioğlu, et al., DECEPTIcON: bridging gaps in in-the-wild deception research. IEEE Trans. Affect. Comput. 16, 3452–3464 (2025).

[5] M. Zhang, L. Yao, W. Yang, Y. Yin, GLFNet: attention mechanism-based global–local feature fusion network for micro-expression recognition. Entropy 27, 1023 (2025).

[6] Y. Zhuo, V.M. Baskaran, L.Y.-K. Wang, R.C.-W. Phan, @LM DeceptionNet: a multimodal approach for efficient transfer learning-based deception detection. Knowl. Based Syst. 318, 113499 (2025).

[7] H. Elbatanouny, A comprehensive analysis of deception detection techniques leveraging machine learning. Expert Syst. Appl. 283, 127601 (2025).

[8] T. Patel, D. Vekariya, Multimodal lie detection using linguistic and visual cues: a fusion of NLP and facial micro-feature analysis. J. Innov. Image Process. 7, 1374–1397 (2025).

[9] S. Wang, et al., Multimodal deception detection: a survey. Mach. Intell. Res. 22, 1–25 (2025).

[10] A.S. Rao, et al., Multimodal machine learning for deception detection using behavioral and physiological data. Sci. Rep. 15, 8943 (2025).

[11] Y. Qi, M. Ibrayim, A. Hamdulla, Attention-based dual-branch network for micro-expression recognition with global-local feature fusion, in Proceedings of the IEEE International Joint Conference on Biometrics (IJCB), Buffalo, NY, USA, September 15-18 (2024), 1–10.

[12] D. Azuma, R. Meléndez, M. Ptaszynski, et al., SVM, BERT, or LLM? A comparative study on multilingual instructed deception detection. AI 6, 239 (2025).

[13] A.J. Hussain, et al., Explainable AI for unraveling the significance of visual cues in high stakes deception detection. IEEE Access 13, 24310–24325 (2025).

[14] R. Thakkar, S. Degadwala, Efficient deception detection using multimodal facial and audio transcript features. Int. J. Sci. Res. Sci. Technol. 12, 102–115 (2025).